

UNIVERSITY OF PIRAEUS

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

DEPARTMENT OF INFORMATICS

PHD THESIS

Thesis Title:	**(English)** **AI Assisted Cyber Crisis Management Exercise Scenario Generation** **(Greek)** **Δημιουργία Σεναρίου Ασκήσεων Διαχείρισης Κυβερνοεπιθέσεων με τη χρήση Τεχνητής Νοημοσύνης**
Student's name-surname:	**ZACHARIS ALEXANDROS**
Father's name:	**ILIAS**
Student's ID No:	**ΠΛΔ2102**
Supervisor:	**Constantinos Patsakis, Professor**

SEPTEMBER 2025

PhD Thesis
was prepared during the Programme of Doctoral Studies of the Department of Informatics of the School of Information and Communication Technologies of the University of Piraeus for the degree of Doctor of Philosophy

3-Member Supervising Committee

Constantinos Patsakis
Professor
University of Piraeus
(Supervisor)

Vasilios Katos
Professor
Bournemouth University
(Member)

Christos Ilioudis
Professor
International Hellenic University
(Member)

PhD Thesis was presented before the 7-Member Examination Committee and approved on July X, 2025

7-Member Examination Committee

Constantinos Patsakis
Professor
Department of Informatics
University of Piraeus

Vasilios Katos
Professor
School of Computing
Bournemouth University

Christos Ilioudis
Professor
Department of Information and Electronic Engineering
International Hellenic University

Efthimios Alepis
Professor
Department of Informatics
University of Piraeus

Dimitrios Vergados
Professor
Department of Informatics
University of Piraeus

Christos Douligeris
Professor
Department of Informatics
University of Piraeus

Emmanouil Magkos
Professor
Department of Informatics
Ionian University

Ευχαριστίες

Μέσα από τις επόμενες γραμμές θα ήθελα να εκφράσω τις ευχαριστίες μου σε όλους εκείνους που συνέβαλαν ουσιαστικά στην πραγματοποίηση και ολοκλήρωση της παρούσας διδακτορικής διατριβής.

Θα ήθελα, πρώτα απ' όλα, να εκφράσω την ευγνωμοσύνη και τις ευχαριστίες μου στον επιβλέποντα και μέντορά μου, Καθηγητή Κωσταντίνο Πατσάκη, για την αμέριστη ηθική υποστήριξη, το συνεχές, αδιάπτωτο ενδιαφέρον και την ακαταπόνητη καθοδήγησή του.

Επιπλέον, θα ήθελα να ευχαριστήσω τους Καθηγητές μου, Χρήστο Ιλιούδη και Βασίλη Κάτο που χρόνια τώρα με παρότρυναν να ακολουθήσω το ταξίδι ενός διδακτορικού στην κυβερνοασφάλεια, κάτι που χωρίς αυτούς ίσως να μην το είχα τολμήσει.

Θα ήθελα να εκφράσω επίσης τις ευχαριστίες μου στον συνάδελφο και φίλο μου Razvan Gavrila, για όλες τις ενδιαφέρουσες συζητήσεις που είχαμε αλλά και την πολύτιμη βοήθεια που μου προσέφερε για την ολοκλήρωση της διατριβής μου.

Τέλος, ευχαριστώ, μέσα από την καρδιά μου, όλους εκείνους που πιστεύουν σε εμένα, την οικογένειά μου και τους φίλους μου, που με ενθαρρύνουν και με στηρίζουν έμπρακτα.

Πειραιάς, Μαΐος 2025

Εκτενής Επιτελική Σύνοψη

Οι Κυβερνοασκήσεις Διαχείρισης Κρίσεων αποτελούν βασικό εργαλείο στην υλοποίηση των στόχων της Στρατηγικής Κυβερνοασφάλειας όλων των κρατών μελών της Ευρωπαϊκής Ένωσης. Ο αυξημένος αριθμός περιστατικών κυβερνοεπιθέσεων με στόχο κρίσιμες υποδομές έχει δημιουργήσει την ανάγκη θεσμοθέτησης ενός πλαισίου διαρκούς εκπαίδευσης και αξιολόγησης της ετοιμότητας των φορέων που διαχειρίζονται κρίσιμες υποδομές ώστε να αποκριθούν έγκαιρα και με ακρίβεια. Ο ρυθμός διενέργειας ασκήσεων κυβερνοασφάλειας αλλά και το περιεχόμενό τους, πρέπει να μπορεί να εξομοιώσει, στο μέτρο που αυτό είναι δυνατό, πραγματικές απειλές που ενδέχεται να επηρεάσουν τις πραγματικές λειτουργίες των οργανισμών υπό εξέταση.

Επιπλέον οι ασκήσεις κυβερνοασφάλειας αποτελούν σημαντικό εργαλείο για την αξιολόγηση της ετοιμότητας των φορέων που συμμετέχουν και τον εντοπισμό των αδυναμιών και της ευπάθειας των συστημάτων τους. Μέσω της προσομοίωσης περιστατικών ασφαλείας παρέχεται η δυνατότητα να αντιμετωπιστούν περιστατικά ασφάλειας σε συνθήκες που αναλογούν σε πραγματικά περιστατικά, με την εφαρμογή σχετικών μέτρων ασφάλειας που έχουν ληφθεί καθώς και συναφών καταρτισθέντων σχεδίων έκτακτης ανάγκης, ώστε οι φορείς να προβούν στις σχετικές βελτιώσεις και επικαιροποιήσεις. Περαιτέρω, με τις ασκήσεις αυτές ενισχύεται η ανταλλαγή πληροφοριών και γνώσεων, η συνεργασία μεταξύ των φορέων που συμμετέχουν, ενώ ενδυναμώνεται παράλληλα η κουλτούρα της συνεργασίας για την αύξηση του επίπεδου Κυβερνοασφάλειας μιας χώρας ή ενός οργανισμού.

Σκοπός αυτής της διδακτορικής διατριβής είναι η δημιουργία μιας μεθοδολογίας που θα αυτοματοποιεί και επιταχύνει την δημιουργία δυναμικού περιεχομένου κυβερνοασκήσεων διαχείρισης κρίσεων με την χρήση τεχνικών μηχανικής εκμάθησης και τεχνητής νοημοσύνης. Η μεθοδολογία αυτή θα επιτρέπει στους διοργανωτές τέτοιου είδους ασκήσεων, ακόμη και με περιορισμένες τεχνικές γνώσεις, να παράγουν δυναμικά, ρεαλιστικά σενάρια παρέχοντας μόνο περιορισμένη αδόμητη πληροφορία όπως: παρόμοια γνωστά περιστατικά επίθεσεων, τεχνικα άρθρα, τεχνικές ή εργαλεία επίθεσης ή τον τύπο της κρίσιμης υποδομής που θα εμπλακεί στην προσομείωση.

Ο εκπαιδευμένος πράκτορας τεχνητής νοημοσύνης αναλαμβάνει να επιλέξει την κατάλληλη κατηγορία απειλών, επιτιθέμενων αλλά και τεχνικών και να προσομοιώσει ποικιλία μονοπατιών επιθέσεων που θα μπορούσαν να υλοποιηθούν με βάση τον τομέα/κατηγορία της κρίσιμης υποδομής υπό εξέταση. Τελικά παράγεται ένα δομημένο τεχνικό σενάριο άσκησης, έτοιμο προς εκτέλεση λαμβάνοντας υπ΄ όψιν τις σχετικές τρέχουσες και μελλοντικές τάσεις των κυβερνοεπιθέσεων που αφορούν τον τομέα/κατηγορία της κρίσιμης υποδομής.

Βασικές Συνεισφορές

Η παρούσα διδακτορική διατριβή εισάγει σημαντικές καινοτομίες στον τομέα της κυβερνοασφάλειας μέσω της ανάπτυξης του πλαισίου AiCEF (AI-powered Cyber Exercise Framework), το οποίο αποτελεί την πρώτη ολοκληρωμένη προσέγγιση για την αυτοματοποίηση της δημιουργίας σεναρίων ασκήσεων κυβερνοασφάλειας με χρήση τεχνητής νοημοσύνης. Οι συνεισφορές της

έρευνας εκτείνονται σε τρεις αλληλένδετους τομείς, προσφέροντας πρωτοποριακές λύσεις σε κρίσιμα προβλήματα της σύγχρονης κυβερνοασφάλειας.

Στον τομέα της παραγωγής ασκήσεων κυβερνοασφάλειας, η διατριβή αναπτύσσει την πρώτη οντολογία σεναρίων ασκήσεων κυβερνοασφάλειας (CESO). Με επίκεντρο τα περιστατικά, εισάγεται μια καινοτόμος προσέγγιση που επιτρέπει την αναγνώριση και περιγραφή μεμονωμένων περιστατικών πριν τη συνένωσή τους σε ευρύτερα σενάρια. Αυτή η οντολογία είναι συμβατή με τις απαιτήσεις των σύγχρονων συστημάτων τεχνητής νοημοσύνης, επιτρέποντας την αποτελεσματική εκπαίδευση και λειτουργία μοντέλων μηχανικής μάθησης. Παράλληλα, αναπτύσσονται εξειδικευμένα μοντέλα μηχανικής μάθησης (MLCESO) που αυτοματοποιούν την εξαγωγή σχετικών οντοτήτων από αδόμητα κείμενα, μετατρέποντας πολύπλοκες αναφορές απειλών σε δομημένες πληροφορίες κατάλληλες για τη δημιουργία ασκήσεων.

Η έρευνα προτείνει επίσης μεθόδους για τον εμπλουτισμό των παραγόμενων σεναρίων με γνωστά πρότυπα δραστηριοτήτων προηγμένων επίμονων απειλών (Advanced Persistent Threats—APTs), βελτιώνοντας σημαντικά τον ρεαλισμό και την εκπαιδευτική αξία των ασκήσεων. Αυτή η προσέγγιση αποτελεί την πρώτη συστηματική μέθοδο ενσωμάτωσης τεκμηριωμένων μοτίβων επίθεσης στη δημιουργία σεναρίων, γεφυρώνοντας το χάσμα μεταξύ θεωρητικής γνώσης και πρακτικής εφαρμογής.

Στον τομέα της ενσωμάτωσης πληροφοριών κυβερνοαπειλών και πρόβλεψης, η διατριβή προτείνει την ενσωμάτωση προηγμένων μηχανισμών πρόβλεψης στη διαδικασία παραγωγής ασκήσεων. Διεξάγεται εκτενής συγκριτική αξιολόγηση διαφόρων αλγορίθμων πρόβλεψης, εστιάζοντας στην ακρίβεια τους για συγκεκριμένους τομείς κρίσιμων υποδομών όπως ορίζονται από την Οδηγία NIS2. Τα ευρήματα αναδεικνύουν τα LSTM και Prophet ως τα αποτελεσματικότερα μοντέλα για την πρόβλεψη κυβερνοαπειλών ανά τομέα, παρέχοντας πολύτιμες κατευθύνσεις για μελλοντικές εφαρμογές. Σημαντική καινοτομία αποτελεί η ανάπτυξη μιας πρωτοποριακής μεθοδολογίας για την πλήρη αυτοματοποίηση της δημιουργίας προτροπών (prompts), η οποία επιτρέπει ακόμη και σε μη έμπειρους οργανωτές ασκήσεων να σχεδιάζουν αποτελεσματικά σενάρια. Αυτή η προσέγγιση γεφυρώνει το κρίσιμο χάσμα εξειδίκευσης που αντιμετωπίζουν πολλοί οργανισμοί, δημοκρατίζοντας την πρόσβαση σε υψηλής ποιότητας εκπαιδευτικές διαδικασίες κυβερνοασφάλειας.

Στον τομέα της δημιουργίας τοπίων κυβερνοαπειλών (Cyber Threat Landscapes), η διατριβή βελτιστοποιεί μοντέλα τεχνητής νοημοσύνης για την εξαγωγή πληροφοριών απειλών, με ιδιαίτερη έμφαση στη βελτίωση της ακρίβειας και αποδοτικότητας. Πραγματοποιείται εκτενής συγκριτική ανάλυση προηγμένων Μεγάλων Γλωσσικών Μοντέλων (LLMs) όπως το Gemini 1.5 Προ και το GPT-4o, καθώς και μοντέλων βασισμένων σε BERT (TRAM, TTPHunter) εφαρμόζοντάς τα στο ίδιο σώμα αναφορών (dataset) κυβερνοασφάλειας για άμεση σύγκριση. Ιδιαίτερα αξιοσημείωτη είναι η εισαγωγή μιας καινοτόμου τεχνικής συμφωνίας βασισμένης σε ψηφοφορία (Voting-Based Agreement) για τη βελτίωση της ακρίβειας των μοντέλων στην εξαγωγή ιδιαίτερα πολύπλοκων χαρακτηριστικών, όπως οι τακτικές, τεχνικές και διαδικασίες MITRE (TTPs). Αυτή η προσέγγιση συνδυάζει τα πλεονεκτήματα διαφορετικών μοντέλων τεχνητής νοημοσύνης, επιτυγχάνοντας υψηλότερη ακρίβεια χωρίς την ανάγκη επανεκπαίδευσης των υποκείμενων μοντέλων.

Η έρευνα αντιμετωπίζει επίσης κρίσιμες προκλήσεις στην εξαγωγή κινήτρων επίθεσης και στην αυτόματη κατηγοριοποίηση απειλών, προτείνοντας υβριδικές μεθόδους που συνδυάζουν την ευελιξία των LLMs με την ειδίκευση των προσαρμοσμένων μοντέλων. Αυτή η προσέγγιση επιτρέπει την αποτελεσματική επεξεργασία μεγάλων όγκων αναφορών απειλών με σημαντικά

μειωμένο κόστος και χρόνο, καθιστώντας την παραγωγή ενημερωμένων τοπίων κυβερνοαπειλών οικονομικά βιώσιμη για οργανισμούς όλων των μεγεθών.

Πειραματική Αξιολόγηση και Αποτελέσματα

Η αποτελεσματικότητα του προτεινόμενου πλαισίου AiCEF έχει επικυρωθεί μέσω εκτενών πειραματικών αξιολογήσεων και περιπτωσιολογικών μελετών που καλύπτουν διαφορετικούς τομείς κρίσιμων υποδομών. Τα αποτελέσματα καταδεικνύουν σημαντικές βελτιώσεις σε πολλαπλές διαστάσεις της διαδικασίας δημιουργίας ασκήσεων κυβερνοασφάλειας. Όσον αφορά την ταχύτητα δημιουργίας σεναρίων, το AiCEF επιτυγχάνει μείωση του απαιτούμενου χρόνου στο μισό σε σύγκριση με παραδοσιακές μεθόδους. Η ποιότητα των παραγόμενων σεναρίων αξιολογήθηκε από ανεξάρτητη ομάδα εμπειρογνωμόνων κυβερνοασφάλειας για διάφορες παραμέτρους όπως ρεαλισμός, τεχνική ακρίβεια και εκπαιδευτική αξία. Τα σενάρια που δημιουργήθηκαν με χρήση του AiCEF έλαβαν το μεγαλυτερο σκορ με μέσο όρο 8/10, βελτιώνοντας κατα μία μονάδα την συνολική ποιότητα σε σχέση με σενάρια κατασκευασμένα με χρήση παραδοσιακών μεθόδων. Ιδιαίτερα εντυπωσιακά είναι τα αποτελέσματα όσον αφορά τη συνάφεια με το σύγχρονο τοπίο απειλών. Τα σενάρια που δημιουργήθηκαν με το AiCEF ενσωματώνουν κατά μέσο όρο 85% των πιο πρόσφατων τεχνικών επίθεσης και τάσεων απειλών, ενώ τα χειροκίνητα σενάρια ενσωματώνουν μόλις 65%. Αυτή η διαφορά οφείλεται στην ικανότητα του συστήματος να επεξεργάζεται και να αναλύει μεγάλους όγκους πρόσφατων δεδομένων απειλών σε πραγματικό χρόνο. Η αξιολόγηση της χρηστικότητας για μη έμπειρους χρήστες αποκάλυψε εξίσου ενθαρρυντικά αποτελέσματα. Οι οργανωτές ασκήσεων χωρίς εξειδικευμένη τεχνική κατάρτιση μπόρεσαν να δημιουργήσουν λειτουργικά σενάρια μετά από μόλις 2-3 ώρες εκπαίδευσης στη χρήση του εργαλείου, ενώ το 92% αυτών δήλωσαν ότι θα επέλεγαν το AiCEF έναντι παραδοσιακών μεθόδων για μελλοντικές ασκήσεις.

Πρακτικές Εφαρμογές και Μελλοντικές Προοπτικές

Το πλαίσιο AiCEF έχει εφαρμοστεί με επιτυχία σε πραγματικές περιπτώσεις χρήσης, συμπεριλαμβανομένης της μοντελοποίησης σύγχρονων κυβερνοσυγκρούσεων και της δημιουργίας σεναρίων για διάφορους τομείς κρίσιμων υποδομών. Η ικανότητα του συστήματος να προσαρμόζεται σε εξελισσόμενες απειλές και να ενσωματώνει νέα δεδομένα καθιστά το AiCEF ιδιαίτερα κατάλληλο για τις απαιτήσεις της σύγχρονης κυβερνοασφάλειας.

Η ανάπτυξη και παρουσίαση συνοδευτικού εργαλείου διευρύνει την προσβασιμότητα και την αναγνωρισιμότητα του πλαισίου στη διεθνή κοινότητα κυβερνοασφάλειας, ανοίγοντας νέους δρόμους για συνεργασία και περαιτέρω ανάπτυξη. Μελλοντικές κατευθύνσεις της έρευνας περιλαμβάνουν την επέκταση του πλαισίου AiCEF για την κάλυψη επιπλέον τομέων κρίσιμων υποδομών, τη βελτίωση των αλγορίθμων πρόβλεψης απειλών με τη χρήση προηγμένων τεχνικών deep learning, και την ενσωμάτωση μεθόδων διατήρησης τησ ιδοτικότητας αλλά και μεθόδων εξηγήσιμης τεχνητής νοημοσύνης για την ενίσχυση της διαφάνειας και εμπιστοσύνης των παραγόμενων σεναρίων.

Σύμφωνα με τα παραπάνω, η παρούσα διδακτορική διατριβή συνεισφέρει σημαντικά στον εκσυγχρονισμό και την αποδοτικοποίηση των διαδικασιών προετοιμασίας για κυβερνοαπειλές,

προσφέροντας ένα ολοκληρωμένο, επιστημονικά τεκμηριωμένο πλαίσιο που μπορεί να αυξήσει δραστικά την ετοιμότητα οργανισμών και κρατών έναντι των συνεχώς εξελισσόμενων κυβερνοκινδύνων.

Abstract

Crisis Management exercises are a key tool in achieving the objectives of the cybersecurity strategy of all European Union member states. The increasing number of cyberattacks targeting critical infrastructures has created the need for a continuous training framework to assess the readiness of entities managing critical infrastructures to respond promptly and accurately. The frequency and content of cybersecurity exercises must be able to simulate, as closely as possible, threats that could affect the actual operations of the targeted organisations.

Furthermore, cybersecurity exercises are an important tool for assessing the preparedness of participating entities and identifying weaknesses and vulnerabilities in their systems. By simulating security incidents, the opportunity is provided to address security incidents in conditions similar to real-life situations, applying relevant security measures and emergency plans already in place. This enables entities to make the necessary improvements and updates. Furthermore, these exercises promote information and knowledge exchange, enhance collaboration between participating entities, and strengthen a culture of cooperation, ultimately raising the level of cybersecurity for a country or organisation.

The purpose of this doctoral thesis is to introduce a novel methodology that automates and accelerates the generation of dynamic content for crisis management cyber exercises using machine learning and artificial intelligence techniques. This methodology will allow organisers of such exercises, with limited technical knowledge, to dynamically produce realistic scenarios by providing only minimal, high-level unstructured information such as: similar known attack incidents, articles, attack techniques or tools, and the type of critical infrastructure involved.

A trained artificial intelligence agent will be responsible for selecting the appropriate category of threats, attackers, and techniques, simulating a variety of attack paths that could be implemented based on the sector/category of the critical infrastructure under consideration. Ultimately, a structured technical exercise scenario will be produced, ready for execution, taking into account relevant current and future trends affecting that particular type of critical infrastructure.

Contents

I Introduction **1**

1 Context, Motivation &Methodology **3**

1.1 Problem Setting 4
1.2 Motivation 4
1.3 Relevance 5
1.4 Contributions of the Thesis 5
1.4.1 Beneficiaries 6
1.5 Literature Review 7
1.5.1 Cybersecurity Exercises 7
1.5.2 Relevant Ontologies,Taxonomies &Frameworks 8
1.5.3 Cybersecurity exercise scenarios and forecasting 9
1.5.4 Forecasting Models 10
1.5.5 AI and ML in Cyber Threat Intelligence 12
1.5.6 Cybersecurity Threat Landscapes 12
1.5.7 Tactics, Techniques, and Procedures Extraction 13
1.5.8 Gaps in the Current Literature 15
1.6 Proposed Framework 17
1.6.1 Main Modules 17
1.6.2 Support Modules &Optimizations 18
1.6.3 Evaluation &Feedback 18
1.7 Research Questions 19
1.8 Positioning of the Research &Sub Domain Focus 20
1.9 Data Sources 22

1.10 Tools . . . 23
1.11 Outline of the Thesis . . . 25
1.12 Acknowledgements . . . 26

2 Publications and Contributions **27**
2.1 Publications of the Compendium and Contributions . . . 27
2.2 Other Publications and Contributions . . . 32

3 Open Problems,Challenges &Future Work **35**
3.1 Open Problems and Challenges . . . 35
3.2 Specific Challenges for this Research . . . 36
3.3 Future Work . . . 36

II AiCEF: An AI-assisted Cyber Exercise Content Generation Framework 39

4 Building an AI-assisted Cyber Exercise Content Generation Framework **41**
4.1 Problem setting and objectives . . . 41
4.2 Framework Overview . . . 44
4.3 Cyber Exercise Scenario Ontology (CESO) . . . 46
4.3.1 Scope of the Ontology . . . 47
4.3.2 Ontologies/taxonomies to be (re)used . . . 48
4.3.3 Scenario Augmented Model . . . 48
4.3.4 Implementing the Ontology . . . 50
4.4 Layer 1: MLCESO &KDb . . . 52
4.4.1 Creating MLCESO pipeline . . . 52
4.4.1.1 Corpus Building . . . 54
4.4.1.2 Corpus Annotation . . . 55
4.4.1.3 Training &Evaluation Using NER . . . 59

4.4.1.4 Knowledge DB (KDb) 61
4.5 Layer 2: Trend analysis &prediction 64
4.5.1 Trend Prediction Module (MLTP) 64
4.6 Layer 3: Exercise generation 66
4.6.1 Incident Generation (IncGen) 66
4.6.2 APT Enhancer 67
4.6.3 Cyber Exercise Generation (CEGEN) 68
4.6.4 Storyline Text Generation 69
4.6.5 Example Use Case 69

5 Evaluation Methodology &Results **77**
5.1 Scenario Content Generation 78
5.2 Scenario Content Evaluation 79
5.2.1 Results Analysis 80

6 Advanced use-case: Modelling a Real Cyber Conflict **89**
6.1 Measuring Effectiveness 89
6.1.1 PHASE 1: Incident Modelling (Pre-Conflict) 90
6.1.2 PHASE 2: Incident Modelling (Post-Conflict) 90
6.1.3 PHASE 3: Comparison 91
6.2 Use-case Implementation 92
6.3 Results &Analysis 93
6.3.1 Public Administration 93
6.3.2 Energy Sector 94
6.3.3 ICT Sector 95
6.3.4 Multi-Sector Exercise Coverage 96
6.4 Findings 97

III Optimizing &Integrating AI-driven Threat Intelligence and Forecasting in the Cyber Security Exercise Content Generation Lifecycle 99

7 Integrating AI-driven Threat Intelligence &Forecasting in the Cybersecurity Exercise Content Generation Lifecycle 101

7.1 Proposed Pipeline . . . 102
7.1.1 Overview . . . 102
7.1.2 Dataset and data preprocessing . . . 104
7.1.3 Text Annotation . . . 104
7.1.4 N-gram Analysis . . . 105
7.1.5 Clustering through Topic Modelling . . . 108
7.1.5.1 Extracted Topics . . . 109
7.1.6 Forecasting, Trend Prediction and Model Evaluation . . . 111
7.1.6.1 Overall Forecasting Results . . . 112
7.2 AI-driven Prompt Generation (Text Synthesis) . . . 113
7.3 Evaluation and Results . . . 116
7.3.1 Evaluation . . . 116
7.3.2 Prompt generation . . . 117
7.3.3 Prompt Evaluation . . . 118
7.3.3.1 Prompt Maturity . . . 118
7.3.3.2 Tag Coverage . . . 119
7.3.4 Trend Forecasting Evaluation . . . 121
7.4 Promising Findings . . . 122

8 Optimising AI Models for Intelligence Extraction in the Life Cycle of Cybersecurity Threat Landscape Generation 133

8.1 Proposed AI Model Comparison Methodology . . . 134
8.1.1 Dataset Selection . . . 135

8.1.2 Dataset Pre-processing . . . 136
8.2 Human Annotation . . . 137
8.3 Model Selection and AI Annotation . . . 139
8.4 Evaluation Metrics &Comparative Analysis . . . 140
8.4.1 Attribute-Specific Insights . . . 141
8.4.1.1 Threat Name Identification . . . 141
8.4.1.2 Motivation Identification . . . 142
8.4.1.3 Attack Type Classification . . . 143
8.4.1.4 Sector Identification . . . 144
8.4.1.5 MITRE TTP Identification . . . 144
8.4.2 Key Performance Findings . . . 145
8.4.3 Resource Comparison Findings . . . 146
8.4.4 Proposed Annotation Workflow . . . 147
8.5 Optimisation through Voting-Based Agreement Methodology . . 148
8.5.1 Simple Majority Voting . . . 148
8.5.2 Weighted Voting . . . 150
8.5.3 Ensemble learning: Stacking Model . . . 150
8.5.4 Voting Results Comparison . . . 152
8.6 Findings . . . 153

9 AiCEF Automation **159**
9.1 Key Features of AiCEF tool . . . 159
9.2 Role of AiCEF in Supporting Research . . . 160

IV Closure **163**

10 Conclusions and Future Work **165**

10.1 Conclusions 166
10.2 Limitations 170
10.3 Future Work 171
10.4 Closing Remarks 174

Appendix ANNEX I - Glossary of Terms **175**

Appendices **175**

Appendix ANNEX II - Useful Code **177**

.1 CESO Ontology Implementation 177
.1.1 Ontology Metadata 177
.1.2 Implementation Details 177
.1.3 Validation and Usage 183
.2 AiCEF Generated Exercise 184

Bibliography **203**

V Publications **217**

List of Tables

1.1 State of the art on MITRE TTP extraction. Notation: A: Accuracy, P: Precision, and R: Recall. 15

1.2 Summary of Datasets used. 22

1.3 Composition of the Main Corpus dataset. 23

4.1 CSE Components to STIX 2.1 objects mapping. 51

4.2 Relationships matrix. 52

4.3 Objects extension matrix. 53

4.4 Training Corpus collection count. 54

4.5 Annotation tag schema with CESO/STIX 2.1 mappings. 57

4.6 Comparative annotation results demonstrating tag assignment consistency. . 58

4.7 Inter-annotator agreement contingency matrix. 58

4.8 NER model performance metrics across entity categories. 60

4.9 Expert evaluation comparison with automated metrics (H: Hit, P: Partial, M: Miss). 61

4.10 Knowledge DB content per source 62

4.11 Annotation example using AiCEF (MLCESO). 72

4.12 Summary of the EnergyPhish Incident 73

4.13 Training topics. 74

5.1 Details of the generated scenarios. 79

5.2 Demographics of the experts. 85

5.3 Scenario Evaluation and Expert Feedback 86

5.4 Turing test results for scenario generation methodology identification. . . . 87

6.1 Sector based Attack Types, Techniques, and Threat Actors 93

7.1 Primary Sector Clusters - Heavily Impacted Industries 106
7.2 Secondary Sector Clusters - Moderately Impacted Industries 107
7.3 Topics and Content 109
7.4 Trend forecasting results for the 2nd Semester of 2022 115
7.5 Sectorial trend forecasting results for 2022. 128
7.6 Demographics of the experts. 129
7.7 Maturity scores 130
7.8 Sectorial trend forecasting results for 2023. 130
7.9 Sectorial forecast results per topic for the year 2023 132

8.1 Cohen's Kappa consistency matrix 138
8.2 Cohen's Kappa values 138
8.3 Classification values per tag. 139
8.4 Model usage per attribute extracted 140
8.5 Performance Metrics for Threat Name Detection 142
8.6 Attacker name True Positive Hit count excluding "UNKNOWN" 142
8.7 Performance Metrics for Motivation Detection 143
8.8 Performance Metrics for Attack Type Detection 143
8.9 Performance Metrics for Sector Detection 144
8.10 Performance Metrics for MITRE TTPs 145
8.11 Annotation resources per task breakdown 146

8.12 Optimisation Methods Comparison: Simple Majority Voting vs. Weighted Voting vs. Stacking Model 153

List of Figures

1.1 Overview of the Key Contributions 20

4.1 Cyber Exercise Structure according to ISO 22398:2013 [37]. 43

4.2 Sequential workflow of AiCEF module interactions and data processing pipeline. 45

4.3 Three-layer architecture of AiCEF showing component relationships and data flow. 45

4.4 The informational (left) and operational (right) layers of CESO. 49

4.5 CESO implementation in OWL, visualised in Protégé. 53

4.6 The MLCESO pipeline. 54

4.7 Incident Coprus generation 55

4.8 Annotation example (using Prodigy UI). 56

4.9 LDA topic visualization and top 30 terms. 63

4.10 Incident Generation report and visualisation. 65

4.11 Training Topics/Objective per month for 2021. 65

4.12 Six (6) month forecast using SARIMA 66

4.13 The workflow of Incident Generation (IncGen). 66

4.14 Incident graph visualisation (STIX 2.1). 67

4.15 Cobalt Group STIX 2.1 graph visualisation 68

4.16 CEGen workflow 69

4.17 IncGen execution flow with intermediate representation steps. 71

4.18 Sample text of an AI-generated exercise. 75

4.19 Sample exercise graph visualisation 76

5.1 Task definition 78

5.2 Overall performance of evaluated scenarios based on total score. 81

5.3 Scenario evaluation parameters 82

5.4 Score range for the Q1-8 of 16 Evaluators. 83

5.5 Expert feedback on exercise planning methodologies. 84

6.1 Scenario evaluation parameters 91

6.2 Public Administration Incidents Coverage Results. 94

6.3 Energy Sector Incidents Coverage Results. 95

6.4 ICT Sector Incidents Coverage Results. 96

6.5 Multi-Sector Exercise Coverage Results. 97

7.1 Proposed trend extraction and prompt generation pipeline overview. 103

7.2 Top 10 sectors impacted between 2019 and 2022 in volume 107

7.3 CV coherence score for different numbers of topics 108

7.4 LDA Topic visualisation example. 111

7.5 Cybersecurity incidents 2019-2022 112

7.6 The form of the template above and an example prompt, below. 124

7.7 STIX 2.1 output generated by AiCEF 125

7.8 Cybersecurity incidents prediction results (2nd Semester 2022) 126

7.9 R^2 Scores (Training, Testing Dataset) per Forecasting Model. 127

7.10 MAE, MAPE, and MAD score per forecasting model 127

7.11 Task definition 129

7.12 Unique annotation tags values per energy landscape report 130
7.13 Unique annotation tag volume comparison 131
7.14 Unique tags of STLs versus Human Experts and AiDPG 131
7.15 Benefit percentage per tag and expert group 132

8.1 Overview of ENISA CTL methodology, source [106]. 134
8.2 Overview of the proposed comparison methodology 136
8.3 F1 Score Comparison on Attribute Detection per Model (Spider Diagram) . 141
8.4 Motivation (%) Model Hit Rate Comparison 155
8.5 True Positive (TP) Comparison for Attack Type 156
8.6 True Positive Hit Rates (%) breakdown by Model per Sector 157
8.7 Proposed hybrid annotation methodology 157
8.8 Distribution of Weights achieving maximum F1 Score (0.5958) in MITRE TTP extraction. 158

9.1 Screenshots of AiCEF Tool Interface 161

Part I

Introduction

Chapter 1

Context, Motivation & Methodology

The rapid evolution of cyber threats poses a significant challenge to organisations tasked with managing critical infrastructure globally. With the increasing reliance on interconnected systems, the consequences of cyberattacks have become more severe, often threatening public safety, economic stability, and national security. As highlighted in the Cybersecurity Strategy of the European Union [1], cybersecurity exercises (CSEs) play an essential role in assessing the readiness of entities responsible for safeguarding critical infrastructures. In line with this strategic focus, the EU has implemented the NIS2 Directive [2], which mandates that Member States designate competent authorities to manage large-scale cybersecurity incidents and crises. These authorities are required to develop national large-scale cybersecurity incident and crisis response plans, incorporating objectives for preparedness and outlining the tasks and responsibilities of cyber crisis management authorities.

Such exercises allow organisations to simulate real-world security incidents, evaluate their current defensive measures, and identify vulnerabilities under realistic conditions. However, traditional methods of designing and generating cyber exercises are time-consuming, require significant expertise, and often lack the adaptability to reflect the dynamic and sector-specific threat landscape. These limitations hinder the ability of organisations to stay prepared for evolving cybersecurity challenges.

1.1 Problem Setting

To respond to the growing complexity of cyber threats, this thesis focuses on the challenge of automating and improving the generation of CSE scenarios. The core issue is the ability to dynamically create realistic and structured scenarios that accurately reflect:

- The current and future threat landscape of a sector, including threat actors and their corresponding Tactics, Techniques, and Procedures (TTPs).
- Realistic incidents and scripted events that mimic potential attacks against organisations within the sectors defined by the NIS2 Directive.
- Detailed artefacts and injects that align with these scenarios to aid participants in understanding and addressing the simulated threats effectively.

Traditional methods heavily depend on human expertise, which can limit scalability, increase costs, and result in scenarios that may not fully represent emerging threats. Additionally, the lack of standardisation in exercise generation often results in inconsistencies in scenario quality, particularly when developed by less experienced exercise planners.

1.2 Motivation

The motivation for this research arises from the urgent need to enhance the preparedness of organisations managing critical infrastructures, with key drivers being the increasing frequency of cyberattacks targeting essential services which necessitates more effective and timely training mechanisms; the need for a structured framework that non-expert planners can use to create high-quality scenarios, reducing dependence on human expertise; and the opportunity to harness advances in machine learning (ML) and artificial intelligence (AI) to introduce automation, accuracy, and adaptability in exercise generation. By addressing these challenges, this research aims to bridge the gap between traditional, resource-intensive methods and the demand for scalable, adaptive solutions, offering measurable improvements in the speed, quality, and relevance of generated scenarios. Current methods for conducting Cybersecurity Exercises (CSEs) are often resource-intensive, time-consuming, and dependent on a limited pool of cybersecurity experts, and traditional methods of scenario generation may

not always capture the most relevant or emerging threats, leading to a gap in the realism and effectiveness of these exercises. Addressing this gap is critical to enhancing the resilience of organisations and ensuring they are prepared to respond effectively to the dynamic threat landscape, with this thesis aiming to introduce a solution that enhances the relevance, efficiency, and scalability of cybersecurity preparedness efforts in critical sectors.

1.3 Relevance

1.4 Contributions of the Thesis

This thesis makes several significant contributions to the field of cybersecurity exercise preparation across three distinct domains: CSE Generation, CTI & Forecasting, and CTL Generation.

In the area of CSE generation, the research develops an AI-powered exercise generation framework that automates the creation of Cybersecurity Exercises (CSEs), producing realistic, sector-specific scenarios based on contemporary threat landscapes. This includes the first incident-centric CSE ontology using a bottom-up approach, methods for enhancing exercises with known APT activity patterns, and a comprehensive comparison between AI/ML and human-powered CSE generation approaches.

For CTI & forecasting, the work advances the integration of AI-driven cyber threat intelligence and forecasting mechanisms into the exercise generation process, substantially enhancing the realism and relevance of the generated exercises. The research evaluates multiple forecasting algorithms to identify those providing the most accurate predictions for sector-specific datasets and introduces a novel methodology for fully automating prompt generation to assist non-expert exercise planners.

In CTL generation, the thesis optimizes AI models specifically for intelligence extraction within the cybersecurity threat landscape generation lifecycle, thereby improving the accuracy and efficiency of extracting actionable insights from diverse data sources. This includes a rigorous comparative analysis of state-of-the-art LLMs and BERT-based models, as well as the introduction of a novel Voting-Based Agreement technique to improve the precision of AI models in extracting challenging attributes such as MITRE TTPs.

The framework introduces improved CSE generation metrics designed to enhance both the speed and quality of exercise creation, with particular emphasis on minimizing dependency on experienced exercise planners while maximizing the relevance of scenarios to real-world threat environments. The research provides a comprehensive comparative evaluation between the proposed AI-driven methodology and traditional human-based approaches, demonstrating the substantial advantages of an automated process and addressing critical gaps in existing cybersecurity exercise preparation practices.

1.4.1 Beneficiaries

Based on the above, the primary beneficiaries of this research include, but are not limited to:

- **Cybersecurity Professionals**: Both experienced and less experienced exercise planners (EPs) will benefit from the proposed ML-powered framework, which reduces reliance on deep expertise while still producing high-quality and relevant scenarios. The framework will also improve the speed of scenario generation, allowing experts to focus on more strategic aspects of such activities.

- **Organisations Managing Critical Infrastructures**: Entities responsible for sectors defined by the NIS2 Directive, such as energy, transport, healthcare, and finance, will benefit from more realistic, timely, and targeted cybersecurity exercises that help them prepare for current and emerging threats.

- **Government Agencies and Policymakers**: The ability to generate sector-specific, AI-driven threat simulations will aid policymakers in designing effective cybersecurity strategies and assessing the resilience of critical sectors to various types of cyber incidents. This is particularly important for ensuring national security and aligning with cybersecurity policies such as the NIS2 Directive.

- **Training and Certification Bodies**: Organisations offering cybersecurity training programs will find this methodology useful for creating realistic, dynamic, and up-to-date exercises that reflect the real-world threat environment. This contributes to better-trained personnel and improved cybersecurity awareness across sectors.

1.5 Literature Review

1.5.1 Cybersecurity Exercises

Cybersecurity Exercises (CSEs), also referred to as Cyber Defense Exercises (CDX) or Crisis Management Exercises (CMEs), have proven to be an effective approach for implementing engaging security awareness training [3, 4]. These exercises are recognized as a highly effective method for delivering an ultimate learning experience [5], helping both individuals and teams with varying expertise to enhance a wide range of information security-related skills. Through exercising, organizations can also identify gaps in their security policies, procedures, and resources [6, 7], which subsequently leads to improvements in awareness training, tools, and policies.

Research in the CSE domain [8] has emphasized the effectiveness of cyber defence competitions or live-attack exercises as superior methods for teaching information security [9, 10]. These approaches assist teams in designing, implementing, managing, and defending computer networks [11, 12, 13, 14, 15]. These findings receive further support from Vigna [16] and Mink [17]. Additional research has examined cyber defence competitions [18, 19] and the most appropriate architecture [20], while Green et al. [21] described tools and techniques for creating active learning experiences. Patriciu and Furtuna [22] outlined several steps and guidelines for CSE design. White [23] presented an alternative approach to live CSEs, sharing lessons learned and offering suggestions to help organizations conduct their own exercises. The literature also includes examinations of how to run CSEs using a service provider model [24].

CSEs serve multiple purposes beyond training: they can generate scientifically valuable datasets for future security research [25, 26] and help uncover hidden risks stemming from weak security policies and procedures [27]. Additionally, CSEs can measure performance against specific standards [28] or evaluate team effectiveness using behavioral assessment techniques [29]. Researchers have conducted experiments using various platforms, such as the RINSE simulator [30] or realistic inter-domain routing experiment platforms [31], to render network behavior.

Addressing the human aspect more specifically, Tobey [32] described Job Performance Modelling (JPM), which utilizes vignettes to enhance cybersecurity talent management through

cyber defence competition design.

The success of a CSE depends heavily on employing a robust scenario. Exercise scenarios must portray worst-case situations that participants can relate to and that are realistic enough to foster seamless engagement. Intuitive scenarios function as powerful tools that can predict future states or situations [5, 3], incorporating issues to be resolved, interactions, and consequences [21, 29] that lead to constructive training experiences.

An exercise scenario provides a sequential, narrative account of a hypothetical incident that catalyzes the exercise and introduces situations designed to inspire responses that demonstrate the exercise objectives [20]. In CSEs, a scenario defines the training environment that guides participants toward fulfilling these objectives [33]. The cybersecurity problem described in a scenario is structured as a Master Scenario Events List (MSEL), which serves as the script for executing the exercise [20]. Although CSE scenario formats may vary [34], the two most common types are outlined scenarios, which provide a general summary of an event's impact on assets [35], and detailed scenarios, which contain exhaustive information sequentially describing the event's impact on specific services or sections of an organization, along with a timeline for restoring key functions [36].

1.5.2 Relevant Ontologies,Taxonomies & Frameworks

A comprehensive review of existing ontologies, taxonomies, frameworks, standards, and formats has been conducted, focusing on their relevance to cybersecurity, with particular emphasis on the representation of key elements within Crisis Management Cyber Exercises (CSEs). This exploration centred on the fundamental building blocks of such exercises, particularly the incidents that are to be simulated during the exercises. The research identified that while a variety of frameworks exist, none fully encapsulates the complexity and dynamic nature of cyber incidents in the context of CSEs. Based on this evaluation, this research concluded that a combination of several established models offers the necessary structure and coverage to develop realistic and comprehensive cyber exercise scenarios. Specifically, the integration of ISO 22398:2013 [37], which outlines guidelines for CSEs, provides a solid foundation for the simulation process. Additionally, the MITRE ATT&CK framework [38], which offers a detailed taxonomy of adversarial tactics and techniques, is invaluable to simulate advanced persistent threats and attack strategies. The Cyber Kill Chain [39] provides a structured view of the stages of a cyber attack, further supporting the simulation of attack sequences. Finally,

MITRE CVE [40], with its extensive repository of known vulnerabilities, enhances the realism of the exercise by enabling the integration of specific vulnerabilities that can be exploited in a simulated attack. Together, these resources form a cohesive set of tools that enable the creation of detailed, realistic, and contextually relevant CSEs, addressing both technical and operational aspects of cyber crisis management.

1.5.3 Cybersecurity exercise scenarios and forecasting

The integration of AI-driven techniques into CSEs has gained significant traction in recent years as organisations seek to enhance their preparedness for emerging threats, enhancing traditional CSE scenario-building methodologies. Successful CSEs are dependent on the effectiveness of their scenarios, which must simulate real-world situations that resonate with participants while maintaining a level of realism that fosters engagement and learning. ENISA acknowledges the benefits of sectorial exercises scenario in their National Exercise - Good Practice Guide [41], an approach recognized by national cybersecurity authorities like ANSSI in their guide on organising CSEs [42].

To generate a useful Master Scenario Events List (MSEL), one needs to successfully detect and classify cyber attacks relevant to the needs of the trainee, as emphasised in [43, 44, 45]. Interestingly, applying advanced ML techniques to predict future cyberattack trends has gained significant momentum in cybersecurity. Long Short-Term Memory (LSTM) networks constitute a specific type of recurrent neural network that has been particularly noted for their ability to learn long-term dependencies and patterns in time-series data, which is characteristic of cyber attack vectors. For instance, [46] demonstrated the efficacy of LSTMs in classifying intrusion detection patterns, emphasising their potential to identify sophisticated cyber threats by analysing historical data patterns. Similar studies to ours can be considered in the cases of [47, 48, 49, 50], where LSTM networks are proposed as a method to forecast cyber incidents with a focus on malware infections.

Parallel to the application of LSTMs, the Neural Prophet model has emerged as a novel approach for time-series forecasting [51]. Although specific studies focusing on the use of this model in cyber attack predictions are scarce, its robustness in detecting complex patterns is essential to predict cyber incidents, as reported in [52] with promising results. In the realm of more traditional statistical methods, ARIMA models have been a mainstay for time series forecasting and have found applications in cyber-security. Sarker et al. [53] discuss integrat-

ing ML approaches like ARIMA in cyber-security, highlighting its effectiveness, especially in scenarios with limited computational resources or where model interpretability is vital. Following the same path, data mining and ML techniques can be applied to CSE scenario development to help develop a mitigation plan for future incident trends and cybersecurity incident timelines as proposed in [54].

Further to analysing existing threats, modelling [55], predicting upcoming cyber incidents [56, 57], and the overall threat trends are equally important. Several studies focus on the short-term prediction of the number or source of attacks to be expected in the following hours or days [58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 56, 70, 71, 72]. Most of these works make predictions in restricted settings; e.g., against a specific entity or organisation for which historical data are available [64, 65, 70]. Forecasting attack occurrences has been attempted with statistical methods, especially when parametric data distributions could be assumed [63, 64], as well as by using ML models [67] and deep learning [73]. Other methods adopt a Bayesian setting and build event graphs suitable to estimate the conditional probability of an attack following a given chain of events [68]. Such techniques rely on libraries of predefined attack graphs to identify the known attacks that are most likely to occur. Research has also covered zero-day attacks [74, 75] with less impressive results.

Almahmoud et al. [61] introduced a machine learning-based proactive approach for long-term cyber-attack prediction, allowing early detection of potential threats. This approach could facilitate cybersecurity experts in prioritising resources and mitigation measures, including training through CSEs, a type of proactive cyber defence [76].

The field of Cyber Threat Landscape (CTL) generation has seen significant advances over the years as a response to the increasing complexity of cyber threats and the need for more efficient and accurate threat analysis [77, 78, 79].

1.5.4 Forecasting Models

Although several forecasting models are available, the primary models considered within the scope of this work, along with their respective limitations, are as follows. SARIMAX (Seasonal Auto Regressive Integrated Moving Average with Exogenous Variables) [80] is a popular statistical model for time-series forecasting. It extends the ARIMA (AutoRegressive Integrated Moving Average) [80] model by incorporating seasonal patterns and exogenous variables. Unlike ARIMA, which effectively analyses and predicts time series data with sta-

tionary or non-stationary properties, SARIMAX captures the relationships between past observations, forecast errors, and external factors to provide more accurate forecasts. However, SARIMAX models assume linearity and stationarity in the data, which may not always be true in real-world scenarios. SARIMAX models do not handle missing data or outliers well, which requires additional preprocessing steps. Interpreting the coefficients and understanding the impact of exogenous variables can be complex.

Exponential Smoothing (ES) [81, 82] is a widely used time series forecasting technique that assigns exponentially decreasing weights to past observations. It is based on the assumption that recent observations have greater significance in predicting future values. Yet, ES may struggle to capture complex non-linear patterns in the data.

Simple Exponential Smoothing (SES) [83] is a specific form of exponential smoothing that only considers the current observation and a single smoothing parameter (alpha). It provides a straightforward method for forecasting by exponentially weighting past observations and updating the forecast based on the weighted average. Nonetheless, SES assumes a constant level and does not capture trends or seasonality, making it less suitable for data with such patterns. Moreover, it may struggle with data that have irregular or changing patterns over time and is more suitable for short-term forecasting rather than long-term predictions.

Prophet [51] is a forecasting framework developed by Facebook's Core Data Science team. It combines the components of additive regression models with flexible seasonality estimation. Prophet incorporates trend components, seasonal effects, and additional regressors to generate forecasts. It also effectively handles outliers and missing data points. However, Prophet assumes that trends and seasonality are additive and follow specific patterns, which may not always hold in diverse datasets. The automatic change-point detection in Prophet might not always accurately identify complex or abrupt changes.

Long Short-Term Memory (LSTM) [84, 85] is a type of recurrent neural network (RNN) architecture commonly used to model and forecast time series data. LSTMs are designed to overcome the limitations of traditional RNNs in capturing long-term dependencies and handling vanishing or exploding gradients. In LSTM models, the network learns to selectively retain or forget information from past observations, allowing it to capture and retain important patterns over longer sequences. On the other hand, training LSTM models requires more data to capture long-term dependencies and avoid over-fitting effectively. LSTM models can be computationally intensive. Hyperparameter tuning for LSTM models can be challenging, as the optimal architecture and parameter settings may vary depending on the specific dataset

and problem.

1.5.5 AI and ML in Cyber Threat Intelligence

Natural Language Processing (NLP) techniques have been extensively applied in CTI analysis, as demonstrated by methods such as Casie [86], Extractor [87], Open-CyKG [88], SecIE [89], and CyberEntRel [90]. These methods leverage advanced NLP models (e.g., BiLSTM, BERT, RoBERTa) to extract actionable insights from text. However, adapting these models to the cyber threat domain requires substantial preprocessing and fine-tuning efforts.

Previous studies have achieved the extraction of cyber-related attributes from CTI reports [91, 92, 93, 94, 95, 96] using NLP to formulate actionable outputs with utilisation in a variety of cybersecurity domains. LLMs have been proven to successfully identify emerging cybersecurity threats, vulnerabilities, and attack trends and enhance existing threat intelligence pipelines [97] as well as detect secrets in source code [98]. Phishing detection and analysis have also been studied in the context of LLMs [99, 100, 101] including deobfuscation of malicious payloads [102].

The application of NLP and LLMs in cyber incident analysis has shown significant promise in enhancing the efficiency and effectiveness of the cyber threat detection process [103], [104]. It enabled analysts to extract, understand, and communicate incident data more comprehensively from both network threat intelligence [103] and social media sources [104]. An innovative GPT-based open-source intelligence pipeline has been developed by Sufi [105], using historical cyber incident reports to extract a seven-dimensional intelligence. The GPT-driven analyses extracted attributes such as Actor Type, Target, Attack Source (i.e., the country from which the attack originated), Attack Destination (i.e., the targeted country), Attack Level, Attack Type, and Attack Timeline.

1.5.6 Cybersecurity Threat Landscapes

Various commercial threat landscapes have been published annually over the last years, focusing on different topics. Threat intelligence and nation-state activity, reports such as CrowdStrike's Global Threat Report[1] and Microsoft's Digital Defense Report[2] provide deep in-

[1] https://www.crowdstrike.com/global-threat-report/
[2] https://www.microsoft.com/en-us/security/business/security-intelligence-report

sights into APTs, cyber espionage, and nation-state-sponsored attacks. Reports that focus on breach analysis and incident patterns, such as the Verizon Data Breach Investigations Report (DBIR)[3], IBM X-Force Threat Intelligence Index[4], and Cisco's Annual Cybersecurity Report[5], offer detailed breakdowns of data breaches, attack vectors, and industry-specific insights.

ENISA publishes an annual Threat Landscape report[6] that identifies and assesses emerging and evolving cybersecurity threats. This report provides an analysis of trends based on observed incidents and developments. In addition, ENISA has made public its own methodology [106], providing structured steps for CTL generation, emphasising the systematic identification, collection, and analysis of electronically stored information (ESI) to assess cyber risks. While robust, this methodology relies heavily on manual processes, which can be rather time-consuming and prone to human errors when applied to real-world scenarios. CTI has emerged as a pivotal aspect of cybersecurity practices, aimed at understanding and anticipating cyber threats for proactive defence [107, 108, 109, 110]. This field encompasses collecting, analysing, and disseminating information about current or potential attacks threatening information systems security [111]. By leveraging CTI, organisations are equipped to identify specific threats and vulnerabilities, enabling them to tailor their security measures more effectively. This intelligence-gathering process draws from a wide array of sources, including historical cyber attack data [112], hacker forums, malware samples, and more, to provide a comprehensive view of the cyber threat landscape, which helps in enhancing the security posture against known threats and aids in predicting and mitigating future cyber attacks.

1.5.7 Tactics, Techniques, and Procedures Extraction

There are various approaches to facilitate the extraction of actionable threat intelligence from reports. For instance, a priority is the accurate extraction of IOCs. To this end, iACE [113] focuses on extracting IOCs, and to achieve this, it uses regular expressions and a custom NER model. Going a step further, ChainSmith [114] classifies the IOCs into the corresponding attack phase, while TIMiner [115] uses word embeddings and syntactic dependencies to

[3] https://www.verizon.com/business/resources/reports/dbir/
[4] https://www.ibm.com/reports/threat-intelligence
[5] https://www.cisco.com/c/en/us/products/security/cybersecurity-reports.html
[6] The latest is published here https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024

extract them and prioritise them according to sector impact.

Moving higher to Bianco's pyramid of pain, the goal is to pass from simple IOCs, e.g., hashes, URLs, and domains, to the tooling and the methods that threat actors use. Given that MITRE's ATT&CK framework [38] has become a common ground for most practitioners and researchers, the goal is to extract TTPs from threat intelligence reports. Indeed, there is a constant improvement in identifying new methods such as ontology, graphs, and keyword matching powered by AI models of various sizes, including LLMs. TTPDrill [116] uses an ontology-based approach with the BM25 weighting method [117] to classify threat actions according to known attack patterns. Both TTPDrill and THREATRAPTOR[118] adopt an unsupervised NLP pipeline based on rule-based and information retrieval techniques. RcATT [119], a machine learning tool, applies Term Frequency–Inverse Document Frequency (TF-IDF) for classification but struggles with word context and synonyms. AttacKG [120] uses graph template matching but faces issues with adjectives describing threats. LADDER [121] enhances threat intelligence by extracting attack patterns with a three-step process. TTPHunter [122] utilises BERT and RoBERTa for contextual understanding but is limited by small datasets. MITRE's TRAM [123] uses SciBERT to improve TTP extraction with TTPXHunter further [124] being one of the latest state-of-the-art models using Domain-specific SecureBERT with Linear Classifier and showcasing improved extraction results.

Purba and Chu [104] and Siracusano et al. [110] utilized GPT-3.5 for tasks ranging from extracting Indicators of Compromise (IoCs) to generating structured CTI formats (e.g., STIX). Liu and Zhan [107] employed ChatGPT to create graphical representations of CTI data. Further advancements were made by Hu et al. [125] and Fengrui and Du [126], who explored the use of pretrained and fine-tuned LLM models. They leveraged GPT-3.5 and ChatGPT for data annotation and augmentation, respectively, to prepare datasets for fine-tuning the LLaMA2-7B model. Specifically, Hu et al. [127] utilised the fine-tuned LLaMA2-7B to construct knowledge graphs, while Fengrui and Du [126] focused on TTP classification. CyNer [128] and TriCTI [129] developed a custom web crawler to retrieve unstructured text from selected trusted websites and manually annotated a subset for evaluation purposes. LLM-TIKG [130] also developed a custom web crawler to collect intelligence from selected platforms, but this study differs from CyNer in that it utilizes an LLM (GPT) for annotation. Finally, LLMCloudHunter [131] harnesses the power of LLMs to perform automated extraction of sigma rules from Cloud-Based CTI, incorporating a TTP extraction module.

Article	Year	Extraction Technique	TTP Extraction	Metrics[7]	Dataset
TTPDrill [116]	2017	TF-IDF, Ontology-based, and improved BM25 similarity rank. Unsupervised NLP	MITRE TTPs	P:0.84 R:0.82	Symantec
Casie [86]	2020	BiLSTM	MITRE TTPs	P:0.77 R:0.77 F1:0.77	CyberWire
rcATT [119]	2020	TF-IDF and ML Models (KNN, Decision Tree, Random Forest, and Extra Tree)	MITRE TTPs	P:0.30 R:0.44 F1:0.30	MITRE Enterprise attack
ThreatRaptor [118]	2021	Unsupervised NLP	Other TTPs	P:0.96 R:0.97 F1:0.96	DARPA TC
TriCTI [129]	2022	Unsupervised NLP using web crawler	Other TTPs	P:0.87 R:0.87 F1:0.87	Labeled IoCs
AttacKG[120]	2022	Entity-based dependency graph and Graph-alignment algorithm	Other TTPs	P:0.782 R:0.860 F1:0.819	DARPA TC
TTPHunter[122]	2023	BERT/RoBERTa followed by Linear Classifier	MITRE TTPs (TOP 50)	P:0.89 R:0.88 F1:0.88	Custom
LADDER[121]	2023	Extract attack phases using a sequence tagging model and map these patterns to TTP using cosine similarity	MITRE TTPs	P:0.65 R:0.63 F1:0.64	Custom
TRAM [123]	2023	SciBERT followed by Linear Classifier	MITRE TTPs (TOP 50)	P:0.94 R:0.94 F1:0.93	Custom
LLM-TIKG [130]	2023	Fine-tuned LLaMA-2-7B	MITRE TTPs	A:0.97 P:0.96 R:0.99 F1:0.98	Custom
TTPXHunter [124]	2024	Domain-specific SecureBERT with Linear Classifier	MITRE TTPs	P:0.97 R:0.96 F1:0.97	Custom
Fengrui and Du [126]	2024	Fine-tuned LLaMA-2-7B	MITRE TTPs	P:0.86 R:0.89 F1:0.87	MITRE ATT&CK STIX
LLMCloudHunter [131]	2024	GPT-4.o	MITRE TTPs	P:0.66 R:0.74 F1:0.69	Custom

Table 1.1: State of the art on MITRE TTP extraction. Notation: A: Accuracy, P: Precision, and R: Recall.

1.5.8 Gaps in the Current Literature

Despite the established maturity and widespread adoption of cyber exercises in the contemporary cybersecurity landscape, significant gaps persist in the literature concerning the impact of AI/ML integration. This thesis aims to fill in the following gaps:

- **Limited use of ML in Exercise Generation:** Recent trends in attack recognition use AI, ML, and NLP tools and techniques to empower their efficiency. However, there needs to be a dedicated methodology that focuses only on CSE scenario generation.

There is a need for a methodologically built and annotated CE corpus that could train multiple algorithms for Cyber Exercise elements. Such a corpus should focus on the syntactic and semantic characteristics of the cyber exercise components and broaden the understanding of the malicious patterns used in cyber incidents that can be reused for CSE material. A similar approach to the one used in building and evaluating an annotated Corpus for automated Recognition attacks has been utilized [132], only this time to extract CSE relevant objects.

- **Lack of AI/ML friendly CSE ontology:** The current literature does not cover the most crucial topic of structuring a CSE using AI, which is a compatible ontology that enables easy annotation and extraction using ML models. This ontology should be easily adaptable to the current evolution and capabilities of modern AI/ML models.

- **Context-Specific and Sectoral Threats:** Existing literature tends to focus on generic cyberattack scenarios that may not fully capture the unique threats faced by specific sectors, such as energy, healthcare, or transportation. The existing literature showcases the potential for AI-driven topic extraction and forecasting techniques to be applied in the identification and prioritisation of both current and future threats in the context of CSEs, minimising their impact upon materialisation [71, 133].

- **AI/ML powered generation pipeline:** the lack of automated and systematically evaluated approaches for generating a CSE while incorporating relevant Cyber Threat Landscape (CTL) reports using ML/AI models is entirely absent. Traditional methods rely heavily on manual work, which is resource-intensive and prone to inconsistencies. Moreover, while various AI/ML models exist for the extraction of cybersecurity threat intelligence, it is still unclear how closely they compete with the human analyst capabilities regarding CSE and CTL generation. These challenges significantly hinder the timely and cost-effective generation of actionable CSEs, a critical need in today's rapidly evolving threat landscape.

- **Measuring the Effectiveness of AI-Driven Exercises:** While AI-based systems are increasingly used for cybersecurity, there is a lack of comprehensive studies evaluating their impact on the effectiveness of CSE generation. Studies on the effectiveness of AI-generated exercises are scarce, particularly in terms of how these exercises improve both the speed and quality of CSE generation, especially for organisations with limited cybersecurity expertise.

1.6 Proposed Framework

AI-powered Cyber Exercise Framework (AiCEF) is a comprehensive system designed to automate and enhance the generation of CSEs by integrating advanced ML techniques with structured ontological representations. It provides a modular approach to scenario creation, ensuring flexibility, scalability, and adaptability to different use cases.

1.6.1 Main Modules

The framework is built upon several interconnected components, each playing a critical role in the exercise generation process.

- **CESO (Cyber Exercise Scenario Ontology):** The foundation of AiCEF, CESO is a structured ontology that describes the various elements of a CSE. It defines the relationships among key components such as threat actors, attack techniques, and sector-specific vulnerabilities, serving as a blueprint for scenario representation.
- **MLCESO (Machine Learning for CESO):** A set of ML models trained to parse unstructured text and extract objects relevant to CESO. These objects include entities such as threat actors, tactics, and vulnerabilities, which are then mapped onto the ontology.
- **IncGen (Incident Generation Module):** This component uses MLCESO-extracted objects to model individual incidents within a CSE. By using CESO, one can create realistic and contextually accurate incidents that can serve as building blocks for broader scenarios.
- **CEGen (Cyber Exercise Generation Module):** Expanding on IncGen, CEGen constructs comprehensive CSEs by assembling incidents into a coherent and structured format. This module ensures that exercises are logically consistent and aligned with the predefined ontology.
- **KDb (Knowledge Database):** A repository storing incidents, extracted objects, and their attributes. The database also includes STIX 2.1 blobs for structured threat intelligence, enabling the framework to reuse and adapt historical data for new scenarios.

By combining these components, AiCEF achieves a high level of automation and accuracy in CSE generation. The integration of CESO ensures that outputs are both machine-readable and human-interpretable, while the ML models enable the system to dynamically adapt to new data sources and evolving threat landscapes. This framework forms the backbone of this thesis, providing the methodology necessary to address the challenges of modern cybersecurity exercises.

1.6.2 Support Modules & Optimizations

An additional AI-driven threat intelligence and forecasting module is developed to ensure the incorporation of current and emerging trends in the CSE development lifecycle (MLTP). To make the system accessible to users with varying levels of expertise, an automated prompt generation methodology (AiDPG) using AI is also incorporated into the framework. This functionality allows exercise planners to provide minimal, high-level inputs—such as examples of known incidents, types of critical infrastructure, or general threat information—and receive detailed, structured scenarios in return. By abstracting away the technical complexities, this feature democratises the scenario generation process, enabling non-experts to utilise the system effectively.

Additional research has been conducted to optimise the usage of AI models for intelligence extraction in the lifecycle of cybersecurity threat landscape generation and CSEs. For this purpose, advanced AI models like Large Language Models (LLMs) and Named Entity Recognition (NER) systems were employed to extract key attributes such as threat actors' names, TTPs. A comparative evaluation of these models ensures that the most effective tools are chosen to provide accurate and actionable intelligence that can later be incorporated into AiCEF.

1.6.3 Evaluation & Feedback

AiCEF is subject to a rigorous case study-based evaluation to assess its performance and relevance. Traditional human-driven methods of scenario generation are compared with the AI-driven approach using key metrics. These metrics include the speed of scenario generation, assessed by measuring the time taken by both methods; the quality of generated scenarios, evaluated through peer review by an ad hoc expert group; and the relevance of the

scenarios to the current threat landscape, assessed by their alignment with emerging trends and sector-specific needs.

Finally, the framework incorporates iterative refinement based on feedback from experts and findings from the case studies. This feedback loop is crucial for improving both the AI models employed and the overall CSE generation process.

1.7 Research Questions

This research is guided by the following core questions:

- How can ML and AI technologies be leveraged to automate the generation of time-efficient and resource-conscious Cybersecurity Exercises (CSEs) for critical infrastructure organizations?
- To what extent does the integration of AI-driven threat intelligence enhance the realism and relevance of CSE scenarios compared to traditional methodologies?
- How can AI-powered systems help non-expert planners develop high-quality CSE scenarios, reducing dependency on scarce cybersecurity expertise?
- What methodologies enable the reliable production of sector-specific threat landscapes using AI models?
- How effectively can AI transform historical cybersecurity incidents into structured, actionable CSE content for training purposes?
- Which forecasting algorithms most accurately predict sector-specific cybersecurity threats, and how can these predictions be systematically incorporated into the CSE generation lifecycle?
- What AI model architectures demonstrate superior performance in extracting critical attributes from Cyber Threat Landscape (CTL) reports for CSE scenario development?
- What are the quantifiable trade-offs between accuracy, efficiency, and scalability when comparing AI-augmented versus manual approaches for CSE generation?
- How do AI-assisted CSE development tools measurably impact the performance of both expert and non-expert Exercise Planners (EPs)?

- Can prompt engineering techniques improve the extraction accuracy of complex cybersecurity attributes (such as MITRE TTPs) without requiring model retraining?

These questions form the foundation for evaluating the proposed framework and its contribution to the field.

1.8 Positioning of the Research & Sub Domain Focus

As previously explained in the contributions section, our research spans three distinct cybersecurity sub-domains as presented in Figure 1.1: **CSE Generation, CTI & Forecasting, and CTL generation**.

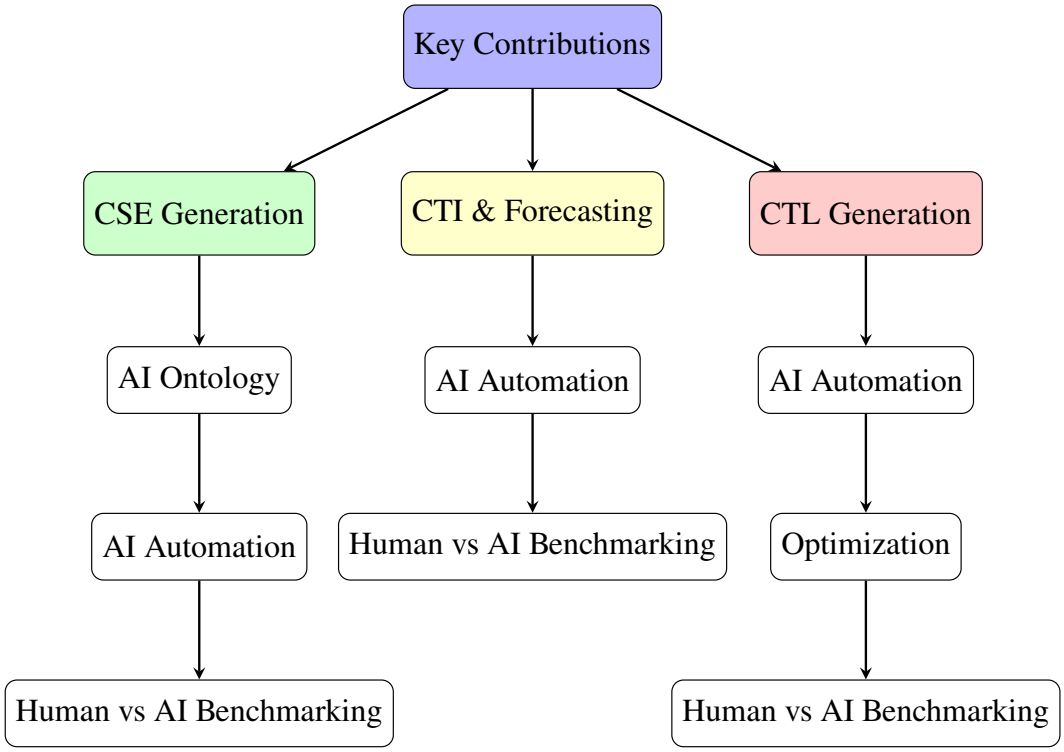


Figure 1.1: Overview of the Key Contributions

The research focus on CSE generation includes the automated representation of cyber incidents in CSE objects, where a CSE is a collection of simulated incidents provided to players in an orchestrated way to achieve the exercise's objectives. These objects can be

single instances or interconnected, forming meaningful relationships that represent the malicious activity of attackers. This work has created the first incident-centric CSE ontology compatible with AI/ML, focusing on using a bottom-up approach that allows one to identify and describe incidents first, so they can be grouped into events and then cover the full generation of CSE scenarios that fit the high-level objectives set. Additionally, extracted objects or relationship graphs can be enhanced utilizing known structures of APT activity to improve and adapt the generated CSE to the requirements of Exercise Planners. This work is the first to propose the extension of known CSE generation methodology to utilize AI/ML for automated analysis and extraction of key attributes and to form meaningful scenarios incorporating fresh sector-specific threat intelligence. Furthermore, this research is the first to compare AI/ML-generated CSEs with CSEs generated using the traditional human-powered methodologies in terms of speed, quality, and efficiency.

Regarding AI-driven threat intelligence and forecasting incorporation in CSE generation, the proposed methodology accurately identifies current and emerging trends with their potential impact in different sectors from any given dataset and incorporates them into a CSE scenario. To achieve this objective, several existing forecasting algorithms have been evaluated using quantifiable criteria to identify those that provide accurate predictions for a given dataset. By comparing the realism and relevance of CSEs created using the proposed methodology versus the traditional human experts' approach, a clear knowledge gap has been identified. To bridge this gap, a novel methodology has been implemented that can fully automate prompt generation, aiding non-expert exercise planners. Surveys based on a case study have been launched to collect quantitative results that help compare the proposed AI-based methodology against the human expert's methods.

Finally, the research focuses on CTL generation for the purpose of enhancing CSEs, including proposing the extension of an existing CTL generation methodology to utilize AI/ML for automated analysis and extraction of key attributes. As part of this research, the performance of state-of-the-art LLMs, including Gemini 1.5 Pro [134, 135] and GPT-4o [136], has been evaluated in conjunction with high-performing BERT-based models like TRAM [123], TTPHunter [122], and custom NER models, applying them to the same corpus of cybersecurity reports. This work has measured and compared the quality, cost, and speed of AI-generated CTL reports with those produced by human experts, providing a rigorous evaluation of the trade-offs involved, with a comparative analysis provided per extracted feature. A novel Voting-Based Agreement technique has been introduced to improve the precision of AI models, particularly to extract challenging attributes such as MITRE TTPs and to test

their effectiveness.

As a result, this work represents a significant advance in the automation of CSE generation, addressing a critical gap in the cybersecurity domain by providing a concrete framework supported by actionable insights for both researchers and practitioners.

1.9 Data Sources

During the research conducted for this thesis, three datasets have been created and used repeatedly to achieve various objectives.

Dataset Name	**#Texts**	**Period**
Main Corpus	4957	2019–2022
Training Corpus	2000	2019–2022
Incident Corpus	1000	2022–2023
TTP Corpus	100	–

Table 1.2: Summary of Datasets used.

The **Main Corpus (MC)** consists of four incident sources from publicly available news outlets, as shown in Table 4.10. These websites contain unique articles previously selected and classified in the Hackmageddon dataset[8], a reputable webpage that describes cybersecurity incidents since 2011. For simplicity, in this work, unique incidents were collected, spanning from 2019-01 to 2022-12 (Table 1.3b), which accounts for 4.957 articles. The collection covered over four years to adhere to the limitations imposed by the time series analysis algorithms to be used in later steps. A subset of the main corpus named **Training Corpus (TC)**, consisting of 2000 articles, was used to train AI models that were later used in AiCEF. The whole dataset was used to populate the KDb and was also used for forecasting purposes as described in Chapter 7.

Further to the main corpus described above a newer support dataset was meticulously curated from open-source cybersecurity outlets covering 12 months, ensuring a comprehensive representation of the evolving threat landscape between 2022-2023. The dataset comprises 1,000 cybersecurity-related texts, averaging 170 words per text and ranging from threat intel-

[8]https://github.com/alexzacharis/AI-driven-Threat-Intelligence-and-Forecasting/blob/main/Hackmageddon%20Dataset.json

Source	# Incidents
bleepingcomputer.com	2431
securityaffairs.co	339
zdnet.com	794
databreaches.net	1393
Total	**4957**

(a) Main Corpus quota per source

Year	# Incidents
2019	1088
2020	1348
2021	1248
2022	1353
Total	**4957**

(b) Main Corpus yearly distribution.

Table 1.3: Composition of the Main Corpus dataset.

ligence reports, security blog posts, to social media posts, comprising the **Incident Corpus (IC)**. These texts were selected to cover a diverse range of cybersecurity threats, including various types of cyberattacks, targeted sectors, underlying motivations, and specific threat actors' behaviour or malware used. The incident corpus has been made publicly available[9] to help future research.

A complementary dataset of 100 texts, averaging 718 words per text, was synthetically generated, only to cover the particular case of MITRE's TTP extraction, comprising the **TTPs Corpus (TTPC)**. The TTPs Corpus has also been made publicly available[10]. The texts were generated using either descriptions or examples provided by MITRE[11] itself when characterising a technique. Using a text generator, two or more techniques were merged simultaneously, resulting in a synthesised report. The experiments conducted indicate that all techniques included in the TTP Corpus (TTPC) were correctly classified at least once by one of the tested models.

1.10 Tools

A variety of tools have been used as part of this research work. While mostly open-source tools were used, in some cases, I opted for paid software in order to achieve better results faster. To address problems that have never been tackled before, I have developed a unique AI-powered Cyber Exercise Framework (AiCEF) tool in an attempt to automate and help

[9] https://github.com/alexzacharis/AICTL/blob/main/MainIncidentCorpus.xlsx
[10] https://github.com/alexzacharis/AICTL/blob/main/TTPCorpus.xlsx
[11] https://attack.mitre.org/docs/enterprise-attack-v15.1/enterprise-attack-v15.1-techniques.xlsx

with the assessment of this work.

Protégé [137] is an open-source ontology development environment and knowledge-base framework developed by Stanford University's School of Medicine. The platform provides a comprehensive suite of tools for creating, editing, and managing ontologies in Web Ontology Language (OWL) format. Protégé offers intuitive graphical interfaces for defining classes, properties, and axioms, while supporting automated reasoning through integrated inference engines such as HermiT and Pellet. The tool enables ontological validation through consistency checking, classification verification, and logical inference testing. For this research, Protégé served as the primary development environment for implementing the Cybersecurity Exercise Scenario Ontology (CESO), facilitating iterative refinement through visual class hierarchy management and automated consistency validation of ontological relationships.

SpaCy[138] is a robust natural language processing (NLP) library designed for efficient handling of various NLP tasks, including tokenization, part-of-speech tagging, dependency parsing, and NER. Its library is optimised for performance, allowing it to process large datasets efficiently. Moreover, SpaCy offers straightforward APIs to create custom models and train them on annotated datasets, making it a versatile tool for NER and other feature extraction tasks. I used SpaCy to train my own NER models using human supervision.

Prodigy (paid version)[139], on the other hand, is an annotation tool built to complement SpaCy by enabling efficient human-in-the-loop workflows for model training. It excels in facilitating the rapid labeling of entities in the text through an intuitive interface and supports active learning, where the model suggests potential annotations for human review. This targeted approach helped us focus on ambiguous or challenging examples, accelerating the annotation process. Prodigy integrated seamlessly with SpaCy and allowed us to iteratively train and refine models while improving accuracy with minimal time investment. Together, SpaCy and Prodigy provided a powerful combination for creating high-quality NER models and extracting features efficiently.

Gensim [140] is an open-source Python library designed for natural language processing (NLP) tasks, with a focus on efficient unsupervised learning techniques i used Gensim to perform topic modeling on a corpus of documents, leveraging its efficient implementation of the Latent Dirichlet Allocation (LDA) algorithm. Gensim enabled us to uncover hidden thematic structures by identifying dominant topics within the dataset, with each topic represented by a distribution of keywords. Additionally, I used its tools to visualize the topic-term distributions, assess topic coherence, and associate documents with their respective topics,

facilitating insightful analysis of the dataset's underlying themes.

The **AI-powered Cyber Exercise Framework (AiCEF) tool** developed as part of this thesis was designed to address critical challenges in the generation of CSE scenarios. Traditional approaches to creating these scenarios often require significant time, effort, and expertise, particularly to ensure realism, timeliness, and relevance to evolving threat landscapes. Moreover, achieving a balance between meeting technical objectives and engaging the target audience is a complex and resource-intensive process. To alleviate these challenges, AiCEF tool was developed to automate and streamline CSE generation, thereby supporting this research and enabling the efficient creation of high-quality scenarios that were used as part of the evaluation strategy.

1.11 Outline of the Thesis

The thesis is structured into five (5) distinct parts.

PART I: Introduction

Chapter 1, provides an introduction to the context, motivation, and methodology underlying this research. This section outlines the problem addressed, offers a comprehensive literature review, establishes the research objectives, and presents the methodology employed to explore the research questions. Additionally, an overview of open problems and future challenges is provided. Chapter 2 follows, detailing the publications and contributions stemming from the research conducted throughout this PhD, highlighting the various articles that were published as a result of this work. Chapter 3 described the open problems and challenges faced, along with proposals for future work.

PART II: AiCEF: An AI-assisted Cyber Exercise Content Generation Framework

This part focuses on the presentation of the AI-Assisted Cyber Exercise Content Generation Framework (AiCEF).

Chapter 4 thoroughly examines the problems encountered and the methodology used to develop this framework from scratch. It includes a detailed discussion of the core ontology developed (CESO) and outlines the individual modules that make up the framework. Chapter 5 presents the evaluation methodology and its corresponding results. Chapter 6 further

explores a realistic use case, demonstrating the application of the framework in modelling a cyber conflict with the objective of creating a realistic CSE scenario.

PART III: Optimizing & Integrating AI-driven Threat Intelligence and Forecasting in the Cybersecurity Exercise Content Generation Lifecycle

This part delves into additional optimisations, specifically the integration of AI-powered intelligence extraction and forecasting within the CSE content generation lifecycle. Chapter 7 discusses the incorporation of AI-driven threat intelligence and forecasting within AiCEF, including the development and evaluation of an AI-driven prompt generation methodology. Chapter 8 focuses on optimising threat intelligence extraction and automating the generation of the Cyber Threat Landscape, proposing tangible improvements in the extraction of key attributes, such as MITRE TTPs, which are essential for CSE development. Chapter 9 introduces the creation of the AiCEF tool, emphasising its key features and contributions to automation.

PART IV: Closure

Chapter 10 presents the key conclusions derived from the objectives initially set while framing the scope for future work.

PART V: Publications

The last part of the thesis annexes all the original publications authored throughout the duration of this research.

1.12 Acknowledgements

The content of this article does not reflect the official opinion of the European Union, its Bodies nor Institutions. Responsibility for the information and views expressed therein lies entirely with the authors.

Chapter 2

Publications and Contributions

The research conducted in the course of the PhD led to the publication of several articles. More precisely, three articles have been published in peer-reviewed journals and one (1) in a conference. Moreover, there was publication to a popular industry conference and a publication of a book chapter in a book.

2.1 Publications of the Compendium and Contributions

This section presents the list of articles this thesis as a compendium of publications is based on. For each, the contributions of the Ph.D. candidate are described:

- **A. Zacharis and C. Patsakis. "AiCEF: an AI-assisted cyber exercise content generation framework using named entity recognition". *International Journal of Information Security*, 22 (5), (2023), pp. 1333–1354**
  - **Overall contributions:** This article presents two major contributions to the field of cybersecurity in the area of CSE development. Firstly, a method for extracting cyber-attack trends of specific sectors and, based on these trends, recommends customised awareness training topics. This is achieved by extracting and clustering cyber-attack attributes using solely Artificial Intelligence. Second, the process of generating content for cyber awareness exercises is automated using

machine learning (ML) techniques. A key innovation is the development of a methodology, accompanied by a set of tools, that enables inexperienced EPs to create fully structured CSE scenarios from free text using the proposed ontology (CESO). In addition to scenario structuring, the system generates cyber exercise content to complement the scenarios and suggests the most relevant training topics. This approach bridges the expertise gap for cybersecurity experts acting as exercise planners by providing automated and easy-to-use tools. The methodology leverages Machine Learning (specifically NER) within the AiCEF framework, a novel CSE scenario generation system. The paper concludes with an evaluation of the proposed methodology and tools, presenting results and offering ideas for future improvements.

  - **Contributions of the PhD candidate:**
    * First author of the article.
    * Conception of the main idea.
    * Creation of a CSE ontology mapped to the STIX 2.1 standard.
    * Implementation of the data collection pipeline.
    * NER model training and data extraction pipeline.
    * Major contributions to the framework and the presentation of the results.
    * Development of the AiCEF tool to automate the exercise generation task.

- **A. Zacharis, V. Katos, and C. Patsakis. "Integrating AI-driven threat intelligence and forecasting in the cyber security exercise content generation lifecycle". *International Journal of Information Security*, (2024), pp. 1–20**

  - **Overall contributions:** This article proposes an AI-driven approach to enhance the development of CSEs by integrating AI-powered sectorial threat landscaping and forecasting into existing frameworks like AiCEF. This allows for the identification of emerging cyber threats and their potential impact in specific sectors, which are then used to generate more accurate and timely CSE scenarios.

    A key innovation is the ability to bridge the gap in creating contextually relevant CSE scenarios, particularly when sector-specific cybersecurity experts or experienced exercise planners are unavailable. By introducing an AI-driven topic extraction model and clustering sector-specific annotation tags, this approach ensures the development of sector-relevant CSEs that reflect current threat trends. The study evaluates forecasting models— SARIMAX, ES, Simple SES, Prophet,

and LSTM—demonstrating that LSTM and Prophet are the most effective for predicting sector-specific cyber threats.

Our methodology significantly improves the maturity of CSE prompts, considerably enhancing the quality of exercise scenarios produced against those generated by exercise experts and sector-specific experts. Furthermore, user surveys comparing traditional CSE generation methods to the AI-driven approach quantify a substantial increase in scenario relevance, based on prompt accuracy and scenario effectiveness for both expert groups. These results emphasise the value of AI-enhanced methodologies in creating more precise and sector-tailored cybersecurity exercise scenarios.

- **Contributions of the PhD candidate:**
  - First author of the article.
  - Conception of the main idea.
  - Forecasting model comparison
  - Expert survey execution and data collection and analysis.
  - Major contributions to the methodology and presentation of the results.

- **A. Zacharis, R. Gavrila, C. Patsakis, and C. Douligeris. "Optimising AI Models for Intelligence Extraction in the Life-cycle of Cybersecurity Threat Landscape Generation".** ***Journal of Information Security and Applications*****, (2025), X–Y**

  - **Overall contributions:** This article evaluates various AI models for automating CTL generation, comparing LLMs like Gemini 1.5 Pro and GPT-4o with specialized models like TRAM, TTPHunter, and custom NER models. The goal is to reduce human workload and costs while ensuring accurate and timely threat intelligence extraction.

    Our work highlights that LLMs excel in general attribute extraction (e.g., Threat Names), while specialised models like AiCEF NER perform better in targeted tasks (e.g., Attack Type, Sector). However, extracting motivation and MITRE TTPs remains challenging, with both AI models and humans struggling.

    Automating the TTP extraction process offers substantial savings. AI models can reduce the cost of annotating large numbers of articles and completing the task much faster than human experts. The proposed methodology, including voting-based and ensemble learning approaches, improved the accuracy and reduced false positive rates without the need for retraining of the models involved.

Overall, our approach significantly boosts efficiency, potentially increasing a cybersecurity expert's performance when using AI-pre-annotated reports. Our methodology provides a faster and more cost-effective solution for the preparation of CTLs.

  - **Contributions of the PhD candidate:**
    * First author of the article.
    * Conception of the main idea.
    * ML model comparison
    * Novel MITRE TTP extraction inception
    * Major contributions to the methodology and the presentation of the results.

- **A. Zacharis, R. Gavrila, C. Patsakis, and D. Ikonomou. "Ai-assisted cyber security exercise content generation: Modeling a cyber conflict".** ***2023 15th International Conference on Cyber Conflict: Meeting Reality (CyCon)*****. IEEE. 2023, pp. 217–238**

  - **Overall contributions:** This work addresses the essential need for effective cyber crisis exercises in the context of cyber conflict preparedness against cyber-attacks targeting critical national functions. Such attacks threaten government and civilian infrastructure, leading to potential damage and loss of life. Our research applies Machine Learning to unstructured information sources to generate structured cyber exercise content.

    We collected publicly available cybersecurity articles to identify future threats and create new exercise scenarios focusing on the example of recent real life cyber conflict with the use of AiCEF.

    Our methodology was evaluated against real-world cyber conflict use cases, modeling pre-conflict and post-conflict scenarios. We showed that AiCEF-generated scenarios align with emerging attack trends, enhancing awareness of upcoming incidents.

    Key findings include increased coverage of training objectives when using historical incidents for scenario development. Finally, exercises covering multiple sectors impacted by the same actors achieved higher objective coverage compared to single-sector scenarios. Overall, we conclude that AiCEF can automate the generation of well-scoped Cybersecurity Exercise scenarios, adapting to evolving conflicts and providing accurate training material with minimal planning expertise.

- **Contributions of the PhD candidate:**
  - First author of the article.
  - Conception of the main idea.
  - Incident collection and analysis
  - Conflict modelling based on a real use case
  - Major contributions to the methodology and the presentation of the results.

- **A. Zacharis, R. Gavrila, and C. Patsakis.** ***AICEF: An AI-Powered Cyber Exercise Content Generation Framework, BLACKHAT ASIA*****. 2023. URL:** https://www.blackhat.com/asia-23/arsenal/schedule/index.html#aicef-an-ai-powered-cyber-exercise-content-generation-framework-31299

  - **Overall Contributions:** The core idea of AiCEF is to harness the intelligence available from online reports, MISP repositories, and threat group activities—such as those cataloged by MITRE—to create relevant and timely cybersecurity exercises. To achieve this, we have developed a specialized ontology, CESO, which extends STIX 2.1. This ontology enables the mapping of data from various sources, including online resources and MISP reports, into a unified, machine-readable graph format.

    These graphs abstract events from the reports, enabling enhanced analysis. The generated graphs can also incorporate additional intelligence, such as detailed threat actor profiles from MITRE, which are mapped using our ontology. This integration addresses potential gaps in the data while allowing users to manipulate the graphs to create custom, unique cybersecurity exercise scenarios.

    Finally, we utilize transformer-based language models, such as GPT, to convert the graph representation into natural language text suitable for cybersecurity exercise scenarios. The AiCEF framework has been tested and validated by a group of cybersecurity exercise experts, with results showing that our tool significantly improves the quality, relevance, and efficiency of creating cybersecurity exercises.

  - **Contributions of the PhD Candidate:**
    - Conceived the core idea of AiCEF [141].
    - Designed and developed the exercise ontology (CESO) to extend STIX 2.1 for cybersecurity exercise modelling.

* Implemented the full pipeline into a practical tool, including the integration of intelligence sources, graph generation, and transformer-based text conversion.
* Conducted rigorous testing to validate the framework's functionality and effectiveness.
* Presented the tool, showcasing its capabilities and engaging with expert feedback.

2.2 Other Publications and Contributions

This section presents the publications that have also been published in the course of this dissertation but are not included in the compendium, as their relevance to the main theme of the thesis is limited.

- **R. Gavrila and A. Zacharis. "Advancements in Malware Evasion: Analysis Detection and the Future Role of AI".** ***Malware: Handbook of Prevention and Detection*****. Ed. by D. Gritzalis, K.-K. R. Choo, and C. Patsakis. Cham: Springer Nature Switzerland, 2025, pp. 275–297. ISBN: 978-3-031-66245-4. DOI:** 10.1007/978-3-031-66245-4_12**. URL:** https://doi.org/10.1007/978-3-031-66245-4_12

  – **Overall Contributions:** This chapter explores the evolution of malware evasion tactics and techniques, with a particular emphasis on evasion in the context of dynamic analysis detection. While the field is extensively researched, our work provides a comprehensive overview that not only consolidates existing knowledge but also identifies potential research opportunities.

  The chapter begins by presenting a historical perspective on malware evolution, highlighting the arms race between malware developers and cybersecurity defenders. It focuses on key anti-analysis techniques such as environment detection, reverse Turing tests, and leveraging artificial intelligence for advanced evasion. We discuss the dual role of AI in enhancing both defensive and offensive strategies, emphasizing its growing impact on the malware landscape.

  Our contribution aims to provide a foundational resource for understanding the current state of malware evasion while offering insights into emerging challenges

and opportunities for future research. This work serves as both an educational primer and a springboard for innovation in this critical area of cybersecurity.

– **Contributions of the PhD Candidate:**
  * Co-authored this book chapter, contributing significantly to the sections on anti-analysis techniques and AI-driven evasion strategies.
  * Conducted an in-depth review of literature on malware evasion tactics, synthesizing insights into the historical evolution and current landscape.
  * Provided analysis and critique of state-of-the-art AI applications in malware development and detection, including identifying gaps for future research.
  * Collaborated on the chapter's structure and presentation, ensuring clarity and coherence in conveying complex technical concepts.

Chapter 3

Open Problems,Challenges & Future Work

3.1 Open Problems and Challenges

The shortage of skilled cybersecurity experts and the widespread lack of awareness in the field continue to pose significant global challenges. With the increasing complexity and frequency of cyber threats, organisations face mounting pressure to ensure their preparedness. CSEs are recognized as an effective tool to address this gap, as they provide a platform for testing organisational resilience and response capabilities.

However, the creation of CSEs is resource-intensive, requiring significant time, expertise, and alignment with the evolving threat landscape. Two critical issues in current CSE development are the timeliness and relevance of the scenarios. As the threat landscape evolves rapidly, scenarios can become outdated, diminishing their effectiveness in preparing organisations for real-world attacks. Ensuring that these exercises remain engaging and aligned with the objectives of diverse stakeholders adds another layer of complexity. These challenges necessitate the development of innovative methodologies to automate and streamline the CSE generation process while maintaining high levels of accuracy and relevance.

3.2 Specific Challenges for this Research

This research seeks to address the broader challenges in the field through the development of the AiCEF. However, several specific challenges were encountered during this work.

The first challenge is automating the generation of realistic and structured CSE scenarios that reflect current or anticipated threats in the absence of a relevant AI/ML friendly ontology. To address this issue, a novel ontology (CESO) was developed, which ensures the standardisation and usability of generated scenarios. This ontology was made the centre of my proposed exercise framework.

Another significant challenge is maintaining the timeliness and relevance of the generated content. As part of this work, a number of historical incidents was converted into an annotated corpus to conduct forecasting and other types of experiments. While AiCEF successfully aligns scenarios with the current threat landscape, operational deployment would benefit from integrating real-time data sources such as live threat feeds and intelligence reports which are mainly paid. This dynamic adaptability remains an area for future enhancement.

The quality of AI-generated outputs also presents a challenge. While current general purpose LLMs and other text synthesis models produce high-quality results, their performance could be further refined by training models exclusively on cybersecurity-specific datasets. I have explored ways to use the power of crowdsourcing, a combination of topic-specific and generic models, in an attempt to tackle this issue.

Additionally, automated incorporation of the updated versions of the descriptions of threat actors and their TTPs would improve the overall depth and utility of the scenarios constructed and make them more relevant.

3.3 Future Work

While this research provides a significant step forward in automating CSE generation, several opportunities for future improvements have been identified.

Utilizing the latest LLMs to enhance the already generated content by this framework is the first step forward. This step can yield immediate results, and its efficacy needs to be

thoroughly evaluated.

Furthermore, automating the creation of popular inject content with the use of AI for common attack categories, such as phishing, could be a valuable extension.

One promising direction to take could be the generation of organisation-level threat landscape automation. By processing diverse datasets, including incident reports, threat alerts, and intelligence feeds, relevant to an organisation, AiCEF could generate more tailor-made scenarios. This would improve its adaptability to evolving threats while focusing on the specific threats applied to each organisation.

Another area for development is the integration of advanced NER techniques to extract context relevant to topics such as attackers motivation, risk level or Cyber Kill Chain phases and linking them to specific injects. This could significantly enhance the framework's functionality.

Improving the training of AI models is another priority. Transitioning from general-purpose generative models to those trained exclusively on cybersecurity corpora would yield more accurate and contextually relevant outputs. This could also improve the technical precision of generated text, better aligning it with the needs of cybersecurity professionals. My work on crowdsourcing multiple models under a voting scheme to resolve the problem of automated TTP extraction is one good example that requires further research.

In addition, further evaluation of AiCEF's impact on non-expert users is needed. Testing the framework with users of varying expertise levels would provide insights into its usability and the quality of its outputs across diverse user profiles.

Finally, the validation of forecasting methodologies can be enhanced by comparing AiCEF-generated predictions with expert insights and real-world trends over extended periods. Collecting and analysing additional incident data will bolster the robustness of the forecasting models and improve their predictive capabilities.

By addressing these challenges and expanding its capabilities, AiCEF has the potential to revolutionize cybersecurity exercise development, making it more accessible, timely, and relevant to the organisations it serves.

Part II

AiCEF: An AI-assisted Cyber Exercise Content Generation Framework

Chapter 4

Building an AI-assisted Cyber Exercise Content Generation Framework

Cyber security exercises are increasingly becoming an integral part of the cybersecurity training landscape [147], providing a hands-on experience to personnel of both public and private organisations worldwide. A CSE, as described in the ISO Guidelines for Exercises[37], is "*a process to train for, assess, practice, and improve performance in an organisation*". ENISA defines a CSE as "*a planned event during which an organisation simulates cyber-attacks or information security incidents or other types of disruptions to test the organisation's cyber capabilities, from being able to detect a security incident to the ability to respond appropriately and minimise any related impact.*" [13]. This chapter presents the steps taken to build a unique AI-assisted Cyber Exercise Content Generation Framework from scratch and deep dive into each component and its use.

4.1 Problem setting and objectives

Creating cybersecurity exercise (CSE) content is a complex process that requires a deep understanding of current threat landscapes, historical incidents faced by entities within specific sectors, and comprehensive knowledge of corresponding attack methodologies. Training employees through simulated incidents provides the most effective method for testing organiza-

tional preparedness and evaluating the effectiveness of established security measures and response procedures. Developing relevant and dynamic CSE scenarios demands specialized expertise and resources that are often unavailable within most organizations.

This work aims to automate the generation of structured CSE scenarios from unstructured information sources, requiring minimal scenario development experience from exercise planners.

The ISO 22398:2013 standard [37] establishes a three-tier methodology for exercise scenario preparation: events, incidents, and injects. Events provide the overarching scenario context (such as "nation-state actor targeting financial infrastructure"). Incidents specify concrete attack vectors and consequences (such as "spear-phishing campaign targeting C-level executives leads to credential compromise"). Injects deliver time-sequenced information packets that drive participant responses (such as "security operations center receives automated alert indicating suspicious file execution on domain controller"). Organizations must ensure scenarios contain only essential information while creating sufficiently stressful environments to test participant capabilities effectively. Events define the general exercise scenario framework, with their number varying based on predetermined objectives. Each event generates specific consequences at the incident level, where multiple incidents may interact and cascade. Injects facilitate communication of events and incidents to participants, providing actionable problems while compelling participants to address consequences and make critical decisions.

Generated scenarios must conform to specifications provided by Exercise Planners (EP), including targeted training objectives, sectoral focus areas, and specific threat vectors of current relevance or anticipated emergence. This framework adopts sectoral classifications defined in Directive (EU) 2022/2555 of the European Parliament and the Council on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS2 Directive) [2], while maintaining compatibility with alternative classification systems.

The Machine Learning-powered Exercise Generation Framework provides three core capabilities: (1) automated generation of structured exercise scenarios that accurately reflect contemporary or emerging threat landscapes, incorporating relevant threat actors and their associated Tactics, Techniques, and Procedures (TTPs); (2) development of realistic scripted events and incidents applicable to organizations across any NIS2-defined sector; and (3) identification and description of appropriate artifacts serving as potential injects within exercise

scenarios.

All generated scenarios adhere to a standardized ontological structure, ensuring machine readability for automated processing while maintaining human readability for practical implementation.

Framework effectiveness is evaluated through three performance indicators: quantitative improvement in cybersecurity exercise development efficiency; qualitative enhancement of exercise quality, particularly beneficial for inexperienced Exercise Planners; and improved alignment between proposed exercise scenarios and current cybersecurity threat landscapes.

Case studies will measure these performance indicators by comparing traditional exercise generation methods against the proposed framework through evaluation conducted by the ENISA Ad-hoc Cyber Awareness Expert Group [1], which will peer review the methodology outputs.

The framework enables inexperienced Exercise Planners to structure comprehensive CSE scenarios from unstructured text following the proposed ontology. Exercise structure follows the traditional Scenario-Events-Incidents-Injects hierarchy established in ISO 22398:2013 as illustrated in Figure 4.1. The system generates additional cyber exercise content to complement scenarios and recommends optimal training topics to enhance organizational preparation for imminent cyber crises.

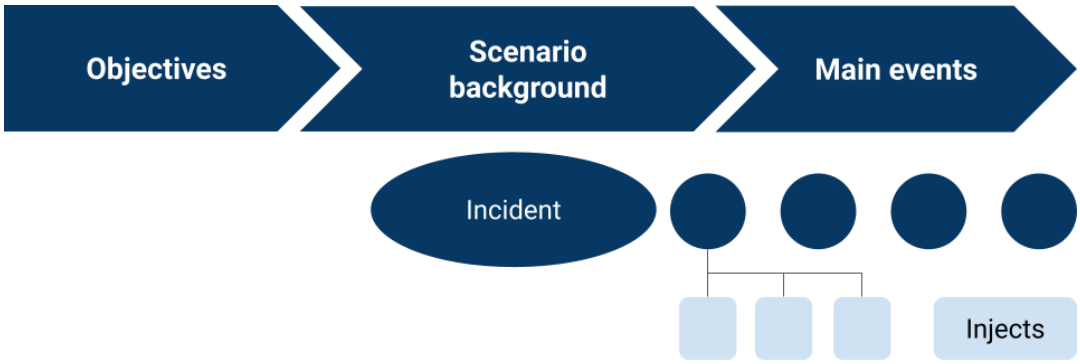


Figure 4.1: Cyber Exercise Structure according to ISO 22398:2013 [37].

[1] https://www.enisa.europa.eu/topics/cybersecurity-education/ad-hoc-working-group-awareness-raising

4.2 Framework Overview

To create the AI-powered Exercise Generation Framework, my work implements AiCEF (AI-powered Cyber Exercise Framework) using Python, developing a modular toolkit that enables Exercise Planners (EPs) across experience levels to generate contemporary and targeted Cybersecurity Exercises (CSEs). The framework operates through six sequential processing stages: (1) data collection from threat intelligence sources, (2) data processing and semantic mapping to cybersecurity ontologies, (3) trend prediction using historical attack patterns, (4) incident generation based on current threat landscapes, (5) enhancement through graph-based relationship modeling, and (6) complete storyline and exercise scenario generation.

AiCEF implements a three-layer architecture as illustrated in Figure 4.3. Layer 1 (Data Foundation) encompasses the Machine Learning Cyber Exercise Scenario Ontology (MLCESO) pipeline and Knowledge Database (KDb), handling raw threat intelligence ingestion, natural language processing, and structured data storage. Layer 2 (Intelligence Analysis) performs trend analysis and prediction using time-series forecasting models to identify emerging threat patterns. Layer 3 (Exercise Synthesis) generates complete exercise scenarios through the Incident Generation (IncGen) and Cyber Exercise Generation (CEGen) modules.

The framework integrates five core components operating across multiple processing layers. The Cyber Exercise Scenario Ontology (CESO) provides a standardized schema defining CSE components, including threat actors, attack vectors, organizational targets, and response procedures. The MLCESO pipeline employs Named Entity Recognition (NER), relationship extraction, and classification models to parse unstructured threat reports and populate CESO-compliant data structures. The Incident Generation module (IncGen) constructs realistic attack scenarios by combining MLCESO-extracted threat actor profiles, tactics, techniques, and procedures (TTPs) with sector-specific vulnerabilities, enhanced through graph-based relationship modeling and Advanced Persistent Threat (APT) behavior simulation. The Cyber Exercise Generation module (CEGen) assembles complete multi-stage exercises incorporating progressive incident escalation, decision points, and inject timing based on ISO 22398:2013 specifications. The Knowledge Database (KDb) maintains persistent storage of processed threat intelligence, incident templates, exercise scenarios, and STIX 2.1-compliant threat objects, enabling historical trend analysis and scenario reuse.

Figure 4.2 presents a sequential workflow that demonstrates component interactions, data flow dependencies, and processing pipelines transforming raw threat intelligence into

structured exercise scenarios. This process visualization enables practitioners to understand system operation and customize individual components according to organizational requirements.

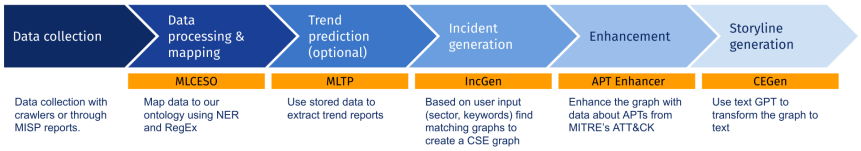


Figure 4.2: Sequential workflow of AiCEF module interactions and data processing pipeline.

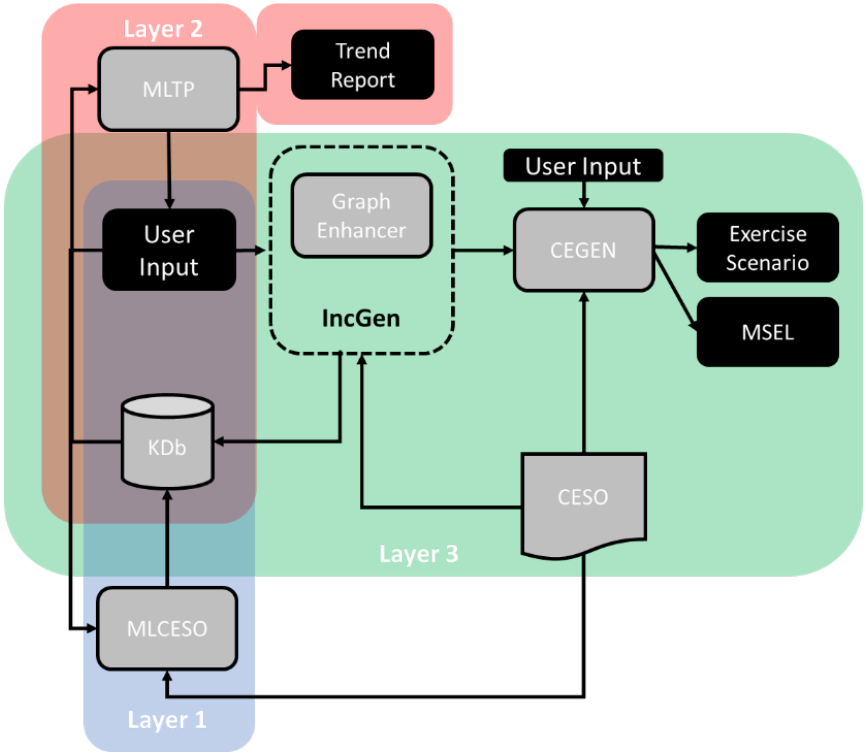


Figure 4.3: Three-layer architecture of AiCEF showing component relationships and data flow.

The modular architecture enables component-level customization, supports integration with existing security information and event management (SIEM) systems, and facilitates adaptation to organization-specific threat models and compliance requirements. The following sections detail the implementation, operational parameters, and integration mechanisms of each component, accompanied by concrete examples demonstrating CSE scenario generation using real-world threat intelligence inputs.

4.3 Cyber Exercise Scenario Ontology (CESO)

This work addresses the critical need for a unified Cybersecurity Exercise (CSE) scenario ontology that enables systematic translation of exercise components while maintaining semantic alignment with established cybersecurity incident representation frameworks. Domain analysis identified fragmented coverage across existing taxonomies: attack classification schemes (MITRE ATT&CK, CAPEC), vulnerability databases (CVE, CWE), sectoral frameworks (NIST CSF, NIS2), and impact assessment models (FAIR, CVSS). These disparate resources required integration into a coherent ontological model that enables Exercise Planners (EPs) to construct semantically consistent and technically accurate CSE scenarios.

The ontology development process addressed three fundamental design challenges: establishing precise ontological scope boundaries, determining optimal reuse strategies for existing cybersecurity taxonomies, and identifying critical ontological terms that bridge theoretical frameworks with practical exercise implementation requirements.

The ontological scope definition employed a competency question methodology involving twelve experienced EPs from government agencies, critical infrastructure operators, and cybersecurity consulting firms. These stakeholders identified essential terminological requirements, including threat actor characterization, attack progression modeling, organizational impact assessment, and response action classification. Interoperability requirements emerged as the primary design constraint, necessitating seamless integration with existing Security Information and Event Management (SIEM) platforms, threat intelligence feeds, and incident response tools. The ontology design prioritized compatibility with established cybersecurity frameworks, including MITRE ATT&CK for tactical analysis, STIX/TAXII for threat intelligence sharing, while ensuring implementability through automated Named Entity Recognition (NER) extraction from unstructured threat reports and vulnerability disclosures.

Domain expert consultation identified prominent cybersecurity ontologies suitable for reuse and extension. The expert evaluation criteria included ontological completeness, community adoption, maintenance support, and semantic precision for representing exercise scenarios.

The systematic ontology development methodology followed established ontological engineering principles:

1. **Scope Definition:** Conducted competency question workshops with Exercise Planners to identify core conceptual requirements, functional boundaries, and integration constraints for CSE scenario representation.

2. **Ontological Reuse Analysis:** Evaluated existing cybersecurity taxonomies including MITRE ATT&CK, CAPEC, CVE/CWE, STIX 2.1 for semantic coverage, structural compatibility, and extension potential within the CSE domain.

3. **Concept Hierarchy Development:** Established primary ontological classes (ThreatActor, AttackVector, Vulnerability, Asset, ImpactMetric, ResponseAction) and defined hierarchical relationships, inheritance structures, and cross-class associations.

4. **Property Specification:** Defined object properties (relationships between concepts), data properties (concept attributes), and constraint axioms governing valid ontological instantiations and logical consistency.

5. **Implementation and Validation:** Implemented the ontology using Web Ontology Language (OWL 2) with the Protégé tool, validated logical consistency through automated reasoning, and tested practical applicability through exercise scenario generation case studies.

4.3.1 Scope of the Ontology

The scope of the defined ontology was to target an efficient and robust way of representing cyber incidents in the context of a CSE. After all, a CSE is a collection of simulated incidents provided to players in an orchestrated way to achieve the exercise's objectives.

The exercise ontology presented is incident-centric, focusing on using a bottom-up approach that allows to identify and describe incidents first so as to group them into events and then cover the full generation of CSE scenarios that fit the high-level objectives set.

The first building block, incidents, are assigned injects and mitigation actions that match the expected scope of the scenario. Injection timing is configured on the attribute level of each object. Moving towards a higher level of the exercise, the scenario is formed. The selected format should allow for the scenario's portability to various existing tools (ex., MISP[2]) and support a decentralised type of CSE execution.

[2]https://www.misp-project.org/

4.3.2 Ontologies/taxonomies to be (re)used

A set of existing ontologies, taxonomies, frameworks, standards, and formats has been explored with relevance to cybersecurity and a focus on the representations of the key element of CSEs from the point of their very building blocks being the incidents to be simulated. My research concluded that a combination of the following would provide the necessary means: ISO 22398 [37], MITRE ATT&CK [38], Cyber Kill Chain [39], MITRE CVE [40], and STIX 2.1 [148].

STIX 2.1 was chosen as the basis for the proposed ontology, which defines a taxonomy of cyber threat intelligence to be extended to cover the need to describe CSE scenarios. ISO 22398 best describes the structure of the cyber exercise components and was used to help repurpose STIX 2.1 to cover the initial scope. The STIX 2.1 model describes an adversary and adversary activities in appropriate data structures by default. STIX Domain Objects cover: Threat Actor; Malware; Tools; Campaign; Intrusion Sets, and Attack Patterns (referencing the Common Attack Pattern Enumeration and Classification taxonomy, CAPEC), perfectly covering what is called incident & injects in the CSE nomenclature. STIX 2.1 supports by default the MITRE ATT&CK, MITRE CVE, and Cyber Kill Chain frameworks, helping achieve my goal for maximum interoperability. Moreover, CTI in a consistent and machine-readable manner, allowing security communities to better understand what computer-based attacks they are most likely to face and anticipate and/or respond to those attacks faster and more effectively.

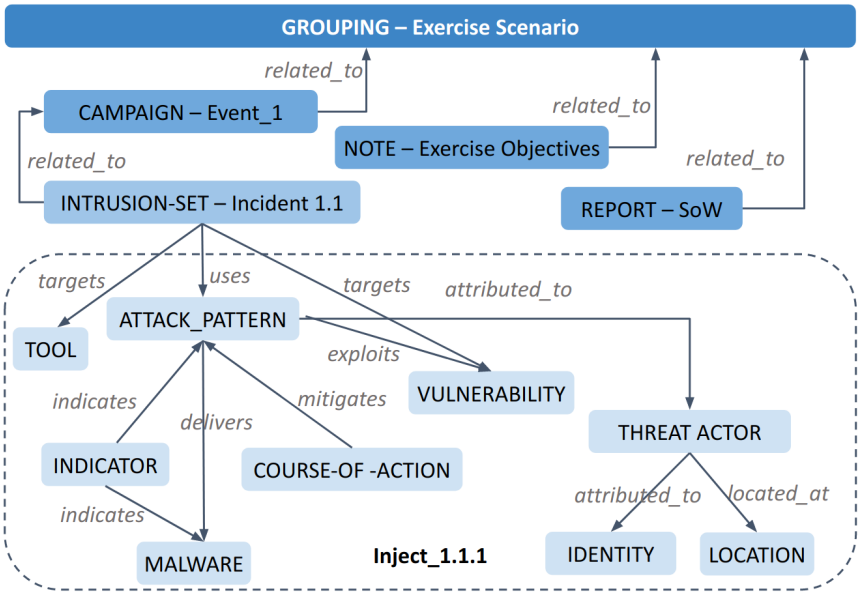
This helps build on top of these communities to reuse existing tools and share CSE scenarios represented in the very same format.

4.3.3 Scenario Augmented Model

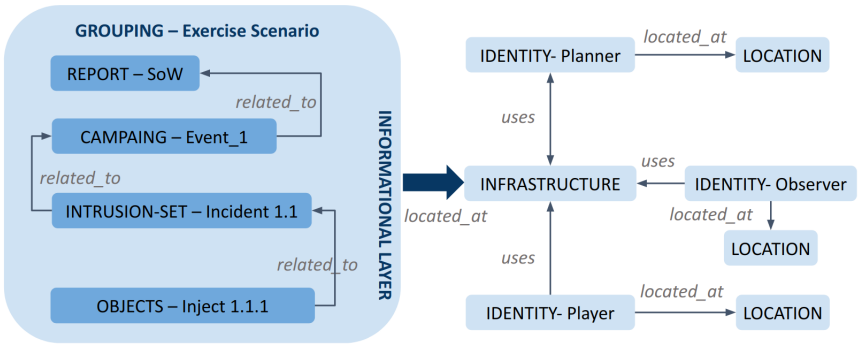
Based on the bottom-up approach, a Scenario Augmented Model (SAM) is proposed in two layers that cover both the Informational and Operational aspects with the same objects but utilise different STIX 2.1 attributes.

The Informational Layer covers the context and main attributes of scenarios. Figure 4.4a describes the key relationships in the informational layer.

The whole exercise is grouped using the Grouping Objects. The object holds informa-



(a) Informational Layer of CESO.



(b) Operational Layer of CESO.

Figure 4.4: The informational (left) and operational (right) layers of CESO.

tion related to the exercise's name, description, and scenario. All Events, Objectives, and their matching objects (Campaign, Note, Report) are related to the Exercise Scenario along with the matching "State of the World (SoW)". The SoW includes details such as the status of various simulated systems and networks, the simulated geopolitical landscape, and any simulated incidents or events that have taken place.

One or more incidents (Intrusion Set) can be related to events. From there, various objects with interlinked dependencies form the Inject in a Course of Actions Instance that refers to all related objects of an attack pattern.

An Inject can contain the following objects: Attack Pattern, Tool, Vulnerability, Indicator, Malware Threat Actor (who is attributed and Identity and is located at a Location), and a Course Of Action. Injects do not have to be related to an Event or Incident. Such examples are the STARTEX or ENDEX [3], which can be represented only with a Course of Action object but are directly related to the Scenario.

The Scenario Operational Layer describes an exercise scenario's execution flow, mainly dealing with injects delivery to the intended recipients. There are two major interrelated parts: (1) the events/injects, which describe the detailed activities of the scenario and expected actions from the participants, and (2) the participants.

The whole scenario, including Events, Incidents, and Injects, is stored in an Infrastructure object, representing the Exercise Platform. This platform is used by EPs (Identity) to design and conduct the exercise, Observers, and Players to interact with the scenario. All participants are located in the same or different locations. The Operational Layer is illustrated in Figure 4.4b.

4.3.4 Implementing the Ontology

Maintaining the structural integrity of CSE components, the following STIX 2.1 objects have been systematically repurposed to represent the main CSE components effectively, providing comprehensive coverage of the Scenario Augmented Model (SAM) along with their corresponding relationships.

Objects: STIX 2.1 defined objects as per specifications (Table 4.1).

Relationships: All relationships are implemented according to the STIX 2.1 relationship object specifications. The relationships enumerated in Table 4.2 represent the semantic edges connecting key objects within the ontological graph structure, although additional relationships may be incorporated as requirements evolve.

Object Extensions: STIX 2.1 objects have been augmented with additional attributes and properties to address the specific requirements of CESO, as detailed in Table 4.3.

OWL Implementation: The complete CESO ontology has been formally implemented

[3] The [START]ing and [END]ing [EX]ercise injects

CSE Component	STIX 2.1 Object
Exercise Details & Scenario Background	Grouping
Objectives	Note
Events	Campaign
Incidents	Intrusion Set
State of the World (SoW)	Report
Injects	Tool, Vulnerability, Threat-Actor, Identity, Location, Attack Pattern, Malware, Indicator, Course-of-action, Observed Data, Malware Analysis, Report
Exercise Platform	Infrastructure
Exercise Participant	Identity, Location

Table 4.1: CSE Components to STIX 2.1 objects mapping.

in Web Ontology Language (OWL) format to ensure machine-readable representation and semantic interoperability with existing cybersecurity frameworks, which can be found under Annex II Section 10.4.1.

The OWL implementation encompasses all classes, object properties, data properties, and relationship specifications defined in the preceding tables, along with additional properties supporting AiCEF-specific functionality, including maturity scoring algorithms, complexity assessments, and training topic categorizations (as seen in Figure 4.5). The ontology includes pre-defined individuals for the six primary training categories (Incident Handling, GDPR Compliance, Cyber Hygiene, Phishing and Social Engineering, Social Media, and BYOD) identified through domain analysis.

This formal specification enables automated reasoning, consistency validation, and seamless integration with ontology management tools such as Protégé, facilitating both human comprehension and machine processing of exercise scenarios.

Source Object	Destination Object	Relationship
Campaign	Grouping	**related_to**
Note	Grouping	
Report	Grouping	
Intrusion-Set	Campaign	
Course-Of-Action	Grouping	
Intrusion-Set	Tool	**targets**
Intrusion-Set	Vulnerability	
Intrusion-Set	Attack-Pattern	**uses**
Identity	Infrastructure	
Attack-Pattern	Threat-Actor	**attributed_to**
Threat-Actor	Identity	
Identity	Location	**located_at**
Attack-Pattern	Malware	**delivers**
Attack-Pattern	Indicator	**indicates**
Indicator	Malware	
Attack-Pattern	Vulnerability	**exploits**
Course-Of-Action	Attack-Pattern	**mitigates**
Course-Of-Action	Vulnerability	

Table 4.2: Relationships matrix.

4.4 Layer 1: MLCESO & KDb

4.4.1 Creating MLCESO pipeline

The creation of the Machine Learning pipeline for CESO object extraction (MLCESO) represents the most critical component of the proposed methodology. This pipeline transforms unstructured cybersecurity text into structured CESO-compliant objects through automated parsing and Named Entity Recognition (NER) techniques. The development process follows a systematic three-phase methodology: **Corpus Building**, **Corpus Annotation**, and **Corpus Evaluation**, each detailed in the subsequent sections.

Upon completion of model training and validation, the operational MLCESO pipeline

Object	Attribute added	Type	Description
Course-of-action	Difficulty (optional)	Integer	An integer from 1 to 5 (1 being easy and five being hard) declaring how difficult a course of action is evaluated to be executed by the player to resolve an incident.
Grouping	Scenario (mandatory)	String	A description that provides more details and context about the exercise scenario.
Identity	Recipient_Group (optional)	String	The name of the recipient group in which players are split to receive different injects.

Table 4.3: Objects extension matrix.

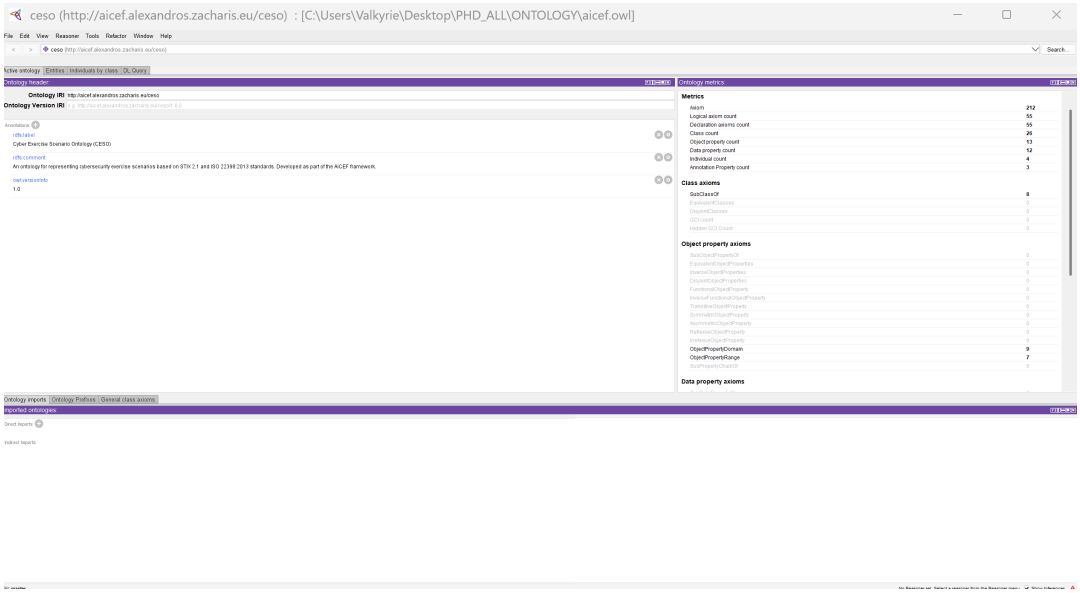


Figure 4.5: CESO implementation in OWL, visualised in Protégé.

processes raw threat intelligence documents and cybersecurity reports to extract relevant entities and relationships. The extracted objects undergo enhancement through graph-based analysis and are subsequently stored in the Knowledge Database (KDb) as interconnected relationship graphs, enabling efficient retrieval and scenario generation. The complete MLCESO processing workflow is illustrated in Figure 4.6, demonstrating the transformation from unstructured text input to structured knowledge representation suitable for automated exercise scenario construction.

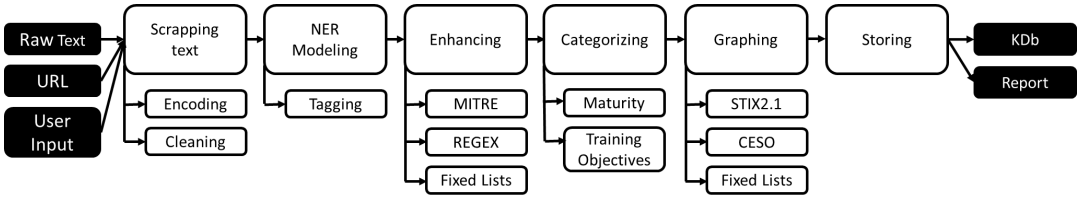


Figure 4.6: The MLCESO pipeline.

4.4.1.1 Corpus Building

Four sources have been identified as the initial input to the corpus (Table 4.4). These websites cover a wide variety of cybersecurity incidents in an article format that dates back many years. For simplicity, in this work, I have collected incidents from 2020-01 till 2022-03, which accounts for 2000 articles. All relevant articles were collected through automated web scraping.

Webpage	**Articles**
`bleepingcomputer.com`	1000
`securityaffairs.co`	150
`zdnet.com`	350
`databreaches.net`	500
Total	**2000**

Table 4.4: Training Corpus collection count.

The raw text was processed using Natural Language Processing (NLP) techniques to form a reduced Incidents Corpus (rIC). Initially, all text was converted to the UTF-8 encoding scheme. Using dictionaries and the Textblob library[4], spelling corrections were performed

[4] https://github.com/sloria/TextBlob

and special characters were removed. Empty lines, specific stopwords, and specific punctuation marks were removed using traditional NLP libraries such as NLTK[5] and spaCy [6]. Moreover, all HTML or other programming codes, URLs, and paths were removed. Any illegal characters were also stripped, and all text was transformed to lowercase. All steps followed to process raw text into IC can be found in Figure 4.7

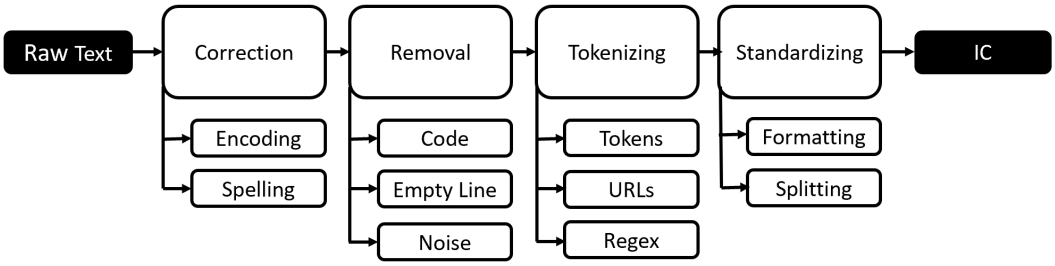


Figure 4.7: Incident Coprus generation

The standard Penn Treebank [149] tokenisation rules were utilised for sentence tokenisation, and finally, standardisation processes were applied to tune the Incidents Text to facilitate annotation. At the end of this step, a corpus composed of Incidents was formed. As discussed, the corpus, from now on referred to as rIC, contains 2000 cybersecurity articles. This accounts for 35.745 sentences containing 819.690 words, leading to a vocabulary of 24594 terms. An example of a corpus line ready for annotation is the following:

```
{"text":"revil sodinokibi ransomware targets chinese users with
    dhl spam"}
```

4.4.1.2 Corpus Annotation

The corpus annotation process implements a systematic six-step methodology designed to transform unstructured cybersecurity text into CESO-compliant tagged datasets. Two cybersecurity domain experts performed manual annotation using standardized guidelines documented in an Annotators Reference Document (ARD), which formally defines entity mappings and interconnection specifications. The annotation platform Prodigy[139] facilitated the labeling process, providing an intuitive interface for tag assignment and quality control (Figure 4.8). Post-annotation validation employed Cohen's Kappa metric to assess inter-

[5] https://www.nltk.org/
[6] https://spacy.io/

annotator agreement, establishing a gold standard corpus for machine learning model training.

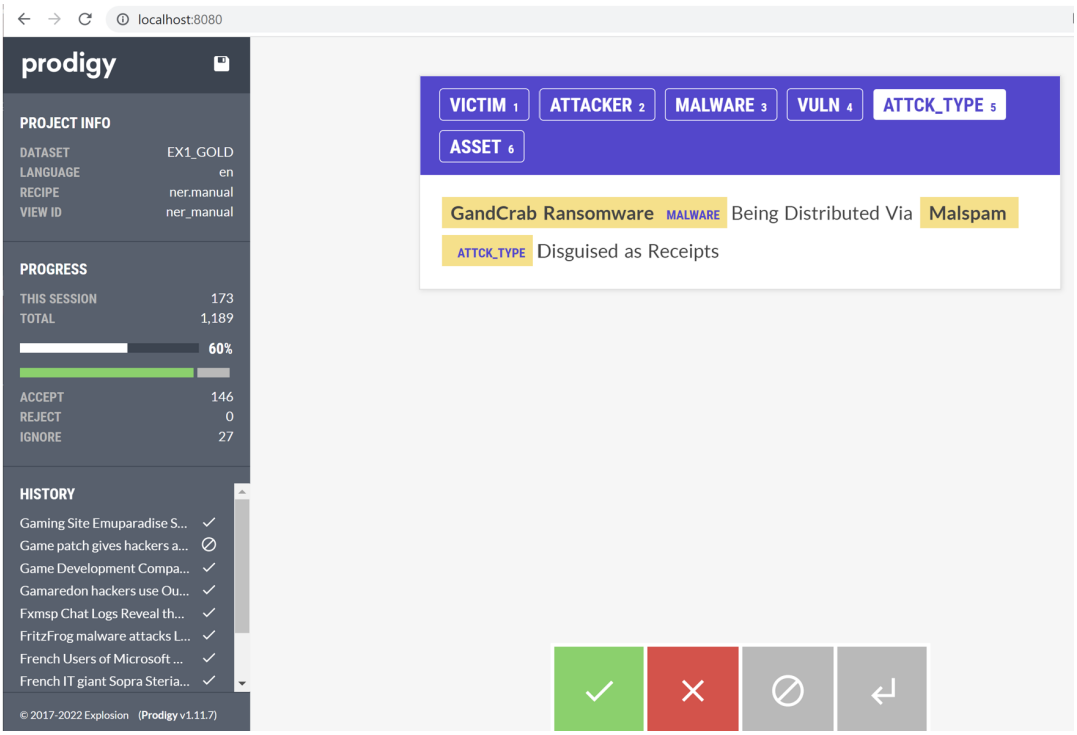


Figure 4.8: Annotation example (using Prodigy UI).

The systematic annotation methodology encompasses the following phases:

Step 1 - Annotation Objectives Definition: The primary objective focuses on creating semantically accurate training targets for Named Entity Recognition (NER) models by assigning contextually appropriate tags to cybersecurity-relevant text spans. Each word sequence receives classification labels drawn from a CESO-derived vocabulary, enabling automated identification of threat actors, attack vectors, and victim characteristics within incident descriptions. The annotation process targets three core entity categories: (i) threat actor identification and attribution, (ii) cyber attack characterization, including techniques and tools, and (iii) victim organization and asset classification.

Step 2 - Formal Model Specification: The annotation framework employs a three-tier classification schema (Attacker, Attack, Victim) with an additional "Other" category for non-relevant terms. This model M comprises a vocabulary of terms T, relationships between terms R, and semantic interpretations I, formally represented as $M = \langle T, R, I \rangle$ where:

- $T = \{$CESO, Attacker, Attack, Victim, Other$\}$
- $R = \{$CESO ::= Attacker|Attack|Victim|Other$\}$
- $I = \{$Attacker = "threat actor entities and attribution data",
  Attack = "malware, techniques, and attack patterns",
  Victim = "targeted sectors, assets, and technologies",
  Other = "contextual terms outside incident scope"$\}$

Step 3 - Annotation Tag Schema: Annotators receive comprehensive documentation specifying ten distinct tag categories mapped to CESO and STIX 2.1 object types (Table 8.3). This standardized schema ensures consistent entity recognition across different cybersecurity incident contexts while maintaining semantic alignment with established threat intelligence frameworks.

Category	Tag	Link to CESO & STIX 2.1
Attacker	ATTACKER_TYPE	Threat Actor Attribute
	ATTACKER_NAME	Threat Actor Attribute, Identity
	ATTACKER_ORIGIN	Location
Attack	MALWARE_TYPE	Malware Attribute
	MALWARE_NAME	Malware Attribute
	ATTACK_TYPE (TECHNIQUE)	Attack Pattern
	VULNERABILITY	Vulnerability
Victim	SECTOR	Identity Attribute, Scenario
	ASSETS	Threat Actor Attribute
	TECHNOLOGY	Tool

Table 4.5: Annotation tag schema with CESO/STIX 2.1 mappings.

Step 4 - Manual Annotation Execution: Two cybersecurity experts independently annotated the 24,594-term corpus using semantic and syntactic analysis to identify entity boundaries and assign appropriate tags. The Prodigy annotation platform facilitated real-time tag validation and consistency checking. Table 4.6 demonstrates the annotation process applied to representative cybersecurity incident descriptions, showing how textual elements map to structured CESO entities.

Step 5 - Gold Standard Corpus Generation: Inter-annotator agreement validation employed Cohen's Kappa coefficient [150] to assess annotation consistency and establish corpus

Category	**Tag**	**Annotator A**	**Annotator B**
Example Text: "qbot malware dropped via context aware phishing campaign infects the energy sector." "russian hacking group claims 1000 windows machines compromised."			
Attacker	ATTACKER_TYPE	hacking group	—
	ATTACKER_NAME	—	—
	ATTACKER_ORIGIN	russian	russian
Attack	MALWARE_TYPE	malware	malware
	MALWARE_NAME	qbot	qbot
	ATTACK_TYPE (TECHNIQUE)	phishing campaign	phishing
	VULNERABILITY	—	—
Victim	SECTOR	energy	energy
	ASSETS	windows machines	windows machines
	TECHNOLOGY	windows	windows

Table 4.6: Comparative annotation results demonstrating tag assignment consistency.

reliability. The metric quantifies agreement beyond chance occurrence using:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{4.1}$$

where p_0 represents observed agreement and p_e denotes expected chance agreement.

Statistical analysis of the annotated corpus containing $N = 24,594$ terms across $m = 4$ categories revealed high inter-annotator consistency: Attacker category (397 agreements), Attack category (1,722 agreements), Victim category (926 agreements), and Other category (21,416 agreements). The contingency matrix in Table 8.1 presents detailed agreement patterns where diagonal elements (x_{ii}) indicate perfect annotator consensus.

Annotator		**Annotator B**				**Total**
	Category	Attacker	Attack	Victim	Other	
A	Attacker	397	10	4	24	**435**
	Attack	13	1,722	8	9	**1,752**
	Victim	10	2	926	15	**953**
	Other	16	10	12	21,416	**21,454**
Total		**436**	**1,744**	**950**	**21,464**	**24,594**

Table 4.7: Inter-annotator agreement contingency matrix.

Quantitative analysis yields an observed agreement of $p_o = \frac{397+1,722+926+21,416}{24,594} = 0.996$ and expected chance agreement of:

$$p_e = \frac{\frac{436\times435}{24,594} + \frac{1,744\times1,752}{24,594} + \frac{950\times953}{24,594} + \frac{21,464\times21,454}{24,594}}{24,594} = 0.768\ (76.8\%)$$

The resulting Cohen's Kappa coefficient of $k = \frac{0.996-0.768}{1-0.768} = \frac{0.228}{0.232} = 0.98$ indicates almost perfect inter-annotator agreement [151], validating the annotation methodology's reliability and establishing a high-quality gold standard corpus for NER model training.

4.4.1.3 Training & Evaluation Using NER

The Named Entity Recognition (NER) model development follows a comprehensive training and evaluation methodology designed to transform the annotated corpus into operational cybersecurity entity extraction capabilities.

Data Preprocessing and Model Architecture: The annotated corpus undergoes preprocessing where each sentence is represented as token-tag pairs, with individual tokens encoded using pre-trained English word embeddings from the spaCy NLP library. This approach leverages transfer learning to capture semantic relationships within cybersecurity terminology while maintaining computational efficiency for real-time processing.

Training Strategy: The training process employs spaCy's neural architecture with customized loss functions optimized for cybersecurity entity recognition. Individual NER models are developed for each CESO object type, enabling specialized learning for distinct entity categories such as threat actors, malware families, and attack techniques. The training protocol utilizes batch-wise gradient descent optimization with iterative refinement, where annotation coverage is progressively expanded and models retrained until achieving F1 scores of approximately 80% across all entity types.

Performance Assessment: Model evaluation employs standard validation protocols using held-out test data not utilized during training phases. The comprehensive assessment reveals strong performance across most entity categories, with precision and recall metrics indicating robust generalization capabilities for cybersecurity text analysis.

Expert Validation Study: To complement automated metrics, an additional validation phase employed two cybersecurity domain experts to manually assess model performance

Category	Tag	Precision	Recall	F1
Attacker	ATTACKER_TYPE	100.00	83.33	90.11
	ATTACKER_NAME	95.29	87.10	91.01
	ATTACKER_ORIGIN	Used Native spaCy LOC tag (no training)		
Attack	MALWARE_TYPE	80.56	76.32	78.38
	MALWARE_NAME	95.29	87.10	91.01
	ATTACK_TYPE (TECHNIQUE)	88.60	87.07	87.83
	VULNERABILITY	87.50	84.00	85.71
Victim	SECTOR	85.84	84.07	84.95
	ASSETS	87.02	89.06	88.03
	TECHNOLOGY	87.60	89.93	88.70

Table 4.8: NER model performance metrics across entity categories.

on 100 previously unseen articles. This empirical evaluation protocol categorizes prediction accuracy using three criteria: **Hit** represents correct tag assignment, **Partial** indicates correct identification with incomplete boundary detection, and **Miss** denotes incorrect or absent tag assignment where entities should be identified.

Performance Analysis and Limitations: The expert validation reveals significant discrepancies between automated F1 scores and real-world performance for several entity types, particularly affecting attacker and malware name recognition. The identification of threat actor names presents inherent challenges due to the dynamic and often ambiguous nature of cybersecurity nomenclature, where attackers frequently employ aliases or operate under multiple identities. Similarly, malware naming conventions lack standardization across threat intelligence sources, creating inconsistencies in entity boundary detection.

Geographic attribution poses additional complexities, as the generic spaCy location model identifies geographical entities without distinguishing between attacker origins, victim locations, or contextually irrelevant geographic references. This limitation necessitates domain-specific training to accurately associate locations with threat actor attribution rather than incident impact zones.

The vulnerability detection model demonstrates particular weakness in recognizing standardized CVE (Common Vulnerabilities and Exposures) identifiers due to their specific alphanumeric format requirements. This limitation can be addressed through hybrid approaches combining NER predictions with regular expression patterns specifically designed to cap-

Category	TAG	Reviewer 1			Reviewer 2			Average			F1
	SCORE TYPE	H	P	M	H	P	M	H	P	M	
Attacker	ATTACKER_-TYPE	90	5	5	90	5	5	90	5	5	90.11
	ATTACKER_-NAME	63	20	17	59	24	17	61	22	17	91.01
	ATTACKER_ORI-GIN	70	25	15	65	30	15	67.5	17.5	15	98
	MALWARE_-TYPE	82	11	7	80	12	8	81	11.5	7.5	78.38
Attack	MALWARE_-NAME	72	14	24	71	14	25	71.5	14	24.5	91.01
	ATTACK_TYPE (TECHNIQUE)	84	11	5	86	9	5	85	10	5	87.83
	VULNERABILITY	75	10	15	79	10	15	77	10	15	85.71
Victim	SECTOR	84	16	0	86	13	1	85	14.5	0.5	84.95
	ASSETS	90	7	3	90	8	2	90	7.5	2.5	88.03
	TECHNOLOGY	90	8	2	86	12	2	88	10	2	88.70

Table 4.9: Expert evaluation comparison with automated metrics (H: Hit, P: Partial, M: Miss).

ture CVE formatting standards, ensuring comprehensive vulnerability identification across diverse threat intelligence sources.

4.4.1.4 Knowledge DB (KDb)

The full MLCESO pipeline aims to generate a usable pool of enriched incidents that can be utilized by other modules later. A set of texts was, for this reason, processed to be used as a baseline. The sources in Table 4.10 were utilized to create the KDb.

Calculating Text Relevance

To ensure relevance, a threshold system was introduced to evaluate the maturity of the parsed articles and NER extracted tags. The scoring system, ranging from 0 to 150, is shown in Algorithm 1. In the implementation, a threshold of 50 was set to consider a text relevant for representing a standalone incident in AiCEF.

Source	Count
bleepingcomputer.com	2431
securityaffairs.co	339
zdnet.com	794
databreaches.net	1393
Total	**4957**

Table 4.10: Knowledge DB content per source

Other techniques to filter out less mature or irrelevant content were also explored as part of this thesis. More specifically, topic modelling with unsupervised machine learning using Spacy for natural language processing and Gensim for topic extraction was utilized.

The whole process can be summarized as follows:

- Step 1: Preprocessing the full corpus of documents by tokenizing and generating a Bag of Words (BoW) representation
- Step 2: Applying Latent Dirichlet Allocation (LDA) to extract and save topics
- Step 3: Visualize results through an intertopic distance map and a ranking of salient terms. (Figure 4.9)
- Step 4: Check if a given text belongs to any of the topics extracted based on the terms included by calculating salient distance.

In a limited experiment, I have analysed one hundred (100) articles, ninety (90) of which were not cybersecurity related and ten (10) being related to a cybersecurity incidents. The results reveal high classification accuracy, achieving **91.11%** for the non-security related documents and **100%** for the security related documents.

Repeating the same experiment using the exact same test document but with a threshold system revealed better scoring, achieving **98.88%** for the non-security-related documents, and I opted to settle for using the threshold system.

Text Enhancements

Two types of enhancements were applied to improve the automatically NER exported tags, namely Regular Expression (REGEX), which is a sequence of characters that defines a

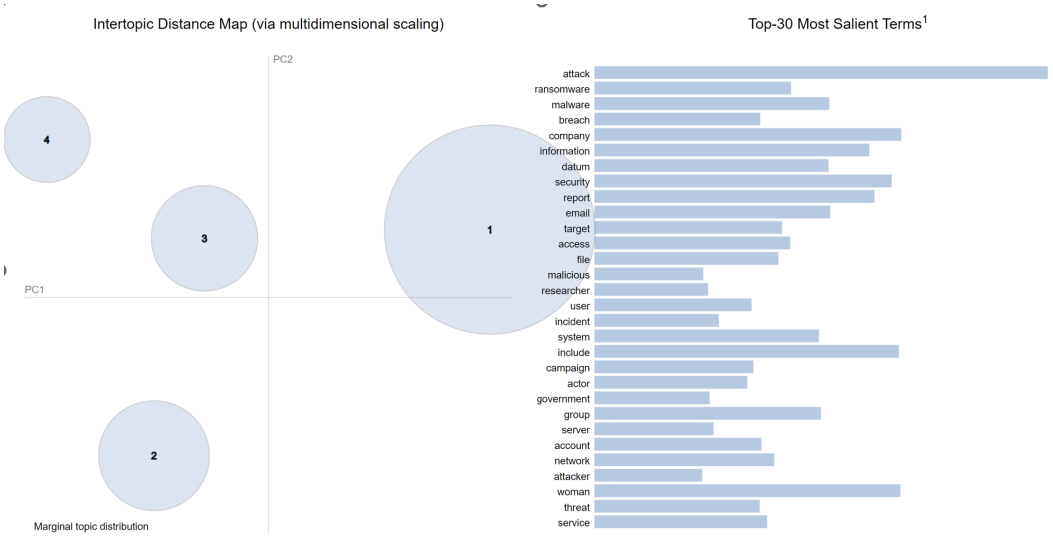


Figure 4.9: LDA topic visualization and top 30 terms.

search pattern, and Hard-coded groups of Strings. Thus, the following tags have been further enhanced:

- Attackers Name: NER + Hardcoded Groups of Strings from MITRE APT list [38]
- Attackers Origin: No NER, Hardcoded Groups of Strings,
- Malware Name: NER + Hardcoded Groups of Strings from MITRE APT list,
- Technique: NER + Hardcoded Groups of Strings from MITRE APT list,
- Vulnerability: NER + CVE REGEX.

These enhancements significantly improved tag detection rates, achieving almost 99% accuracy for vulnerability identification. Based on analysis of the most prominent extracted tags, the tag groups in Table 4.13 were assigned to training topics meta-tags to help categorize text for later use in exercise scenario building. To visualize the training topic extraction capabilities, articles were categorized by training topics and objectives per month for 2021 (Figure 4.11). Each analyzed article can receive multiple training objective tags when incidents span different domains. Tag clustering techniques could be implemented in the future to improve topic extraction quality by identifying relationships between entities and reducing classification ambiguity. The MLCESO pipeline generates comprehensive reports that

Algorithm 1 Compute the maturity of a parsed text.

Require: Set of Tags T
  $maturity \leftarrow 0$
  if $Attacker_Type \in T$ OR $Attack_Type \in T$ **then**
    $maturity \leftarrow 50$
    if $Vulnerability \in T$ **then** $maturity \leftarrow maturity + 15$
    else $maturity \leftarrow maturity - 10$
    end if
    if $Malware \in T$ **then** $maturity \leftarrow maturity + 15$
    else $maturity \leftarrow maturity - 10$
    end if
    if $Attack_Type \in T$ **then**
      $maturity \leftarrow maturity + 15$
      if $Attacker_Type \in T$ **then**
        $maturity \leftarrow maturity + 50$
        if $Technology \in T$ **then** $maturity \leftarrow maturity + 10$
        end if
        if $Sector \in T$ **then** $maturity \leftarrow maturity + 10$
        end if
        if $Assets \in T$ **then** $maturity \leftarrow maturity + 10$
        end if
        if $Attackers_Origin \in T$ **then** $maturity \leftarrow maturity + 10$
        end if
      end if
    end if
  end if
  return $maturity$

transform cybersecurity articles into structured intelligence for exercise development. A visualization of the generated report is shown in Figure 4.10, while a complete example of an analyzed article stored in the Knowledge Database is presented in Table 4.11.

4.5 Layer 2: Trend analysis & prediction

4.5.1 Trend Prediction Module (MLTP)

The Trend Prediction Module provides EP with valuable information by analysing the KDb and extracting trends based on predetermined Training Topics (Table 4.13) to generate a

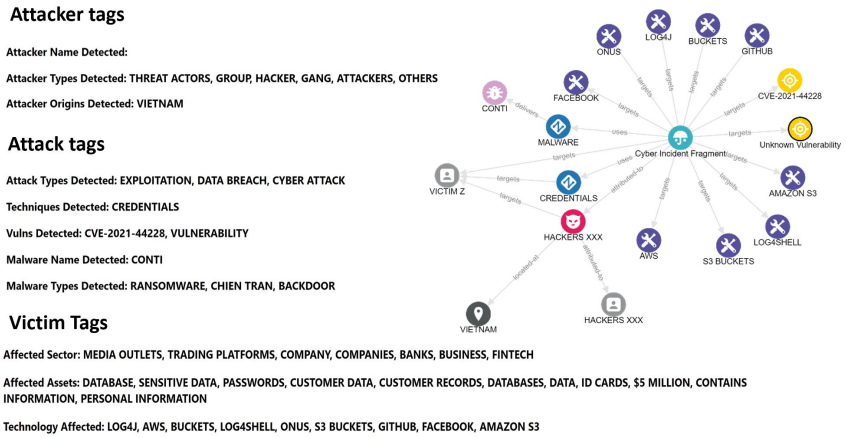


Figure 4.10: Incident Generation report and visualisation.

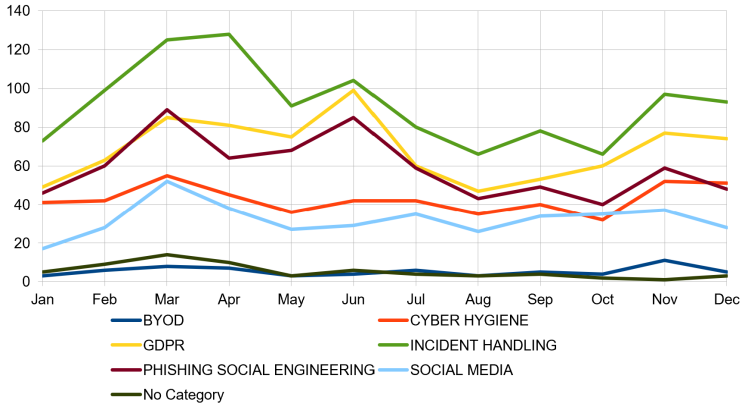


Figure 4.11: Training Topics/Objective per month for 2021.

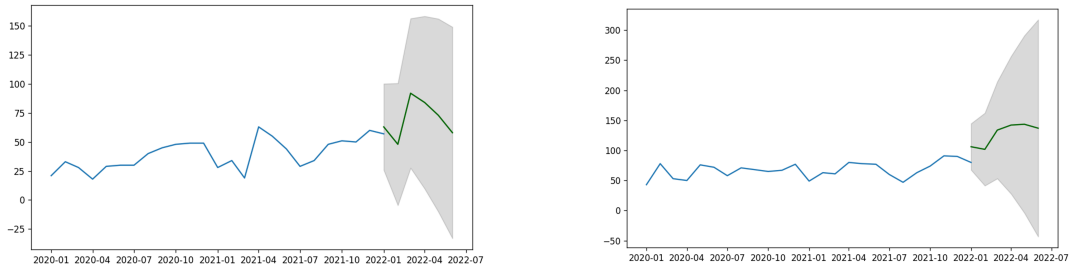
report. The MLTP process consists of three steps:

1. Receiving as input any subset of incidents placed in KDb.
2. Extracting incident statistics based on specified sector and Training Topics
3. Performing Time Series Analysis to plot and calculate future trends for a specific Attack Type and/or Training Topic (Figure 4.12).

In my initial implementation, I chose the SARIMA[7] equation to represent the trends in the

[7]SARIMA is Seasonal ARIMA, or simply put, ARIMA with a seasonal component. ARIMA is a statistical analysis model that uses time series data to predict future trends

existing KDb of 2970 articles as represented in Table 4.10.



(a) Forecast for Ransomware Attack Type (b) Forecast for GDPR Training Topic

Figure 4.12: Six (6) month forecast using SARIMA

Further exploration of methods to boost the capabilities of MLTP can be found in PART III (Chapter 7: Optimizing & Integrating AI-driven Threat Intelligence and Forecasting in the Cybersecurity Exercise Content Generation Lifecycle).

4.6 Layer 3: Exercise generation

4.6.1 Incident Generation (IncGen)

Incident creation is the most important step of the scenario generation procedure and consists of several steps to achieve maximum customisation (Figure 4.13). It assumes that a KDb is already in place. All of the steps can be automated, generating a variety of Incidents from which a Planner can choose to fit most.

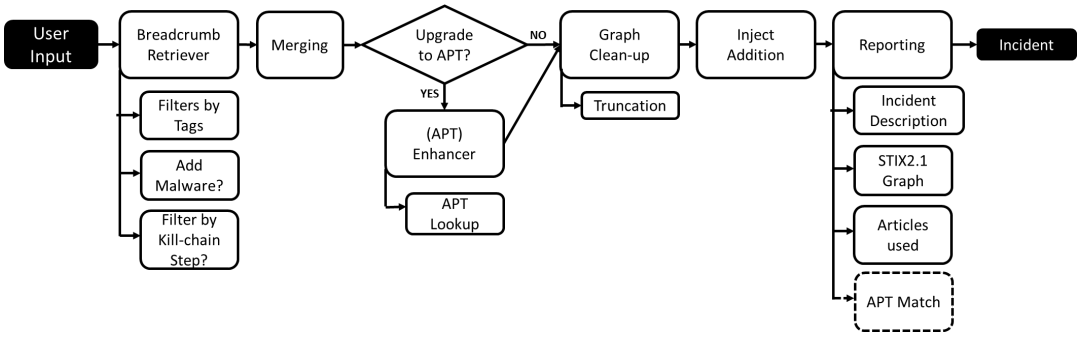


Figure 4.13: The workflow of Incident Generation (IncGen).

The EP can choose to provide specific text or articles for conversion to Incidents or rely

on a dynamic generation based on filtering parameters and a search of the existing KDb. Incidents can be enhanced with activity simulating TTPs of known APT actors using the APT Enhancer module.

An output report (Table 4.12) and visualisation (using stixview[8] library) of IncGen can be seen in Figure 4.14.

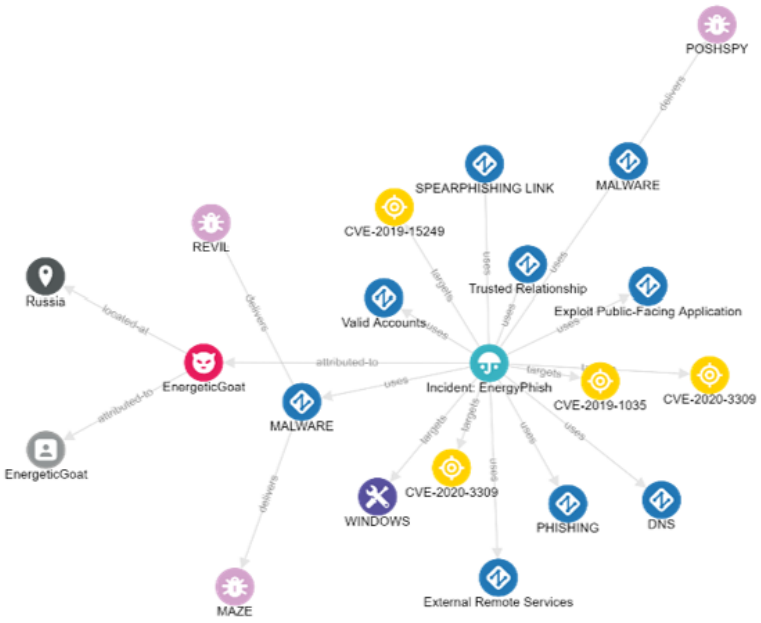


Figure 4.14: Incident graph visualisation (STIX 2.1).

4.6.2 APT Enhancer

To simulate the activity of APT groups, a STIX 2.1 structure was created for each actor using MITRE ATT&CK Groups [9]. Various attributes of the APT group were programmatically extracted from the available STIX 2.1 repository[10] to populate the database as: Name, Description, Origin, Software, Techniques, CVEs, and Malware.

A unique STIX 2.1 graph was also generated per APT group for comparison and en-

[8] https://github.com/traut/stixview
[9] https://attack.mitre.org/groups/
[10] https://github.com/mitre-attack/attack-stix-data/

hancement purposes. An example of a known APT group can be seen in Figure 4.15.

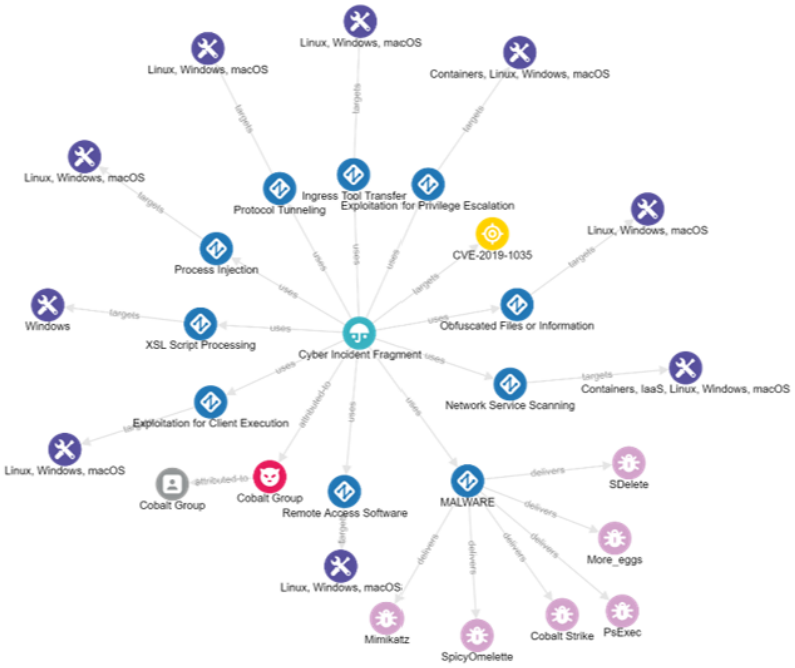


Figure 4.15: Cobalt Group STIX 2.1 graph visualisation

During the incident enhancement, the graph generated after the incident text analysis is compared to all known graphs of APT actors, and the most similar is proposed for enhancement. The similarity score, based on a set of weighted properties and ranging from 0 to 100, is calculated using the STIX 2.1 Python API[11]. In AiCEF, the EP can entirely or partially merge the draft incident graph with that of known APT actors. The EP also has the opportunity to merge the initial graph with that of any APT group of his/her liking.

4.6.3 Cyber Exercise Generation (CEGEN)

The exercise generation flow can be broken down into several steps as seen in (Figure 4.16). The CEGEN module interacts both with the user who provides input in various steps and the IncGEn module recursively to in order to incorporate a number of incidents in the exercise.

[11] https://stix2.readthedocs.io/en/latest

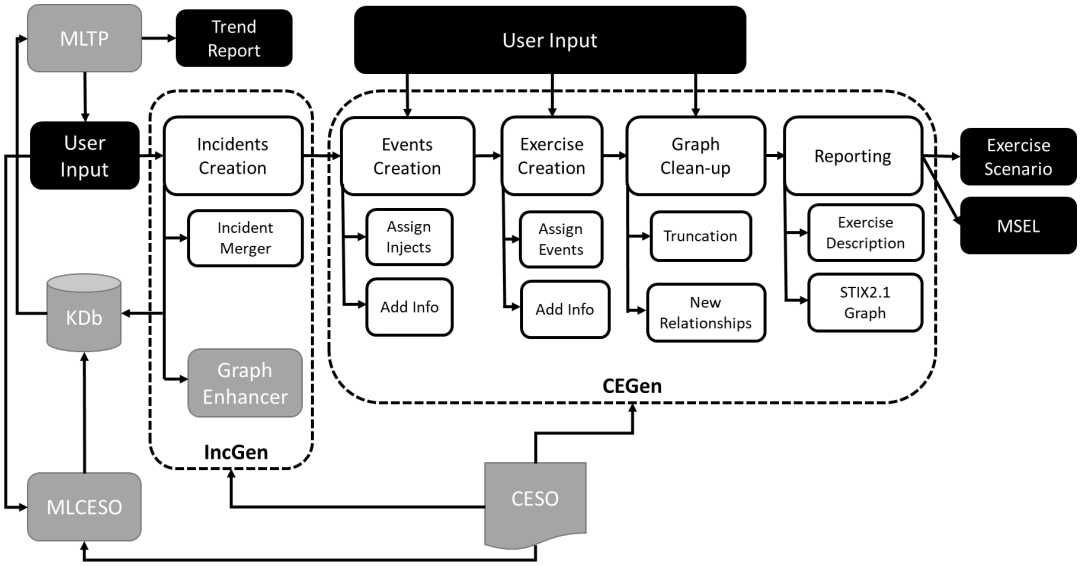


Figure 4.16: CEGen workflow

Using the CEGEN module will lead to the generation of an Exercise Scenario Summary along with a State of the World (SoW) skeleton and the Master Scenario Event List (MSEL) tree skeleton, consisting of the name and description of injects.

4.6.4 Storyline Text Generation

The Storyline Text Generator (STG) creates synthetic text based on predefined input. Initially (2022) Generative Pre-trained Transformer 2 (GPT-2)[12], an AI large-scale unsupervised language model was used to create coherent paragraphs of text from small pieces of text input. I later opted for more modern LLMs like GPT4.o [136] that excel in the task. Prompt examples and outputs can be found in Chapter 5.

4.6.5 Example Use Case

The operational workflow of AiCEF demonstrates its practical application through a systematic exercise generation process that transforms unstructured threat intelligence into comprehensive cybersecurity training scenarios.

[12] https://openai.com/blog/better-language-models/

Initial Data Population: Exercise Planners begin by populating the Knowledge Database (KDb) with cybersecurity incidents relevant to their organizational context and training objectives. The MLCESO pipeline automatically processes these incidents, extracting entities according to the CESO ontology and generating interconnected graph representations that capture threat actor relationships, attack vectors, and target characteristics.

Scenario Initiation: When developing a new cybersecurity exercise, Exercise Planners provide AiCEF with specific keywords reflecting their training focus, such as "ransomware healthcare" or "phishing financial sector." These keywords serve as semantic filters to identify relevant content within the knowledge base.

Trend Analysis Support: AiCEF's trend prediction module analyzes historical incident patterns to generate comprehensive reports highlighting emerging threats and attack methodologies relevant to the specified exercise objectives. This temporal analysis enables Exercise Planners to design scenarios that reflect current threat landscapes and anticipate future attack vectors.

Content Retrieval and Graph Generation: The system performs semantic search across the KDb using the provided keywords, retrieving the most relevant incident articles and their associated entity graphs. These graphs contain structured representations of threat actors, attack patterns, vulnerabilities, and organizational impacts, forming the foundation for scenario development.

Graph Enhancement and Customization: Exercise Planners can enhance the generated incident graphs through two primary mechanisms: merging with known Advanced Persistent Threat (APT) group profiles to incorporate realistic adversary behaviors, and filtering content according to specific Cyber Kill Chain phases to focus training on particular attack stages such as reconnaissance, weaponization, or command and control.

Inject Population and Finalization: The enhanced incident graph provides the structural foundation for inject generation, where specific training stimuli are mapped to graph nodes representing attack events, system compromises, or response triggers. This process transforms the abstract graph representation into actionable exercise content ready for deployment.

The complete incident graph generation workflow, from keyword input through final inject population, is illustrated in Figure 4.17, demonstrating the systematic transformation of raw threat intelligence into structured exercise scenarios.

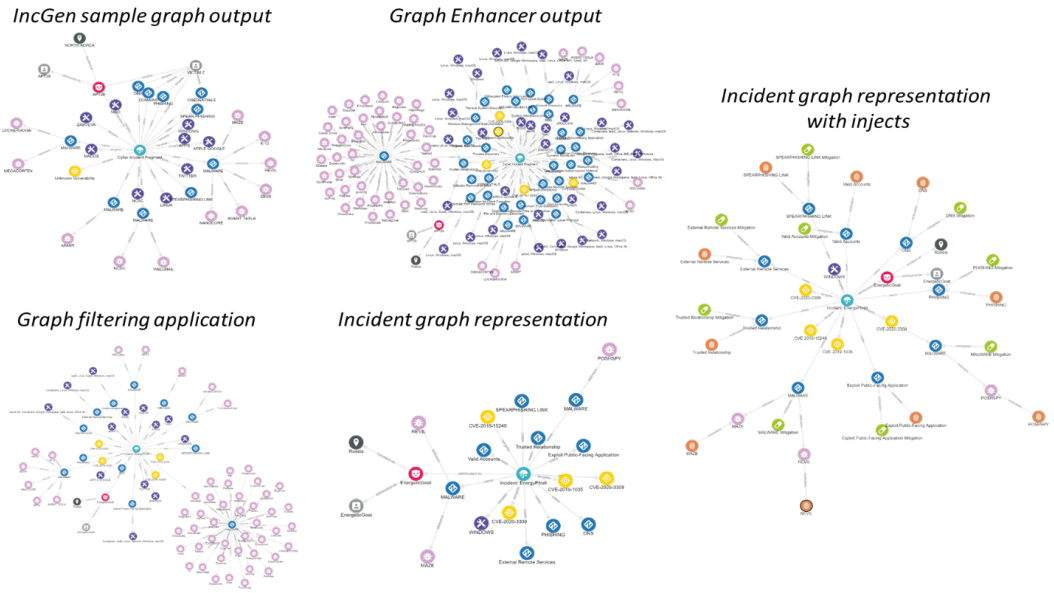


Figure 4.17: IncGen execution flow with intermediate representation steps.

This process is repeated multiple times to generate the number of wanted incidents for a specific CSE. The EP follows the CEGen flow to compile a full exercise and generate a scenario (Figure 4.18) and Exercise graph (Figure 4.19). A fully functional STIX2.1 formatted exercise generated using AiCEF can be found in the Annex II Section 10.4.2.

The whole framework has been programmatically implemented in a single tool to ease the automation of the exercise generation process and is presented in Chapter 9.

Field	Value
id	2246
Name	202203-2021-0153-58be06f8-f18c-4e1b-a243-9ecb13442636
URL	https://www.bleepingcomputer.com/news/security/bluenoroff-hackers-steal-crypto-using-fake-metamask-extension/
Text	BlueNoroff hackers steal crypto using fake MetaMask extension...
Month	1
Year	2022
Maturity	150
MetaTags	LINUX, SCRIPTS, HACKS, POWERSHELL, CRYPTOCURRENCY, THREAT ACTOR, PASSWORD, MALWARE, EXCEL, WORD, USER CREDENTIALS, CREDENTIALS, POWERSHELL, CHROME EXTENSION, WINDOWS, HACKERS, FILES, ATTACKERS, NORTH KOREA, BROWSER, BROWSER EXTENSIONS, CVE-2017-0199, DATA BREACH, TEMPLATE INJECTION, WALLET, FINANCIAL, GROUP, VULNERABILITY, BLUENOROFF, THREAT ACTORS, DATA
Attacker_Name	BLUENOROFF
Attacker_Type	HACKERS, GROUP, THREAT ACTOR
Threat_Actor_Type	crime-syndicate
Attack_Type	DATA BREACH
Technique	CREDENTIALS, POWERSHELL
Vuln	TEMPLATE INJECTION, VULNERABILITY, CVE-2017-0199
Malware_Name	
Malware_Type	SCRIPTS, MALWARE
Attacker_Origin	NORTH KOREA
Sector	FINANCIAL
Assets	FILES, WALLET, USER CREDENTIALS, CRYPTOCURRENCY, PASSWORD, DATA
Technology	WINDOWS, LINUX, EXCEL, WORD, BROWSER, POWERSHELL, BROWSER EXTENSIONS, CHROME EXTENSION

Table 4.11: Annotation example using AiCEF (MLCESO).

Field	Details
Incident Title	EnergyPhish
Description	ENERGYCORP affected by cyber-attacks.
Threat Actor Simulated	APT29
Common Techniques	SPEARPHISHING LINK, DOMAINS, CREDENTIALS, DNS
Kill Chain Steps Identified	• Weaponization • Social Engineering • Initial Access • Reconnaissance • Delivery
Technologies Targeted	WINDOWS
Malware Arsenal	REVIL, MAZE, POSHSPY
Resources Used	• https://www.bleepingcomputer.com/news/security/us-govt-warns-of-ransomware-attacks-on-pipeline-operations/ • https://www.bleepingcomputer.com/news/security/cyber-espionage-campaign-targets-renewable-energy-companies/ • https://www.bleepingcomputer.com/news/security/russian-hackers-target-covid-19-vaccine-research-with-custom-malware/ • https://www.bleepingcomputer.com/news/security/hackers-target-eu-commission-covid-19-cold-chain-supply-orgs/ • https://www.bleepingcomputer.com/news/security/us-shares-info-on-north-korean-malware-used-to-steal-cryptocurrency/ • https://www.bleepingcomputer.com/news/security/us-warns-of-increased-cyberattacks-against-k-12-distance-learning/
Object Name	PHISHSPY.json

Table 4.12: Summary of the EnergyPhish Incident

Training Topic	**Tags**
INCIDENT HANDLING	MALWARE, RANSOMWARE, APT, CYBER, ATTACK, WEBSITE, HACKER, EXPLOIT, ZERO-DAY
GDPR	PRIVACY, DATA LEAKAGE, PERSONAL DATA, EXFILTRATION, CLOUD, SENSITIVE DATA, DATA, GOOGLE DRIVE, AWS, MEDICAL DATA, PASSPORT
CYBER HYGIENE	PASSWORD, ACCOUNT, USERNAME, LOGIN, ACCOUNTS, FILES, CREDENTIALS
PHISHING & SOCIAL ENGINEERING	PHISHING, SCAM, FRAUD, VISHING, IMPERSONATION, BEC, EMAIL, GMAIL
SOCIAL MEDIA	FACEBOOK, TWITTER, LINKEDIN, META, INSTAGRAM
BYOD	MOBILE, ANDROID, IOS, LAPTOP, IOT, GOOGLE PLAY

Table 4.13: Training topics.

State of the World (SoW)

Description: **STG Input**:: '*State sponsored Threat actor EnergeticGoat has historically targeted government organizations, non-government organizations (NGOs), think tanks, military, IT service providers, health technology and research, and telecommunications providers. With this latest attack, EnergeticGoat attempted to target the Energy Sector*'
STG Output: "State sponsored Threat actor EnergeticGoat has historically targeted government organizations, non-government organizations (NGOs), think tanks, military, IT service providers, health technology and research, and telecommunications providers. With this latest attack, EnergeticGoat attempted to target the Energy Sector in South Europe. More specifically the following targets have been identified: 1) EU energy firms that have received investments from the Energysubsidiary, 2) SA government units and departments who receive funding from the Energysubsidiary, 3) the SA energy regulator, 4) the Energy Sector's main lobbying and public relations organization, 5) the Energy Sector's largest customer, the Energysubsidiary, and 6) the Energy Sector's corporate regulator, the Energy Regulator of SA (ERA). Energy Sector entities and their customers and regulators are usually among the least likely organizations to be attacked by EnergeticGoat.
For example, the Energy Regulator of SA (ERA) is not a high-profile organization. It is a largely unregulated regulatory body that provides general oversight of the energy sector and serves as a key regulator and enforcement body of relevant energy laws in EU. Another example would be the Energy Corporation (Zesa). Zesa is a state-funded enterprise that provides essential infrastructure and related services to its citizens, including the electricity, gas, and broadband services that power large swaths in EU.
The attackers may have targeted other Energy Sector actors, such as Eskom or the SA Energy Institute (SAEI). However, it is thought that EnergeticGoat is specifically targeting Zesa and other organizations as a result of the threat actor's perceived inability to target the SA Energy Sector as a whole. EnergeticGoat's choice to target a South African energy firm with its latest attack is indicative of this perception."

Scenario Summary

EXERCISE NAME: Energy Test
EXERCISE OBJECTIVES: Test detection of Attacks and Incident Handling Procedures
EVENT 1: EnergyWipeOut ### Description: **STG Input**:: '*State sponsored Threat Actor EnergeticGoat is believed to be behind cyber-attacks against Energy Corp. The attackers used PHISHING & DATA DESTRUCTION techniques to*'
STG Output: "State sponsored Threat Actor EnergeticGoat is believed to be behind cyber-attacks against Energy Corp. The attackers used PHISHING & DATA DESTRUCTION techniques to infect the system with malware. In addition to the malware used, the attackers also used other tools to get around the security countermeasures of ENERGYCORP. It is understood that it will take at least a few days for the authorities to identify the attackers as they have only just informed the company about the attack. A spokesperson for the company said the company did not have any details as to the motive behind the attack as it was still investigating. "We are working with law enforcement to bring this to the most serious level. The situation is under investigation and we cannot comment any further than that"
*** Incident 1.1: EnergyPhish *** Description: **STG Input**:: '*EnergeticGoat deployed Phishing attacks against Energy Corp. Malicious word documents were sent*'
STG Output:"EnergeticGoat deployed Phishing attacks against Energy Corp. Malicious word documents were sent to Energy Corp. employees in an attempt to obtain personal information. On October 26, 2022, a new employee of Energy Corp., received a suspicious email containing a malicious word document. The email identified as 'malware' was in fact a phishing attack with the subject 'Vendors of Energy Corp. under Federal Investigation"

Figure 4.18: Sample text of an AI-generated exercise.

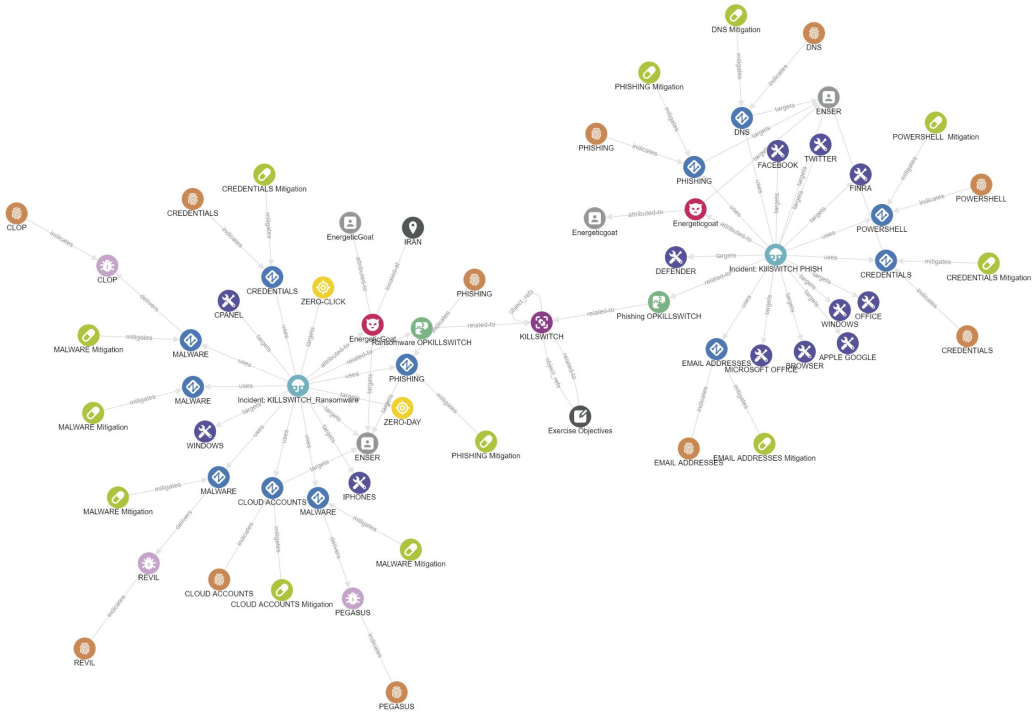


Figure 4.19: Sample exercise graph visualisation

Chapter 5

Evaluation Methodology & Results

To evaluate this work, a case study was developed and while field experts were used to measure the effectiveness of the proposed framework and methodology. This chapter outlines the steps taken to effectively assess the capabilities of this approach.

- **Step 1: Scenario Content Generation:** A group of exercise planners with varying expertise levels was tasked with generating the same exercise scenario using traditional methods and the AiCEF methodology and tools. Their performance was monitored for timeliness, effectiveness, creativity, and the methodology used.
- **Step 2: Content Evaluation:** The generated reports were anonymized and presented to a group of evaluators. These evaluators graded the complexity, technical depth, and richness of lessons learned from the generated subset of exercise scenarios based on the objectives and KPIs set through a questionnaire.
- **Step 3: Results Collection and Analysis:** The results of this process were evaluated against the previously established KPIs to estimate:
  - Improved speed in cyber exercise scenario generation (quantitative) using AiCEF.
  - Enhanced quality in cyber exercise scenario generation (qualitative) for inexperienced planners using AiCEF.
  - Increased relevance of proposed cyber exercise scenarios to the current threat landscape (qualitative) using AiCEF.

5.1 Scenario Content Generation

Four EPs were selected to individually generate a CSE scenario according to the provided high-level exercise requirements and specifications, see Figure 7.11. All EPs have deep knowledge of cybersecurity, and their skill sets resemble that of a CISO. Two EPs were senior cybersecurity exercise planners with more than 10 years of experience in the task. A Scripted Exercise Planner (SEP) was also programmed to be used as part of the evaluation, using different parameters and flows of the AiCEF methodology and toolset.

Generate a CSE scenario for a cyber awareness exercise by filling in a provided Scenario Template. The CSE should include two events consisting of 1 incident each. All incidents should be accompanied by a short description of indicative injects to be sent to players. At least 3 inject descriptions per incident should be provided.
The company which will use the scenario is an **Energy Service Provider** and all its Employees can be potential Players.
The exercise should last between 2-4 hours and can include technical artefacts for analysis.
The two main objectives of the exercise are:

1. Provide awareness to employees regarding Phishing Attacks
2. Provide awareness to employees regarding Ransomware Attacks

These objectives can be updated and more can be added. The Scenario Development task will be timed and should not last more than 4 hours

Figure 5.1: Task definition

The EPs were briefly introduced to the basics of developing CSE scenarios and were provided with a course on using AiCEF and the accompanying tools. All EPs were provided with the same Scenario Template (ST) to fill in as an output of their task.

The provided ST had the following generic structure:

- Section 1: Storyline (SoW)
- Section 2: Scenario & MSEL
- Section 3: Scenario Analysis
- Section 4: Resources Used

Detailed instructions were provided on the expected content per paragraph to all involved planners to streamline the information of the generated reports and create homogeneous outputs to be evaluated in the later step.

As a result, five complete exercise scenarios were generated, as shown in Table 5.1, based on the following grouping of EPs:

- Group 1: Experienced EP
- Group 2: Novice EP
- Group 3: Experienced EP supported by AiCEF SEP
- Group 4: Novice EP supported by AiCEF SEP
- Group 5: AiCEF SEP

Eval_Tag	**Explanatory Name Tag**	**Exercise Expertise**	**AiCEF**	**Duration**	**TAGS**
ExSc1	Sc1:Exp(erienced)Hum(an)	YES	NO	2h 00min	Google, MITRE
ExSc2	Sc2:Nov(ice)Hum(an)	NO	NO	2h 35min	Google
ExSc3	Sc3:Exp(erienced)Hum(an)&AI	YES	YES	1h 20min	Online Resources
ExSc4	Sc4:Nov(ice)Hum(an)&AI	NO	YES	2h 10 min	Google
ExSc5	Sc5:AiCEF	N/A	YES	20min	-

Table 5.1: Details of the generated scenarios.

5.2 Scenario Content Evaluation

To evaluate the generated scenarios, an anonymous online survey was conducted from September 1 to September 30, 2022, targeting cybersecurity professionals with diverse expertise and backgrounds. The evaluation process deliberately recruited participants from different cyber awareness and exercise planning groups across various sectors and nationalities to minimize

potential bias. The study specifically engaged the ENISA Ad-hoc Cyber Awareness Expert Group, achieving a response rate of 66% with 16 expert participants from the invited pool. This sample represents significant domain expertise, with participants selected through independent criteria established by ENISA rather than the research team, thereby reducing selection bias and enhancing evaluation credibility.

The expert demographics reveal substantial diversity across multiple dimensions (Table 7.6). Geographic representation spans twelve European countries, with participants distributed across critical sectors including government (6), ICT (4), energy (2), and specialized domains such as law enforcement and critical infrastructure. Experience levels range from novice practitioners handling small-scale exercises to senior professionals managing large, cross-border European exercises, providing comprehensive coverage of the exercise planning community.

The evaluation instrument comprised 11 questions organized into three distinct assessment categories. Eight questions focused on scenario quality evaluation using standardized metrics for relevance, complexity, technical depth, and usability. Two questions implemented a Turing test methodology to assess whether evaluators could distinguish between human-generated and AI-assisted content. Additional questions captured demographic information and gathered feedback for future system improvements. All scenarios were presented using anonymous "Eval_Tag" identifiers without revealing authorship or generation methodology, ensuring unbiased evaluation.

The scenario assessment framework employed eight evaluation criteria with corresponding scoring ranges detailed in Table 5.3a. These metrics encompass critical aspects of exercise scenario quality, including objective alignment, technical accuracy, and practical usability for training purposes.

5.2.1 Results Analysis

The comprehensive evaluation reveals significant performance advantages for AI-assisted scenario generation approaches, providing valuable insights into AiCEF's capabilities and areas for improvement while offering a deeper understanding of contemporary exercise planning methodologies.

The quantitative analysis demonstrates clear superiority of hybrid human-AI approaches,

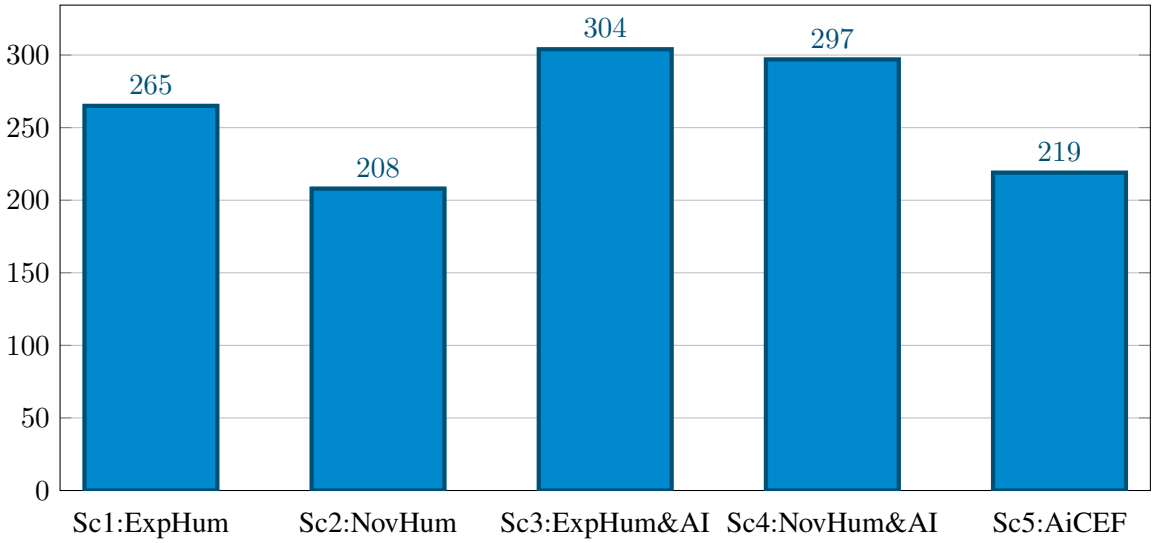


Figure 5.2: Overall performance of evaluated scenarios based on total score.

with both AI-assisted scenarios (Sc3:ExpHum&AI and Sc4:NovHum&AI) achieving the highest overall scores (Figure 5.2). The hybrid methodology enables even novice Exercise Planners to outperform experienced professionals using traditional methods, indicating substantial capability enhancement through AI assistance. Remarkably, the fully automated AiCEF-generated scenario (Sc5:AiCEF) achieved performance comparable to novice human planners, suggesting significant potential for autonomous scenario generation.

Detailed performance analysis across individual evaluation criteria reveals specific strengths and limitations of the AiCEF approach. The automated system excelled in resource relevance, event alignment, and technical depth, demonstrating strong capability for generating technically accurate and well-sourced content. However, limitations emerged in threat actor characterization, scenario complexity, and incident-objective alignment, areas where raw AI-generated content occasionally contained inconsistencies or contextual misalignments.

The human-AI collaborative approach addresses these limitations effectively, with expert curation enabling substantial improvements in content quality and coherence. Most significantly, AI-assisted experts completed scenario development 33% faster than traditional methods while achieving superior quality scores, demonstrating both efficiency and effectiveness gains.

The Turing test results provide compelling evidence of AI content quality, with evaluators

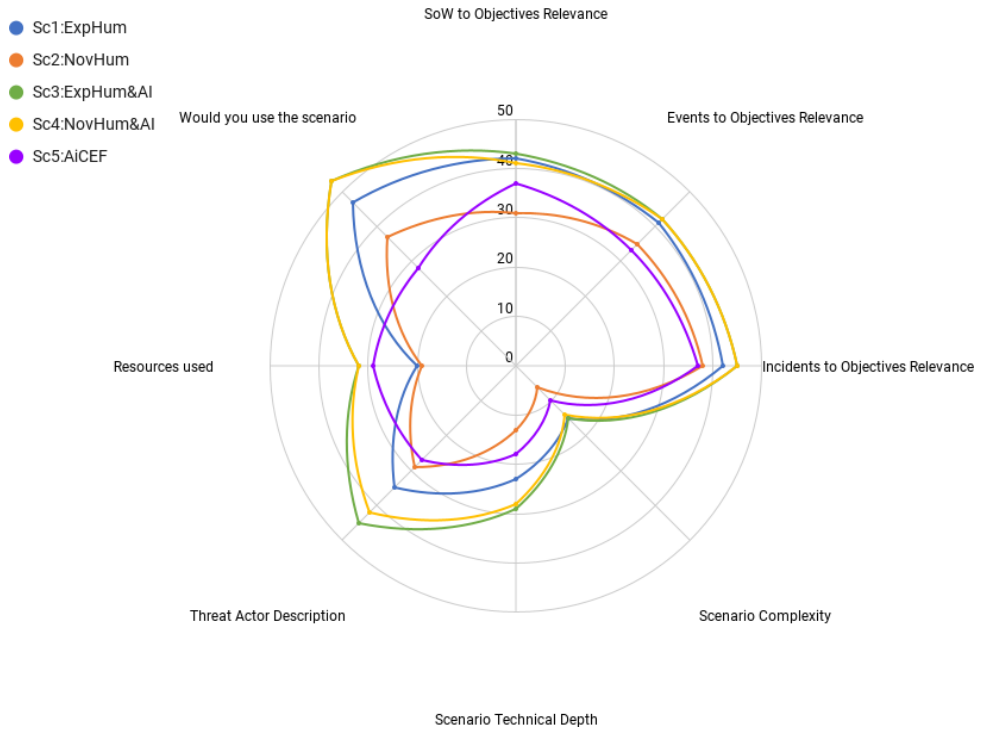
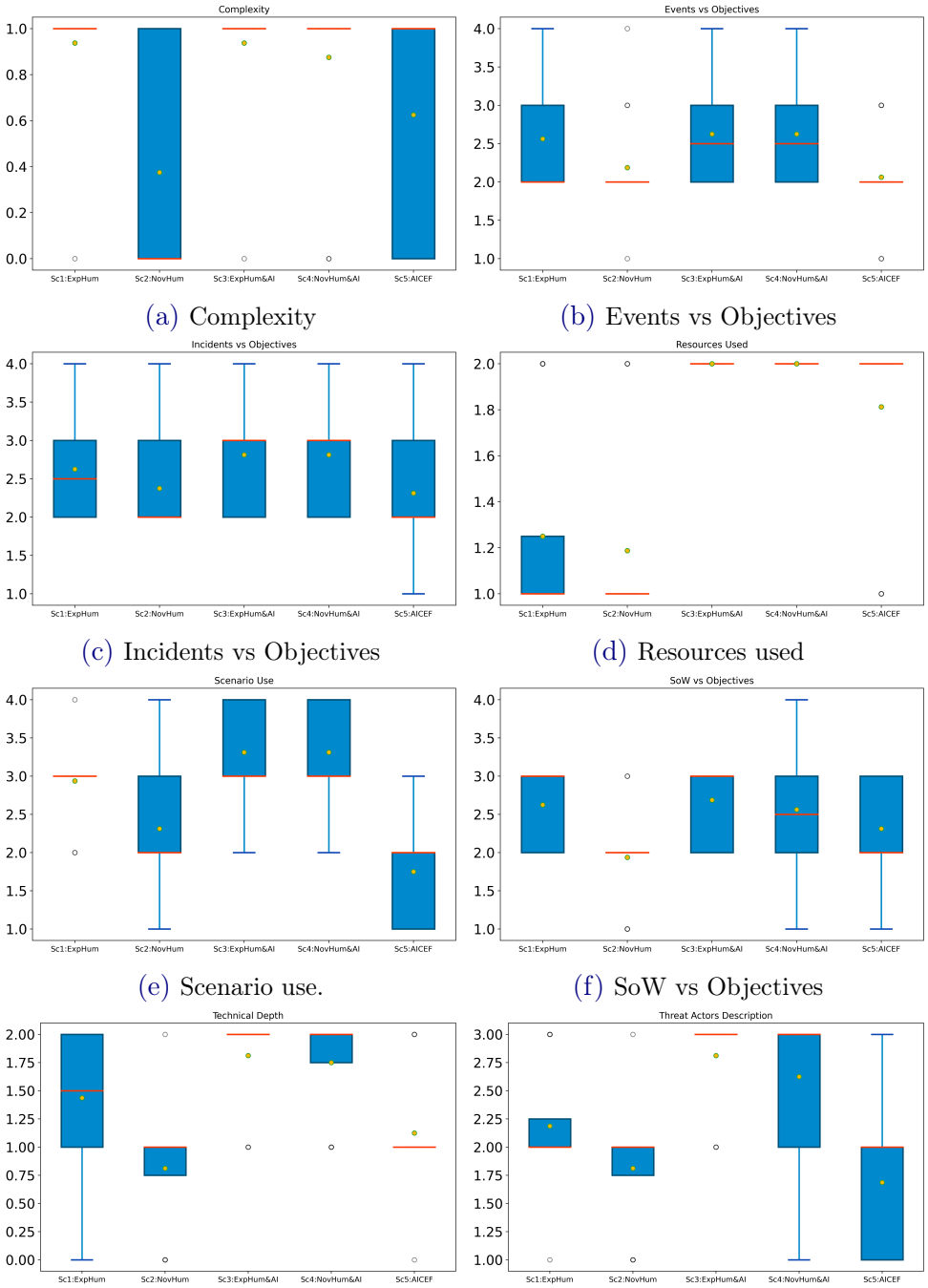


Figure 5.3: Scenario evaluation parameters

frequently unable to distinguish purely AI-generated scenarios from human-created content (Table 5.4). This finding suggests that automated generation capabilities approach human-level quality perception, validating the technical approach while highlighting opportunities for further refinement.

Expert feedback on exercise planning methodologies reveals valuable insights into current practices and tool requirements. When defining exercise scope and objectives, practitioners primarily rely on known incident analysis and risk assessment (Figure 5.5), while scenario content development draws heavily from news sources and technical reports. These findings align well with AiCEF's design philosophy of leveraging contemporary threat intelligence for scenario generation.

The economic impact assessment reveals substantial potential for efficiency improvements, with experts reporting an average of 53 hours invested in scenario content development using traditional methods. This baseline establishes clear targets for AI-assisted tools to demonstrate value through time reduction while maintaining or improving quality standards. Current tool usage patterns indicate reliance on basic resources such as web search engines,



(a) Complexity

(b) Events vs Objectives

(c) Incidents vs Objectives

(d) Resources used

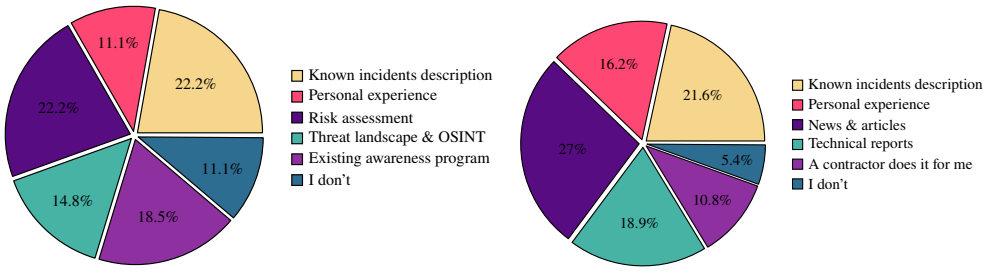
(e) Scenario use.

(f) SoW vs Objectives

(g) Technical depth

(h) Threat actors' description

Figure 5.4: Score range for the Q1-8 of 16 Evaluators.



(a) Experts' responses to "How do you define the scope/objectives of the exercise?" question.

(b) Experts' responses to "How do you define the scenario content?" question.

Figure 5.5: Expert feedback on exercise planning methodologies.

cybersecurity news websites, and general productivity software, suggesting significant opportunities for specialized AI-powered exercise generation platforms.

Countries	#
Greece	2
Austria	1
Italy	1
Belgium	2
Poland	1
Spain	1
Romania	1
Portugal	1
Czechia	1
Netherlands	2
France	2
Finland	1

(a) Countries of origin of the experts

Sector	#
Governmental	6
Energy	2
ICT	4
Critical Infra.	1
Law enforcement	1
Education	1
Other	1

(b) Sector that experts are working in

Seniority	#
Novice (small exercises)	5
Medium (medium scale exercises for a few years)	3
Expert (EU level, cross-country exercises)	4
Senior (large-sized exercises)	1
None	3

(c) Seniority self-assessment.

Table 5.2: Demographics of the experts.

Table 5.3: Scenario Evaluation and Expert Feedback

(a) Scenario Evaluation Questions and Scores

Evaluation Question	Score (Range)
How do you evaluate the relevance of the State of the World text to the Objectives of the Exercise?	(0–4)
How do you evaluate the relevance of the selected Events to the Objectives of the exercise?	(0–4)
How do you evaluate the relevance of the selected Incidents to the Objectives of the exercise?	(0–4)
How do you evaluate the Complexity of the Scenario?	(0–1)
How do you evaluate the Technical Depth of the Scenario?	(0–2)
How do you evaluate the Threat Actor's description?	(1–3)
How do you evaluate the used resources?	(0–2)
Would you be willing to use this Scenario based on the task description?	(0–4)

(b) Expert Feedback Questions

Question
How was the scenario generated?
How skilled was the planner?
How much time did you invest in the Scenario Content Development?
How do you define the scope/objectives of the exercise?
How do you define the scenario content?
What tools did you use to create the scenario or define the objectives, if any?

Scenario	Human	AiCEF & Human	AiCEF
Sc1:ExpHum	1	13	2
Sc2:NovHum	9	2	5
Sc3:ExpHum&AI	0	9	7
Sc4:NovHum&AI	1	10	5
Sc5:AiCEF	6	5	5

Table 5.4: Turing test results for scenario generation methodology identification.

Chapter 6

Use case: Modelling a Cyber Conflict

In order to explore the relevance and timeliness of generated scenarios using AiCEF based on incidents that comprise a real and recent conflict, a unique use case was devised. Modern conflicts are combined with electronic warfare, disinformation campaigns, anti-satellite attacks, and precision-guided munitions based on freshly harvested information from social media. Cyber operations can also be used for political impact by disrupting finance, energy, transportation, and government services [152]. This multifaceted threat landscape creates significant challenges for cyber exercise planners, particularly those with limited experience or subject matter expertise, in developing scenarios that are both realistic and pedagogically effective. By grounding scenario generation in real conflict data, this use case addresses the critical need for more authentic and relevant cybersecurity training that reflects the actual threat environment practitioners will encounter.

6.1 Measuring Effectiveness

To measure the effectiveness of AI-generated cybersecurity incidents for use in CSE scenarios, sets of incidents that occurred before and during a cyber conflict (Pre-Conflict and Post-Conflict sets) are collected and divided into sector-specific pools. The Pre-Conflict pools are then compared to the actual incidents that took place (covered in the Post-Conflict pool) to evaluate the value of the methodology and determine the best strategy for generating CSE

incidents with AI.

As *"Conflict"* one can consider a fight between armed troops leading to war issued officially by governments/states.

My approach creates distinct incident pools with varying levels of sophistication and targeting. Pre-Conflict pools include Pool 0, containing randomly selected incidents from pre-conflict reports, Pool 1 with relevant and targeted incidents from pre-conflict publications, Pool 2 featuring merged relevant reports to create richer incidents, and Pool 3 containing merged reports enhanced with known threat actor tactics, techniques, and procedures. The Post-Conflict pool consists of two targeted incidents generated from conflict-period reports, plus one merged incident combining multiple relevant sources.

6.1.1 PHASE 1: Incident Modelling (Pre-Conflict)

This phase focuses on collecting and processing incidents that occurred before the conflict began. First, relevant pre-conflict incidents are identified from various sources, and their text is extracted. The extracted text is then parsed using MLCESO and stored in the knowledge database with PRE-CONFLICT meta tags. A trend analysis is performed on the parsed incidents to identify the most prominent attack types, techniques, and threat actors, filtered by affected sectors. Finally, incidents are extracted from the database or synthetically generated to form four different pools per sector, each with varying levels of sophistication ranging from basic random selection to enhanced incidents incorporating known threat actor behaviors.

6.1.2 PHASE 2: Incident Modelling (Post-Conflict)

This phase handles the collection and processing of incidents that occurred during the actual conflict. Post-conflict incidents are identified from various resources, and their text is extracted. These incidents are parsed using MLCESO and stored in the knowledge database with POST-CONFLICT meta tags. The phase concludes by generating a pool of incidents for each affected sector, including two high-quality individual incidents and one comprehensive merged incident that combines multiple relevant sources.

6.1.3 PHASE 3: Comparison

This phase calculates the average coverage percentage of Post-Conflict pool versus the Pre-Conflict pool sets to identify the best strategy in CSE creation using AiCEF. The comparison function extracts object attributes, including attack patterns, tools, vulnerabilities, malware, and location names, from both Pre-Conflict and Post-Conflict incidents. Coverage is calculated through a systematic comparison between incident pairs, measuring how well Pre-Conflict pools predict Post-Conflict incidents. Total coverage combines individual coverage scores for each object type, while average total coverage provides the overall effectiveness measure across all incident comparisons.

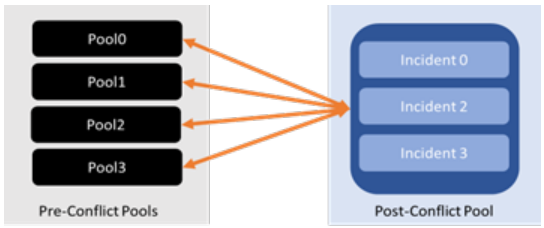


Figure 6.1: Scenario evaluation parameters

The comparison function used to calculate coverage between modelled incidents can be summarised as follows:

$$\text{TotalCoverage}(i) = \frac{\text{AttackPattern_Coverage} + \text{Tool_Coverage} + \text{Vulnerability_Coverage} + \text{Malware_Coverage} + \text{Location_Coverage}}{5} \quad (6.1)$$

Where an example coverage of a Vulnerability Object between two incidents can be calculated as:

$$\text{Coverage} = \frac{\text{Number of PreConflictPool}(i) \cdot \text{Inc0[Vulnerability] list items in PostConflictPool}(i)}{\text{PostConflictPool}(i) \cdot \text{Inc0[Vulnerability] list size}} \times 100 \quad (6.2)$$

Then, the AverageTotalCoverage of all `PreConflictPool` Incidents per `PostConflictPool` Incident is calculated as:

$$\text{AverageTotalCoverage} = \frac{\sum_{i=0}^{n} \text{TotalCoverage}(i)}{n} \quad (6.3)$$

6.2 Use-case Implementation

The Russian-Ukrainian cyber conflict was selected as a test case to assess whether generated scenarios match actual attack trends and demonstrate AiCEF's ability to provide targeted cybersecurity awareness training.

The research analyzed cyber incidents from this conflict using historical data modelled into Pre-Conflict and Post-Conflict incident sets. These sets contained verified cyber attacks against Ukraine documented in publications and specialized security sources. Each entry was manually reviewed for relevance before inclusion, resulting in 46 articles for the Pre-Conflict set and 14 for the Post-Conflict set, classified across three key sectors, namely Public Administration, Energy, and ICT.

The Pre-Conflict incidents occurred between January 2020 and January 2022, with the conflict's official start date considered February 24, 2022. No February 2022 incidents were included in the pre-conflict analysis to maintain clear temporal separation.

Two exercise scenarios were developed for comparison purposes: one based on the Pre-Conflict incident set and another on the Post-Conflict set. This approach enables evaluation of the training coverage achieved by each method.

The trend analysis of the parsed incidents revealed interesting patterns in attack types, techniques, and threat actors across the affected sectors, as shown in Table 6.1.

For actors such as APT28, Sandworm, Turla, and APT29, the findings were verified against the CCDCOE study of 2021 [152] and for ACTINIUM, against Microsoft's January 2022 report [153].

The pools of the following size were then formatted per sector ([Pool0(40), Pool1(6), Pool2(40), Pool3(40)]) with 40 articles generated per Pool2-3 using AiCEF.

As part of Phase 2 of modelling Post-Conflict incidents, 14 Articles have been selected covering the three sectors of interest for the period of March 2022 - June 2022.

Two, rich in objects, incidents (Incident1-2) were generated per Sector. Both were merged into a single Incident (Incident0). After performing all possible comparisons in phase 3, both at the incident and exercise level, the coverage results were retrieved and are presented in the

Sector	Attack Types & Techniques	Threat Actor
Public Administration	DOS, DDOS, MALWARE, DATA EXFILTRATION, PHISHING, DEFACEMENT	APT28, SANDWORM, TURLA, APT29
Energy	WIPER, RANSOMWARE, ESPIONAGE	STRONTIUM, KRYPTON, GAMAREDON, BLACKENERGY, TURLA, FANCY BEAR, SOFACY, NOBELIUM, UNC2452, PAWN STORM, SEDNIT
ICT	DOS, DDOS, WIPER, MALWARE	ACTINIUM, NOBELIUM, STRONTIUM

Table 6.1: Sector based Attack Types, Techniques, and Threat Actors

next section.

6.3 Results & Analysis

The results are broken down based on the affected sector, to help us analyze more granularly the effectiveness of the approach and pinpoint the most effective scenario generation strategy.

6.3.1 Public Administration

For the Public Administration Sector, which was the most impacted in the conflict examined in terms of attack volume, I was able to determine the following, as seen in Figure 6.2.

- Pre-Conflict Pool 2 Incidents covered with an improved rate of an average of 20% of the three Post-Conflict Incidents selected. Overall, the coverage achieved was close

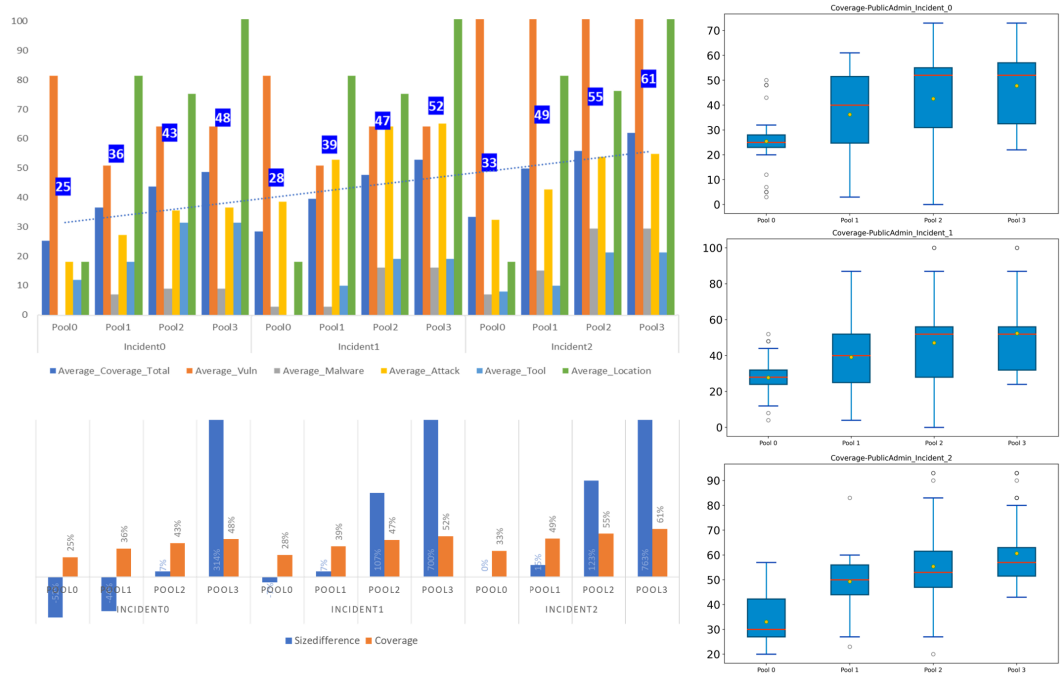


Figure 6.2: Public Administration Incidents Coverage Results.

to 50%, meaning that attacks used during the conflict were not similar to the ones performed before the conflict.

- APT enhancement considerably improved the coverage, but the complexity of the Incidents and the overall exercise generated rose drastically, providing little training benefit.

6.3.2 Energy Sector

The Energy Sector is considered the most critical to target during conflict. Taking its importance into account, I concluded that sufficient training coverage was achieved. More specifically and as depicted in Figure 6.3:

- Pre-Conflict Pool 2 Incidents covered, with an improved rate of an average of 17%, the three Post-Conflict Incidents. Overall, the coverage achieved for Pool 2 was close to 74%, meaning that attacks used during the conflict were similar to the ones performed

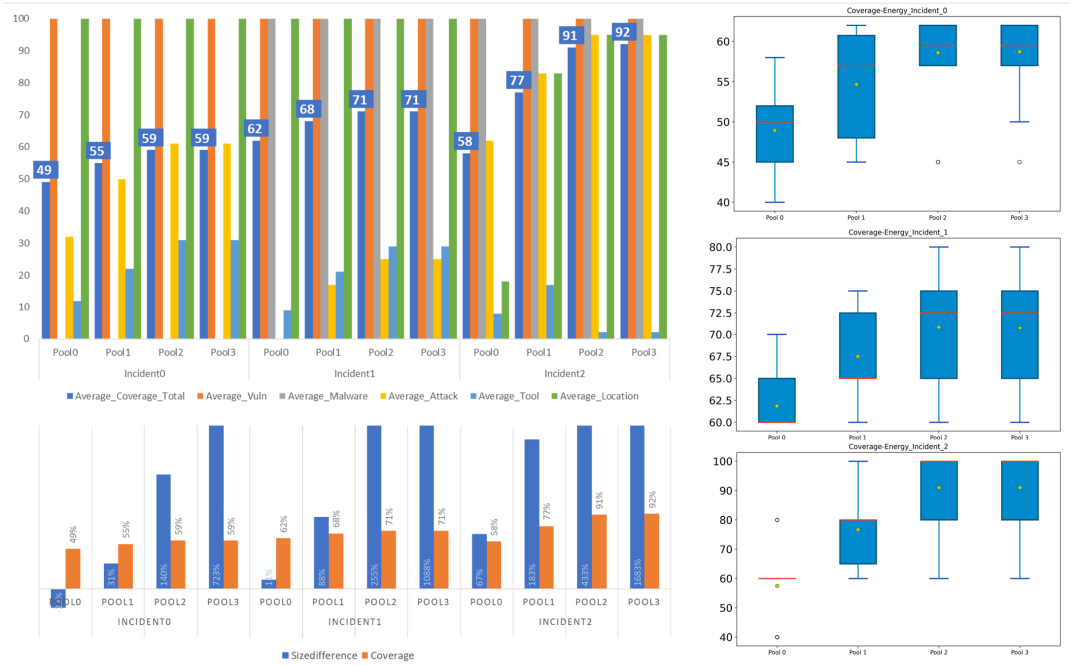


Figure 6.3: Energy Sector Incidents Coverage Results.

before the conflict, providing a good space for training and exercises to improve awareness.

- Although APT enhancement did not improve the coverage generated, incidents in both Pools 2 and 3 managed to cover 100% for the Post-Conflict Incident 2 with a constant rate of over 60% for all 40 incidents generated.

- Malware prediction, although low in the automatic statistics extracted, seen below, can be considered satisfactory upon manual analysis. As an example, AiCEF proposed INDUSTROYER malware for exercise, whereas in reality, Post-Conflict Incidents in the energy sector were populated by INDUSTROYER v.2.0, a malware of the same family.

6.3.3 ICT Sector

For the ICT Sector (Figure 6.4), Pre-Conflict Pool 2 Incidents covered, with an improved rate of an average of 26%, the three Post-Conflict Incidents.

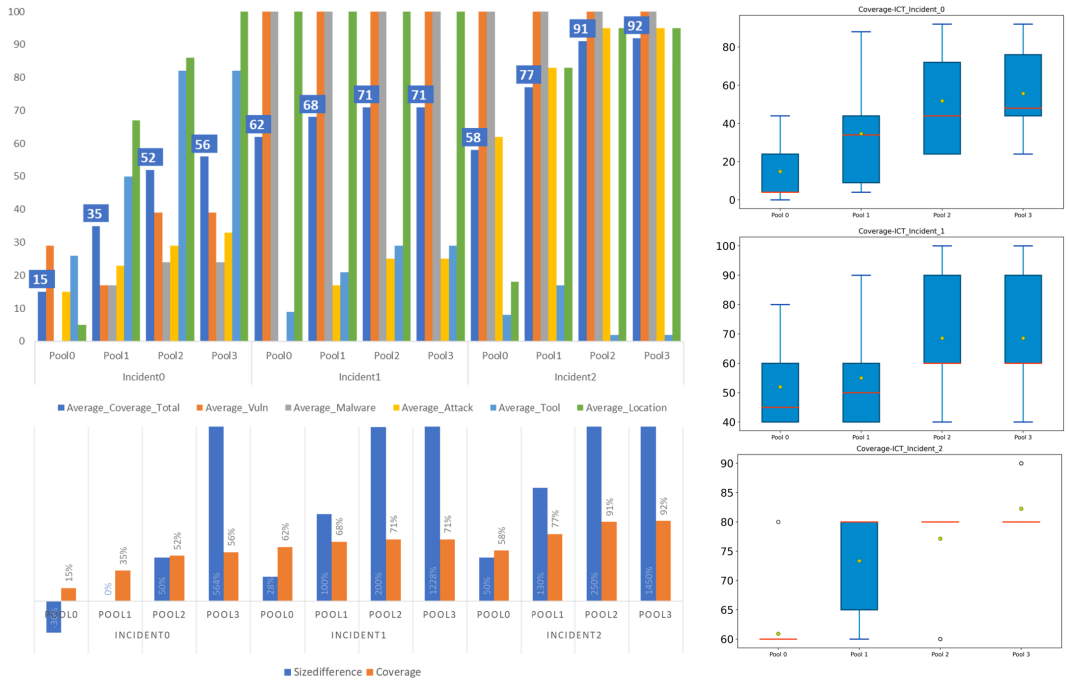


Figure 6.4: ICT Sector Incidents Coverage Results.

Overall, the coverage achieved for Pool 2 was close to 72% meaning that attacks used during the conflict were similar to the ones performed before the conflict.

Although APT enhancement significantly improved coverage, the incidents generated in both pools 2 and 3 managed to cover 100% of the Post-Conflict Incident 1 with an average of 70% for the 40 generated incidents.

6.3.4 Multi-Sector Exercise Coverage

Based on observations comparing Pre-Conflict Incidents with Post-Conflict Incidents across three Sectors, it became clear that Pre-Conflict Pool 2 (comprising relevant merged incidents) can achieve the highest coverage with a justifiable difference in graph size compared to Post-Conflict Incidents. I decided to make a final comparison, but now at the exercise level, by creating a CSE consisting of one event with three incidents, one for each sector named Targeted Exercise. The incident of each sector was a merge of all Incidents of the Pre-Conflict Pool 1 (removing duplicate objects). I then created a Post-Conflict exercise by simply modelling a CSE consisting of one event with three incidents being the Incident 0's of the Post-Conflict

Pool of each sector. The exercise bundles were compared (Figure 6.5), and the overall average total coverage achieved was 81% with a 304% size graph difference, meaning that the proposed exercise graph was three times bigger than the Post-Conflict exercise based on real incidents.

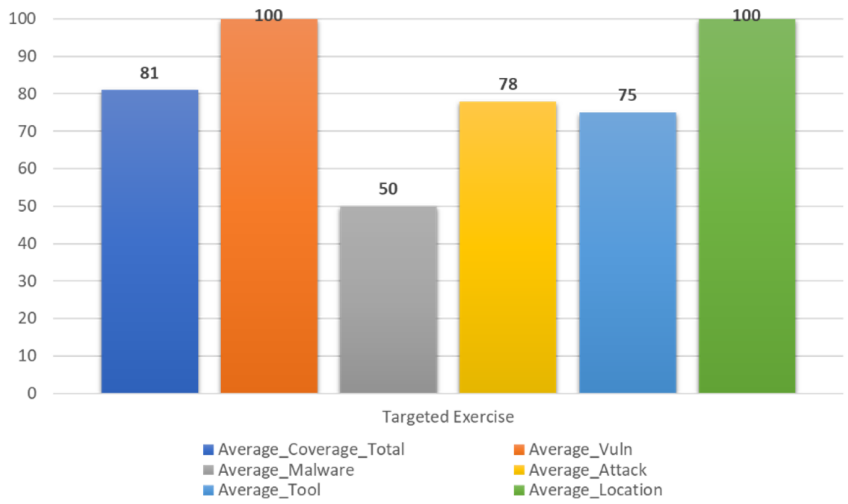


Figure 6.5: Multi-Sector Exercise Coverage Results.

Specific vulnerabilities used by the attackers along with the fact that attackers used unknown vulnerabilities, were predicted by AiCEF at 100%. Attack techniques used in the incidents along with tools used by the attackers, were also well detected, providing ground for a well-scoped exercise generation.

6.4 Findings

Based on the methodology selected and the use case explored I have reached some valuable findings on how to create insightful CSEs that can prepare a country or a sector-specific organisation ahead of a cyber conflict.

Given the initially set objectives for this use case I was able to:

- Map real-world threat information using CESO;
- Generate insightful CSEs that can prepare a country or a sector-specific organisation

ahead of a cyber conflict.

- Identify the best strategy for using known and relevant incidents to predict and train on incidents that will materialise in the future.

Following a statistically proven strategy, one can achieve the best coverage of training objectives in the expectation of future conflicts. The results of this study show that the Pre-Conflict Pool 1 incident generation strategy performs better than a random generation of sector-specific incidents. Using past incidents as a model for exercises can help achieve better coverage of exercise objectives up to an average of +13%.

The Pre-Conflict Pool 2 incident generation strategy, which involves merging multiple resources to describe the same incident, is the most effective in creating incidents using AiCEF, achieving an average of +21% overall coverage of the exercise objective set. The APT enhancement strategy (Pre-Conflict Pool 3) can achieve better overall coverage results but with a high complexity impact. In the future, more innovative methods should be used to enhance a graph based on known actors in a more effective manner.

Furthermore, exercises that cover multiple sectors affected by the same actors can achieve a higher coverage (81%) of training objectives than those that cover a single sector (Public Admin: 5%, Energy: 74%, ICT: 72%).

As a final remark, it can be stated with confidence that AiCEF can automate the generation of well-scoped CSE scenarios to prepare a specific sector or country ahead of a cyber conflict. Scenarios can be adapted in real-time as the conflict progresses, leveraging the power of AI to provide accurate training material with minimal cyber exercise planning expertise.

Part III

Optimizing & Integrating AI-driven Threat Intelligence and Forecasting in the Cyber Security Exercise Content Generation Lifecycle

Chapter 7

Integrating AI-driven Threat Intelligence & Forecasting in the Cybersecurity Exercise Content Generation Lifecycle

Throughout the previous chapters, the challenges of generating context-relevant CSE scenarios using the power of AI have been highlighted. Drawing on the extensive experience in planning and conducting complex CSEs, I have identified a critical need for scenarios that are tailored to specific sectors and incorporate the latest trends in cyber threats.

Although threat intelligence and forecasting have been utilized in CSE design in the past, this work introduces a novel, fully automated methodology that seamlessly integrates forecasting into CSE scenario generation. This innovative approach enables organizations to produce highly realistic and sector-specific scenarios, thereby enhancing the relevance and impact of CSEs. When integrated into exercise generation frameworks such as the AiCEF, this methodology addresses key limitations of traditional approaches, offering a scalable and efficient solution.

This chapter highlights two major enhancements to the proposed framework:

- The integration of AI-driven sector-specific threat intelligence into the CSE generation process.

- The generation of actionable threat forecasts to improve the prompts used for creating CSE scenarios within the framework.

This chapter evaluates the realism and relevance of scenarios generated using the proposed methodology in comparison to those developed by human experts. The evaluation includes the implementation of an automated prompt generation system designed to be accessible to non-expert exercise planners. Surveys conducted as part of a case study assess the perceived relevance and effectiveness of AI-driven scenarios compared to traditional, expert-designed approaches. Quantitative analyses compare the methodology against human expertise and established threat landscapes as benchmarks.

The findings, supported by expert feedback, reveal that integrating the proposed AI-driven methodology into the CSE generation process significantly enhances the realism and relevance of the scenarios. This innovation enables organizations to address real-world cyber threats with improved preparedness and confidence.

7.1 Proposed Pipeline

7.1.1 Overview

This section outlines the proposed pipeline for analyzing cybersecurity incident datasets using the AiCEF framework, specifically the MLCESO module. The pipeline generates AI-driven prompts based on identified trends, prioritizing sector-specific needs through the proposed AI-driven Prompt Generation (AiDPG) module.

The threat intelligence collection approach follows steps from the ENISA methodology [106] to create sectorial threat intelligence (STI) reports, with the key difference of using AI exclusively for data processing instead of human analysts. Sector references align with the NIS2 Directive [2], though any classification system can be applied. The pipeline consists of five distinct steps as illustrated in Figure 7.1.

The pipeline assumes access to a comprehensive, representative dataset of unique cybersecurity incidents in text format spanning several years.

1. **Text annotation using AiCEF**

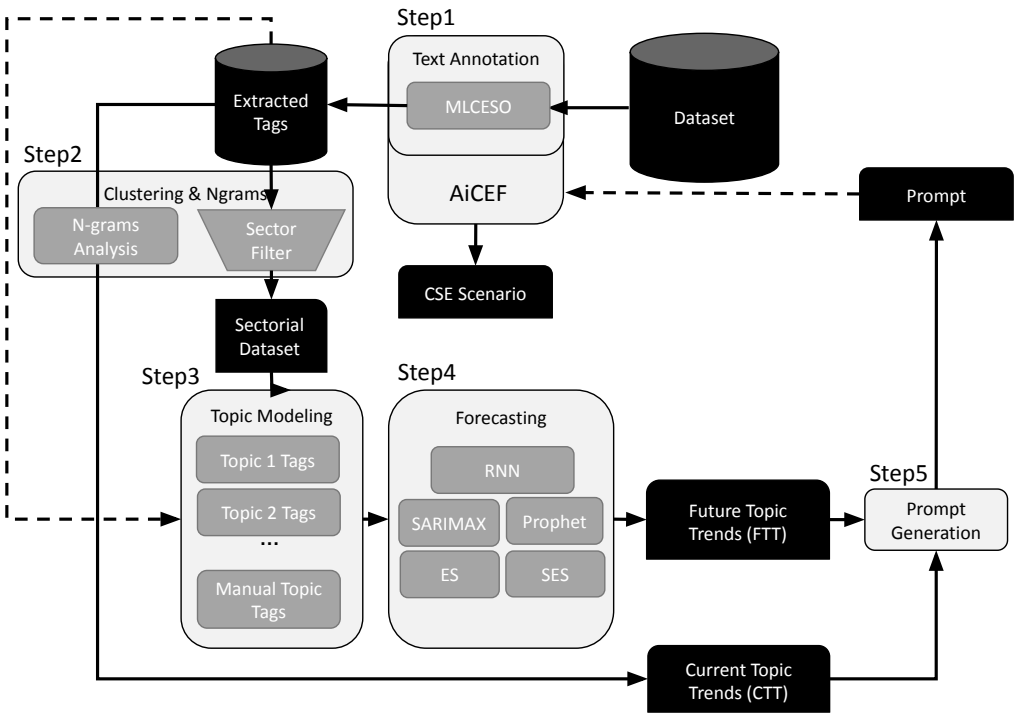


Figure 7.1: Proposed trend extraction and prompt generation pipeline overview.

The MLCESO module preprocesses and annotates text, extracting exercise-related objects including attacker types, attack types, and exploited vulnerabilities.

2. **Clustering and N-Gram Analysis**
   Incidents are clustered into sectors based on extracted word tags to identify similarities and enable detailed N-gram analysis of grouped incidents.

3. **Topic Modelling**
   The LDA algorithm extracts topics relevant to cybersecurity exercise scenarios. Given LLMs' exceptional performance across domains, they are expected to automate these steps efficiently in the future.

4. **Forecasting**
   Occurrence analysis captures incident frequency, patterns, and trends over time to identify current and future developments, patterns, anomalies, and emerging trends within each sector. Multiple algorithms including SARIMAX, ES, SES, Prophet, and LSTM are evaluated using R-squared (R^2) [154], mean absolute error (MAE) [154], mean

absolute deviation (MAD) [155] and mean absolute percentage error (MAPE) [156] metrics to identify optimal performance.

5. **Prompt Generation**
   Exercise prompts represent detailed cybersecurity exercise scenarios containing essential elements for drafting exercises. Actionable prompts are generated through the AiDPG methodology, using identified trends to address cybersecurity challenges while prioritizing training topics based on sector-specific requirements.

7.1.2 Dataset and data preprocessing

As shown in Table 4.10, four incident sources have been identified as input to generate the corpus. These websites contain unique articles previously selected and classified in the Hackmageddon dataset[1], covering unique incidents from 2019-01 to 2022-12 (Table 1.3b), which accounts for 4.957 articles. The above form the Main Corpus, described in detail in Section 1.10.

All relevant articles were collected through automated web scraping. Initially, all texts were converted to the UTF-8 encoding scheme. Using dictionaries and the Textblob library[2], spelling was corrected and special characters were removed. Empty lines, specific stopwords, and punctuation marks were removed using traditional NLP libraries such as NLTK[3] and spaCy [4]. In addition, all HTML and other programming codes, URLs, and paths were removed.

The standard Penn Treebank [149] tokenisation rules were used for sentence tokenisation to tune the incidents' text and facilitate annotation. At the end of this step, a corpus composed of timestamped incidents was formed.

7.1.3 Text Annotation

To annotate the comprehensive dataset, pre-trained Named Entity Recognition models (Table 4.8) from AiCEF's MLCESO module were employed to automatically extract and classify

[1] https://www.hackmageddon.com
[2] https://github.com/sloria/TextBlob
[3] https://www.nltk.org/
[4] https://spacy.io/

cybersecurity entities using the predefined tag schema detailed in Table 8.3. The structured annotation results are systematically stored in the Knowledge Database (KDb), with representative examples of the stored information presented in Table 4.11.

7.1.4 N-gram Analysis

The application of Named Entity Recognition for keyword extraction establishes the foundation for comprehensive N-gram analysis, a computational technique designed to identify meaningful patterns and insights from textual cybersecurity data. This methodology involves analyzing sequences of N consecutive words (N-grams) that occur frequently within the corpus, enabling the identification of prevalent attack patterns, threat actor behaviors, and sector-specific vulnerabilities.

Within the context of cybersecurity incident processing, N-gram analysis provides valuable intelligence regarding dominant attack types, tactical approaches, adversary nomenclature, malware families, targeted assets, exploited vulnerabilities, and affected industry sectors. This analytical approach facilitates a deeper understanding of relationships and emerging trends within my cybersecurity incident database.

The analysis encompasses a comprehensive examination of the top-ten terms across seven key categories (Attack Types, Techniques, Attackers, Malware Types, Assets, Vulnerabilities, and Sectors) spanning four years of accumulated data within the KDb. Results undergo clustering processes to group semantically similar phrases where contextually appropriate, revealing co-occurrence patterns that indicate relationships between different threat categories. For instance, frequent co-occurrence of N-grams such as "APT3" and "PlugX malware" suggests specific correlations between advanced persistent threat groups and their preferred malware arsenals.

Sector Impact Analysis: Following analysis of 4,957 cybersecurity incidents, distinct clustering patterns emerged representing heavily impacted industry sectors. The sector classification reveals both primary targets (Table 7.1) experiencing frequent attacks and secondary sectors (Table 7.2) with notable but less frequent compromise incidents.

This sectoral analysis reveals the distribution of cybersecurity threats across critical infrastructure and commercial sectors, with ICT and Technology demonstrating the highest terminological diversity, reflecting the sector's broad attack surface and varied organizational

Table 7.1: Primary Sector Clusters - Heavily Impacted Industries

Sector	Associated Terms
Commerce	COMMERCE, COMPANY, COMPANIES, BUSINESS, BUSINESSES, RETAIL SECTOR, ENTERPRISE, ENTERPRISES, ORGANIZATIONS, ORGS, RETAIL, FIRMS, FIRM
ICT & Technology	E-COMMERCE, GAMING, INFRASTRUCTURE, ITECH, ELECTRONICS, IT SERVICES, INTERNET SERVICE PROVIDERS, MARKETPLACE, ONLINE SERVICES, INFORMATION TECHNOLOGY, CLOUD HOSTING, SERVICE PROVIDER, HOSTING, TECHNOLOGY, TECH, CLOUD STORAGE, CLOUD STORAGE SERVICES, NETWORK OPERATOR, IT, TELECOMS, TELCO, TELECOMMUNICATIONS PROVIDER, TELECOMMUNICATIONS, TELECOM, CLOUD SERVICES
Finance	BANKS, STOCK EXCHANGE, BANK, FINANCE, FINANCIAL, BANKING SECTOR
Healthcare	HEALTHCARE, CLINICS, CLINIC, HOSPITAL, MEDICAL PROVIDERS, HEALTHCARE SYSTEM, HEALTH, MEDICAL CENTER, MEDICAL, PHARMACEUTICAL
Education	EDUCATION, SCHOOLS, SCHOOL, UNIVERSITY, UNIVERSITIES, COLLEGE, EDUCATION SECTOR, EDUCATIONAL, HIGHER EDUCATION
Government	GOVERNMENT, PARLIAMENT, MINISTRY, MINISTRIES, GOVT, GOVTS, GOVERNMENTS, GOVERNMENT AGENCIES, POLITICAL, MUNICIPALITIES, ELECTION, CITY, CITIES, FEDERAL AGENCIES

structures. The comprehensive term clustering enables accurate sector identification within incident reports while accommodating the varied nomenclature used across different threat intelligence sources.

According to the analysis dataset from 2019 to 2022, the ten largest sector clusters affected are represented per yearly occurrence in total volume in Figure 7.2. Since some incidents could have impacted more than one sector, they appear in multiple sectors affected.

Following the same process for all other labels: Assets, Vulnerabilities, Attack Types, Attacker Names, Malware Types, and Malware Names, one can derive the respective threat intelligence. This information is then stored in the KDb per sector, providing us with the

Table 7.2: Secondary Sector Clusters - Moderately Impacted Industries

Sector	Associated Terms
Energy	ENERGY, ENERGY SECTOR, NUCLEAR, OIL, GAS, POWER, PETROL, POWER SUPPLY
Defence	DEFENCE, DEFENSE, LAW ENFORCEMENT, LAW ENFORCEMENT AGENCIES, LAW ENFORCEMENT AGENCY, POLICE, MILITARY
Aerospace	SPACE, AEROSPACE, AVIATION
Media	MEDIA, MEDIA OUTLETS, NEWS AGENCY, NEWSPAPER, RADIO STATION
Maritime	MARITIME, PORT, SHIP, SHIPS
Food	FOOD, FOOD CHAIN

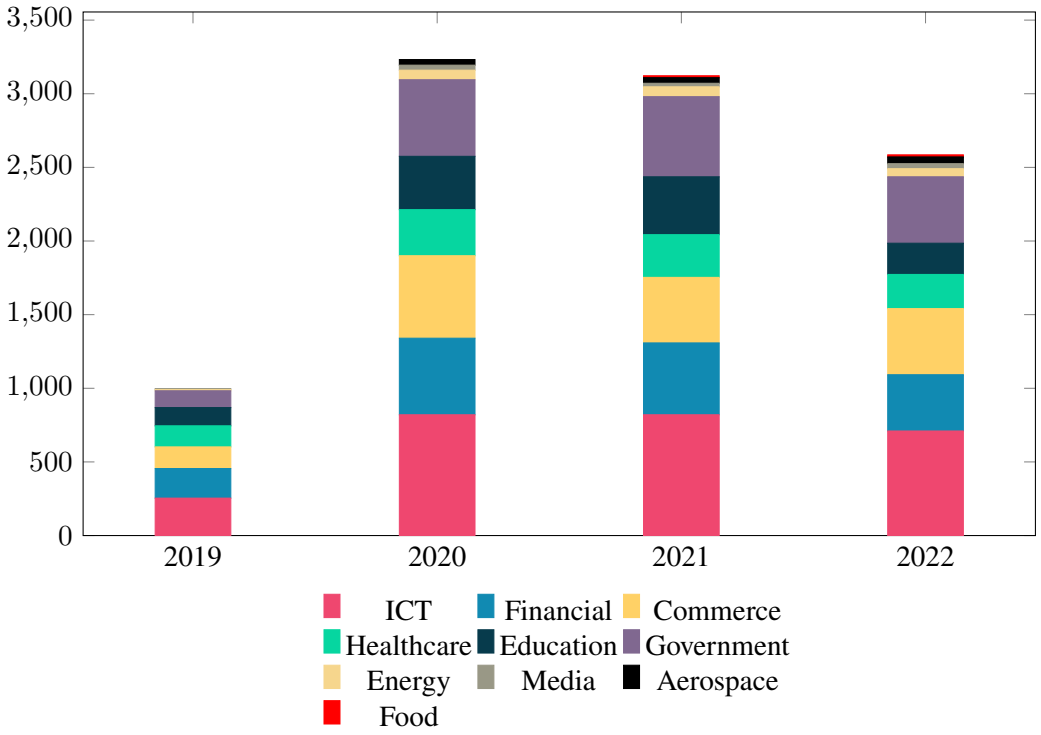


Figure 7.2: Top 10 sectors impacted between 2019 and 2022 in volume

Sectorial Threat Intelligence (STI) needed to enrich CSE scenarios.

7.1.5 Clustering through Topic Modelling

Topic modeling through key phrase cluster analysis provided essential capabilities for extracting meaningful semantic patterns from large cybersecurity text corpora. This analytical approach enabled me to identify thematic structures within incident reports, facilitating automated categorization and trend analysis across my dataset.

While several topic modeling approaches exist, including BERTopic and Non-negative Matrix Factorization (NMF), I opted for the Latent Dirichlet Allocation (LDA) approach based on specific advantages for cybersecurity corpus analysis [157, 158, 159]. LDA represents a generative probabilistic model that assumes documents comprise mixtures of latent topics, with each topic characterized by probability distributions over vocabulary terms. This approach scales effectively with large datasets, making it particularly suitable for analyzing extensive cybersecurity incident collections. Furthermore, LDA's transparent probabilistic interpretation enables a clearer understanding of topic-document relationships, facilitating more effective integration with exercise generation workflows, as in our case.

Successful LDA implementation required the determination of the optimal number of topics through coherence score analysis using CV coherence metrics [160]. The coherence score quantifies topic quality on a normalized scale from zero to one, with higher values indicating stronger semantic cohesion within topic clusters. Iterative evaluation across varying topic numbers revealed that five topics represent the optimal configuration for my dataset, as demonstrated in Figure 7.3, ensuring meaningful differentiation.

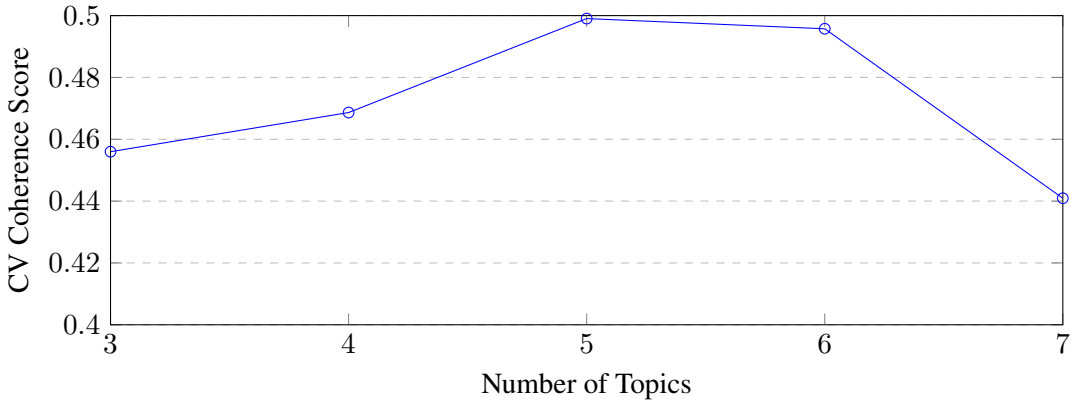


Figure 7.3: CV coherence score for different numbers of topics

The LDA model was visualised using the pyLDAvis[5] library to gain insights and better interpret the topics. This library generates an interactive visualisation representing the topics and their relationships, aiding in topic exploration and analysis.

7.1.5.1 Extracted Topics

Once I trained my LDA model and settled for five topics, I extracted a set of unique words per topic, allowing me to categorise them with specific tags, reported in Table 7.3.

Unique words	**Tag**
HEALTH, RECORD, RECORDS, PHI, PI, PERSONAL, DATA, BIOMETRIC, SSNS, BREACH, DATABASE, ACCOUNT, DETAILS, IDENTIFICATION, FILES	Personal Data Breaches
PHISHING, SPEAR-PHISHING, SPEAR, WALLET, ETHEREUM, SCAM, SIM, SWAPPING, SPOOF, IMPERSONATION, LOGINS, EMAIL, MEDIA, ACCOUNT, LINK	Phishing & social engineering
DDOS, DOS, EXPLOITATION, INFRASTRUCTURE, WINDOWS, LINUX, WEB, INTERNET, DENIAL, SSH, NETWORK, IP, SERVER, ROUTER, CHAIN, SUPPLY, SHELL, INTERNET, PROXY, DNS	Network and System Exploitation
CREDIT, CARD, CARDS, CREDENTIALS, DATA, BANKING, PASSWORD, PAYMENT, DATABASE, SKIMMING, SKIMMER, LOGIN, KEYLOGGING, LINK, COOKIE, DEBIT	Payment Card Fraud & Online Theft
ACCESS, REMOTE, RAT, TROJAN, EXTORTION, RANSOMWARE, TROJANS, VIRUS, SCREEN, VOICE, VIDEO, CLIENT, SCREENSHOTS, ZERO-DAY, O-DAY, MACOS, IOS, ANDROID, SPYING, PHOTO, STEGANOGRAPHY	RATs, Ransomware, Spyware

Table 7.3: Topics and Content

Based on the identified topics content (Table 7.3), I have labelled them based on the context of cybersecurity incidents as follows:

1. Personal Data Breaches: Topic 1, comprising keywords such as "HEALTH," "RECORDS," "DATA," and "BREACH," suggests a focus on personal data breaches. This topic indicates a potential threat to sensitive personal information, including medical records, identification

[5] https://github.com/bmabey/pyLDAvis

details, and biometric data. The presence of terms like "ACCOUNT" and "FILES" further highlights the risk of unauthorised access and disclosure of personal data.

2. Phishing & Social Engineering Scams: Topic 2, characterised by terms like "PHISHING," "SPEAR-PHISHING," "SPOOF," and "IMPERSONATION," points towards phishing and social engineering scams. These tactics involve deceptive methods to trick individuals into divulging their confidential information, such as login credentials and account details. Including terms like "EMAIL" and "LINK" suggests the exploitation of communication channels for fraudulent purposes.

3. Network & System Exploitation: Topic 3, encompassing terms like "DDOS," "EXPLOITATION," "NETWORK," and "SERVER," indicates a focus on network and system exploitation. This topic highlights the potential threats of denial-of-service (DDoS) attacks, the exploitation of vulnerabilities in network infrastructure, and unauthorised access to servers and systems. The presence of terms like "WINDOWS," "LINUX," and "WEB" suggests a broad scope of potential targets.

4. Payment Card Fraud & Online Theft: Topic 4, including keywords such as "CREDIT CARD," "CREDENTIALS," "SKIMMING," and "PAYMENT," aligns with payment card fraud and online theft. This topic signifies the risks associated with unauthorised access to credit card information, compromising banking credentials, and malicious activities targeting online payment systems. The inclusion of terms like "DATABASE" and "KEYLOGGING" suggests potential data breaches and the interception of sensitive financial information.

5. Remote Access Trojans (RATs), Ransomware, Spyware: Topic 5, characterised by terms like "REMOTE ACCESS TROJANS," "RANSOMWARE," and "SPYWARE," indicates the presence of sophisticated malware threats. This topic encompasses various malicious activities, including remote control of infected systems, ransomware attacks, and unauthorised surveillance through spyware. Terms like "ANDROID," "IOS," and "MACOS" further emphasise the potential risks across different operating systems and devices.

A visualisation of the extracted topics can be seen in Figure 7.4

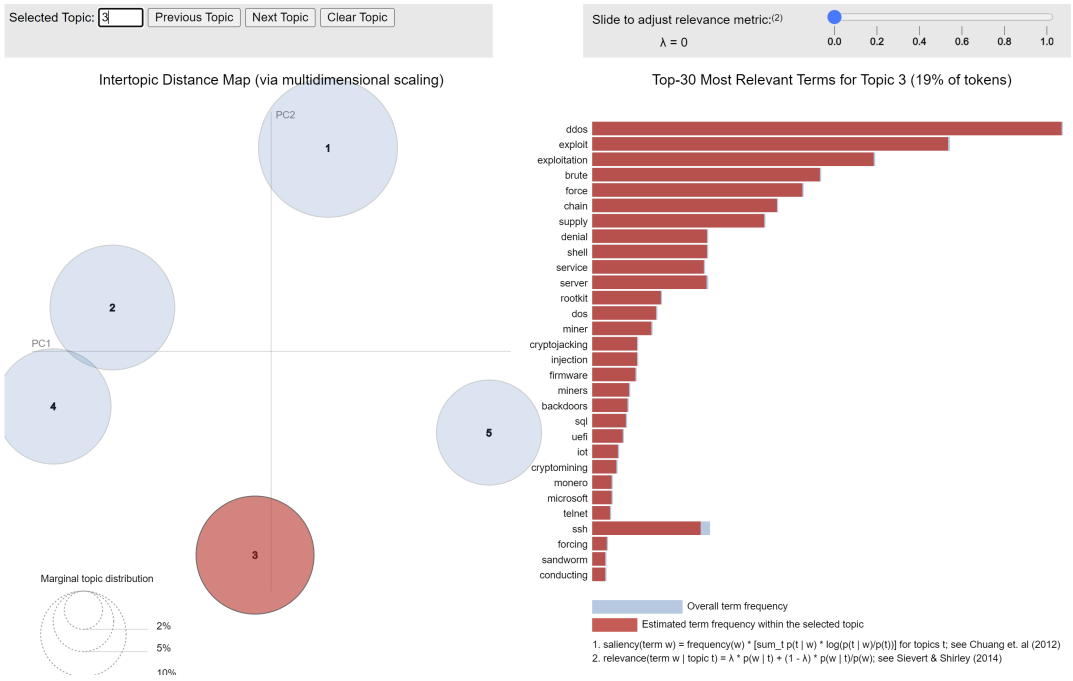


Figure 7.4: LDA Topic visualisation example.

7.1.6 Forecasting, Trend Prediction and Model Evaluation

As already discussed, I have tested various time series analysis methodologies, such as SARIMAX, Exponential Smoothing (ES), Simple Exponential Smoothing (SES), Prophet, and Long Short-Term Memory (LSTM) modelling using Recurrent Neural Networks (RNN), following recommendations of previous research findings [61]. Nonetheless, the goal was not to perform an exhaustive comparison of all possible time series analysis methodologies but to assess whether time-series analysis can support the effort of generating relevant cybersecurity scenarios with a limited incident dataset (of up to four years in this case) of cybersecurity incidents at hand. To this end, I performed trend prediction using data for three years (2019-2021) of the available dataset to predict the values for the fourth year (2022).

For model evaluation, this research will use the following metrics: Mean Absolute Error (MAE), Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), and R-squared (R2).

7.1.6.1 Overall Forecasting Results

Given the dataset of incidents covering 2019 to 2022, described in Table 1.3b, I extracted the monthly statistics per topic and calculated the volume fluctuation per month, visualised in Figure 7.5.

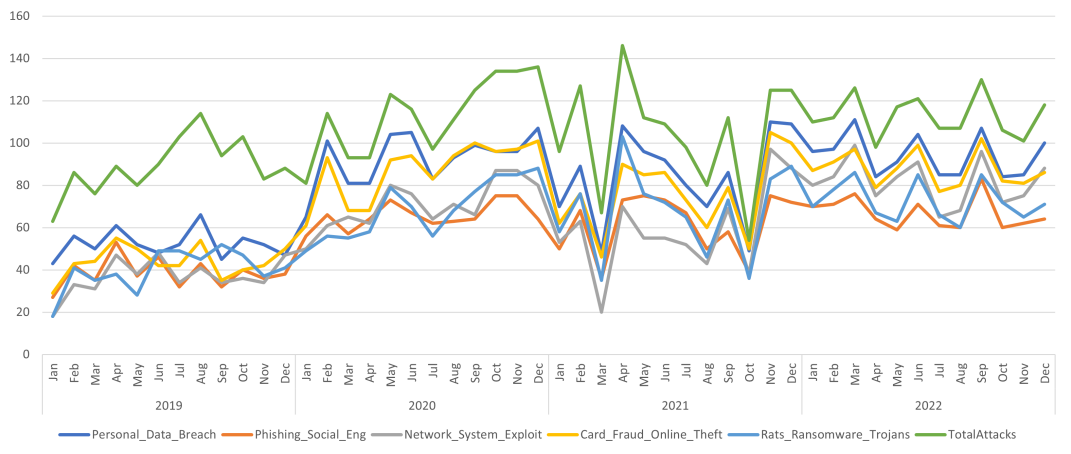


Figure 7.5: Cybersecurity incidents 2019-2022

The trend prediction methodology employs multiple time series forecasting models to analyze cyber attack volume patterns and predict future incident trends. A comprehensive evaluation of forecasting approaches was conducted using cyber attack time series data, with model effectiveness assessed against the specific requirements of cybersecurity exercise planning and threat landscape analysis.

My forecasting evaluation follows a systematic data partitioning approach, allocating 90% of the available dataset for model training and reserving 10% for validation testing. This split ensures sufficient historical data for pattern learning while maintaining adequate test samples for performance assessment. The training dataset encompasses incident statistics from 2019 through the first semester of 2022, enabling models to capture multi-year patterns and seasonal variations in cybersecurity threat activity.

The forecasting models target the prediction of incident volumes for the second semester of 2022, with validation performed against actual reported incidents from the corresponding period. Rather than pursuing exact numerical predictions for individual months, the methodology focuses on trend direction identification, recognizing that precise volume forecasting presents significant challenges given the stochastic nature of cybersecurity incidents and ex-

ternal factors influencing attack patterns.

My prediction framework employs a five-category trend classification system that captures meaningful directional changes in incident volumes: *upward (↑)*, *slightly upward (↗)*, *stable (-)*, *slightly downward (↘)*, and *downward (↓)*. This categorical approach provides actionable intelligence for exercise planning while accommodating the inherent uncertainty in cybersecurity trend prediction.

Trend determination employs a comparative analysis of consecutive period averages, calculating the difference between six-month incident volume averages to establish directional patterns (Algorithm 2). The six-month periodization balances temporal granularity with statistical stability, reducing noise from short-term fluctuations while maintaining sensitivity to meaningful trend shifts. This approach enables the identification of sustained patterns that inform strategic exercise planning and threat preparedness initiatives.

Algorithm 2 Trend Value Labelling

```
Given TrendValue, ranging from negative to positive.
if TrendValue ≥ 6 then
    Label ← "Upward"                          ▷ Trend is upward
else if 3 < TrendValue < 6 then
    Label ← "Slightly upward"                 ▷ Trend is slightly upward
else if −3 ≤ TrendValue ≤ 3 then
    Label ← "Stable"                          ▷ Trend is stable
else if −6 < TrendValue < −3 then
    Label ← "Slightly downward"               ▷ Trend is slightly downward
else
    Label ← "Downward"                        ▷ Trend is downward
end if
```

7.2 AI-driven Prompt Generation (Text Synthesis)

The prompt generation process focuses on creating textual prompts based on a set of specific requirements by mixing user input with Current Topic Trends (CTTs) and Future Topic Trend Predictions (FTTPs) extracted from previous steps. The user provides a dataset of incidents and the sector of interest (optional) as input. The user can also define the forecasting model to be used. We then extract the following information from the dataset based on the user input:

- **Current Topic Trends (CTTs)**: The top ten keywords in volume for the following tag categories: [ATTACK TYPE], [ATTACKER NAME], [VULNERABILITY], [ASSETS], [TECHNOLOGY], [MALWARE NAME], [MALWARE TYPE] for a given [SECTOR] over the course of a year. This is achieved by executing steps 1 and 2 of this pipeline.
- **Future Topic Trend Predictions (FTTPs)**: A set of keywords that stem from previously defined topics (Table 7.3) and that exhibit a clear upward trend based on the proposed trend prediction methodology by using the user selected forecasting algorithm. This is achieved by executing steps 1, 2, 3 and 4 of this pipeline.

A textual prompt template consisting of four sections (with fixed structure along with dynamic sections) is then used to generate prompts that will form the basis of a CSE scenario, consisting of the fifth and final step of this pipeline.

An example of a prompt template can be seen in Figure 7.6

The resulting output can then be edited or used as input to AiCEF, and a CSE incident will be generated, ready to be incorporated in an exercise scenario, as visualised in Figure 7.7. This pipeline can produce meaningful, sectorial-focused prompts and scenarios with trend prediction incorporated seamlessly into the prompt without the need for cybersecurity expertise from the user.

Figure 7.8 maps the predicted values per month for the second semester of 2022 against the real volume of incidents, while Figures 7.9 and 7.10 present the various model evaluation metrics.

Then, using this dataset, I decided to forecast the trends for 2022 for all topics (Table 7.4) and then performed a dedicated forecast by filtering the dataset per sector. I focused my efforts on four sectors, namely energy, healthcare, government, and finance, as seen in Table 7.5, using the forecasting algorithms side by side to ease the comparison of the results. This approach can help generate tailored leads and forecasts and, as a result, more focused CSE scenarios per sector.

Four forecasting methods, namely LSTM, Prophet, ES, and SES, returned positive R^2 values, with LSTM being the fittest algorithm, closely followed by Prophet. The R^2 values of ES and SES, while positive, show that these models do not explain the variability of the response data around the mean; therefore, the predictions cannot be considered accurate. On

	Actual Trend 2022	SARIMAX	Prophet	ES	SES	LSTM
All data-topic agnostic	-	↓	↓	↓	↓	-
RATS, ransomware, trojans	↘	↓	-	-	↘	-
Network_System_Exploit	↓	-	↓	↓	↓	↓
Phishing, Social_Eng	↘	↑	↑	↑	↘	↘
Personal_-Data_Breach	↘	↘	↓	↑	↘	↘
Card_-Fraud_Online_Theft	↘	-	↓	↑	-	-

Table 7.4: Trend forecasting results for the 2nd Semester of 2022

the contrary, while not ideal, the range of R^2 values of both LSTM and Prophet implies a good fitness of the models. Finally, the negative values of SARIMAX clearly illustrate that this method is unsuitable for this dataset.

Based on the fact that LSTM had low MAE, MAPE, and MAD scores, it can be considered the most suitable algorithm for this task. This is validated by the fact that LSTM was one of the best in trend prediction when forecasting future volumes of the five topics per sector, as presented in Table 7.5. Similarly, Prophet had low MAE, MAPE, and MAD scores and satisfactory trend predictions when forecasting future volumes for the five topics per sector. Therefore, it can also be considered suitable for the task. SARIMAX, beyond the negative R^2 values, had high MAE, MAPE, and MAD values, validating that it cannot be used for accurate forecasting or trend prediction in this dataset.

From the other two methods with low R^2 values, ES had high MAE, MAPE, and MAD scores, which were mapped in the poor performance when forecasting future incident volumes for the five topics per sector. Finally, even if SES did not fit well based on the R^2 values, it had the lowest MAE, MAPE, and MAD scores, and its performance was very satisfying

for forecasting future incident volumes for the five topics per sector. One can interpret the above as an indication that, while the predictions were not very accurate, the model was able to fit enough to identify the trends. For the remaining steps in the proposed methodology, I will use the top three performing algorithms (Prophet, LSTM, SES) for further evaluation and comparison, while dismissing ES and SARIMAX due to their low performance in the selected metrics. When implementing this methodology in real-life scenarios, the use of a single forecasting algorithm is adequate to provide the desired predictions.

7.3 Evaluation and Results

7.3.1 Evaluation

To evaluate and measure the effectiveness of the proposed framework, a comprehensive case study was developed focusing on the energy sector. The energy sector was selected due to its critical designation in both NIS and NIS2 Directives [2], its high impact on interconnected industries when compromised, and its unique threat landscape that attracts specific threat actors with distinct motivations and tactics, techniques, and procedures (TTPs).

My evaluation was conducted using a three-phase approach. The first phase involved prompt generation, where exercise planners with varying expertise levels generated textual inputs for the AiDPG module. This phase employed two expert groups: cybersecurity exercise development specialists and sector-specific professionals with limited exercise planning experience. This mixed-expertise approach aligns with recommendations from national cybersecurity agencies such as ANSII[6], which advocate combining sectoral knowledge with exercise planning expertise to create technically valuable and contextually relevant training scenarios.

Four distinct prompt sets were generated through this process:

- Prompt Set 1: The Exercise experts' prompt set, created by the experts in CSE development.
- Prompt Set 2: The Sector experts' prompt set, created by experts in the sector examined

[6] https://www.ssi.gouv.fr/uploads/2021/09/anssi-guide-organising_a_cyber_crisis_management_exercise-v1.0.pdf

(ENERGY).

- Prompt Set 3: The Human experts prompt set, which is the combination of sets 1 & 2.
- Prompt Set 4: The AiDPG prompt set was generated purely using this methodology, focusing on the sector of interest (ENERGY) with forecasting enabled.

AiDPG (no forecasting) prompt set was also generated as a subset of Prompt Set 4 and will be used only for limited comparisons against other prompt sets.

The second phase conducted a lexicological analysis of all prompt sets to identify keyword patterns and potential biases. AiCEF annotation process generated five distinct tag sets corresponding to each prompt category plus a baseline derived from Sectorial Threat Landscapes (STLs). These tag sets enable comparative analysis to assess prompt maturity and evaluate coverage of sector-oriented threat landscapes between methodology-generated and human-expert-generated content.

The final phase validated trend forecasting capabilities through expert survey methodology. Two expert groups evaluated the forecast accuracy against professional expertise and domain intuition, providing qualitative validation of the prediction framework's effectiveness.

The subsequent analysis presents detailed results from each evaluation phase, demonstrating the framework's performance across multiple assessment criteria.

7.3.2 Prompt generation

Twenty-six cybersecurity experts were selected to generate five (5) individual CSE scenario prompts each, according to the high-level requirements and specifications, as illustrated in Figure 7.11.

All experts have cybersecurity backgrounds, and their skill sets resemble those of a Chief Information Security Officer (CISO) or similar. For the cyber exercise expert group, I invited several planners with high expertise in CSE management, mixed ethnicity, and focus sectors. More precisely, I invited members of a Cyber Awareness Expert Group [7], from which 12 exercise planners responded. The sectorial experts group had a cybersecurity background

[7] https://www.enisa.europa.eu/topics/cybersecurity-education/ad-hoc-working-group-awareness-raising

for the energy sector (Members of the European Network of Transmission System Operators for Gas or European Network of Transmission System Operators for Electricity) and acted as CISOs for their organisations. In total, 14 sectorial experts responded. The demographic distribution of all experts is illustrated in Table 7.6.

Not all experts provided the maximum number of five (5) prompts as requested. Therefore, I collected fifty-five (55) prompts per group, totaling 110 human prompts. All experts provided at least three prompts.

The prompt specs were input in the proposed methodology, and an AiDPG prompt set was generated, focusing on the energy sector, concluding the prompt generation task.

7.3.3 Prompt Evaluation

To evaluate the sets of generated prompts, keyword extraction using AiCEF was applied.

Two types of evaluation metrics were used:

- Prompt maturity, using AiCEF's maturity calculation algorithm, to evaluate the richness of the prompt in terms of exercise-related objects.
- Tag coverage of the human and AI-assisted tag sets against that of a sectorial baseline. This metric helps evaluate the relevance of the prompts generated.

7.3.3.1 Prompt Maturity

Based on AiCEF's maturity calculation algorithm, the maximum maturity score for any text can be 150 points, indicating that the text analysed has the maximum richness of information, covering all MLCESO tags in Table 8.3.

Table 7.7 illustrates the average maturity score per group of prompts. The average maturity score of AiDPG (150) (even with no forecasting (135)) clearly exceeds the scores of any of the two human expert groups.

Indeed, these scores provide sufficient proof that the sector-focused AiDPG methodology can help human CSE planners improve their prompts and effectively exercise scenarios. More

precisely, a 50.4% improvement for the exercise experts and an improvement of 66.67% for sector experts can be achieved in terms of prompt maturity.

7.3.3.2 Tag Coverage

After analysing and comparing the accuracy of the lexicological elements of the prompt sets generated by exercise experts, sector experts (energy experts), and their combination, along with the AiDPG methodology, a baseline set of tags was defined. The baseline tag set was generated by extracting tags (see Table 8.3) from acknowledged STL reports. More precisely, three STLs were used to evaluate the proposed methodology: the Deloitte Oil and Gas (Energy) Threat Landscape [8], the Energy Sector 2022 Threat Landscape by CITALID and Sekoia.io [9] and the Electric Avenue report by BlueViv & Neurosoft[10]. These threat landscapes cover the broader energy sector, using cybersecurity incidents from as early as 2017 up to early 2023. The unique tags per report are visualised in Figure 7.12.

The analysis of the following set of tags is visualised in Figure 7.13.

- "Exercise experts" Tag set
- "Sector experts" Tag set
- "Human Experts" Tag set
- "AiDPG Energy" Tag set
- "Energy Threat Landscapes" Tag set

Finally, the overlapping unique tags for "ENERGY THREAT LANDSCAPES", "HUMAN EXPERTS", and "AiDPG ENERGY" are represented in Figure 7.14, providing a better visual representation of the coverage between AI-generated and human tag sets.

To verify the superiority of AiDPG over human-expert-generated prompts, the Jaccard similarity was used to calculate the similarity between two sets of unique phrases based on the number of common phrases and the total number of distinct phrases in the sets.

[8] https://www2.deloitte.com/us/en/pages/risk/articles/oil-and-gas-sectors-threat-landscape.html
[9] https://blog.sekoia.io/the-energy-sector-2022-cyber-threat-landscape/
[10] https://neurosoft.gr/the-electric-avenue-an-overview-of-the-energy-sectors-threat-landscape/

The Jaccard similarity coefficient (JSC) is defined as the size of the intersection of the sets divided by the size of the union of the sets.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where:

J(A, B) represents the Jaccard similarity coefficient (JSC) between sets A and B.

$|A \cap B|$ represents the size (cardinality) of the intersection of sets A and B.

$|A \cup B|$ represents the size (cardinality) of the union of sets A and B.

After comparing the Jaccard similarity, it is clear that the prompts generated with the AiDPG methodology are more relevant when compared to the ones generated by human experts, when compared to the set generated when combining all energy threat landscapes:

- JSC for "AiDPG Energy" and "Energy Threat Landscapes" with value: 0.22
- JSC for "Human Experts" and "Energy Threat Landscapes" with value: 0.19
- JSC for "Sector Experts" and "Energy Threat Landscapes" with value: 0.16
- JSC for "Exercise Experts" and "Energy Threat Landscapes" with value: 0.17

To identify which group of experts would benefit more from the AiDPG methodology, I calculated the intersection of unique annotation tag categories per group and measured the number of tags present in both "Energy Threat Landscapes" and "AiDPG Energy" but not in the individual human-generated sets. By doing this, I assumed that, at a minimum, the content of the "Energy Threat Landscapes" that is also part of the "AiDPG Energy" set can improve the prompt and become more relevant.

Let A represent the set "Energy Threat Landscapes", B represent the set "AiDPG Energy", and C represent a human-generated set. The number of tags in the intersection of A and B but not in C can be calculated as follows:

$$\text{Benefit} = \|A \cap B\| - \|A \cap B \cap C\|$$

$$\text{BenefitPercentage} = \frac{\text{IntersectionCount}}{\|A\|} \times 100\%$$

Using the formula above for the annotation tags in Table 8.3, one can calculate the individual benefit percentages that the two human groups of experts could achieve using this methodology, as shown in Figure 7.15.

The results underline the effectiveness of this methodology since both expert groups can benefit from at least 18.83% more relevant tags proposed by AiDPG, not selected by them but present in the "Energy Threat Landscapes" set.

7.3.4 Trend Forecasting Evaluation

To compare the sectorial trend forecasting results, I conducted a survey with a pool of specialized planners. The survey aimed to collect insights on the prioritization of cyber threats and assets/resources based on both established threat landscape data and personal research, with a focus on preparing for incidents anticipated to occur in the coming months in the Energy sector. I used the ENISA Threat Landscape 2022[11] to extract eight prime cyber threats and assets affected, linking them directly to the five types of incidents.

Cyber Threats: Ransomware, Malware, Disinformation/ misinformation, Threats against availability - (DOS, DDOS), Phishing & Social engineering threats, Threats against data (data breaches & leaks), Supply-chain attacks, Physical threats with security impact.

Assets Affected: Company Data, Email or other Credentials, Personal Information, Website, Network or other Infrastructure, Sensitive Personal Information, Monetary Funds, Reputation.

I asked the planners to provide feedback on the following questions:

1. What is your trend prediction on the overall volume of cyber incidents in 2023 compared to 2022 affecting the energy sector?

[11] https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

2. What is your trend prediction for the following types of incidents in 2023 compared to 2022 affecting the energy sector?

    - (Personal) Data Breaches
    - Phishing & Social Engineering
    - Network & System Exploitation
    - Payment Card Fraud & Online Theft
    - Remote Access Trojans (RATs), Ransomware, Spyware

The possible answers were: *upward (↑)*, *slightly upward (↗)*, *stable (-)*, *slightly downward (↘)*, and *downward (↓)*.

I then implemented the proposed trend forecasting methodology for energy-related incidents for 2023 using the top three scoring models (Prophet, SES, and LSTM). A comparison of the results against the expectations of the two expert groups can be found in Tables 7.8 and 7.9, where both the normalized forecasted score (as per Algorithm 2) and trend is presented.

Regarding forecasting the overall sectorial incident trends, the two groups of humans agreed that there would be a "Slightly Upward" trend in 2023, a result also captured by the LSTM model. When comparing human foresight on sectorial incident trends per topic, it is clear that the two expert groups do not align. Focusing more on the foresight of sectorial experts, one can spot that both Prophet and LSTM are making trend predictions that are close to those of humans. The results are satisfying in that, in most cases, the methodology did not produce the opposite results to the humans' foresight. Yet, this methodology is based on incident volumes and cannot replicate human foresight consisting of broader knowledge aspects that can contribute to a change trends like geopolitics. Only the actual data of cyber incidents against the energy sector in 2023 can prove whether the trend forecast or human insight was more accurate in the predictions made, a task to be performed in future work.

7.4 Promising Findings

The proposed approach successfully identified and quantified the gap in context-relevant CSE scenarios generated by sectorial cybersecurity experts and seasoned exercise planners. More precisely, through the AI-driven topic extraction model and by clustering sector-specific annotation tags, I was able to compare various forecasting models and accurately identify

emerging (sectorial) cyber threats for a given set of cybersecurity incidents. My experiments illustrated that for the given dataset dataset, LSTM and Prophet were the two fittest algorithms.

Regarding prompt maturity, the sector-focused AiDPG methodology can indeed assist human CSE planners in improving their prompts and effectively the exercise scenarios. More specifically, a 50.4% improvement for the exercise experts and a 66.67% improvement for the sector experts can be achieved in terms of prompt maturity, generating not only richer but also more sector-focused scenarios.

After calculating the Jaccard similarity, it is clear that the prompts generated by the AIDPG methodology are more relevant to the set generated when combining all energy threat landscapes than the set generated by all human experts combined:

- "AiDPG Energy" and "Energy Threat Landscapes" with value: 0.22
- "HUMAN Experts" and "Energy Threat Landscapes" with value: 0.19.

Even more, by leveraging the user surveys to measure the perceived relevance of CSE with AI-driven threat intelligence compared to traditional exercise generation, I managed to quantify the value of this methodology. The results underline the effectiveness of this methodology since both expert groups surveyed, Exercise Experts and Sector Experts, can improve by at least 18.83%, the prompt and effective scenario relevance to a specific sector.

Limitations of our approach have been identified regarding the volume and time distribution of past incidents needed to form accurate forecasting. For Prophet specifically, three years of observations are the minimum data to predict a fourth year in sequence. Further to the above, forecasting algorithms cannot replace human foresight based on the broader knowledge of a domain that leads human experts to expect trend changes.

Prompt Template:

- **Section 1: Attack Type and Sector**
  - Example: ATTACK TYPE in the SECTOR.
- **Section 2: Trending TTPs and Impact based on STI**
  - Example: Attackers ATTACKER NAME used VULNERABILITY to affect ASSETS.
  - The malware employed is MALWARE NAME MALWARE TYPE.
  - Technologies used or abused: TECHNOLOGY.
- **Section 3: Future Trending Topics** (*This section can be omitted if not enough data are available to perform forecasting)
  - Use NUM of the following trending elements: SECTOR TOPIC[0] ITEMS
- **Section 4: Styling and Formatting**
  - Start the prompt with: ["*Generate a cybersecurity incident describing an* "]. Prompt size: [200] words max

Example Prompt:

Generate a cybersecurity incident describing a BREACH in the ENERGY SECTOR. Attackers CONTI used CVE-2022-24002 to affect PASSWORDS, DATA. The malware employed is CONTI RANSOMWARE. Technologies used or abused: WINDOWS. The following trending elements were used in the incident: PHISHING, SPEAR-PHISHING, MALWARE.

Figure 7.6: The form of the template above and an example prompt, below.

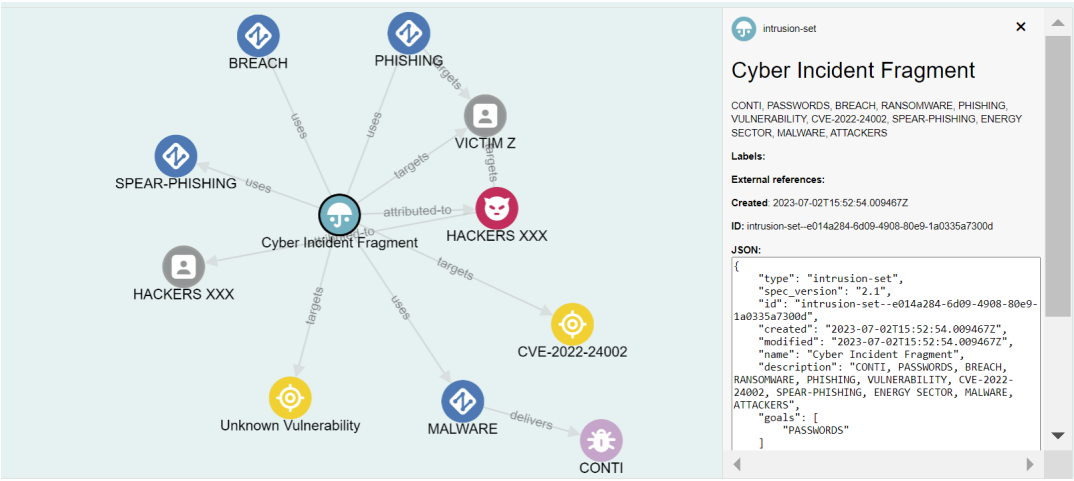


Figure 7.7: STIX 2.1 output generated by AiCEF

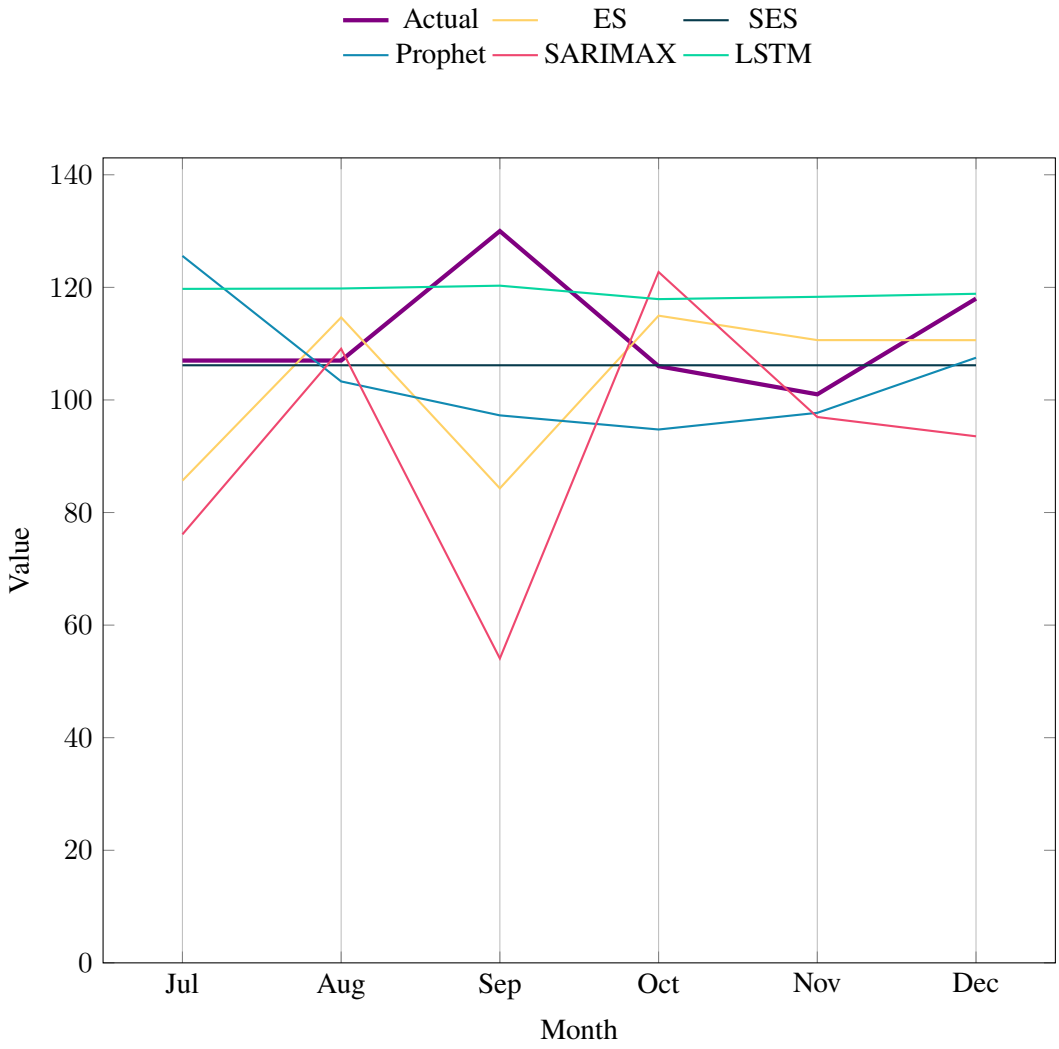


Figure 7.8: Cybersecurity incidents prediction results (2nd Semester 2022)

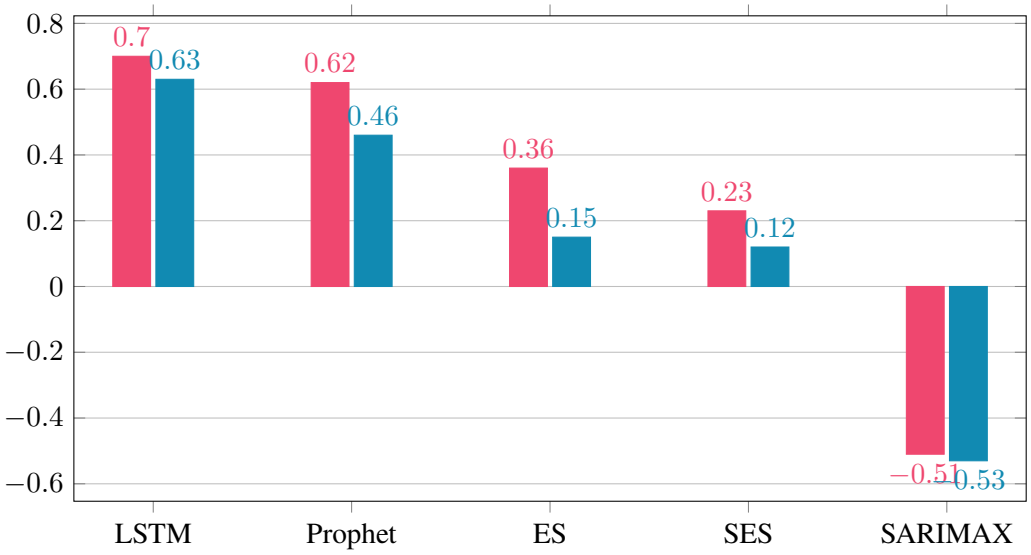


Figure 7.9: R^2 Scores (Training, Testing Dataset) per Forecasting Model.

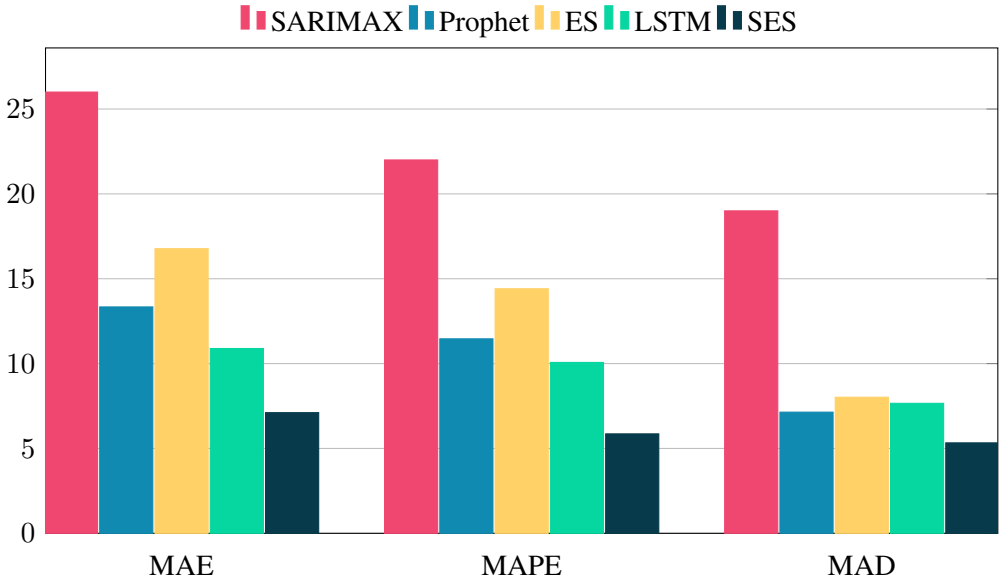


Figure 7.10: MAE, MAPE, and MAD score per forecasting model

Sector	Actual Trend 2022	LSTM	Prophet	SARIMAX	ES	SES
Energy						
Sectorial Incidents	-	-	-	↗	↗	-
Personal_Data_Breach	-	-	-	↗	↗	-
Phishing_Social_Eng	-	-	-	-	-	-
Netwok_System_Exploit	-	-	↗	↗	↗	-
Card_Fraud_Online_Theft	-	-	↗	↗	↗	-
Rats_Ransomware_Trojans	-	-	-	-	↗	-
Healthcare						
Sectorial Incidents	-	-	-	↗	-	-
Personal_Data_Breach	-	-	-	↗	-	-
Phishing_Social_Eng	-	-	-	-	-	-
Network_System_Exploit	-	-	-	↗	↗	-
Card_Fraud_Online_Theft	-	-	-	↗	↗	-
Rats_Ransomware_Trojans	-	-	-	-	-	-
Government						
Sectorial Incidents	↓	-	-	-	↓	-
Personal_Data_Breach	↓	-	-	-	↓	-
Phishing_Social_Eng	↓	-	-	-	↓	↓
Network_System_Exploit	-	-	-	↑	↗	-
Card_Fraud_Online_Theft	↓	-	-	-	↓	↓
Rats_Ransomware_Trojans	↘	-	-	↗	↓	↓
Financial						
Sectorial Incidents	↓	-	↘	↓	↓	↓
Personal_Data_Breach	↓	-	↘	↓	↓	↓
Phishing_Social_Eng	↓	↓	↘	↓	↓	↓
Network_System_Exploit	-	-	-	-	-	-
Card_Fraud_Online_Theft	↓	-	↘	↓	↓	↓
Rats_Ransomware_Trojans	↓	-	↓	↓	↓	↓

Table 7.5: Sectorial trend forecasting results for 2022.

Prompt Specs: Generate a CSE prompt for a cyber awareness exercise. The prompt should not be longer than 600 words and should include as many technical details as possible related to a cybersecurity incident. The company which will use the scenario is an **Energy Service Provider** and all its employees can be potential Players. The incident generated should focus on the existing threat landscape and should train employees to prepare for current or future threat trends relevant to the designated sector in an exercise that will take place in 2023.

Figure 7.11: Task definition

Country	**#**	**Country**	**#**
France	3	Luxembourg	1
Netherlands	3	Lithuania	1
Czechia	2	Austria	1
Spain	2	Italy	1
Belgium	2	Finland	1
Greece	2	Ireland	1
Denmark	2	Portugal	1
Croatia	1	Poland	1
Germany	1	**Total**	**26**

(a) Countries of origin of the experts

Sector	**#**
Energy	14
Public Administration	5
ICT	2
Transport	2
Law Enforcement	2
Education	1

(b) Sector that experts are working in

Table 7.6: Demographics of the experts.

Label Name	Prompts	Maturity Score
Sector Experts	55	50,41
Exercise Experts	55	73,94
AiDPG (no forecasting)	55	135
AiDPG	55	150

Table 7.7: Maturity scores

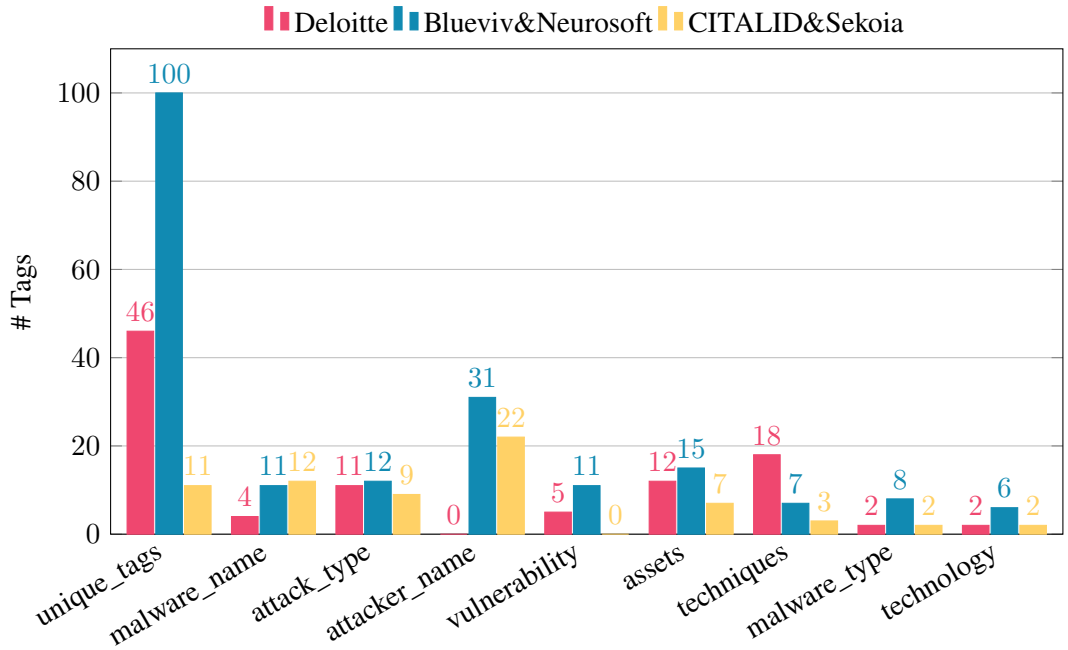


Figure 7.12: Unique annotation tags values per energy landscape report

Forecast Type	Overall Energy Attacks trend
Prophet	7, Upward Trend (↑)
SES	0, Stable (-)
LSTM	3.17, Slightly Upward (↗)
SECTOR EXPERTS	3.9, Slightly Upward (↗)
EXERCISE EXPERTS	3.75, Slightly Upward (↗)

Table 7.8: Sectorial trend forecasting results for 2023.

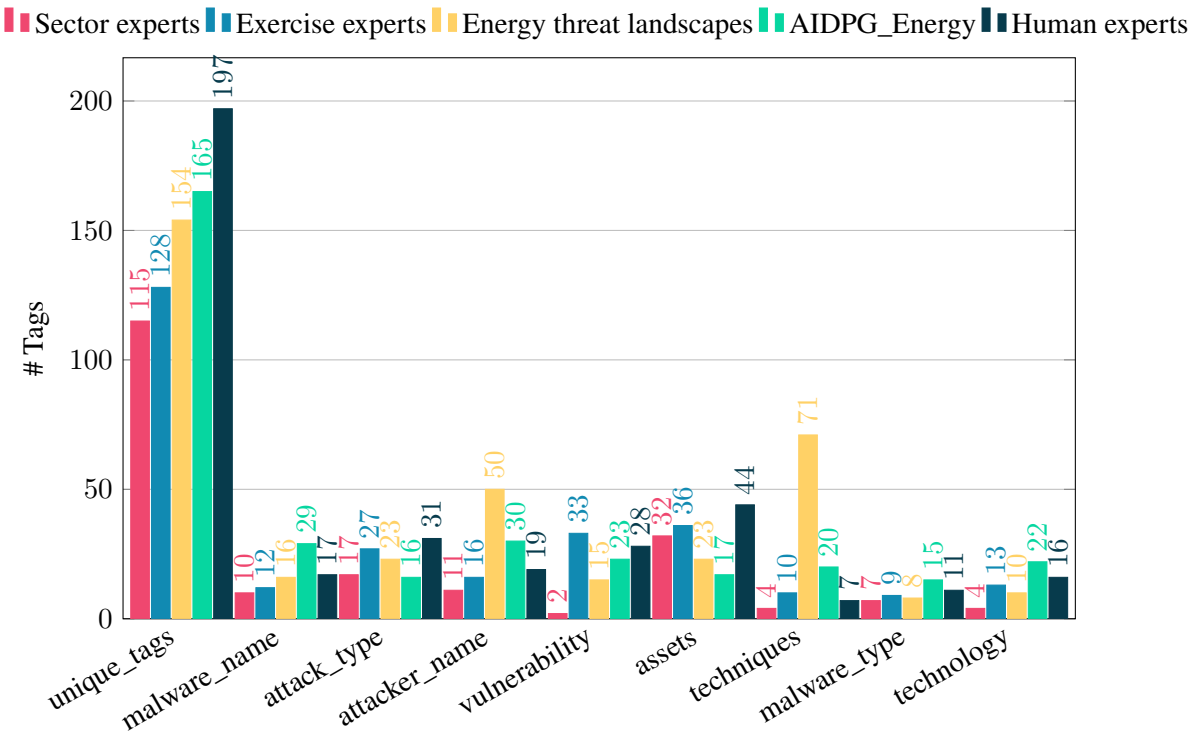


Figure 7.13: Unique annotation tag volume comparison

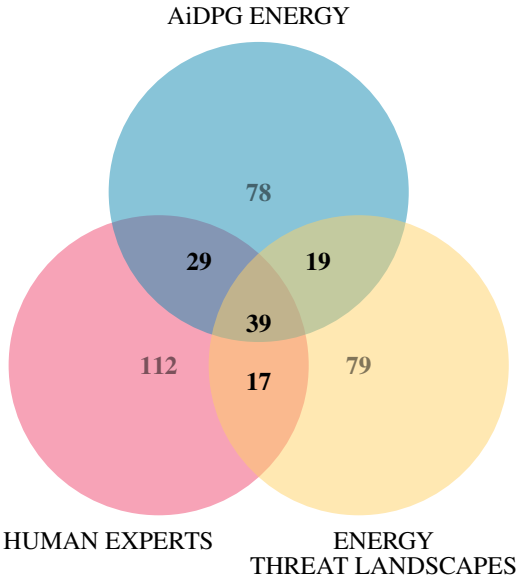


Figure 7.14: Unique tags of STLs versus Human Experts and AiDPG

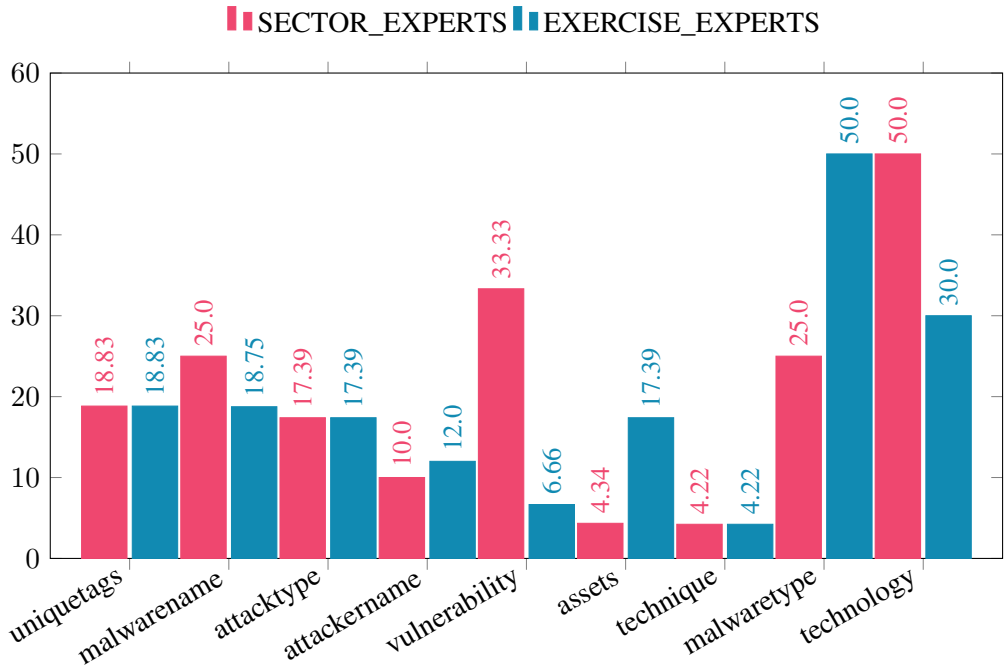


Figure 7.15: Benefit percentage per tag and expert group

Forecast Type	Card Fraud & Online Theft	Network System Exploit	Personal Data Breach	Phishing & Social Eng	Rats, Ransomware, Trojans
Prophet	4, ↗	9, ↑	7, ↑	5, ↑	4, ↗
SES	1, -	3, -	0, -	0, -	0, -
LSTM	2.1, -	3.5, ↗	3.1, ↗	1.9, -	2.4, -
SECTOR EXPERTS	2.71, -	4.29, ↑	3, -	4.64, ↑	4.15, ↑
EXERCISE EXPERTS	2.08, ↘	3.58, ↗	2.41, ↘	4.75, ↑	4.58, ↑

Table 7.9: Sectorial forecast results per topic for the year 2023

Chapter 8

Optimising AI Models for Intelligence Extraction in the Life Cycle of Cybersecurity Threat Landscape Generation

To support my initial framework for building relevant and timely cybersecurity exercises, I had to find a way to extract accurate threat intelligence and build custom threat landscapes to use as part of my scenarios.

A major focus was placed on improving the accuracy of extracting a crucial component of an exercise scenario from existing incidents, specifically the MITRE TTPs. This effort led to the proposal of a novel technique that can be adopted by experts in the threat intelligence field.

Current cybersecurity threat landscape (CTL) report generation relies predominantly on manual analysis by security experts, creating significant bottlenecks in threat intelligence processing. This approach is time-consuming, costly, and susceptible to human error, limiting the scalability of threat analysis operations. While machine learning and artificial intelligence models have demonstrated capabilities in cybersecurity feature extraction, their effectiveness in generating comprehensive CTL reports remains underexplored.

The complexity of CTL attributes presents substantial challenges for automated extraction. These attributes require deep contextual understanding and domain expertise, making them difficult to identify even for experienced human analysts.

Building on the established importance of accurate CTLs in cybersecurity exercise development, this chapter investigates automated approaches for CTL generation using artificial intelligence methodologies. The research addresses three fundamental questions regarding the automation of threat landscape analysis.

First, this work examines whether AI-driven methodologies can effectively automate and enhance CTL report generation beyond traditional manual approaches. Second, it identifies the most suitable AI model architectures for extracting complex attributes typically found in threat landscape reports. Finally, it evaluates the performance trade-offs between accuracy, efficiency, and scalability when comparing AI-based systems against conventional manual analysis methods for CTL generation.

8.1 Proposed AI Model Comparison Methodology

To better scope the flow of this work, this work follows the steps of the well-defined ENISA CTL methodology (see Figure 8.1) and pinpoint the part of the methodology in which AI models can contribute the most, potentially replacing human effort.

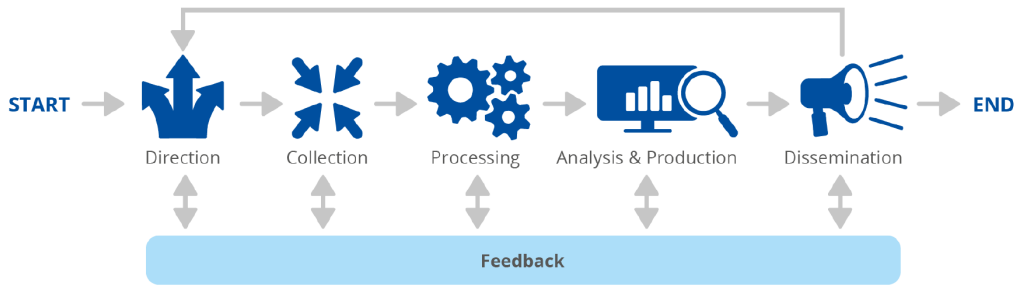


Figure 8.1: Overview of ENISA CTL methodology, source [106].

The output of the ENISA methodology is a set of actionable, timely, and accurate insights based on a collection of timestamped resources. It focuses on at least one sector and tries to answer questions like:

- Who are the potential adversaries and their motivations?
- What are the most common attack vectors and threats?
- What techniques are adversaries using to achieve their goals?

Using as a reference the content of the ENISA methodology templates ([106], B ANNEX: Reference Templates, Page 36), I decided to focus my effort on recreating insights relevant to five key attributes: (1) Attack Type, (2) Motivation, (3) Threat Name, (4) Sector Affected and (5) TTPs. Other attributes usually referenced in CTLs are Indicators of Compromise (IoCs) and weaponised vulnerabilities, but I decided to limit the scope to the five attributes mentioned above, as these cannot be easily extracted using pattern-based methods. For instance, a weaponised vulnerability could be referred to as a CVE, which is easy to extract with, e.g., regular expressions; nevertheless, a source report may describe the vulnerability, especially when there is no CVE for it. Similar approaches can be applied to hashes and URLs; nevertheless, one needs to be aware of the context of the URL, as it can have multiple interpretations in a report.

This work starts with *collecting* the right dataset, then *processing* it using human analysts and different types of AI models to identify which of them better fits the assigned task. The results of this *analysis* can then determine which model or combination of models can be used to *produce* core parts of a CTL (Figure 8.2). My work does not examine the automation of text generation based on the provided analysis results, a task that generative AI can also automate, but it can be considered beyond the scope of this thesis.

8.1.1 Dataset Selection

The core dataset used for this study comprises 1,000 cybersecurity-related texts, ranging from threat intelligence reports, security blog posts, to social media posts, comprising the Incident Corpus (IC) in English.

Due to the fact that two of the models selected, namely TRAM and TTPHunter, for this task had a limitation on detecting only the top 50 MITRE techniques, I have decided to generate an additional dataset focusing on these techniques to make the evaluation against LLMs fair. The texts were generated using either descriptions or examples provided by the MITRE

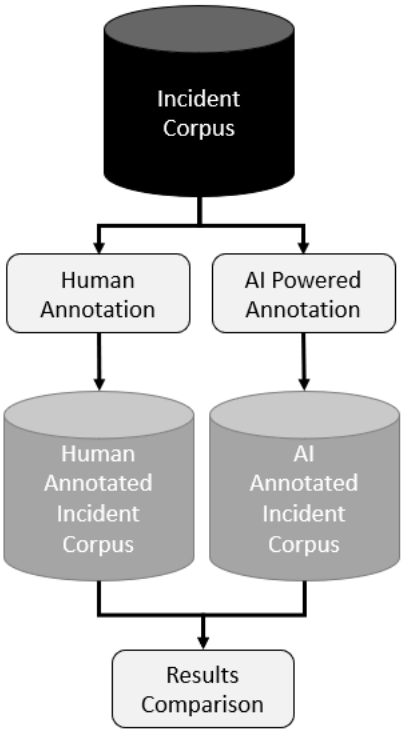


Figure 8.2: Overview of the proposed comparison methodology

ATT&CK dataset for enterprise techniques v15.1 [1] when characterising a technique. Using a text generator, I merged two or more techniques simultaneously, resulting in a synthesised report. The complementary dataset of 100 texts, was synthetically generated only to cover the particular case of MITRE's TTP extraction, comprising the TTPs Corpus (TTPC) in English. Both datasets are described in detail in Section 1.10.

8.1.2 Dataset Pre-processing

All text found in both corpora was processed using Natural Language Processing (NLP) techniques. Initially, all text was converted to the UTF-8 encoding scheme. Using dictionaries

[1] https://github.com/alexzacharis/AICTL/blob/main/MITRE/enterprise-attack-v15.1-techniques.xlsx

and the Textblob library[2], I corrected spelling errors and removed special characters. Empty lines, specific stopwords, and specific punctuation marks were removed using traditional NLP libraries such as NLTK[3] and spaCy [4]. Moreover, all HTML or other programming codes were removed.

The standard Penn Treebank [149] tokenisation rules were used for sentence tokenisation, and finally, standardisation processes were applied to adjust the text to facilitate annotation. An example of a corpus text ready for annotation is the following:

```
{
  "text": "The National Institute of Cybersecurity Incibe has
     detected a phishing campaign impersonating WiZink bank."
}
```

8.2 Human Annotation

A reliable benchmark was established through manual annotation by two cybersecurity experts serving as Senior Cyber Threat Analysts. Each text underwent a comprehensive review, categorization, and annotation by both experts to ensure consistency and accuracy in the ground truth dataset. The expert classifications provided the gold standard against which AI model performance was evaluated. Following the annotation process, inter-annotator agreement was assessed using Cohen's Kappa metric [161] to validate the reliability of the human-generated ground truth. The final gold standard version of the annotated corpus was produced after achieving satisfactory agreement levels between the two expert annotators.

Each incident in IC was classified according to the ENISA taxonomy [162] using three categories of five attributes/tags describing the Attacker, the Attack and the Victim, which were scoped for extraction. The agreed values are in Table 8.3. The IAA was validated using Cohen's Kappa [150]. The formula used is defined as follows:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{8.1}$$

where p_0 expresses the relative observed agreement, and p_e is the hypothetical probability of chance agreement.

[2] https://github.com/sloria/TextBlob
[3] https://www.nltk.org/
[4] https://spacy.io/

Cohen's Kappa was calculated separately for all five attributes. Table 8.1 shows an example of the contingency matrix for the motivation attribute, with its Cohen Kappa calculated to 0.98.

Annotator		**B**						**Total**
	Category	Destruction	Disruption	Espionage	Financial Gain	Ideology	Unknown	
A	Destruction	165	1	0	0	0	0	**166**
	Disruption	1	315	0	1	0	0	**317**
	Espionage	0	0	94	1	0	1	**96**
	Financial Gain	0	1	1	362	1	0	**365**
	Ideology	0	1	1	1	25	0	**28**
	Unknown	0	0	0	0	0	28	**28**
Total		166	318	96	365	26	29	**1000**

Table 8.1: Cohen's Kappa consistency matrix

A summary of the calculation of Cohen's kappa values for the respective five attributes is depicted in Table 8.2.

Minor differences in classification emerged for *Motivation* and *Attack Type* attributes but were resolved through secondary review, with both annotators reaching consensus. More substantial disagreements were identified in the classification of TTPs and *Sector* attributes. Although most of the discrepancies were resolved during the second examination phase, certain cases presented legitimate ambiguity in which multiple classifications could be considered valid. To address this issue, a merging approach was implemented allowing incidents to receive multiple attribute values when both expert classifications were deemed accurate. This solution accommodates the inherent complexity of cybersecurity incidents that may legitimately span multiple categories within these attribute domains. The completion of this annotation process produced the *Human Annotated Incident Corpus*, serving as the authoritative ground truth dataset for subsequent AI model evaluation and comparison.

Attribute	**Cohen's Kappa Value**
Threat Name	100
Motivation	0.98
Attack Type	0.92
TTPs	0.81
Sector	0.83

Table 8.2: Cohen's Kappa values

Category	Tag-Attribute	Values
Attacker	Threat_Name	Any ATTACKER NAME or MALWARE NAME or NONE
	Motivation	DESTRUCTION, DISRUPTION, ESPIONAGE, FINANCIAL GAIN, IDEOLOGY, UNKNOWN
Attack	Attack_type	DATA, DOS/DDOS/RDOS, MALWARE, MISINFORMATION/DISINFORMATION, RANSOMWARE, SUPPLY CHAIN ATTACK, WEB THREATS,bZERO DAY, SOCIAL ENGINEERING THREATS, UNKNOWN
	Tactics, Techniques and Procedures	TTP ID [38]
Victim	Sector	CHEMICALS, DEFENSE, DRINKING WATER, EDUCATION, ENERGY, FINANCIAL, FOOD, GOVERNMENTAL, HEALTHCARE, ICT, LAW ENFORCEMENT & MILITARY, MANUFACTURING, MEDIA AND COMMUNICATION, POSTAL/COURIER, RETAIL, SERVICES, SPACE, TARGETED INDIVIDUAL, TRANSPORT, WASTE MANAGEMENT, WASTE WATER, UNKNOWN

Table 8.3: Classification values per tag.

8.3 Model Selection and AI Annotation

Multiple LLMs were selected for evaluation, including Gemini 1.5 Pro [135] and GPT-4o [136] (both paid versions), alongside BERT-based models TRAM [123] and TTPHunter [122] trained for specific extraction tasks. GPT-4o has demonstrated effectiveness in threat intelligence extraction [131], while TRAM and TTPHunter represent publicly available state-of-the-art models for this domain. Custom NER models developed within AiCEF were employed as baseline alternatives, trained on domain-specific cybersecurity data, and optimized for resource efficiency and privacy preservation. These models offer significant advantages for organizations processing sensitive threat intelligence that cannot be shared with propri-

etary online LLMs. Each model independently categorized cybersecurity texts according to five predefined attributes, with model usage summarized in Table 8.4. This process produced the *AI Annotated Incident Corpus* for subsequent performance evaluation against human expert annotations.

Attribute	Gemini 1.5 Pro	GPT-4o	TTPHunter	TRAM	AiCEF NER
Threat Name	✓	✓	-	-	✓
Motivation	✓	✓	-	-	-
Attack Type	✓	✓	-	-	✓
MITRE TTPs (*Top 50)	✓	✓	✓	✓	-
Sector	✓	✓	-	-	✓

Table 8.4: Model usage per attribute extracted

8.4 Evaluation Metrics & Comparative Analysis

Performance assessment employed precision, recall, and F1-score metrics, supplemented by qualitative error analysis to understand model misclassifications where AI outputs deviated from human expert categorization. Based on F1 scores in Figure 8.3, key findings emerged: Gemini 1.5 Pro performed well on Threat Name (0.9549) and Motivation (0.7450) but struggled with MITRE TTPs (0.5484), requiring more domain-specific knowledge. GPT-4o achieved the highest Threat Name score (0.9666) but showed weakness in Motivation (0.6598) and MITRE TTPs (0.4828). TTPHunter, despite being fine-tuned for TTP detection, performed poorly on MITRE TTPs (0.2213), highlighting the importance of corpus-specific validation. Training corpus differences can significantly impact performance regardless of model specialization. TRAM achieved moderate MITRE TTPs detection (0.4465), outperforming TTPHunter. AiCEF's NER model excelled in Attack Type (0.8929) and Sector (0.6411) detection but underperformed elsewhere, demonstrating strength in targeted extraction but limited general capability. High performance resulted from training on the exact taxonomy used in the CTL experiment. Overall, LLMs (Gemini 1.5 Pro, GPT-4o) offered broader generalization but faltered in specialized domains like MITRE TTP extraction. Task-specific models (AiCEF NER, TRAM) excelled in narrow attributes but lacked versatility across broader detection tasks.

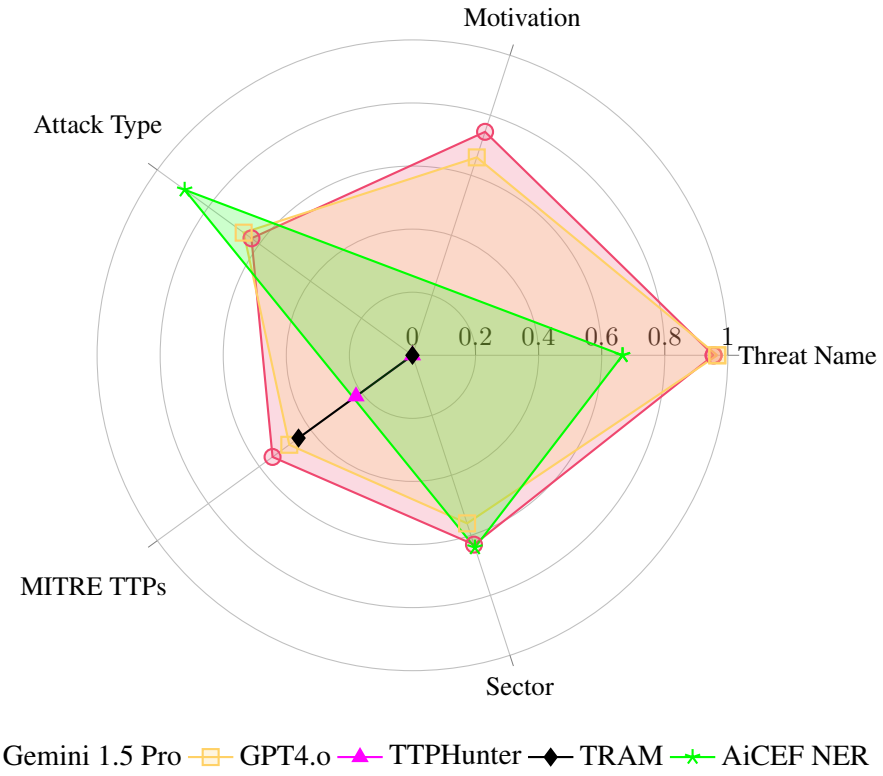


Figure 8.3: F1 Score Comparison on Attribute Detection per Model (Spider Diagram)

8.4.1 Attribute-Specific Insights

Diving deeper into the statistics collected during the model annotation phase, a series of valuable insights regarding the performance of each type of model used can be identified.

8.4.1.1 Threat Name Identification

Table 8.5 presents Threat Name Detection performance across three models using standard classification metrics. Gemini 1.5 Pro achieved perfect precision (1.0) and strong recall (0.914), resulting in an F1 score of 0.9549 despite missing 86 cases. GPT-4o outperformed with 935 TPs and 0 FPs, achieving 1.0 precision and superior recall (0.9349), yielding the highest F1 score (0.9666). AiCEF NER showed limitations with 713 TPs and 465 FPs, resulting in low precision (0.6111) and moderate recall (0.7309), producing a lower F1 score (0.6662). This underperformance may stem from training data limited to pre-2023 incidents,

highlighting the need for model updates. Analysis reveals interesting patterns in threat group identification (Table 8.6). All models consistently identified prominent groups like NoName057(16), demonstrating reliability for well-known actors. However, the LockBit ransomware group showed discrepancies: humans identified 66 cases, AiCEF NER found 63, GPT-4o detected 64, while Gemini split classifications into "LockBit ransomware group" (50) and "LockBit" malware (18), reflecting classification ambiguity found in the literature. RansomHouse and LeakBase appeared only in human annotations, absent from all AI models. Notably, human experts could not annotate threat names in 352 articles where attackers were not clearly indicated, demonstrating the inherent complexity of threat actor identification in cybersecurity text.

Threat Name (1000)	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1 Score**
Gemini 1.5 Pro	914	0	86	1	0.914	0.9549
GPT-4o	935	0	65	1	0.9349	0.9666
AiCEF NER	713	465	269	0.6111	0.7309	0.6662

Table 8.5: Performance Metrics for Threat Name Detection

Attacker Name	**Human Hits**	**Gemini 1.5 Pro**	**GPT-4o**	**AiCEF NER**
NoName057(16)	110	111	107	110
LockBit	66	50 (LockBit ransomware group) / 18 (LockBit)	64	63
Anonymous Sudan	35	34	35	-
Play ransomware group	18	24	16	-
BianLian	11	11	11	-
Royal ransomware group	8	-	8	-
RansomHouse	8	-	-	-
LeakBase	8	-	-	-
BlackCat ALPHV	7	6 (BlackCat/ALPHV ransomware group)	7	-

Table 8.6: Attacker name True Positive Hit count excluding "UNKNOWN"

8.4.1.2 Motivation Identification

Table 8.7 compares Motivation detection performance between two LLMs, as no BERT or NER models were available for this task. Gemini 1.5 Pro achieved balanced performance with precision of 0.7631 and recall of 0.7276, yielding an F1 score of 0.7450. GPT-4o showed lower precision (0.6074) but comparable recall (0.7201), resulting in a reduced F1 score of 0.6598 due to higher false positive rates. Although both models demonstrated reasonable performance, the results indicate that Motivation detection cannot be reliably assigned exclusively to AI models without human oversight.

Motivation (1000)	TP	FP	FN	Precision	Recall	F1 Score
Gemini 1.5 Pro	622	193	233	0.7631	0.7276	0.7450
GPT-4o	548	354	213	0.6074	0.7201	0.6598

Table 8.7: Performance Metrics for Motivation Detection

This dataset reveals substantial discrepancies in model performance among various categories (see Figure 8.4). Gemini 1.5 Pro outperforms GPT-4o in most categories, but GPT-4o demonstrates strong performance in detecting 'Disruption' related events, although it exhibits a significant decline in accuracy for 'Destruction', achieving a **0%**hit rate. This indicates that GPT-4o may have constraints in managing narrower or more specific threat classifications. Moreover, the analysis conducted shows that Gemini 1.5 Pro performs subpar for the used dataset.

8.4.1.3 Attack Type Classification

Table 8.8 presents Attack Type Detection performance across three models. Gemini 1.5 Pro achieved a precision of 0.6632 and a recall of 0.6017, yielding an F1 score of 0.6311. Low recall indicates missed attack types, while moderate false positives affect precision. GPT-4o showed balanced precision and recall (0.6616), resulting in a slightly better F1 score (0.6616) than Gemini, although both models struggled with false positives and missed detections. AiCEF NER significantly outperformed with high precision (0.9499) and recall (0.8414), achieving the best F1 score (0.8929). This superior performance, particularly in precision due to low false positive rates, results from training on the same taxonomy used for CTL construction.

Attack Type (990)	TP	FP	FN	Precision	Recall	F1 Score
Gemini 1.5 Pro	595	302	395	0.6632	0.6017	0.6311
GPT-4o	655	335	335	0.6616	0.6616	0.6616
AiCEF NER	833	44	157	0.9499	0.8414	0.8929

Table 8.8: Performance Metrics for Attack Type Detection

All findings are illustrated in Figure 8.5. It is worth noting that all models performed well in identifying categories like "Malware" and "Ransomware." For example, both Gem-

ini 1.5 pro and GPT-4o achieved hit rates above **90%** in the "Ransomware" category. This consistency highlights the models' proficiency in handling common and well-defined cyber threats. On the contrary, the "Intrusion" category exposed significant weaknesses, particularly in Gemini 1.5 pro, where the hit rate was only **25%**, coupled with a false positive count of **45**. This indicates that the model may be over-relying on specific indicators that do not generalise well across different intrusion scenarios.

8.4.1.4 Sector Identification

Table 8.9 summarizes Sector detection performance across three models. Gemini 1.5 Pro achieved a precision of 0.6659 and a recall of 0.6000, yielding an F1 score of 0.6312. Moderate performance was affected by missed detections and false positives. GPT-4o showed lower precision (0.5477) and similar recall (0.5737), resulting in the lowest F1 score (0.5604) among the three models due to high false positive rates. AiCEF NER demonstrated higher precision (0.7382) but lower recall (0.5670), achieving an F1 score of 0.6411. While showing decent performance, it missed a significant number of actual sectors. Gemini 1.5 Pro achieved the highest F1 score, followed by AiCEF NER, which exhibited better precision but suffered from higher false negatives. All models struggled with recall, but Gemini 1.5 Pro demonstrated the most balanced performance for this task.

Sector (990)	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F1 Score**
Gemini 1.5 Pro	578	290	384	0.6659	0.6000	0.6312
GPT-4o	533	440	396	0.5477	0.5737	0.5604
AiCEF NER	556	197	424	0.7382	0.5670	0.6411

Table 8.9: Performance Metrics for Sector Detection

Going into more detail on the TP rates per sector, an overview can be seen in Figure 8.6, where a pattern is clearly visible. All models struggle to categorise the Retail industry, while GPT-4o and AiCEF NER also score low in sectors like Healthcare, Finance, and Energy.

8.4.1.5 MITRE TTP Identification

Table 8.10 presents MITRE TTPs detection performance across four models: Gemini 1.5 Pro, GPT-4o, TTPHunter, and TRAM. All models focused on detecting the Top 50 MITRE TTPs

using the TTPC, with this limitation enforced for BERT model training and applied at the prompt level for LLMs to ensure comparable results. The task was treated as multi-labeling to simulate real-world reports containing multiple TTPs. Gemini 1.5 Pro achieved the highest precision (0.324) and recall (0.6389), yielding an F1 score of 0.43. Despite moderate performance, false positives affected overall effectiveness. GPT-4o showed lower precision (0.276) but higher recall (0.69), resulting in an F1 score of 0.394. High false positive rates hindered true positive identification. TTPHunter demonstrated poor performance with very low precision (0.1426) and recall (0.3214), producing an F1 score of 0.1982 due to high false positive and false negative rates. TRAM achieved a precision of 0.3204 and a recall of 0.4232, yielding an F1 score of 0.3667, indicating subpar performance similar to Gemini but with a lower recall.

MITRE TTPs (345)	TP	FP	FN	Precision	Recall	F1 Score
Gemini 1.5 Pro	220	239	125	0.4792	0.6384	0.5484
GPT-4o	238	386	107	0.3806	0.6899	0.4828
TTPHunter	111	550	234	0.1675	0.3217	0.2213
TRAM	146	164	199	0.4710	0.4232	0.4465

Table 8.10: Performance Metrics for MITRE TTPs

In MITRE TTPs detection, Gemini 1.5 Pro performs best, particularly in recall, but it suffers from a high false positive rate, leading to a moderate F1 score. GPT-4o shows slightly better detection capability with more true positives, but also struggles with precision. TTPHunter exhibits significantly poorer performance than TRAM, with TTPHunter having the lowest metrics overall.

Although TRAM has a lower detection rate than GPT-4o, it produces fewer false positives, leading to very similar F1 scores. All models face challenges in balancing precision and recall, with high false positive rates being a critical issue. This highlights the difficulty in accurately detecting MITRE TTPs as a task overall.

8.4.2 Key Performance Findings

For *Threat Name detection*, both LLMs (Gemini 1.5 Pro and GPT-4o) demonstrated high precision and recall, indicating effective identification with minimal false positives, making them suitable for real-time intelligence extraction in CTL creation.

In *Motivation* detection, moderate performance was achieved overall, with Gemini 1.5 Pro outperforming GPT-4o in precision and F1 score. Although LLMs can generally identify threat actor motivations effectively, high false positive rates suggest their use as a first step in hybrid classification schemes where humans verify proposed annotations.

Then, for *Attack Type* detection, AiCEF's NER outperformed both LLMs, demonstrating the effectiveness of task-specific models for precise attribute extraction in CTL intelligence gathering. For *Sector* detection, both LLMs performed poorly in precision and recall, indicating significant effectiveness gaps. AiCEF NER showed better precision but faced recall challenges, highlighting improvement opportunities. AI models would benefit from more diverse and representative training data to improve sector identification as defined in threat landscapes.

Finally, for *MITRE TTP detection*, all models struggled with low detection rates and high false positives. TRAM performed relatively well compared to LLMs, offering a robust and time-efficient solution for analysts while limiting exposure of threat-related information to public LLM providers. Based on these findings, both LLMs and NER models can serve as first-step pre-tagging in hybrid classification schemes requiring human verification. Local MITRE TTP detection remains challenging and requires further enhancement. To address this, a novel methodology is proposed that relies on model crowd-sourcing rather than re-training.

8.4.3 Resource Comparison Findings

	Human Expert	**Gemini 1.5 Pro**	**GPT-4o**	**TRAM**	**ThreatHunter**	**AiCEF NER**
Avg. Time (min) per Text						
Threat Name - Motivation - Attack Type - Sector	10	0.02	0.02	n/a	n/a	0.01
MITRE TTPs	15	0.3	0.3	1	0.1	n/a
Avg. Cost (€) per Text						
Threat Name - Motivation - Attack Type - Sector	n/a	0.023	0.065	n/a	n/a	0.003
MITRE TTPs	n/a	0.023	0.14	0.003	0.003	n/a

Table 8.11: Annotation resources per task breakdown

During this research, I attempted to estimate potential organizational savings from automating CTL intelligence extraction using AI. While these findings (Table 8.11) are preliminary and should be treated cautiously, they provide useful indicators of potential financial

and time savings if such approaches improve. I recorded annotation times and found that extracting four attributes (Threat Name, Motivation, Attack Type, and Sector) requires approximately 10 minutes per text for reading, tagging, and achieving annotator agreement. TTP extraction requires an additional 15 minutes due to the inclusion of richer threat information and technical details. For AI models, basic tag extraction time is negligible, while TTP extraction is computationally intensive, requiring up to 1 minute for longer texts. Cost calculations vary significantly based on expert seniority and country rates. For 1,000 articles, human annotation of five attributes is estimated at above €10,000. While exact predictions are beyond this work's scope, this indicates realistic improvement ranges. Using GPT-4o, tag extraction costs approximately €65 (€0.065 per text), while TTP extraction costs €140 (€0.14 per text). Total extraction from 1,000 articles using GPT-4o would cost €205. Gemini 1.5 Pro analysis was covered by a monthly Premium subscription (€23 flat rate), extracting all attributes from 1,000 texts at €0.023 per article. More articles could be analyzed at the same rate, reducing per-text costs. Local models required approximately €3 for virtual machine setup and package installation, or €0.003 per article. Again, more texts could be analyzed under the same flat rate. Expert training time, tag verification, and fine-tuning opportunities should be considered for accurate savings estimates. Nevertheless, my limited tests showed annotation time decreased from 10 to 2 minutes per text for non-TTP attributes when checking pre-annotated tags.

8.4.4 Proposed Annotation Workflow

Taking into account the above, the use of AI models for intelligence extraction for CTL generation can be represented in a hybrid workflow, as seen in Figure 8.7. Specific types of AI models perform better than others based on the attribute annotated and should be preferred, as indicated by the results collected. Attributes like Threat Name and Attack Type can be annotated purely by AI models, while others like Motivation, Sector, and MITRE TTPs still require human curation. AI can be utilised as an initial annotation prior to or in parallel with human annotation and play a supportive role in the decision-making of the categorisation of these attributes.

8.5 Optimisation through Voting-Based Agreement Methodology

Wanting to contribute further to the improvement of the extraction precision of a key attribute for the success of cybersecurity scenarios, the MITRE TTPs, I performed several experiments using combinations of existing AI models.

Based on the findings presented above, I decided to test whether a classic systematic voting mechanism [163] could ensure consensus between multiple AI models, denoted as $M_1, M_2, ...M_n$, to resolve the problem of high false positives observed in the MITRE TTP extraction task. My proposed method uses voting to determine the final consensus output. For the first experiment, I decided to test two algorithms: one based on simple majority voting (Algorithm 3) and one based on weighted voting (Algorithm 4), where each model is assigned a weight based on its historical performance, accuracy, or confidence in its predictions. Finally, I experimented with ensemble learning, a machine learning technique that aggregates two or more models to produce better predictions. In other words, an ensemble model combines several individual models to make more accurate predictions than a single model alone [164].

For all methods, let's assume that:

- All models are pre-trained and exhibit competence in the specified domain.
- Each model produces outputs represented as one or more MITRE TTP labels.
- For the algorithm to work, at least three models should vote.

8.5.1 Simple Majority Voting

In the simple majority method, the goal is to aggregate the outputs of the models to find a consensus decision. The label is accepted if N models agree on a label based on a user-defined majority *Threshold*. Let's set the threshold to **Threshold**=**2**, where at least two models agree out of three models available, and **Threshold**=**3**, where three models agree out of four models available.

Algorithm 3 The simple majority voting method.

Input: Set of AI models $\{M_1, M_2, \ldots, M_n\}$
Input: Test instance X
Input: User-defined threshold T
Output: Aggregated consensus output C
for each model M_i in $\{M_1, M_2, \ldots, M_n\}$ **do**
    Get output O_i from $M_i(X)$
end for
Initialise a count dictionary count for TTP labels
for each output O_i **do**
    Increment count of O_i in count
end for
Initialise an empty list accepted_labels
for each TTP label C_j in count **do**
    if count of $C_j \geq T$ **then**
        Add C_j to accepted_labels
    end if
end for
if accepted_labels is empty **then**
    Return: No consensus found
end if
Find the TTP label C_{max} with the highest count in count from accepted_labels
Set $C \leftarrow C_{max}$
return C

8.5.2 Weighted Voting

For the weighted voting method, each model's strengths in coverage and precision were considered when assigning prediction weights. A trial-and-error approach tested all possible weighted combinations with weights between 0.5 and 5 for groups of three or more models. The threshold was set at five and above ($\geq$5) to identify optimal F1 scores.

For the three-model weighted voting system (Gemini 1.5 Pro - GPT-4o - TRAM), testing all weight combinations at Threshold=5 yielded 305 combinations scoring above 0.5484. The maximum F1 score achieved was 0.5924.

Optimal weight ranges per model were identified as: Gemini 1.5 Pro: [5], GPT-4o: [0.5-4.5], TRAM: [0.5-4.5], where GPT-4o and TRAM combined scores should total at least five.

For the four-model weighted voting system (Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter), testing all weight combinations at threshold T=5 yielded 1,881 combinations scoring above 0.5484. The maximum F1 score achieved was 0.5958.

Optimal weight ranges per model were: Gemini 1.5 Pro: [5], GPT-4o: [0.5-4], TRAM: [0.5-4], TTPHunter: [0.5-4], where GPT-4o, TRAM, and TTPHunter combined scores should total five or above.

Based on individual model F1 performance, the following weighting strategy was adopted: maximum weight of 5 for Gemini 1.5 Pro to leverage its superior F1 score, with balanced weights of 0.5-4.5 distributed among GPT-4o, TRAM, and TTPHunter to benefit from their contributions.

Weight combination distribution analysis achieving maximum F1 score using weighted voting is presented in Figure 8.8. The optimal threshold for the proposed method is Threshold=5.

8.5.3 Ensemble learning: Stacking Model

In my final attempt, I combined individual model predictions as features and trained a meta-model to make final predictions. This meta-model learns optimal weighting and aggregation of each model's outputs. Each pre-trained model (Gemini, GPT, TRAM, and TTPHunter)

Algorithm 4 The weighted voting method.

Input:
  M: Set of models $\{M_1, M_2, \ldots, M_n\}$
  W: Set of corresponding weights for each model $\{w_1, w_2, \ldots, w_n\}$
  $T_{predicted}$: Set of techniques predicted by each model M_i
  $Threshold$: Voting threshold (e.g., 5)
  T: Set of all possible techniques
Output:
  T_{final}: Set of final predicted techniques
Initialise:
  $Votes[t] \leftarrow 0, \forall t \in T$ ▷ Initialise vote count for each technique
for each model $M_i \in M$ **do**
    for each predicted technique $t \in T_{predicted}[M_i]$ **do**
      $Votes[t] \leftarrow Votes[t] + w_i$ ▷ Add weight of model M_i to the technique t
    end for
end for
$T_{final} \leftarrow \emptyset$ ▷ Initialise the set of final predicted techniques
for each technique $t \in T$ **do**
    if $Votes[t] \geq Threshold$ **then**
      $T_{final} \leftarrow T_{final} \cup \{t\}$ ▷ Include t in the final prediction set if vote exceeds threshold
    end if
end for
return T_{final}

generated individual predictions, with these outputs forming a new dataset representing the ensemble's feature space. Then, I trained a simple classifier on this dataset to predict final TTP labels. The implementation was adapted for multi-label classification using the Multi Output Classifier[5] wrapped around a Logistic Regression model to handle multi-label output. I collected predictions from all four models for each dataset sample, using them as input features to the meta-model. The meta-model was trained on actual ground truth labels to learn optimal combination strategies for base model predictions.

8.5.4 Voting Results Comparison

Upon reviewing Table 8.10, the proposed methodology surpasses the maximum F1 score of the best-performing model (Gemini 1.5 Pro, F1 score: 0.5484) and reduces false positives below the average rate of combined models (Average FP (3 models): 263, Average FP (4 models): 334.75), representing an optimization to the intelligence extraction process of CTL generation.

The objective was achieved by testing all possible combinations of four available models (Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter) for both voting methods and training a stacking model. Table 8.12 showcases successful combinations that improved the previous best F1 score without retraining existing models while reducing false positive rates.

Weighted voting proved to be the most effective among traditional voting methodologies. The four-model combination (Gemini 1.5 Pro—GPT-4o—TRAM—TTPHunter) achieved a 4.74% increase.

Simple majority voting achieved a 3.11% F1 score increase using only the 3-model combination (Gemini 1.5 Pro - GPT-4o - TRAM), reducing false positives by 32 units. Simple majority voting for four models failed to improve the F1 score, potentially indicating that the fourth model (TTPHunter) degraded performance.

Ensemble learning, specifically a newly trained stacking model based on predictions from all four models (Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter) using ninety (90) samples to predict 10 out of the available dataset, yielded satisfactory results with an F1 score of 0.5714, translating to a 2.30% increase.

[5] https://scikit-learn.org/dev/modules/generated/sklearn.multioutput.MultiOutputClassifier.html

Given a larger annotated dataset, a more robust model can be trained to extract tags for the MITRE TTP attribute.

Model Combination	Precision	Recall	F1 Score	Improvement
Simple Majority Voting				
Gemini 1.5 Pro - GPT-4o - TRAM	0.4979	0.6812	0.5795	+3.11%
Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter	0.4649	0.6957	0.5585	+1.01%
Weighted Voting				
Gemini 1.5 Pro - GPT-4o - TRAM	0.5010	**0.7246**	0.5924	+4.40%
Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter	0.5194	0.6986	**0.5958**	**+4.74%**
Stacking Model				
Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter	**1.0000**	0.4000	0.5714	+2.30%

Table 8.12: Optimisation Methods Comparison: Simple Majority Voting vs. Weighted Voting vs. Stacking Model

8.6 Findings

Automating the intelligence extraction process using AI models presents significant potential for reducing costs and time spent compared to human annotation. While human annotation of a dataset of 1,000 articles would cost some thousands of euros, AI models can drastically reduce these costs to as low as €200 under specific subscription models. Additionally, the time savings are substantial as AI models almost instantly complete the basic tag extraction. In contrast, TTP extraction, which is more computationally intensive, is still much faster than manual methods. With further refinement and integration of hybrid approaches, organizations could potentially reduce the costs of CTL generation by more than half while significantly accelerating the process.

The presented approach can potentially automate and improve a cybersecurity expert's overall throughput when preparing organizational threat landscapes. It is estimated that by leveraging the proposed methodology, a cybersecurity expert presented with an AI pre-annotated report could boost their performance by up to 80% (see Table 8.11). This performance enhancement depends on error-rate tolerance levels and specific organizational objectives, with more sophisticated models potentially required for comprehensive control identification across relevant TTPs.

Wanting to contribute to this direction, I proposed a novel approach focusing on MITRE TTP extraction, leveraging existing models without retraining, and demonstrating promising results and significant potential for future research. Using voting-based methods and ensem-

ble learning, I outperformed the highest F1 score achieved by an individual model (Gemini 1.5 Pro with an F1 score of 0.5484) on this rather complex extraction topic. Specifically, a weighted voting combination of four models (Gemini 1.5 Pro, GPT-4.o, TRAM, TTPHunter) achieved a 4.74% improvement of the F1 score, dropping the FP rate by 16 points and improving TP by 21 points. The simple majority voting method also delivered a 3.11% F1 score improvement using the three-model combination, reducing the false positive rate by 32 units. Additionally, a stacking model trained on a small dataset showed a 2.30% increase in F1 score, with the potential for further optimization as more data becomes available. The proposed methodology can be extended to any combination of LLM and BERT-based models and easily extended to cover the extraction of all MITRE TTPs.

The improved MITRE TTP extraction capabilities directly enhance the AiCEF framework's ability to generate more sophisticated and realistic cybersecurity exercise scenarios. Enhanced TTP identification enables better threat actor behavior mapping and more accurate attack sequence modeling within the incident generation module (IncGen). This improvement facilitates better integration with the APT enhancer component, allowing exercise planners to create training scenarios that test specific defensive capabilities against well-documented attack techniques as discussed in previous chapters.

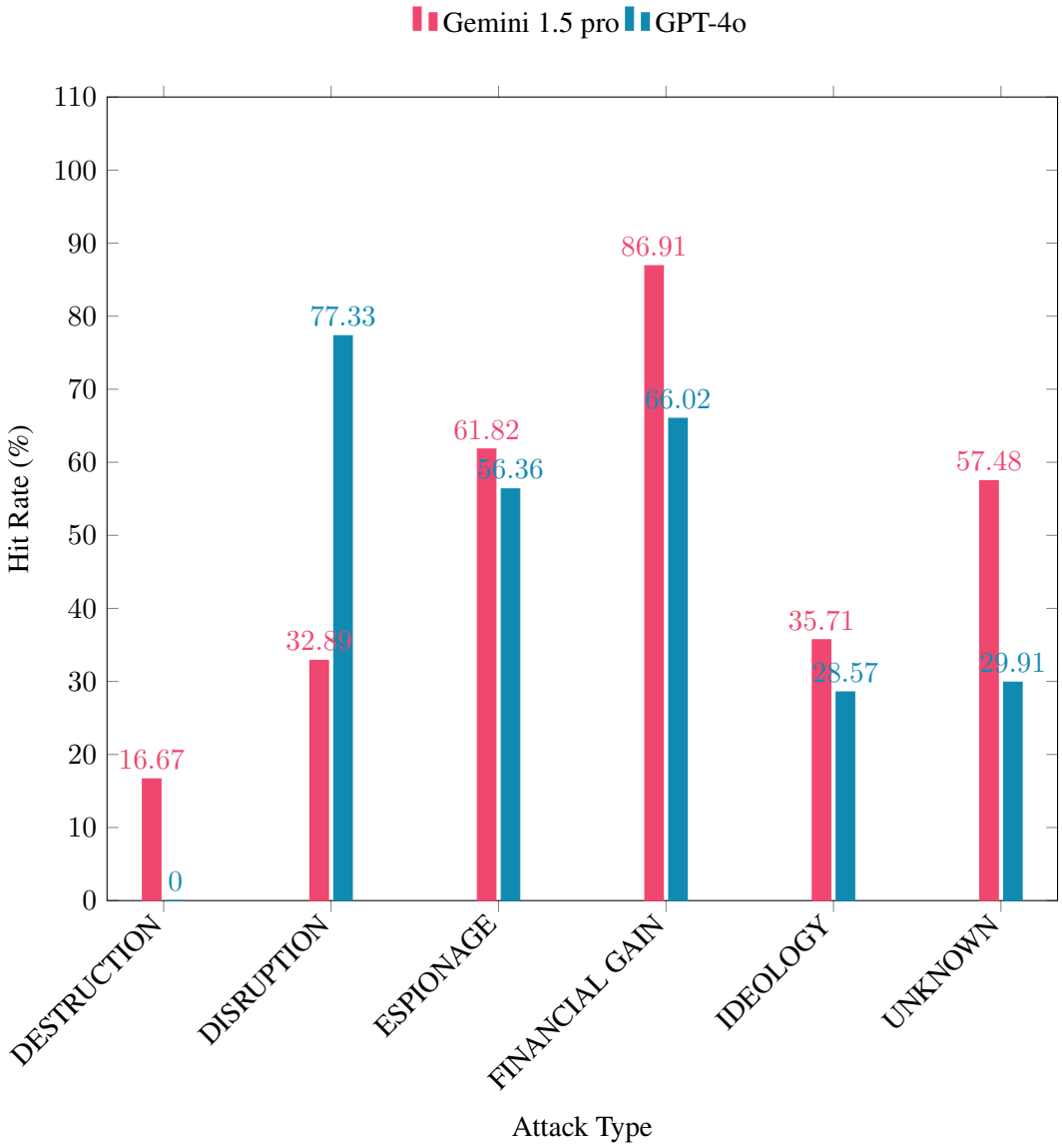


Figure 8.4: Motivation (%) Model Hit Rate Comparison

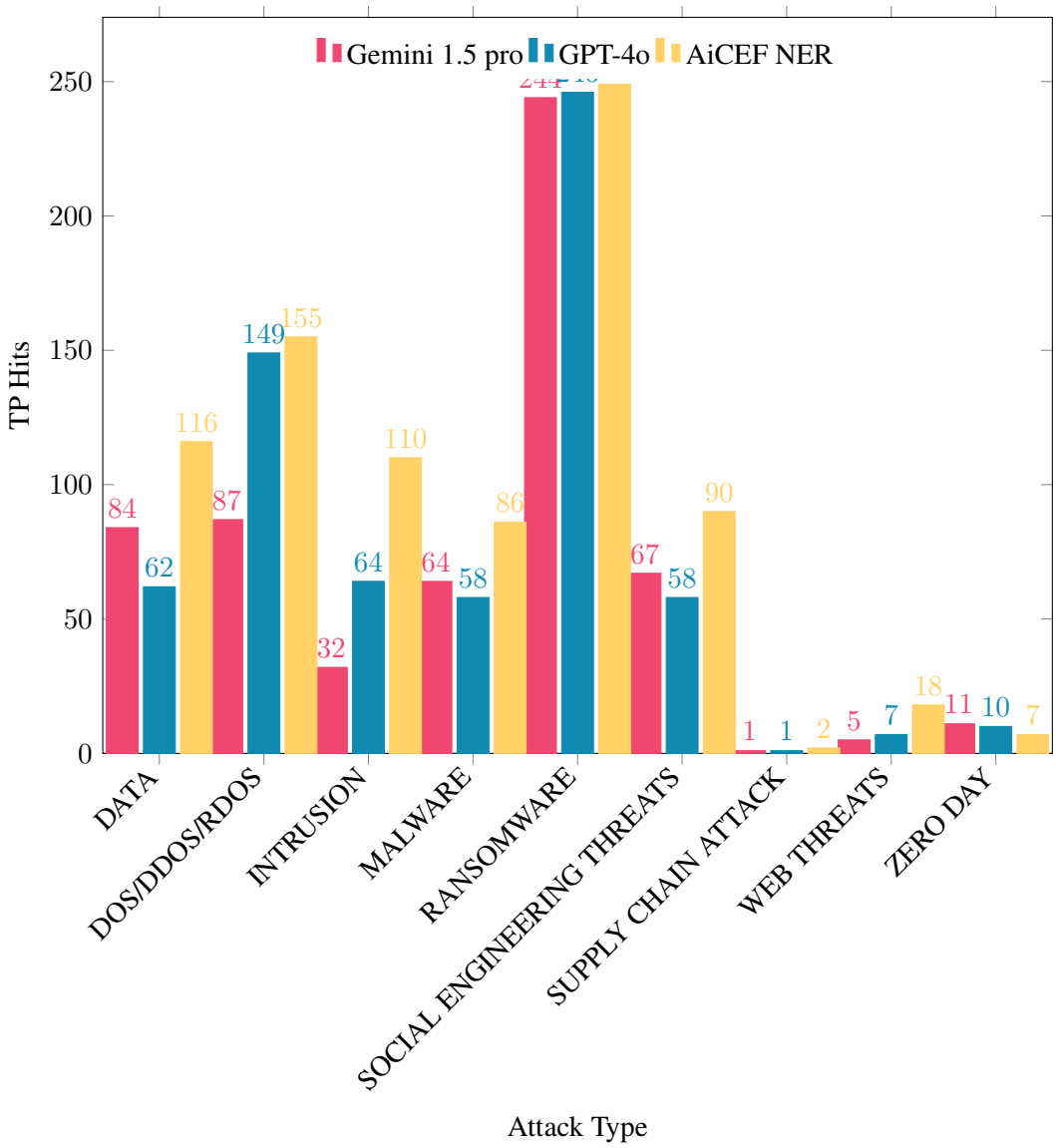


Figure 8.5: True Positive (TP) Comparison for Attack Type

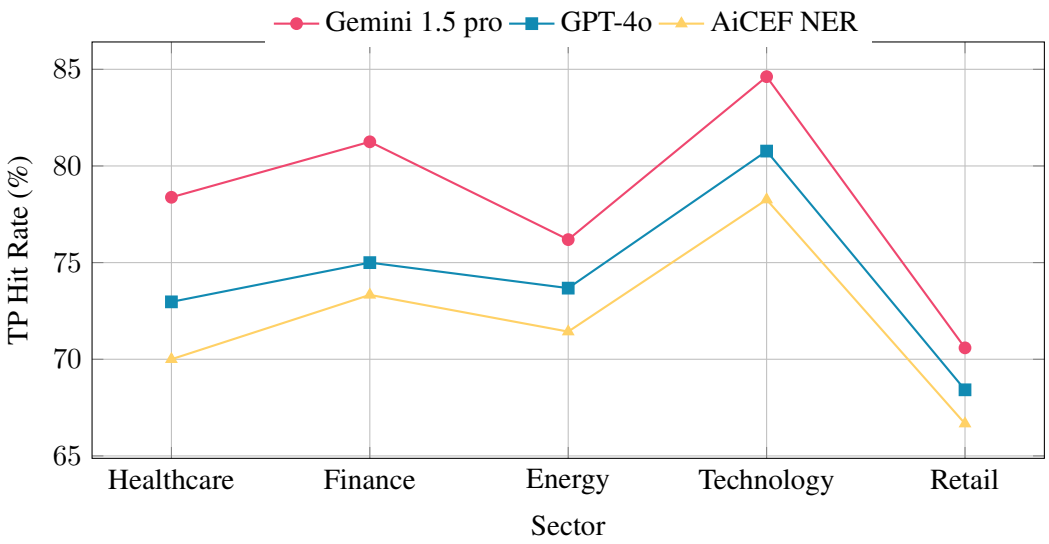


Figure 8.6: True Positive Hit Rates (%) breakdown by Model per Sector

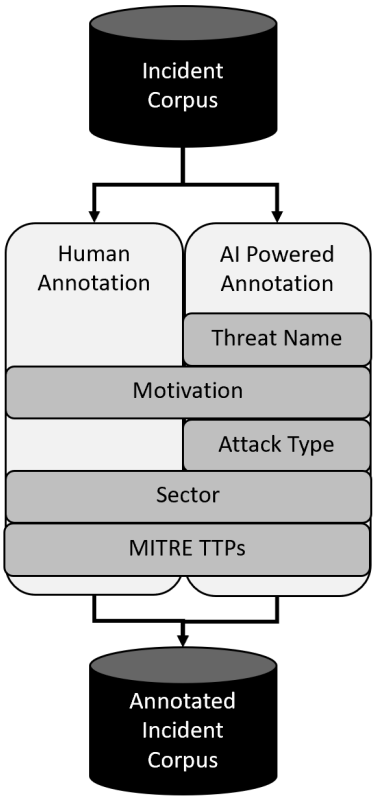


Figure 8.7: Proposed hybrid annotation methodology

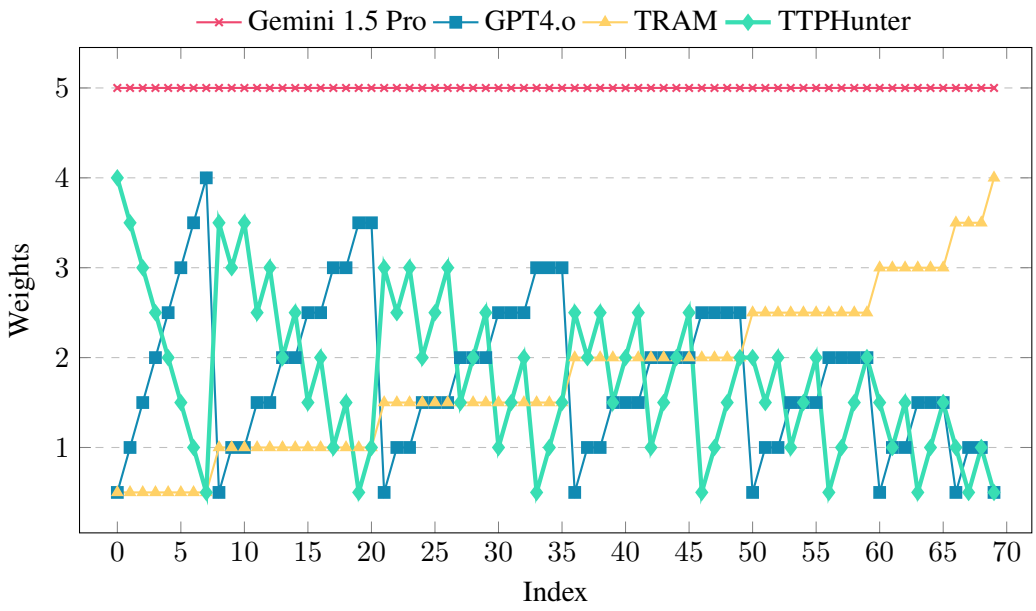


Figure 8.8: Distribution of Weights achieving maximum F1 Score (0.5958) in MITRE TTP extraction.

Chapter 9

AiCEF Automation

The proposed AI-powered cyber exercise framework has been automated using custom-made software as part of this research. The motivation behind creating the AiCEF tool was to explore the use of tools to speed up the data analysis, but also to provide the ground for the evaluation of the proposed methodology by experts.

9.1 Key Features of AiCEF tool

The primary purpose of the AiCEF tool is to support the research presented in this thesis by providing a robust platform for automating the creation of realistic and up-to-date cyber exercise content. The framework includes the following features developed as part of this research:

- **Ontology-Based Framework**: CESO serves as the foundational structure for AiCEF tool, ensuring that all generated scenarios are logically consistent, machine-readable, and human-interpretable. Text is analysed using the AI models developed to extract and store interesting cybersecurity attributes.
- **AI-Driven Incident Generation**: AiCEF tool enables the automated creation of incidents based on TTPs. It specifically models scenarios derived from the activities of 125

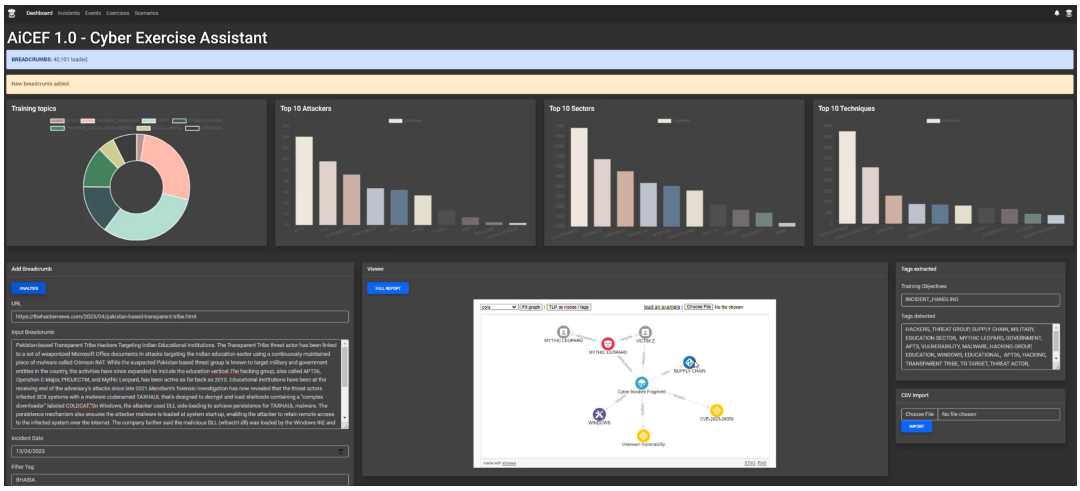
known Advanced Persistent Threat (APT) groups, ensuring high fidelity to real-world threat dynamics.

- **Multi-Format Output**: The framework produces output in both **STIX 2.1** (structured JSON format) for machine-readability and **PDF** or **XLS** for human-readable exercise documentation.
- **User-Friendly Accessibility**: AiCEF tool simplifies the CSE generation process, allowing users with minimal technical expertise to input high-level information and produce complete and realistic scenarios with minimal effort.
- **Technology Stack**:
  - Built using **Python** and **EEL** for modular and flexible development.
  - Employs the open-source library **Stixview** to enable visualization and interaction with structured threat intelligence.
  - Stores extracted and generated data in a **MySQL** database, ensuring scalability and efficient data management.

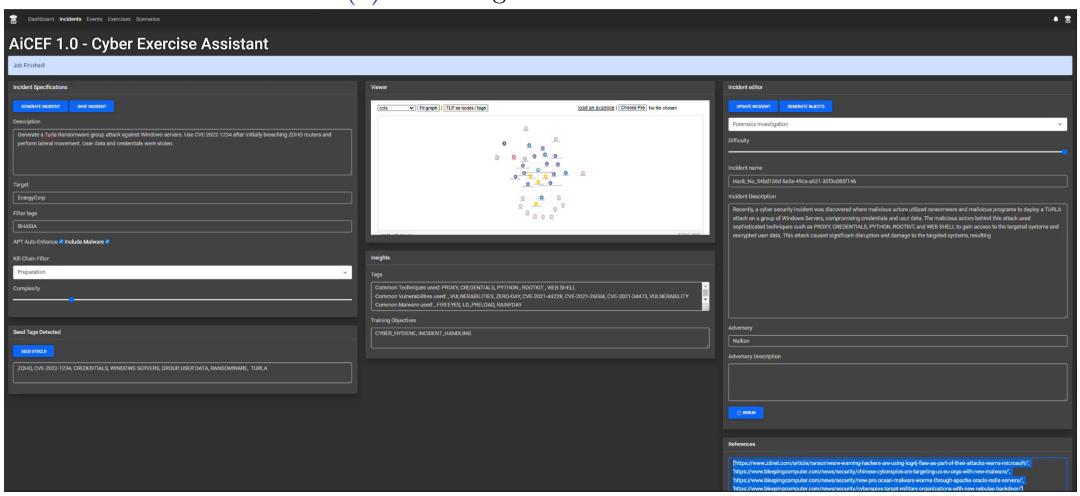
9.2 Role of AiCEF in Supporting Research

AiCEF tool was crucial in facilitating the research objectives outlined in this thesis. By automating the generation of cyber exercise scenarios, the framework enabled the iterative development, testing, and refinement of methodologies for ontology-driven scenario modelling. The structured outputs produced by the AiCEF tool were essential in evaluating the effectiveness of the proposed approach, particularly when comparing traditional, human-led methods with AI-powered techniques. Furthermore, its ability to model realistic scenarios at scale provided a reliable platform for conducting case studies and analysing key performance indicators, such as scenario relevance, quality, and generation speed.

In conclusion, the AiCEF tool represents a significant contribution to both the practical and theoretical dimensions of cybersecurity research. Its integration of advanced AI techniques, structured ontologies, and user-centric design ensures that it not only supports the goals of this thesis, but also provides a foundation for future advancements in the generation of cyber exercises.



(a) Knowledge Database Preview



(b) Visualization of Incident Data in STIX 2.1 Format.

Figure 9.1: Screenshots of AiCEF Tool Interface

Part IV

Closure

Chapter 10

Conclusions and Future Work

This chapter concludes the thesis providing a comprehensive overview of the findings and significant contributions achieved through the development and evaluation of the AiCEF framework, marking a milestone in the application of artificial intelligence and machine learning techniques to cybersecurity exercise development. The research presented has successfully addressed fundamental challenges in the field while establishing new methodological foundations that advance the state-of-the-art in the generation of automated cybersecurity training scenarios.

The work demonstrates how AI-powered approaches can revolutionize traditional cybersecurity exercise planning by introducing systematic automation, enhanced accuracy, and unprecedented efficiency gains. Through rigorous evaluation involving expert practitioners across multiple sectors and comprehensive performance analysis, this research has validated the practical utility and theoretical soundness of intelligent exercise generation methodologies.

The conclusions presented herein not only validate the research hypotheses, but also reveal unexpected insights that expand our understanding of how AI can augment human expertise in cybersecurity education. These findings provide a solid foundation for future researchers and practitioners seeking to leverage advanced computational techniques in cybersecurity training development, offering both theoretical frameworks and practical implementation strategies.

Furthermore, this chapter identifies emerging research directions that hold significant promise for advancing the field beyond the current work. The proposed avenues for future research address both immediate opportunities for enhancement and longer-term challenges that could shape the evolution of cybersecurity exercise development in an increasingly complex threat landscape, such as the current one.

10.1 Conclusions

This research has systematically investigated the application of AI methodologies to enhance cybersecurity exercise scenario generation, successfully addressing the inherent limitations of traditional human-centric approaches. The primary objectives of evaluating AI-driven methods for improving scenario relevance, timeliness, and quality while reducing expert workload have been comprehensively achieved through rigorous experimentation and evaluation.

The proposed AiCEF framework has bridged critical gaps in automated scenario generation, providing evidence-based solutions that significantly enhance cybersecurity preparedness. Through systematic integration of machine learning techniques, ontological frameworks, and intelligent automation, this work has established new paradigms for cybersecurity exercise development that benefit both novice and experienced practitioners.

The empirical findings reveal substantial improvements across multiple evaluation dimensions, with the framework consistently outperforming traditional methods in scenario quality, development efficiency, and threat landscape alignment. These achievements validate the research hypotheses while demonstrating the practical viability of AI-assisted cybersecurity training development. The following sections present the key contributions and breakthrough results that collectively establish new benchmarks for the field.

1. **Novel Ontological Framework and Automated Exercise Generation:**

   A foundational contribution of this research is the development of the Cyber Exercise Scenario Ontology (CESO), which represents the first comprehensive semantic framework specifically designed for cybersecurity exercise representation. This innovative ontology, combined with the automated exercise generation capabilities, has established a new paradigm for systematic and scalable cybersecurity training development.

- The CESO ontology successfully integrates established cybersecurity frameworks (STIX 2.1, MITRE ATT&CK, ISO 22398:2013) into a unified semantic model, enabling machine-readable representation of exercise scenarios while maintaining compatibility with existing threat intelligence systems.
- The ontology's formal implementation in Web Ontology Language (OWL) enables automated reasoning, consistency validation, and seamless integration with ontology management tools, facilitating both human comprehension and machine processing of exercise scenarios.
- The AiCEF framework demonstrates the practical application of this ontological foundation by enabling fully automated exercise generation from a single prompt to a STIX 2.1 formatted exercise, transforming unstructured threat intelligence into complete, structured cybersecurity training scenarios following ISO 22398:2013 standards.

2. **Enhanced Scenario Relevance and Quality:**

   The integration of artificial intelligence with human expertise has yielded unprecedented improvements in scenario quality and relevance. This groundbreaking approach has redefined the standards for cybersecurity exercise development by demonstrating that AI-human collaboration can surpass traditional methods across multiple evaluation criteria.

   - The hybrid approach of combining AI and human expertise consistently outperformed traditional human-only planning methods. This included surpassing the performance of seasoned planners in certain aspects.
   - AiCEF excelled in parameters such as resource relevance, event significance, and technical depth when generating scenarios as compared to those produced by human experts.
   - Integration of realistic behaviour of threat actors was a game changer in the way scenarios were generated by AiCEF. This would have not been possible without the improved extraction of data and forecasting using the proposed methodologies,

3. **Threat Modelling and Objectives Coverage:**

   One of the most significant achievements of this research lies in the comprehensive threat modeling capabilities integrated into the AiCEF framework. The system's ability to accurately simulate known threat actors while generating multi-sector scenarios

with exceptional training objective coverage represents a paradigm shift in how cybersecurity exercises can be designed and implemented.

- In real-life simulation of threat actors, designing complex exercise scenarios while taking into consideration incidents of multiple sectors affected, can reach up to an 81% coverage of training objectives and subsequently incidents that could materialize. This coverage is significantly higher than when designing narrow single-sector focused exercises (Public Administration: 5%, Energy: 74%, ICT: 72%). While a by-product of the main research, this finding can pave the way for planners to better strategize their content generation to be more effective.
- Source selection of information to use as input to the proposed framework is key to the magnitude of the coverage of foreseen incidents materializing for the generated scenarios. Scenario generation strategies which integrated multiple sources describing the same incident from different angles (ex. technical, operational), demonstrated the highest objective coverage improvement of +21%. This finding leads to another improvement in the scenario construction process, prompting planners to collect as much information as possible for incidents they would like to simulate and not count only on single references.

4. **Efficiency Gains:**

   The efficiency improvements achieved through AI-assisted exercise planning represent a transformative advancement for cybersecurity professionals. These results demonstrate that AI can not only accelerate the development process but also democratize access to high-quality scenario generation capabilities.

   - Exercise planners using AI tools achieved scenario creation times 33.33% faster than traditional methods, demonstrating the potential for significant time savings without compromising quality.
   - The most impressive finding, though, was the fact that novice planners using AI tools managed to match and in some cases outperform experienced planners with no AI support. This fact highlights the ability of AI to bridge expertise gaps and democratize scenario generation, allowing for more non-experts to be able to perform cybersecurity exercises, raising resilience for organizations globally.

5. **AI-Driven Forecasting and Threat Intelligence Analysis:**

The development of sophisticated AI-driven threat intelligence extraction and forecasting capabilities marks a major breakthrough in cybersecurity automation, further contributing to the improvement of the exercise generation lifecycle. These innovations have established new methodologies for threat analysis that significantly outperform traditional manual approaches while maintaining high accuracy standards.

- AI-driven threat intelligence extraction achieved cost and time reductions of over 50% compared to manual methods, with ensemble methods showing promise for improving accuracy in extracting complex attributes like MITRE TTPs. These findings further indicate that the use of AI can not only reduce costs and time but also improve the quality of scenarios within a single step.
- In an attempt to benchmark the evolution of AI tools that could be potentially used, this research provides one more interesting finding. Ensemble learning and voting-based approaches without retraining existing models yielded satisfactory results in the task of automated extraction of MITRE TTPs. More specifically, a weighted voting scheme utilizing the combination of Gemini 1.5 Pro, GPT-4.o, TRAM, and TTPHunter models achieved a 4.74% increase in F1 score, a 16-point reduction in false positives, and a 21-point increase in true positives. This constitutes an improvement not only for the CTI extraction field but also for the quality of the exercises generation lifecycle.
- Given enough historical data, forecasting of future incident trends at a sectorial level was possible, further improving the scenario quality and making the generated exercise more relevant and current. Among forecasting methods, Long Short-Term Memory (LSTM) and Prophet models proved most effective for identifying emerging sectorial threats that can be leveraged when designing CSE scenarios.

6. **CSE Prompt & Scenario Quality:**

   The innovative AiDPG methodology has established new standards for prompt generation and scenario quality assessment in cybersecurity exercise development. These achievements demonstrate the potential for AI to enhance both the technical accuracy and practical relevance of training scenarios across diverse industry sectors.

   - The proposed AiDPG methodology facilitated a 50.4% improvement in prompt maturity for exercise experts and 66.67% for sector experts, ensuring that generated scenarios align more closely with sector-specific threats and objectives. It

should be noted that the prompts generated by AiDPG can be used as input to any AI solution other than AiCEF, allowing future planners to leverage emerging systems that keep their prompts as current and rich as possible when it comes to the threat landscape.

- Jaccard similarity calculations further verified that AI-generated scenarios powered by AiDPG, were more aligned with sectorial threat landscapes than those generated by their human counterparts.
- Surveys indicated that exercise and sector experts perceived an 18.83% improvement in scenario relevance when leveraging AI-driven methods. This number can be significantly improved by incorporating the latest advancements in AI (ex., latest LLMs)

These comprehensive achievements establish AiCEF as a novel framework that has successfully addressed longstanding challenges in cybersecurity exercise development while opening new avenues for future research and practical applications in the field.

Despite these advancements, limitations remain. AI tools require extensive datasets for accurate and realistic scenario building or trend forecasting. The extraction of certain attributes, critical for developing realistic scenarios, such as motivation and MITRE TTPs, remains challenging. Furthermore, while AI models streamline content generation and intelligence extraction, human oversight remains crucial to address nuanced context and mitigate false positives, steering the process in the desired direction.

10.2 Limitations

While this research has made significant contributions to the field of automated cybersecurity exercise generation, several limitations have been identified that present opportunities for future work and continued improvement.

Dataset Requirements

The framework's forecasting capabilities are constrained by data availability and quality requirements. Accurate trend forecasting necessitates more than three years of historical data, which may not be readily available for all sectors or threat categories. Additionally,

the limited availability of comprehensive sectorial incident datasets poses challenges for developing robust, sector-specific threat models. The varying quality and consistency of data across different sources further complicates the extraction and analysis processes, potentially affecting the reliability of generated scenarios.

Attribute Extraction Challenges

Despite advances in natural language processing and machine learning, the extraction of certain critical attributes remains challenging. Complex cybersecurity concepts such as attacker motivation and specific MITRE TTPs require nuanced understanding that current automated systems struggle to capture completely. Context-dependent attributes, which are essential for creating realistic and relevant exercise scenarios, often require human expertise to interpret correctly. These limitations highlight the ongoing need for human oversight in the automated generation process.

Generalization Limitations

The framework's ability to generalize across diverse organizational contexts presents another area for improvement. While sectorial threat modeling provides valuable insights, it may not fully capture all organizational nuances, particularly for entities with unique operational characteristics or hybrid sector affiliations. Furthermore, the framework's performance varies across different threat landscapes, and cultural and regional threat variations are not comprehensively addressed in the current implementation. These limitations suggest that additional customization mechanisms may be necessary to ensure optimal performance across diverse deployment scenarios.

These limitations, while significant, represent clear directions for future research and development. Addressing these challenges through enhanced data collection methodologies, improved AI techniques, and expanded evaluation frameworks will further strengthen the utility and applicability of automated cybersecurity exercise generation systems.

10.3 Future Work

This research marks a significant advancement in automating CSE generation, but also reveals areas for improvement, which, if pursued, could greatly enhance the utility and impact of the proposed framework.

Leveraging Advanced LLMs

One of the most immediate steps forward is the integration of the latest LLMs into the framework to enhance the generated content. These models, with their advanced contextual understanding and the ability to generate nuanced text, could significantly improve the quality and relevance of the content produced by AiCEF. Evaluating the efficacy of these integrations through rigorous testing and user feedback will be crucial to ensuring measurable improvements. This step not only provides a path for short-term advancements but also sets the stage for iterative refinement of the proposed framework. Sufficient effort should be given in exploring the use of local LLMs in an attempt to identify advantages in terms of privacy of the analysed information and disadvantages in terms of quality and use of resources.

Automating Inject Content Generation for Common Attack Categories

A valuable extension of this work is the automation of inject creation for widely recognized attack categories, such as phishing, ransomware, or insider threats. By utilizing AI to generate inject content (including technical artefacts) tailored to these categories, organizations can streamline the process of creating realistic and impactful cybersecurity exercises. Automating the inject creation process would provide a practical and scalable approach to addressing common threat vectors, enhancing both training effectiveness and operational readiness.

Organization-Level Threat Landscape Automation

Another promising direction involves automating the generation of organization-specific threat landscapes. By processing private datasets—including incident reports, threat alerts, and intelligence feeds—AiCEF could create highly tailored scenarios that reflect the unique challenges faced by individual organizations. Such customization would improve the adaptability of the framework, enabling it to focus on evolving threats that are most relevant to each organization's operational context. This advancement would also support a more proactive approach to cybersecurity by aligning training exercises with real-world threat dynamics. Serious effort should be put in anonymization of data and potentially exploration of hybrid methods while utilizing online and local AI models.

Enhancing Contextual Understanding with AI

Integrating advanced AI techniques into AiCEF represents another critical opportunity for improvement. These techniques could be used to extract and analyse context-specific

information related to attackers' motivations, risk levels, or phases of the Cyber Kill Chain, and link them directly to relevant injects. The advancement of LLMs can surely support such research, but the use of custom topic-specific AI models should also be explored. By incorporating such granular contextual insights, the framework's functionality and relevance would be significantly enhanced, leading to more precise and impactful scenario generation.

Training AI Models on Domain-Specific Data

Improving the training of AI models used within AiCEF is a high priority. Transitioning from general-purpose generative models to those trained specifically in cybersecurity-related corpora would produce outputs that are more accurate, technically precise, and contextually relevant. Such an approach could better align the generated content with the needs of cybersecurity professionals. For instance, this current work on crowdsourcing multiple models under a voting scheme to address the challenges of automated TTP extraction is a promising step in this direction and warrants further research.

Evaluating Usability Across Diverse User Profiles

AiCEF's usability and impact on users with varying levels of expertise also require further evaluation. Testing the framework with both novice and expert users would provide valuable insights into its accessibility, the quality of its outputs, and its effectiveness across different user profiles. This feedback could guide iterative refinements, ensuring that AiCEF meets the needs of a broad audience while maintaining its utility for advanced cybersecurity professionals.

Enhancing Forecasting Methodologies

Finally, the robustness of AiCEF's forecasting capabilities can be improved by validating its predictions against expert assessments and real-world trends over extended periods. By collecting and analysing additional incident data, the framework's predictive models can be fine-tuned to provide more accurate and reliable insights. Such efforts would bolster the credibility of AiCEF as a tool for forecasting emerging threats and guiding strategic cybersecurity planning.

By addressing these challenges and seizing opportunities for improvement, AiCEF has the potential to revolutionize the development of cybersecurity exercises. Its enhanced capabilities would make cybersecurity training more accessible, timely, and relevant, empowering organizations to better prepare for an ever-evolving threat landscape. Ultimately, these ad-

vancements would contribute to a more secure digital ecosystem, reinforcing the importance of innovation in the cybersecurity domain.

Automating cybersecurity exercise generation

While the proposed approach is ideal for tabletop scenarios, a hands-on approach could be more beneficial in terms of understanding the practical impacts and addressing the technical challenges and knowledge gaps. Since AiCEF produces machine-readable represantations of cyber security exercises, the automation of generating the actual cyber exercises, e.g., creating the proper virtual machines, installing the proper software and frameworks, could be to some extent automated. An obvious solution to this direction seems to be again through LLMs; however, there are several challenges in this direction to verify that the generated exercise would meet the requirements and expectations in terms of adherence to the scenario, but also functionality of the delivered environment.

10.4 Closing Remarks

This research underscores the transformative potential of AI in CSE scenario generation, offering a framework that enhances efficiency, relevance, and quality while bridging expertise gaps. The proposed methodologies demonstrate that AI tools are not merely supplementary but integral to the future of cybersecurity exercise planning. Future research should focus on refining AI models for specialized tasks, expanding datasets for improved forecasting, and integrating hybrid approaches to achieve even greater accuracy and adaptability. Through continued development, AI can empower organizations to prepare more effectively for the evolving threat landscape, ultimately bolstering cybersecurity resilience.

ANNEX I - Glossary of Terms

After Action Report (AAR) A detailed report generated after a cybersecurity exercise to evaluate the performance, outcomes, and lessons learned.

AiCEF AI-assisted Cyber Exercise Framework, a framework used for generating cyber exercise scenarios.

APT Advanced Persistent Threat, a cyber attack in which an attacker gains unauthorized access to a network and remains undetected for an extended period.

Blue Team A group responsible for defending systems and responding to attacks during a cybersecurity exercise, typically composed of internal staff.

CSE Crisis Management Cyber Exercise, a simulated event to test cybersecurity readiness and response.

Event A discrete occurrence within a cyber exercise scenario, such as a breach, attack, or failure, intended to test a specific response capability.

Exercise Planner An individual or team responsible for designing, organizing, and managing the logistics and content of a cybersecurity exercise.

Injects Pre-planned events, actions, or information that are introduced during a cyber exercise to simulate real-world conditions and test participants' responses.

Incident An occurrence of an event that has an actual or potential impact on the security of a system or organization, often triggering a response from incident response teams.

Hotwash A debriefing session immediately after the exercise to gather feedback, discuss lessons learned, and identify areas for improvement.

MSEL (Master Scenario Event List) A document used to guide the flow of a simulation, detailing the timing, objectives, injects, and events that occur during an exercise.

Red Team A group of ethical hackers or cybersecurity experts tasked with simulating adversarial attacks on systems to test defenses and response protocols.

Scenario A simulated sequence of events in a cyber exercise designed to test specific capabilities, such as incident response or system recovery.

Simulation A controlled, artificial environment used to mimic real-world cyber attacks and responses for the purpose of training or testing.

Threat Intelligence Information that organizations gather and analyze to understand potential cyber threats, often used in exercise scenarios to create realistic attack models.

Threat Landscape The overall view of an organization's exposure to cyber threats, including the types of attacks, potential attackers, vulnerabilities, and external factors that may impact cybersecurity.

White Team A group responsible for observing and facilitating the exercise, ensuring that objectives are met, and that the exercise runs smoothly.

ANNEX II - Useful Code

.1 CESO Ontology Implementation

The Cyber Exercise Scenario Ontology (CESO) has been implemented in Web Ontology Language (OWL) format to ensure machine-readable representation and semantic interoperability with existing cybersecurity frameworks. The complete ontology specification is provided below, defining all classes, properties, relationships, and training topic individuals required for automated exercise scenario generation within the AiCEF framework.

.1.1 Ontology Metadata

The CESO ontology is defined with the namespace `http://aicef.alexandros.zacharis.eu/ceso#` and includes comprehensive documentation linking each component to its corresponding STIX 2.1 object and ISO 22398:2013 exercise structure. Version 1.0 implements all relationships from Table 4.2 and object mappings from Table 4.1.

.1.2 Implementation Details

The OWL implementation includes:

- **Object Properties**: Implementing all 12 relationship types from the relationships matrix

- **Data Properties**: Supporting AiCEF-specific attributes including maturity scoring, difficulty levels, and training topic classifications
- **Class Hierarchy**: Mapping CSE components to STIX 2.1 objects with proper inheritance structures
- **Named Individuals**: Pre-defined training topic instances for the six categories identified

Listing 1: CESO Ontology OWL Implementation

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://aicef.alexandros.zacharis.eu/ceso#"
     xml:base="http://aicef.alexandros.zacharis.eu/ceso"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:ceso="http://aicef.alexandros.zacharis.eu/ceso#"
     xmlns:stix="http://stix.mitre.org/stixCommon-1#">
    <owl:Ontology rdf:about="http://aicef.alexandros.zacharis.eu/ceso">
        <rdfs:label>Cyber Exercise Scenario Ontology (CESO)</rdfs:label>
        <rdfs:comment>An ontology for representing cybersecurity exercise scenarios based on STIX 2.1 and
              ISO 22398:2013 standards. Developed as part of the AiCEF framework.</rdfs:comment>
        <owl:versionInfo>1.0</owl:versionInfo>
    </owl:Ontology>

    <!--
    ///////////////////////////////////////////////////////////////////
    //
    // Object Properties
    //
    ///////////////////////////////////////////////////////////////////
     -->

    <!-- http://aicef.alexandros.zacharis.eu/ceso#attributedTo -->
    <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#attributedTo">
        <rdfs:label>attributed to</rdfs:label>
        <rdfs:comment>Links an attack pattern to the threat actor responsible for it, or threat actor to
             identity</rdfs:comment>
    </owl:ObjectProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#delivers -->
    <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#delivers">
        <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#AttackPattern"/>
        <rdfs:range rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Malware"/>
        <rdfs:label>delivers</rdfs:label>
        <rdfs:comment>Indicates that an attack pattern delivers malware</rdfs:comment>
    </owl:ObjectProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#exploits -->
    <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#exploits">
        <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#AttackPattern"/>
        <rdfs:range rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Vulnerability"/>
        <rdfs:label>exploits</rdfs:label>
        <rdfs:comment>Links an attack pattern to the vulnerability it exploits</rdfs:comment>
    </owl:ObjectProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#hasEvent -->
    <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#hasEvent">
        <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#ExerciseScenario"/>
        <rdfs:range rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Event"/>
        <rdfs:label>has event</rdfs:label>
        <rdfs:comment>Links an exercise scenario to its constituent events</rdfs:comment>
    </owl:ObjectProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#hasIncident -->
    <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#hasIncident">
        <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Event"/>
        <rdfs:range rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Incident"/>
        <rdfs:label>has incident</rdfs:label>
        <rdfs:comment>Links an event to its constituent incidents</rdfs:comment>
    </owl:ObjectProperty>

```

```
62      <!-- http://aicef.alexandros.zacharis.eu/ceso#hasInject -->
63      <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#hasInject">
64          <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Incident"/>
65          <rdfs:range rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Inject"/>
66          <rdfs:label>has inject</rdfs:label>
67          <rdfs:comment>Links an incident to its constituent injects</rdfs:comment>
68      </owl:ObjectProperty>
69
70      <!-- http://aicef.alexandros.zacharis.eu/ceso#hasObjective -->
71      <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#hasObjective">
72          <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#ExerciseScenario"/>
73          <rdfs:range rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Objective"/>
74          <rdfs:label>has objective</rdfs:label>
75          <rdfs:comment>Links an exercise scenario to its training objectives</rdfs:comment>
76      </owl:ObjectProperty>
77
78      <!-- http://aicef.alexandros.zacharis.eu/ceso#indicates -->
79      <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#indicates">
80          <rdfs:label>indicates</rdfs:label>
81          <rdfs:comment>Links attack patterns to indicators, or indicators to malware</rdfs:comment>
82      </owl:ObjectProperty>
83
84      <!-- http://aicef.alexandros.zacharis.eu/ceso#locatedAt -->
85      <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#locatedAt">
86          <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Identity"/>
87          <rdfs:range rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Location"/>
88          <rdfs:label>located at</rdfs:label>
89          <rdfs:comment>Indicates the physical or logical location of an identity</rdfs:comment>
90      </owl:ObjectProperty>
91
92      <!-- http://aicef.alexandros.zacharis.eu/ceso#mitigates -->
93      <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#mitigates">
94          <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#CourseOfAction"/>
95          <rdfs:label>mitigates</rdfs:label>
96          <rdfs:comment>Indicates what attack patterns or vulnerabilities a course of action addresses</
                rdfs:comment>
97      </owl:ObjectProperty>
98
99      <!-- http://aicef.alexandros.zacharis.eu/ceso#relatedTo -->
100     <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#relatedTo">
101         <rdfs:label>related to</rdfs:label>
102         <rdfs:comment>General relationship connecting campaigns, notes, reports, intrusion sets, and
                courses of action to grouping objects</rdfs:comment>
103     </owl:ObjectProperty>
104
105     <!-- http://aicef.alexandros.zacharis.eu/ceso#targets -->
106     <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#targets">
107         <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Incident"/>
108         <rdfs:label>targets</rdfs:label>
109         <rdfs:comment>Indicates what tools or vulnerabilities an intrusion set targets</rdfs:comment>
110     </owl:ObjectProperty>
111
112     <!-- http://aicef.alexandros.zacharis.eu/ceso#uses -->
113     <owl:ObjectProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#uses">
114         <rdfs:label>uses</rdfs:label>
115         <rdfs:comment>Indicates what attack patterns intrusion sets use, or what infrastructure
                identities use</rdfs:comment>
116     </owl:ObjectProperty>
117
118     <!--
119     //////////////////////////////////////////////////////////////////////
120     //
121     // Data Properties
122     //
123     //////////////////////////////////////////////////////////////////////
124      -->
125
126     <!-- http://aicef.alexandros.zacharis.eu/ceso#complexity -->
127     <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#complexity">
128         <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#integer"/>
129         <rdfs:label>complexity</rdfs:label>
130         <rdfs:comment>Complexity level of an exercise component (1-5 scale)</rdfs:comment>
131     </owl:DatatypeProperty>
132
133     <!-- http://aicef.alexandros.zacharis.eu/ceso#cveId -->
134     <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#cveId">
135         <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Vulnerability"/>
136         <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
137         <rdfs:label>CVE ID</rdfs:label>
138         <rdfs:comment>Common Vulnerabilities and Exposures identifier</rdfs:comment>
139     </owl:DatatypeProperty>
140
141     <!-- http://aicef.alexandros.zacharis.eu/ceso#description -->
```

```
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#description">
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
        <rdfs:label>description</rdfs:label>
        <rdfs:comment>Textual description of an entity</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#difficulty -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#difficulty">
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#integer"/>
        <rdfs:label>difficulty</rdfs:label>
        <rdfs:comment>Difficulty level for executing a course of action (1-5 scale)</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#duration -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#duration">
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#duration"/>
        <rdfs:label>duration</rdfs:label>
        <rdfs:comment>Time duration for an exercise component</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#maturityScore -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#maturityScore">
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#integer"/>
        <rdfs:label>maturity score</rdfs:label>
        <rdfs:comment>Maturity score calculated by AiCEF (0-185 scale)</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#name -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#name">
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
        <rdfs:label>name</rdfs:label>
        <rdfs:comment>Name identifier for an entity</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#recipientGroup -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#recipientGroup">
        <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Identity"/>
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
        <rdfs:label>recipient group</rdfs:label>
        <rdfs:comment>Group designation for exercise participants</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#scenario -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#scenario">
        <rdfs:domain rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#ExerciseScenario"/>
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
        <rdfs:label>scenario</rdfs:label>
        <rdfs:comment>Detailed scenario description providing context</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#sector -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#sector">
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
        <rdfs:label>sector</rdfs:label>
        <rdfs:comment>Industry sector classification (NIS2 compliant)</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#timing -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#timing">
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#dateTime"/>
        <rdfs:label>timing</rdfs:label>
        <rdfs:comment>Scheduled time for inject delivery or event occurrence</rdfs:comment>
    </owl:DatatypeProperty>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#trainingTopic -->
    <owl:DatatypeProperty rdf:about="http://aicef.alexandros.zacharis.eu/ceso#trainingTopic">
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
        <rdfs:label>training topic</rdfs:label>
        <rdfs:comment>Training focus area (e.g., INCIDENT_HANDLING, GDPR, PHISHING)</rdfs:comment>
    </owl:DatatypeProperty>

    <!--
    ///////////////////////////////////////////////////////////////////////
    //
    // Classes
    //
    ///////////////////////////////////////////////////////////////////////
     -->

    <!-- http://aicef.alexandros.zacharis.eu/ceso#AttackPattern -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#AttackPattern">
        <rdfs:label>Attack Pattern</rdfs:label>
        <rdfs:comment>A type of Tactics, Techniques, and Procedures (TTP) that describes ways that
```

```
            adversaries attempt to compromise targets</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#CourseOfAction -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#CourseOfAction">
        <rdfs:label>Course of Action</rdfs:label>
        <rdfs:comment>A response, countermeasure, or defensive action taken to mitigate or remediate
            threats</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Event -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Event">
        <rdfs:subClassOf rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Campaign"/>
        <rdfs:label>Event</rdfs:label>
        <rdfs:comment>High-level description of an exercise scenario component that groups related
            incidents. Maps to STIX 2.1 Campaign object.</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Campaign -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Campaign">
        <rdfs:label>Campaign</rdfs:label>
        <rdfs:comment>STIX 2.1 Campaign object used for Events</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#ExerciseComponent -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#ExerciseComponent">
        <rdfs:label>Exercise Component</rdfs:label>
        <rdfs:comment>Abstract base class for all exercise-related components</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#ExercisePlatform -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#ExercisePlatform">
        <rdfs:subClassOf rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Infrastructure"/>
        <rdfs:label>Exercise Platform</rdfs:label>
        <rdfs:comment>Technical infrastructure used to conduct cybersecurity exercises. Maps to STIX 2.1
            Infrastructure object.</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#ExerciseParticipant -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#ExerciseParticipant">
        <rdfs:subClassOf rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Identity"/>
        <rdfs:label>Exercise Participant</rdfs:label>
        <rdfs:comment>Individuals participating in cybersecurity exercises. Maps to STIX 2.1 Identity and
             Location objects.</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#ExerciseScenario -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#ExerciseScenario">
        <rdfs:subClassOf rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Grouping"/>
        <rdfs:label>Exercise Scenario</rdfs:label>
        <rdfs:comment>Complete cybersecurity exercise scenario containing events, incidents, and injects.
             Maps to STIX 2.1 Grouping object for Exercise Details and Scenario Background.</
            rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Grouping -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Grouping">
        <rdfs:label>Grouping</rdfs:label>
        <rdfs:comment>STIX 2.1 Grouping object used for Exercise Details and Scenario Background</
            rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Identity -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Identity">
        <rdfs:label>Identity</rdfs:label>
        <rdfs:comment>Individuals, organizations, or groups involved in cybersecurity exercises</
            rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Incident -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Incident">
        <rdfs:subClassOf rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#IntrusionSet"/>
        <rdfs:label>Incident</rdfs:label>
        <rdfs:comment>Specific cybersecurity incident or attack scenario within an exercise event. Maps
            to STIX 2.1 Intrusion Set object.</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#IntrusionSet -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#IntrusionSet">
        <rdfs:label>Intrusion Set</rdfs:label>
        <rdfs:comment>STIX 2.1 Intrusion Set object used for Incidents</rdfs:comment>
    </owl:Class>

```

```
    <!-- http://aicef.alexandros.zacharis.eu/ceso#Indicator -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Indicator">
        <rdfs:label>Indicator</rdfs:label>
        <rdfs:comment>Observable phenomena that suggest the presence of a cybersecurity threat</
            rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Infrastructure -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Infrastructure">
        <rdfs:label>Infrastructure</rdfs:label>
        <rdfs:comment>Systems, networks, or platforms supporting cybersecurity exercises</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Inject -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Inject">
        <rdfs:subClassOf rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#ExerciseComponent"/>
        <rdfs:label>Inject</rdfs:label>
        <rdfs:comment>Timed information input delivered to exercise participants to drive responses. Can
            be composed of multiple STIX 2.1 objects: Tool, Vulnerability, Threat-Actor, Identity,
            Location, Attack Pattern, Malware, Indicator, Course-of-action, Observed Data, Malware
            Analysis, Report.</rdfs:comment>
    </owl:Class>

    <!-- Additional STIX 2.1 objects that can be part of Injects -->

    <!-- http://aicef.alexandros.zacharis.eu/ceso#ObservedData -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#ObservedData">
        <rdfs:label>Observed Data</rdfs:label>
        <rdfs:comment>STIX 2.1 Observed Data object - can be used as Inject component</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#MalwareAnalysis -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#MalwareAnalysis">
        <rdfs:label>Malware Analysis</rdfs:label>
        <rdfs:comment>STIX 2.1 Malware Analysis object - can be used as Inject component</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Location -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Location">
        <rdfs:label>Location</rdfs:label>
        <rdfs:comment>Geographic or logical location relevant to cybersecurity exercises</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Malware -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Malware">
        <rdfs:label>Malware</rdfs:label>
        <rdfs:comment>Malicious software used in cyber attacks</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Objective -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Objective">
        <rdfs:subClassOf rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Note"/>
        <rdfs:label>Objective</rdfs:label>
        <rdfs:comment>Training or assessment goal of a cybersecurity exercise. Maps to STIX 2.1 Note
            object.</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Note -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Note">
        <rdfs:label>Note</rdfs:label>
        <rdfs:comment>STIX 2.1 Note object used for Objectives</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#StateOfTheWorld -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#StateOfTheWorld">
        <rdfs:subClassOf rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Report"/>
        <rdfs:label>State of the World</rdfs:label>
        <rdfs:comment>Background context and initial conditions for an exercise scenario. Maps to STIX
            2.1 Report object.</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Report -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Report">
        <rdfs:label>Report</rdfs:label>
        <rdfs:comment>STIX 2.1 Report object used for State of the World (SoW) and as potential Inject
            component</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#ThreatActor -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#ThreatActor">
        <rdfs:label>Threat Actor</rdfs:label>
        <rdfs:comment>Individual, group, or organization conducting malicious cyber activities</
            rdfs:comment>
```

```
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Tool -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Tool">
        <rdfs:label>Tool</rdfs:label>
        <rdfs:comment>Software, hardware, or technique used in cyber attacks or defenses</rdfs:comment>
    </owl:Class>

    <!-- http://aicef.alexandros.zacharis.eu/ceso#Vulnerability -->
    <owl:Class rdf:about="http://aicef.alexandros.zacharis.eu/ceso#Vulnerability">
        <rdfs:label>Vulnerability</rdfs:label>
        <rdfs:comment>Weakness in systems, networks, or processes that can be exploited</rdfs:comment>
    </owl:Class>

    <!--
    ///////////////////////////////////////////////////////////////////
    //
    // Individuals
    //
    ///////////////////////////////////////////////////////////////////
     -->

    <!-- Training Topic Individuals -->
    <owl:NamedIndividual rdf:about="http://aicef.alexandros.zacharis.eu/ceso#IncidentHandling">
        <rdf:type rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Objective"/>
        <ceso:name>Incident Handling</ceso:name>
        <ceso:trainingTopic>INCIDENT_HANDLING</ceso:trainingTopic>
        <ceso:description>Training focused on malware, ransomware, APT, cyber attacks, website security,
            hacker activities, exploits, and zero-day vulnerabilities</ceso:description>
    </owl:NamedIndividual>

    <owl:NamedIndividual rdf:about="http://aicef.alexandros.zacharis.eu/ceso#GDPRCompliance">
        <rdf:type rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Objective"/>
        <ceso:name>GDPR Compliance</ceso:name>
        <ceso:trainingTopic>GDPR</ceso:trainingTopic>
        <ceso:description>Training on privacy, data leakage, personal data protection, exfiltration,
            cloud security, and sensitive data handling</ceso:description>
    </owl:NamedIndividual>

    <owl:NamedIndividual rdf:about="http://aicef.alexandros.zacharis.eu/ceso#CyberHygiene">
        <rdf:type rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Objective"/>
        <ceso:name>Cyber Hygiene</ceso:name>
        <ceso:trainingTopic>CYBER_HYGIENE</ceso:trainingTopic>
        <ceso:description>Training on password security, account management, usernames, login procedures,
            and credential protection</ceso:description>
    </owl:NamedIndividual>

    <owl:NamedIndividual rdf:about="http://aicef.alexandros.zacharis.eu/ceso#PhishingAndSocialEngineering
        ">
        <rdf:type rdf:resource="http://aicef.alexandros.zacharis.eu/ceso#Objective"/>
        <ceso:name>Phishing and Social Engineering</ceso:name>
        <ceso:trainingTopic>PHISHING_SOCIAL_ENGINEERING</ceso:trainingTopic>
        <ceso:description>Training on phishing, scams, fraud, vishing, impersonation, business email
            compromise, and email security</ceso:description>
    </owl:NamedIndividual>

</rdf:RDF>
```

.1.3 Validation and Usage

The ontology has been validated for logical consistency using Protégé's built-in reasoners (HermiT and Pellet). The implementation ensures compatibility with STIX 2.1 specifications while extending the model to support cybersecurity exercise-specific requirements. This OWL file serves as the foundational schema for the AiCEF framework's knowledge representation and automated scenario generation capabilities.

.2 AiCEF Generated Exercise

This section contains an example output of a cyer security exercise generated using the AiCEF tool. The exercise is stored as a STIX2.1 objects and can be easily previewed by online tools like: https://oasis-open.github.io/cti-stix-visualization/

Listing 2: STIX2.1 output of AiCEF- Modeling a CSE


```
{
    "type": "bundle",
    "id": "bundle--7e9f5a03-24dd-47d7-8c0e-7ae348cc1f07",
    "objects": [
        {
            "type": "grouping",
            "spec_version": "2.1",
            "id": "grouping--84e4d88f-44ea-4bcd-bbf3-b2c1c320bcb3",
            "created": "2023-04-20T11:24:46.729683Z",
            "modified": "2023-04-20T11:25:20.022277Z",
            "name": "BLACKHAT_ASIA_EX",
            "description": "A group of hackers used the TURLA virus to exploit weaknesses in an
                 organization's user data and credentials, allowing them to gain access to the system and
                  install ransomware. This ransomware enabled the hackers to access and leak the data and
                  personal information of the victims, posing a major threat to the organization. In
                 order to protect their data and systems, the organization must remain vigilant and
                 constantly monitor their systems for any malicious activity.",
            "context": "['suspicious-activity']",
            "object_refs": [
                "grouping--84e4d88f-44ea-4bcd-bbf3-b2c1c320bcb3"
            ]
        },
        {
            "type": "note",
            "spec_version": "2.1",
            "id": "note--9fa20812-d8ac-4af2-8d3e-6b5a21d512a3",
            "created": "2023-04-20T11:24:46.732678Z",
            "modified": "2023-04-20T11:24:46.732678Z",
            "abstract": "Exercise Objectives",
            "content": "Test cyber preparedness in case of targeted cyber attacks.",
            "object_refs": [
                "grouping--84e4d88f-44ea-4bcd-bbf3-b2c1c320bcb3"
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--36139584-2d74-40e1-995c-2a5f5a2cce6d",
            "created": "2023-04-20T11:24:46.733683Z",
            "modified": "2023-04-20T11:24:46.733683Z",
            "relationship_type": "related-to",
            "source_ref": "note--9fa20812-d8ac-4af2-8d3e-6b5a21d512a3",
            "target_ref": "grouping--84e4d88f-44ea-4bcd-bbf3-b2c1c320bcb3"
        },
        {
            "type": "campaign",
            "spec_version": "2.1",
            "id": "campaign--95ab80b3-b15f-4643-a9ea-a3ad11aeff79",
            "created": "2023-04-20T11:20:51.295509Z",
            "modified": "2023-04-20T11:21:08.41177Z",
            "name": "BLACKHAT_ASIA_EV1",
            "description": "In this case, the malicious TURLA virus was used to exploit weaknesses in
                 group user data and credentials, resulting in the installation of ransomware on the
                 system."
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--4e4d903a-28f3-4624-b1ba-eb5eae2fad5e",
            "created": "2023-04-19T15:18:23.272243Z",
            "modified": "2023-04-19T15:18:23.272243Z",
            "name": "WINDOWS - Forensics Main",
            "description": "Conduct all the appropriate actions in order to collect evidence for the
                 following system or infrastructure: WINDOWS"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
```

```
            "id": "relationship--d3e6ef38-1efe-49ad-afb0-bfcdb18eced2",
            "created": "2023-04-19T15:18:23.272243Z",
            "modified": "2023-04-19T15:18:23.272243Z",
            "relationship_type": "mitigates",
            "source_ref": "course-of-action--4e4d903a-28f3-4624-b1ba-eb5eae2fad5e",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
        {
            "type": "threat-actor",
            "spec_version": "2.1",
            "id": "threat-actor--8c3a5475-76eb-470a-8953-b364fef84676",
            "created": "2023-04-19T14:48:29.297822Z",
            "modified": "2023-04-19T15:16:01.347606Z",
            "name": "TURLA",
            "threat_actor_types": [
                "crime-syndicate"
            ],
            "goals": [
                "DATA, PASSWORD, CRYPTOCURRENCY, USER CREDENTIALS, FILES, WALLET"
            ]
        },
        {
            "type": "threat-actor",
            "spec_version": "2.1",
            "id": "threat-actor--8c3a5475-76eb-470a-8953-b364fef84676",
            "created": "2023-04-19T14:48:29.297822Z",
            "modified": "2023-04-19T15:16:01.347606Z",
            "name": "TURLA",
            "threat_actor_types": [
                "crime-syndicate"
            ],
            "goals": [
                "DATA, PASSWORD, CRYPTOCURRENCY, USER CREDENTIALS, FILES, WALLET"
            ]
        },
        {
            "type": "identity",
            "spec_version": "2.1",
            "id": "identity--a08a65d0-442b-4934-b306-2ed0db6fe5b0",
            "created": "2023-04-19T14:48:29.297822Z",
            "modified": "2023-04-19T15:15:10.377742Z",
            "name": "Attacker"
        },
        {
            "type": "intrusion-set",
            "spec_version": "2.1",
            "id": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703",
            "created": "2023-04-19T14:48:29.297822Z",
            "modified": "2023-04-19T15:16:01.347607Z",
            "name": "BLACKHAT_ASIA_INC1",
            "description": "A recent cybersecurity incident involved a malicious actor using the TURLA
                 virus to target Windows Servers. This virus exploited weaknesses in the group user data,
                  as well as credentials stored on the system. By exploiting these weaknesses, the hacker
                  was able to gain access to the system and install ransomware.\n\nVarious techniques
                 were then used in order to gain control of the Windows Servers. These included
                 credentials stealing, use of PowerShell, brute force, as well as scheduled tasks.",
            "goals": [
                "DATA, PASSWORD, CRYPTOCURRENCY, USER CREDENTIALS, FILES, WALLET"
            ]
        },
        {
            "type": "note",
            "spec_version": "2.1",
            "id": "note--b1d37ba5-83bc-4b3f-8d84-4bfbb43e8fde",
            "created": "2023-04-19T15:18:23.273511Z",
            "modified": "2023-04-19T15:18:23.273511Z",
            "abstract": "Training_Objectives",
            "content": "CYBER_HYGIENE, INCIDENT_HANDLING",
            "object_refs": [
                "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--17fa3b5e-7058-496e-9ab0-27561e76d5a8",
            "created": "2023-04-19T15:18:23.274255Z",
            "modified": "2023-04-19T15:18:23.274255Z",
            "relationship_type": "related-to",
            "source_ref": "note--b1d37ba5-83bc-4b3f-8d84-4bfbb43e8fde",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
        {
```

```
            "type": "note",
            "spec_version": "2.1",
            "id": "note--a8e93413-43d2-4cbb-b173-af65f771dbd6",
            "created": "2023-04-19T15:18:23.27525Z",
            "modified": "2023-04-19T15:18:23.27525Z",
            "abstract": "Resources",
            "content": "['https://www.bleepingcomputer.com/news/security/qnap-warns-of-ech0raix-
                ransomware-attacks-roon-server-zero-day/', 'https://www.bleepingcomputer.com/news/
                security/bluenoroff-hackers-steal-crypto-using-fake-metamask-extension/', 'https://www.
                bleepingcomputer.com/news/security/jenkins-projects-confluence-server-hacked-to-mine-
                monero/', 'https://www.bleepingcomputer.com/news/security/synology-warns-of-malware-
                infecting-nas-devices-with-ransomware/', 'https://securityaffairs.co/wordpress/118306/
                digital-id/kubernetes-clusters-teamtnt.html']",
            "object_refs": [
                "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--41633607-7147-4582-b513-584151d9731c",
            "created": "2023-04-19T15:18:23.27525Z",
            "modified": "2023-04-19T15:18:23.27525Z",
            "relationship_type": "related-to",
            "source_ref": "note--a8e93413-43d2-4cbb-b173-af65f771dbd6",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
        {
            "type": "report",
            "id": "report--7a7402d7-a6df-4bb2-8c27-26479c8e81e1",
            "created": "2023-04-19T15:18:25.362Z",
            "modified": "2023-04-19T15:18:25.362Z",
            "name": "News Item 0",
            "description": "\n\nEnergyCorp was the latest victim of a Turla group ransomware attack.
                Sensitive user data and Windows servers were targeted, and credentials were stolen
                through CVE-2022-1234. Zoho was used to demand ransom to recover the data. #CyberAttack
                #Ransomware #Zoho",
            "published": "2023-04-19T18:18:25.362663Z",
            "object_refs": [
                "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--1705a3e2-e57b-479a-b165-3b6cd30335de",
            "created": "2023-04-19T15:18:25.363664Z",
            "modified": "2023-04-19T15:18:25.363664Z",
            "relationship_type": "related-to",
            "source_ref": "report--7a7402d7-a6df-4bb2-8c27-26479c8e81e1",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
        {
            "type": "report",
            "id": "report--d37b2123-b77f-4a25-b64f-574406039c83",
            "created": "2023-04-19T15:18:27.698Z",
            "modified": "2023-04-19T15:18:27.698Z",
            "name": "News Item 1",
            "description": "\n\nEnergyCorp was recently hit with a #cyberattack from a #Turla group that
                compromised #WindowsServers and stole #userdata. CVE-2022-1234 was exploited to gain #
                credentials for #ZOHO and a #ransomware was deployed. An organization's worst nightmare!
                ",
            "published": "2023-04-19T18:18:27.698699Z",
            "object_refs": [
                "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--b16be571-3e50-4c33-b244-a6f10a6f9957",
            "created": "2023-04-19T15:18:27.698699Z",
            "modified": "2023-04-19T15:18:27.698699Z",
            "relationship_type": "related-to",
            "source_ref": "report--d37b2123-b77f-4a25-b64f-574406039c83",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
```

```
        {
            "type": "report",
            "id": "report--d550a94b-5dd8-4e1f-bbe4-6a1234c9b6cc",
            "created": "2023-04-19T15:18:30.402Z",
            "modified": "2023-04-19T15:18:30.402Z",
            "name": "News Item 2",
            "description": "\n\n\"Cyber attack group, TURLA, exploits CVE-2022-1234 to steal EnergyCorp
                 user data. Using stolen ZOHO credentials, they breach Windows servers and deploy
                 ransomware to encrypt user data. EnergyCorp is now at the mercy of TURLA's demands for a
                  ransom.",
            "published": "2023-04-19T18:18:30.402236Z",
            "object_refs": [
                "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--83ad9051-2ac4-41c3-b387-8c4d44521892",
            "created": "2023-04-19T15:18:30.403234Z",
            "modified": "2023-04-19T15:18:30.403234Z",
            "relationship_type": "related-to",
            "source_ref": "report--d550a94b-5dd8-4e1f-bbe4-6a1234c9b6cc",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
        {
            "type": "report",
            "id": "report--ce8b4fb1-d335-4188-8d07-58c0536bba07",
            "created": "2023-04-19T15:18:34.582Z",
            "modified": "2023-04-19T15:18:34.582Z",
            "name": "Email 1",
            "description": " \n\nEmail 1:\n\nFrom: Helpdesk\nTo: CEO\nSubject: Recent Cyber Attack on
                 EnergyCorp\n\nDear CEO,\n\nI wanted to make you aware of a recent cyber attack on
                 EnergyCorp involving the malicious ransomware, TURLA. The attackers seemed to have
                 targeted our Windows Servers and gained access to user data, group information and
                 credentials. We believe the attack was made through a CVE-2022-1234 vulnerability and
                 suspect ZOHO may",
            "published": "2023-04-19T18:18:34.582069Z",
            "object_refs": [
                "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--3b57e985-94c6-4414-a26b-e5e924d75a98",
            "created": "2023-04-19T15:18:34.583084Z",
            "modified": "2023-04-19T15:18:34.583084Z",
            "relationship_type": "related-to",
            "source_ref": "report--ce8b4fb1-d335-4188-8d07-58c0536bba07",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
        {
            "type": "identity",
            "spec_version": "2.1",
            "id": "identity--32bcbd3d-f9e8-4679-9045-3e47f5e37de7",
            "created": "2023-04-19T14:48:29.297822Z",
            "modified": "2023-04-19T15:15:10.378738Z",
            "name": "EnergyCorp",
            "roles": [
                "FINANCIAL, COMPANIES, POINTS"
            ],
            "external_references": [
                {
                    "source_name": "News Article Source",
                    "url": "https://www.bleepingcomputer.com/news/security/bluenoroff-hackers-steal-
                         crypto-using-fake-metamask-extension/"
                }
            ]
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--5f08e28f-169c-42bc-a051-d9776a6dec08",
            "created": "2023-04-19T15:18:34.584087Z",
            "modified": "2023-04-19T15:18:34.584087Z",
            "name": "CVE-2017-0199 Vulnerability Used",
```

```
            "description": "Research and collect evidence for the following Vulnerability: CVE-2017-0199.
                More info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE-2017-0199"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--6f35d24b-c3e2-479f-aa12-35f62f9ea149",
            "created": "2023-04-19T15:18:34.585115Z",
            "modified": "2023-04-19T15:18:34.585115Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703",
            "target_ref": "course-of-action--5f08e28f-169c-42bc-a051-d9776a6dec08"
        },
        {
            "type": "location",
            "spec_version": "2.1",
            "id": "location--d6a16cf4-c101-40fd-94a9-0826bd5527c1",
            "created": "2023-04-19T14:48:29.30282Z",
            "modified": "2023-04-19T14:48:29.30282Z",
            "name": "CHINA",
            "latitude": 48.8566,
            "longitude": 2.3522
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--e97fb560-5fb0-4f2e-8733-4ba73d4f29c4",
            "created": "2023-04-19T15:18:34.586089Z",
            "modified": "2023-04-19T15:18:34.586089Z",
            "name": "XMRIG Mitigation",
            "description": "Conduct all the appropriate actions in order to mitigate the following attack
                pattern: XMRIG"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--36b376f2-d99d-4302-b1cb-ef86c7ea1b7a",
            "created": "2023-04-19T15:18:34.586089Z",
            "modified": "2023-04-19T15:18:34.586089Z",
            "relationship_type": "mitigates",
            "source_ref": "course-of-action--e97fb560-5fb0-4f2e-8733-4ba73d4f29c4",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
        {
            "type": "indicator",
            "spec_version": "2.1",
            "id": "indicator--11b82a61-d77e-456d-89cd-7135c9ad92b6",
            "created": "2023-04-19T15:18:34.587085Z",
            "modified": "2023-04-19T15:18:34.587085Z",
            "name": "XMRIG",
            "description": "This file is part of XMRIG. Download artefact here:http://localhost:8000/STIX
                /OBJECTLIBRARY/Ex/testexercise/artifacts/XMRIG0.zip",
            "indicator_types": [
                "malicious-activity"
            ],
            "pattern": "[ file:hashes.'x_foo_hash' = 'aaaabbbbccccddddeeeeffff0123457890']",
            "pattern_type": "stix",
            "pattern_version": "2.1",
            "valid_from": "2022-01-01T00:00:00Z"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--ac0fe9ba-9939-4eca-b2b1-808bb8755fa3",
            "created": "2023-04-19T15:18:34.600084Z",
            "modified": "2023-04-19T15:18:34.600084Z",
            "relationship_type": "indicates",
            "source_ref": "indicator--11b82a61-d77e-456d-89cd-7135c9ad92b6",
            "target_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--c188ff45-c3d0-4d99-896d-27beaac2ad24",
            "created": "2023-04-19T15:18:34.600084Z",
            "modified": "2023-04-19T15:18:34.600084Z",
            "name": "ZERO-DAY Vulnerability Used",
            "description": "Research and collect evidence for the following Vulnerability: ZERO-DAY. More
                info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=ZERO-DAY"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
```

```
            "id": "relationship--a7e5e4b9-65ad-4d06-a88a-a77a3fa24139",
            "created": "2023-04-19T15:18:34.601086Z",
            "modified": "2023-04-19T15:18:34.601086Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703",
            "target_ref": "course-of-action--c188ff45-c3d0-4d99-896d-27beaac2ad24"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--12d60784-9a50-4ece-b2c2-5b1c7b0461e0",
            "created": "2023-04-19T15:18:34.601086Z",
            "modified": "2023-04-19T15:18:34.601086Z",
            "name": "CVE-2021-26084 Vulnerability Used",
            "description": "Research and collect evidence for the following Vulnerability: CVE
                -2021-26084. More info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE
                -2021-26084"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--f6b3a3e3-1647-4ff8-b40c-9b4df14f4e90",
            "created": "2023-04-19T15:18:34.601086Z",
            "modified": "2023-04-19T15:18:34.601086Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703",
            "target_ref": "course-of-action--12d60784-9a50-4ece-b2c2-5b1c7b0461e0"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--87d3d589-6ce2-4d67-80de-bb0eb15ef31a",
            "created": "2023-04-19T15:18:34.602084Z",
            "modified": "2023-04-19T15:18:34.602084Z",
            "name": "CVE-2020-3309 Vulnerability Used",
            "description": "Research and collect evidence for the following Vulnerability: CVE-2020-3309.
                 More info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE-2020-3309"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--65a8b9cd-e331-4767-b568-8e232b716c84",
            "created": "2023-04-19T15:18:34.602084Z",
            "modified": "2023-04-19T15:18:34.602084Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703",
            "target_ref": "course-of-action--87d3d589-6ce2-4d67-80de-bb0eb15ef31a"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--04250639-829c-4e7b-804d-e4eb711bdf73",
            "created": "2023-04-19T14:48:29.298818Z",
            "modified": "2023-04-19T14:48:29.298818Z",
            "relationship_type": "attributed-to",
            "source_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703",
            "target_ref": "threat-actor--8c3a5475-76eb-470a-8953-b364fef84676"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--138461d9-c978-4ff4-a07f-9e90f9c0e65c",
            "created": "2023-04-19T14:48:29.298818Z",
            "modified": "2023-04-19T14:48:29.298818Z",
            "relationship_type": "attributed-to",
            "source_ref": "threat-actor--8c3a5475-76eb-470a-8953-b364fef84676",
            "target_ref": "identity--a08a65d0-442b-4934-b306-2ed0db6fe5b0"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--98af9970-b7db-4593-98fe-e2d4ff3fa701",
            "created": "2023-04-19T14:48:29.298818Z",
            "modified": "2023-04-19T14:48:29.298818Z",
            "relationship_type": "targets",
            "source_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703",
            "target_ref": "identity--32bcbd3d-f9e8-4679-9045-3e47f5e37de7"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--59a41e19-f686-4f2a-9585-8f0118a7de56",
            "created": "2023-04-19T14:48:29.298818Z",
```

```
            "modified": "2023-04-19T14:48:29.298818Z",
            "relationship_type": "targets",
            "source_ref": "threat-actor--8c3a5475-76eb-470a-8953-b364fef84676",
            "target_ref": "identity--32bcbd3d-f9e8-4679-9045-3e47f5e37de7"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--cf34db27-9356-44d7-88aa-a6b80820c57f",
            "created": "2023-04-19T14:48:29.30282Z",
            "modified": "2023-04-19T14:48:29.30282Z",
            "relationship_type": "located-at",
            "source_ref": "threat-actor--8c3a5475-76eb-470a-8953-b364fef84676",
            "target_ref": "location--d6a16cf4-c101-40fd-94a9-0826bd5527c1"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--13ac34ed-fb0a-4e1a-9b53-2cb30016114d",
            "created": "2023-04-20T11:20:51.374224Z",
            "modified": "2023-04-20T11:20:51.374224Z",
            "relationship_type": "related-to",
            "source_ref": "intrusion-set--c03fa2e3-b421-4916-91e1-95c9cd2a1703",
            "target_ref": "campaign--95ab80b3-b15f-4643-a9ea-a3ad11aeff79"
        },
        {
            "type": "threat-actor",
            "spec_version": "2.1",
            "id": "threat-actor--3deed99f-6859-4386-8e2a-0a28aaacb337",
            "created": "2023-04-19T12:13:38.50924Z",
            "modified": "2023-04-20T11:01:23.068577Z",
            "name": "TURLA",
            "threat_actor_types": [
                "crime-syndicate"
            ],
            "goals": [
                "USER CREDENTIALS, PRIVATE INFORMATION, EMAIL, ACCOUNT CREDENTIALS, DATA, ADDRESSES"
            ]
        },
        {
            "type": "threat-actor",
            "spec_version": "2.1",
            "id": "threat-actor--3deed99f-6859-4386-8e2a-0a28aaacb337",
            "created": "2023-04-19T12:13:38.50924Z",
            "modified": "2023-04-20T11:01:23.068577Z",
            "name": "TURLA",
            "threat_actor_types": [
                "crime-syndicate"
            ],
            "goals": [
                "USER CREDENTIALS, PRIVATE INFORMATION, EMAIL, ACCOUNT CREDENTIALS, DATA, ADDRESSES"
            ]
        },
        {
            "type": "identity",
            "spec_version": "2.1",
            "id": "identity--7e3bbdf1-5e27-43f8-b792-2e797f3c4d88",
            "created": "2023-04-19T12:13:38.50924Z",
            "modified": "2023-04-20T10:59:57.088399Z",
            "name": "Attacker"
        },
        {
            "type": "intrusion-set",
            "spec_version": "2.1",
            "id": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0",
            "created": "2023-04-19T12:13:38.508055Z",
            "modified": "2023-04-20T11:01:23.068938Z",
            "name": "BLACKHAT_ASIA_INC2",
            "description": "\n\nCybersecurity incidents are on the rise, and targeted attacks on
                businesses and organizations are becoming increasingly more malicious. Vulnerabilities
                in networks, systems, software, and applications pose a major risk, as cyber criminals
                are leveraging phishing tactics to gain access and launch malicious file attacks. Spear
                phishing campaigns are becoming more sophisticated, and hackers are now using stolen
                credentials, SSH, proxies and domains to gain access to organizational networks and
                easily acquire email accounts.",
            "goals": [
                "USER CREDENTIALS, PRIVATE INFORMATION, EMAIL, ACCOUNT CREDENTIALS, DATA, ADDRESSES"
            ]
        },
        {
            "type": "note",
            "spec_version": "2.1",
            "id": "note--08ed1400-9254-414c-90b0-928db03fb097",
```

```
            "created": "2023-04-20T11:01:38.988727Z",
            "modified": "2023-04-20T11:01:38.988727Z",
            "abstract": "Training_Objectives",
            "content": "PHISHING_SOCIAL_ENGINEERING",
            "object_refs": [
                "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--2ed6926b-e4e1-41ca-b6ea-0b1f3ea9df42",
            "created": "2023-04-20T11:01:38.988727Z",
            "modified": "2023-04-20T11:01:38.988727Z",
            "relationship_type": "related-to",
            "source_ref": "note--08ed1400-9254-414c-90b0-928db03fb097",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "note",
            "spec_version": "2.1",
            "id": "note--a220db49-6593-4986-81dd-801716b48459",
            "created": "2023-04-20T11:01:38.989724Z",
            "modified": "2023-04-20T11:01:38.989724Z",
            "abstract": "Resources",
            "content": "['https://www.bleepingcomputer.com/news/security/north-korean-hackers-target-
                defense-industry-with-custom-malware/', 'https://www.zdnet.com/article/north-korean-
                hackers-launch-rokrat-trojan-in-campaigns-against-the-south/', 'https://www.zdnet.com/
                article/fbi-phishing-emails-are-spreading-this-sophisticated-malware/', 'https://www.
                bleepingcomputer.com/news/security/microsoft-code-sign-check-bypassed-to-drop-zloader-
                malware/', 'https://www.bleepingcomputer.com/news/security/ikea-email-systems-hit-by-
                ongoing-cyberattack/', 'https://www.bleepingcomputer.com/news/security/phishing-attacks-
                use-qr-codes-to-steal-banking-credentials/']",
            "object_refs": [
                "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--aff9be9c-a59d-4e3c-9851-a2df1aa4f516",
            "created": "2023-04-20T11:01:38.989724Z",
            "modified": "2023-04-20T11:01:38.989724Z",
            "relationship_type": "related-to",
            "source_ref": "note--a220db49-6593-4986-81dd-801716b48459",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "report",
            "id": "report--9b6b49d2-ea1f-49e4-a5d2-794fc28bf487",
            "created": "2023-04-20T11:01:41.565Z",
            "modified": "2023-04-20T11:01:41.565Z",
            "name": "News Item 0",
            "description": "\n\nAlert: #EnergyCorp was a victim of a targeted attack, exploiting multiple
                vulnerabilities & phishing tactics. Time to stay vigilant & bolster cybersecurity
                measures to protect against future attacks. #cybersecurity #cyberattack #targetedattacks
                #vulnerabilities #phishing",
            "published": "2023-04-20T14:01:41.56513Z",
            "object_refs": [
                "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--c43e33bb-71aa-4884-b01e-b5881b9593af",
            "created": "2023-04-20T11:01:41.566084Z",
            "modified": "2023-04-20T11:01:41.566084Z",
            "relationship_type": "related-to",
            "source_ref": "report--9b6b49d2-ea1f-49e4-a5d2-794fc28bf487",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "report",
            "id": "report--a660bb8a-2d8a-4498-876c-bf1aa712ceb7",
            "created": "2023-04-20T11:01:46.071Z",
            "modified": "2023-04-20T11:01:46.071Z",
            "name": "News Item 1",
            "description": "\n\nEnergyCorp was recently the victim of a targeted cyber attack exploiting
                several vulnerabilities. The attack was likely due to a phishing scam that tricked users
```

```
                into surrendering their sensitive account information. #CyberAttack #TargetedAttacks #
                Vulnerabilities #Phishing",
            "published": "2023-04-20T14:01:46.071143Z",
            "object_refs": [
                "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--d7451236-f4ce-4141-8d7b-944cb2e747ec",
            "created": "2023-04-20T11:01:46.072138Z",
            "modified": "2023-04-20T11:01:46.072138Z",
            "relationship_type": "related-to",
            "source_ref": "report--a660bb8a-2d8a-4498-876c-bf1aa712ceb7",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "report",
            "id": "report--78a3db9a-43d8-46f8-85d8-3c1271ca9863",
            "created": "2023-04-20T11:01:49.552Z",
            "modified": "2023-04-20T11:01:49.552Z",
            "name": "News Item 2",
            "description": "\n\nCyber Attack Targeted at EnergyCorp Exposes Vulnerabilities Through
                Phishing. EnergyCorp, a multinational energy corporation, was the target of a malicious
                cyber attack that exposed numerous vulnerabilities through phishing. .",
            "published": "2023-04-20T14:01:49.55226Z",
            "object_refs": [
                "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--dc6634fb-70d1-43de-9254-b63d8d08c908",
            "created": "2023-04-20T11:01:49.553257Z",
            "modified": "2023-04-20T11:01:49.553257Z",
            "relationship_type": "related-to",
            "source_ref": "report--78a3db9a-43d8-46f8-85d8-3c1271ca9863",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "report",
            "id": "report--8a7134e0-db09-442e-9789-35079641ad98",
            "created": "2023-04-20T11:01:56.516Z",
            "modified": "2023-04-20T11:01:56.516Z",
            "name": "Email 1",
            "description": "\n\nFrom: Helpdesk\nTo: CEO\nSubject: Recent Cyber Attack on EnergyCorp\n\
                nDear CEO,\n\nWe have recently become aware of a targeted cyber attack on EnergyCorp.
                This attack was likely enabled by exploiting vulnerabilities in our systems, as well as
                potentially through phishing tactics. We are currently investigating the attack and will
                 provide updates as we learn more.\n\nAt the moment, we have taken measures to secure
                our systems while the investigation is ongoing. We will keep",
            "published": "2023-04-20T14:01:56.516438Z",
            "object_refs": [
                "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--b3b503ea-fd64-4c50-80a4-7185bb3d33e1",
            "created": "2023-04-20T11:01:56.51748Z",
            "modified": "2023-04-20T11:01:56.51748Z",
            "relationship_type": "related-to",
            "source_ref": "report--8a7134e0-db09-442e-9789-35079641ad98",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "identity",
            "spec_version": "2.1",
            "id": "identity--4dddd7d0-df9b-4a1a-86af-26417a540358",
            "created": "2023-04-19T12:13:38.50924Z",
            "modified": "2023-04-20T10:59:57.090265Z",
```

```
            "name": "EnergyCorp",
            "roles": [
                "ENTERPRISE, BANKING, IT SECTOR"
            ],
            "external_references": [
                {
                    "source_name": "News Article Source",
                    "url": "https://www.bleepingcomputer.com/news/security/microsoft-code-sign-check-bypassed-to-drop-zloader-malware/"
                }
            ]
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--2dcc8a3a-5d48-42bd-8337-613c94d770d4",
            "created": "2023-04-20T11:01:56.518473Z",
            "modified": "2023-04-20T11:01:56.518473Z",
            "name": "DATA BREACH Mitigation",
            "description": "Conduct all the appropriate actions in order to mitigate the following attack pattern: DATA BREACH"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--9ff72e18-6315-4acd-a83f-f4ff02b2df54",
            "created": "2023-04-20T11:01:56.519462Z",
            "modified": "2023-04-20T11:01:56.519462Z",
            "relationship_type": "mitigates",
            "source_ref": "course-of-action--2dcc8a3a-5d48-42bd-8337-613c94d770d4",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "indicator",
            "spec_version": "2.1",
            "id": "indicator--bef8bd87-28f9-4acf-8f2e-815766ef010b",
            "created": "2023-04-20T11:01:56.520466Z",
            "modified": "2023-04-20T11:01:56.520466Z",
            "name": "DATA BREACH",
            "description": "This file is part of DATA BREACH. Download artefact here:http://localhost:8000/STIX/OBJECTLIBRARY/Ex/testexercise/artifacts/DATA BREACH0.zip",
            "indicator_types": [
                "malicious-activity"
            ],
            "pattern": "[ file:hashes.'x_foo_hash' = 'aaaabbbbccccddddeeeeffff0123457890']",
            "pattern_type": "stix",
            "pattern_version": "2.1",
            "valid_from": "2022-01-01T00:00:00Z"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--eee9dd43-82ce-46e3-8b7f-c48c6a9b796e",
            "created": "2023-04-20T11:01:56.556352Z",
            "modified": "2023-04-20T11:01:56.556352Z",
            "relationship_type": "indicates",
            "source_ref": "indicator--bef8bd87-28f9-4acf-8f2e-815766ef010b",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--672c9d18-0ed6-4b2a-9f48-874e8ff877e0",
            "created": "2023-04-20T11:01:56.557344Z",
            "modified": "2023-04-20T11:01:56.557344Z",
            "name": "COMPROMISE Mitigation",
            "description": "Conduct all the appropriate actions in order to mitigate the following attack pattern: COMPROMISE"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--21cb0ca3-d317-420b-bdb1-bbcb0152b265",
            "created": "2023-04-20T11:01:56.557344Z",
            "modified": "2023-04-20T11:01:56.557344Z",
            "relationship_type": "mitigates",
            "source_ref": "course-of-action--672c9d18-0ed6-4b2a-9f48-874e8ff877e0",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "indicator",
            "spec_version": "2.1",
            "id": "indicator--814f434e-ce08-45a6-af85-bcc2a1f3bc41",
```

```
            "created": "2023-04-20T11:01:56.557344Z",
            "modified": "2023-04-20T11:01:56.557344Z",
            "name": "COMPROMISE",
            "description": "This file is part of COMPROMISE. Download artefact here:http://localhost:8000/STIX/OBJECTLIBRARY/Ex/testexercise/artifacts/COMPROMISE1.zip",
            "indicator_types": [
                "malicious-activity"
            ],
            "pattern": "[ file:hashes.'x_foo_hash' = 'aaaabbbbccccddddeeeeffff0123457890']",
            "pattern_type": "stix",
            "pattern_version": "2.1",
            "valid_from": "2022-01-01T00:00:00Z"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--e178f70a-c6f5-461b-b882-83af010f8594",
            "created": "2023-04-20T11:01:56.560337Z",
            "modified": "2023-04-20T11:01:56.560337Z",
            "relationship_type": "indicates",
            "source_ref": "indicator--814f434e-ce08-45a6-af85-bcc2a1f3bc41",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--6d755cdf-89f5-4367-abe0-3dd4af949972",
            "created": "2023-04-20T11:01:56.561334Z",
            "modified": "2023-04-20T11:01:56.561334Z",
            "name": "TWITTER Forensics",
            "description": "Conduct all the appropriate actions in order to collect evidence for the following system or infrastructure: TWITTER"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--615469b8-8d0d-4c0e-88e6-b0338e733ccd",
            "created": "2023-04-20T11:01:56.561334Z",
            "modified": "2023-04-20T11:01:56.561334Z",
            "relationship_type": "mitigates",
            "source_ref": "course-of-action--6d755cdf-89f5-4367-abe0-3dd4af949972",
            "target_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--28917d9f-eb44-44d9-8127-a212f33abdeb",
            "created": "2023-04-19T12:13:38.50924Z",
            "modified": "2023-04-19T12:13:38.50924Z",
            "relationship_type": "attributed-to",
            "source_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0",
            "target_ref": "threat-actor--3deed99f-6859-4386-8e2a-0a28aaacb337"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--0df75a8b-374b-40f5-a740-be70e38f2f30",
            "created": "2023-04-19T12:13:38.50924Z",
            "modified": "2023-04-19T12:13:38.50924Z",
            "relationship_type": "attributed-to",
            "source_ref": "threat-actor--3deed99f-6859-4386-8e2a-0a28aaacb337",
            "target_ref": "identity--7e3bbdf1-5e27-43f8-b792-2e797f3c4d88"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--2455d499-e27c-4034-ac60-d828a54d9ef7",
            "created": "2023-04-19T12:13:38.50924Z",
            "modified": "2023-04-19T12:13:38.50924Z",
            "relationship_type": "targets",
            "source_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0",
            "target_ref": "identity--4dddd7d0-df9b-4a1a-86af-26417a540358"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--82f065f0-58e4-4941-9b29-8d87f360d625",
            "created": "2023-04-19T12:13:38.510237Z",
            "modified": "2023-04-19T12:13:38.510237Z",
            "relationship_type": "targets",
            "source_ref": "threat-actor--3deed99f-6859-4386-8e2a-0a28aaacb337",
            "target_ref": "identity--4dddd7d0-df9b-4a1a-86af-26417a540358"
        },
```

```
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--9ca19a1c-5ed3-41a6-8521-20bf10afc9ea",
            "created": "2023-04-20T11:20:51.405172Z",
            "modified": "2023-04-20T11:20:51.405172Z",
            "relationship_type": "related-to",
            "source_ref": "intrusion-set--651f1f5c-f852-4592-9dec-5db521a046f0",
            "target_ref": "campaign--95ab80b3-b15f-4643-a9ea-a3ad11aeff79"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--24557e0d-5b8e-484d-9582-98871d32b468",
            "created": "2023-04-20T11:24:46.785282Z",
            "modified": "2023-04-20T11:24:46.785282Z",
            "relationship_type": "related-to",
            "source_ref": "campaign--95ab80b3-b15f-4643-a9ea-a3ad11aeff79",
            "target_ref": "grouping--84e4d88f-44ea-4bcd-bbf3-b2c1c320bcb3"
        },
        {
            "type": "campaign",
            "spec_version": "2.1",
            "id": "campaign--b9ee53d9-8311-49f0-bb55-fb3e33b5dde2",
            "created": "2023-04-20T11:23:38.330009Z",
            "modified": "2023-04-20T11:23:48.98785Z",
            "name": "BLACKHAT_ASIA_EV2",
            "description": "As a result, the data and personal information of the victims were leaked and
                 the attackers were able to access sensitive information stored in the system. This
                cybersecurity incident highlights the need for organizations to remain vigilant in
                protecting their systems and data from attackers."
        },
        {
            "type": "threat-actor",
            "spec_version": "2.1",
            "id": "threat-actor--aef14e61-89d9-4247-86f5-395a61adf696",
            "created": "2023-04-19T12:16:26.480352Z",
            "modified": "2023-04-20T11:03:20.69561Z",
            "name": "Attacker",
            "threat_actor_types": [
                "crime-syndicate"
            ],
            "goals": [
                "USER ACCOUNTS, INFORMATION, PERSONAL INFORMATION, PASSWORD, BITCOIN, EMAIL,
                     INFRASTRUCTURE, CREDENTIALS, PASSWORDS"
            ]
        },
        {
            "type": "threat-actor",
            "spec_version": "2.1",
            "id": "threat-actor--aef14e61-89d9-4247-86f5-395a61adf696",
            "created": "2023-04-19T12:16:26.480352Z",
            "modified": "2023-04-20T11:03:20.69561Z",
            "name": "Attacker",
            "threat_actor_types": [
                "crime-syndicate"
            ],
            "goals": [
                "USER ACCOUNTS, INFORMATION, PERSONAL INFORMATION, PASSWORD, BITCOIN, EMAIL,
                     INFRASTRUCTURE, CREDENTIALS, PASSWORDS"
            ]
        },
        {
            "type": "threat-actor",
            "spec_version": "2.1",
            "id": "threat-actor--aef14e61-89d9-4247-86f5-395a61adf696",
            "created": "2023-04-19T12:16:26.480352Z",
            "modified": "2023-04-20T11:03:20.69561Z",
            "name": "Attacker",
            "threat_actor_types": [
                "crime-syndicate"
            ],
            "goals": [
                "USER ACCOUNTS, INFORMATION, PERSONAL INFORMATION, PASSWORD, BITCOIN, EMAIL,
                     INFRASTRUCTURE, CREDENTIALS, PASSWORDS"
            ]
        },
        {
            "type": "threat-actor",
            "spec_version": "2.1",
            "id": "threat-actor--aef14e61-89d9-4247-86f5-395a61adf696",
            "created": "2023-04-19T12:16:26.480352Z",
            "modified": "2023-04-20T11:03:20.69561Z",
```

```
            "name": "Attacker",
            "threat_actor_types": [
                "crime-syndicate"
            ],
            "goals": [
                "USER ACCOUNTS, INFORMATION, PERSONAL INFORMATION, PASSWORD, BITCOIN, EMAIL,
                    INFRASTRUCTURE, CREDENTIALS, PASSWORDS"
            ]
        },
        {
            "type": "identity",
            "spec_version": "2.1",
            "id": "identity--14abc8b6-392a-45df-a51f-8e3709bd0086",
            "created": "2023-04-19T12:16:26.480352Z",
            "modified": "2023-04-20T11:03:20.696607Z",
            "name": "Attacker"
        },
        {
            "type": "intrusion-set",
            "spec_version": "2.1",
            "id": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "created": "2023-04-19T12:16:26.480352Z",
            "modified": "2023-04-20T11:05:04.429032Z",
            "name": "BLACKHAT_ASIA_INC3",
            "description": "\nRecently, a cybersecurity incident resulted in the leak of sensitive data
                and personal information. Attackers used emails and malicious files to access data
                stored on the system. They then used MetaTags like DISINFORMATION, PPI and CREDENTIALS
                to exploit system vulnerabilities. \nThe attackers used a variety of techniques to
                infiltrate the system, such as a web shell, credentials masquerading, malicious domains
                and proxies.",
            "goals": [
                "USER ACCOUNTS, INFORMATION, PERSONAL INFORMATION, PASSWORD, BITCOIN, EMAIL,
                    INFRASTRUCTURE, CREDENTIALS, PASSWORDS"
            ]
        },
        {
            "type": "note",
            "spec_version": "2.1",
            "id": "note--9b0fb542-cf51-44ee-9ae4-0ba91ba69f45",
            "created": "2023-04-20T11:04:13.219731Z",
            "modified": "2023-04-20T11:04:13.219731Z",
            "abstract": "Training_Objectives",
            "content": "CYBER_HYGIENE, GDPR",
            "object_refs": [
                "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--dbc4660c-c017-4df3-ba96-b7b302932f8a",
            "created": "2023-04-20T11:04:13.220722Z",
            "modified": "2023-04-20T11:04:13.220722Z",
            "relationship_type": "related-to",
            "source_ref": "note--9b0fb542-cf51-44ee-9ae4-0ba91ba69f45",
            "target_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
        },
        {
            "type": "note",
            "spec_version": "2.1",
            "id": "note--9e16f79c-59d5-4207-87ea-7e86dfe485d2",
            "created": "2023-04-20T11:04:13.220722Z",
            "modified": "2023-04-20T11:04:13.220722Z",
            "abstract": "Resources",
            "content": "['https://www.bleepingcomputer.com/news/security/govt-hackers-impersonate-hr-
                employees-to-hit-israeli-targets/', 'https://www.zdnet.com/article/ransomware-warning-
                hackers-are-using-log4j-flaw-as-part-of-their-attacks-warns-microsoft/', 'https://www.
                zdnet.com/article/cisa-warns-of-apt-actors-exploiting-newly-identified-vulnerability-in-
                manageengine-adselfservice-plus/', 'https://www.bleepingcomputer.com/news/security/
                phishing-attacks-use-qr-codes-to-steal-banking-credentials/', 'https://securityaffairs.
                co/wordpress/118306/digital-id/kubernetes-clusters-teamtnt.html', 'https://www.zdnet.com
                /article/microsoft-warns-of-current-nobelium-phishing-campaign-impersonating-usaid/', '
                https://www.bleepingcomputer.com/news/security/new-cring-ransomware-hits-unpatched-
                fortinet-vpn-devices/', 'https://www.bleepingcomputer.com/news/security/jenkins-projects
                -confluence-server-hacked-to-mine-monero/', 'https://www.bleepingcomputer.com/news/
                security/synology-warns-of-malware-infecting-nas-devices-with-ransomware/']",
            "object_refs": [
                "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
```

```
            "id": "relationship--f9217b2c-6327-49d3-baf8-e97bd50226af",
            "created": "2023-04-20T11:04:13.221729Z",
            "modified": "2023-04-20T11:04:13.221729Z",
            "relationship_type": "related-to",
            "source_ref": "note--9e16f79c-59d5-4207-87ea-7e86dfe485d2",
            "target_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
        },
        {
            "type": "report",
            "id": "report--92684e47-796c-4f09-935d-3b32972006a8",
            "created": "2023-04-20T11:04:17.010Z",
            "modified": "2023-04-20T11:04:17.010Z",
            "name": "News Item 0",
            "description": "\n\n#EnergyCorp was the victim of a #cyberattack that included emails, #
                disinformation, and the release of personal data, #PPI, and sensitive credentials.
                Attackers have gained access to sensitive information and must be stopped! #databreach #
                security",
            "published": "2023-04-20T14:04:17.010025Z",
            "object_refs": [
                "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--f7679896-e2e4-42e9-a76c-4f1114b96f1c",
            "created": "2023-04-20T11:04:17.011024Z",
            "modified": "2023-04-20T11:04:17.011024Z",
            "relationship_type": "related-to",
            "source_ref": "report--92684e47-796c-4f09-935d-3b32972006a8",
            "target_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
        },
        {
            "type": "report",
            "id": "report--5fbdf91a-62dc-441f-bc03-1608ddd4f4e1",
            "created": "2023-04-20T11:04:19.366Z",
            "modified": "2023-04-20T11:04:19.366Z",
            "name": "News Item 1",
            "description": "\n\nEnergyCorp was the victim of a cyber attack where their personal data,
                credentials, and sensitive information were compromised. Attackers attempted to spread
                disinformation by taking and selling consumer PPI, leaving emails vulnerable to further
                attacks. #CyberAttack",
            "published": "2023-04-20T14:04:19.366461Z",
            "object_refs": [
                "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--98fcc649-26ae-4c65-a900-4b6518da5e14",
            "created": "2023-04-20T11:04:19.367465Z",
            "modified": "2023-04-20T11:04:19.367465Z",
            "relationship_type": "related-to",
            "source_ref": "report--5fbdf91a-62dc-441f-bc03-1608ddd4f4e1",
            "target_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
        },
        {
            "type": "report",
            "id": "report--41b7bf51-85e7-49d1-a7bd-d96b1c029904",
            "created": "2023-04-20T11:04:24.179Z",
            "modified": "2023-04-20T11:04:24.179Z",
            "name": "News Item 2",
            "description": "\n\n\"EnergyCorp, the US energy provider, has become the latest victim of a
                sophisticated cyber attack targeting emails, personal data, and sensitive credentials.
                Attackers used a sophisticated network of disinformation to steal personal information
                and potentially protected data such as PPI. The full extent of the attack is still under
                 investigation, but EnergyCorp has warned customers to be aware of this new threat. .",
            "published": "2023-04-20T14:04:24.17959Z",
            "object_refs": [
                "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
            ],
            "labels": [
                ""
            ]
        },
        {
```

```
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--3d80a1c8-163d-4d4b-85d6-7ffcadc3f938",
            "created": "2023-04-20T11:04:24.180545Z",
            "modified": "2023-04-20T11:04:24.180545Z",
            "relationship_type": "related-to",
            "source_ref": "report--41b7bf51-85e7-49d1-a7bd-d96b1c029904",
            "target_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
        },
        {
            "type": "report",
            "id": "report--2a50ff65-6a05-4a23-8c39-1e88e0553180",
            "created": "2023-04-20T11:04:30.527Z",
            "modified": "2023-04-20T11:04:30.527Z",
            "name": "Email 1",
            "description": "\n\nEMAIL1\nFrom: Helpdesk\nTo: ICT\nSubject: Cyber attack at EnergyCorp\n\
                nDear ICT,\n\nWe have received reports of a recent cyber attack against EnergyCorp. The
                attack involved the unauthorized access and retrieval of sensitive data, such as emails
                and personal data, as well as credentials, such as PPI and other usernames and passwords
                .\n\nAs we work to identify the attackers and secure our systems, it is imperative that
                we",
            "published": "2023-04-20T14:04:30.527856Z",
            "object_refs": [
                "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
            ],
            "labels": [
                ""
            ]
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--9e31befc-ea3e-4d8c-bafb-f4d79a14f52d",
            "created": "2023-04-20T11:04:30.528857Z",
            "modified": "2023-04-20T11:04:30.528857Z",
            "relationship_type": "related-to",
            "source_ref": "report--2a50ff65-6a05-4a23-8c39-1e88e0553180",
            "target_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
        },
        {
            "type": "identity",
            "spec_version": "2.1",
            "id": "identity--898adce6-e915-4a56-ac55-0f648490eadd",
            "created": "2023-04-19T12:16:26.480352Z",
            "modified": "2023-04-20T11:03:20.698602Z",
            "name": "EnergyCorp",
            "roles": [
                ""
            ],
            "external_references": [
                {
                    "source_name": "News Article Source",
                    "url": "https://www.bleepingcomputer.com/news/security/synology-warns-of-malware-
                        infecting-nas-devices-with-ransomware/"
                }
            ]
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--1dc3282f-d9da-4388-a12c-7d591d90b752",
            "created": "2023-04-20T11:04:30.529855Z",
            "modified": "2023-04-20T11:04:30.529855Z",
            "name": "STEALTHWORKER Mitigation",
            "description": "Conduct all the appropriate actions in order to mitigate the following attack
                pattern: STEALTHWORKER"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--34ee973a-02ab-4c4c-b4e7-61b8a1942250",
            "created": "2023-04-20T11:04:30.529855Z",
            "modified": "2023-04-20T11:04:30.529855Z",
            "relationship_type": "mitigates",
            "source_ref": "course-of-action--1dc3282f-d9da-4388-a12c-7d591d90b752",
            "target_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
        },
        {
            "type": "indicator",
            "spec_version": "2.1",
            "id": "indicator--3f2e1c47-7e09-4ed0-9a8f-2cc1c54d3670",
            "created": "2023-04-20T11:04:30.530862Z",
            "modified": "2023-04-20T11:04:30.530862Z",
```

```
            "name": "STEALTHWORKER",
            "description": "This file is part of STEALTHWORKER. Download artefact here:http://
                localhost:8000/STIX/OBJECTLIBRARY/Ex/testexercise/artifacts/STEALTHWORKER0.zip",
            "indicator_types": [
                "malicious-activity"
            ],
            "pattern": "[ file:hashes.'x_foo_hash' = 'aaaabbbbccccddddeeeeffff0123457890']",
            "pattern_type": "stix",
            "pattern_version": "2.1",
            "valid_from": "2022-01-01T00:00:00Z"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--faf3b99a-41d5-4220-9100-9c462366fd89",
            "created": "2023-04-20T11:04:30.53584Z",
            "modified": "2023-04-20T11:04:30.53584Z",
            "relationship_type": "indicates",
            "source_ref": "indicator--3f2e1c47-7e09-4ed0-9a8f-2cc1c54d3670",
            "target_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--1876a312-500b-46a3-9472-1a345cff41e8",
            "created": "2023-04-20T11:04:30.536834Z",
            "modified": "2023-04-20T11:04:30.536834Z",
            "name": "CVE-2021-26084 Vulnerability Used",
            "description": "Research and collect evidence for the following Vulnerability: CVE
                -2021-26084. More info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE
                -2021-26084"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--ebcef4fc-a96a-4a6f-9b70-f4e4f363e133",
            "created": "2023-04-20T11:04:30.536834Z",
            "modified": "2023-04-20T11:04:30.536834Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "target_ref": "course-of-action--1876a312-500b-46a3-9472-1a345cff41e8"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--d3578126-6796-4134-bda0-0907ed00b684",
            "created": "2023-04-20T11:04:30.538334Z",
            "modified": "2023-04-20T11:04:30.538334Z",
            "name": "CVE-2021-44228 Vulnerability Used",
            "description": "Research and collect evidence for the following Vulnerability: CVE
                -2021-44228. More info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE
                -2021-44228"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--31c4fd5f-8df2-468b-ac95-428e6dccb30e",
            "created": "2023-04-20T11:04:30.538839Z",
            "modified": "2023-04-20T11:04:30.538839Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "target_ref": "course-of-action--d3578126-6796-4134-bda0-0907ed00b684"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--646b45ae-c154-47b5-95d4-f76943563ab5",
            "created": "2023-04-20T11:04:30.539785Z",
            "modified": "2023-04-20T11:04:30.539785Z",
            "name": "CVE-2018-13379 Vulnerability Used",
            "description": "Research and collect evidence for the following Vulnerability: CVE
                -2018-13379. More info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE
                -2018-13379"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--91fc4d57-9751-4e01-8d9a-de93ea998552",
            "created": "2023-04-20T11:04:30.540782Z",
            "modified": "2023-04-20T11:04:30.540782Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "target_ref": "course-of-action--646b45ae-c154-47b5-95d4-f76943563ab5"
```

```
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--a8d0a888-6707-4af2-8298-d6854d40fb5d",
            "created": "2023-04-20T11:04:30.541819Z",
            "modified": "2023-04-20T11:04:30.541819Z",
            "name": "CVE-2019-5591 Vulnerability Used",
            "description": "Research and collect evidence for the following Vulnerability: CVE-2019-5591.
                 More info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE-2019-5591"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--413e8936-e745-4130-a4ee-d873b71643de",
            "created": "2023-04-20T11:04:30.541819Z",
            "modified": "2023-04-20T11:04:30.541819Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "target_ref": "course-of-action--a8d0a888-6707-4af2-8298-d6854d40fb5d"
        },
        {
            "type": "course-of-action",
            "spec_version": "2.1",
            "id": "course-of-action--534801f2-7b5e-4aa6-a5e2-6c8adbf3a4c2",
            "created": "2023-04-20T11:04:30.542818Z",
            "modified": "2023-04-20T11:04:30.542818Z",
            "name": "CVE-2020-12812 Vulnerability Used",
            "description": "Research and collect evidence for the following Vulnerability: CVE
                 -2020-12812. More info here: https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=CVE
                 -2020-12812"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--388ea17e-a23a-476d-982a-e8ddab91de57",
            "created": "2023-04-20T11:04:30.542818Z",
            "modified": "2023-04-20T11:04:30.542818Z",
            "relationship_type": "exploits",
            "source_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "target_ref": "course-of-action--534801f2-7b5e-4aa6-a5e2-6c8adbf3a4c2"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--651e0631-828f-4546-83ae-a9c4af39a770",
            "created": "2023-04-19T12:16:26.481349Z",
            "modified": "2023-04-19T12:16:26.481349Z",
            "relationship_type": "attributed-to",
            "source_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "target_ref": "threat-actor--aef14e61-89d9-4247-86f5-395a61adf696"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--95a1b69d-8ea1-485e-a83f-57419d24c811",
            "created": "2023-04-19T12:16:26.481349Z",
            "modified": "2023-04-19T12:16:26.481349Z",
            "relationship_type": "attributed-to",
            "source_ref": "threat-actor--aef14e61-89d9-4247-86f5-395a61adf696",
            "target_ref": "identity--14abc8b6-392a-45df-a51f-8e3709bd0086"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--59e1d8f0-2056-4cc1-a11e-43a10216b2f5",
            "created": "2023-04-19T12:16:26.481349Z",
            "modified": "2023-04-19T12:16:26.481349Z",
            "relationship_type": "targets",
            "source_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "target_ref": "identity--898adce6-e915-4a56-ac55-0f648490eadd"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--a196c6ff-572c-4c26-8378-04d3100ecb58",
            "created": "2023-04-19T12:16:26.481349Z",
            "modified": "2023-04-19T12:16:26.481349Z",
            "relationship_type": "targets",
            "source_ref": "threat-actor--aef14e61-89d9-4247-86f5-395a61adf696",
            "target_ref": "identity--898adce6-e915-4a56-ac55-0f648490eadd"
        },
        {
```

```
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--bff6c957-002c-438d-9b26-ca462ba88c50",
            "created": "2023-04-20T11:23:38.358911Z",
            "modified": "2023-04-20T11:23:38.358911Z",
            "relationship_type": "related-to",
            "source_ref": "intrusion-set--040ef41b-dc59-4e0d-9250-ae6eb8a7aca0",
            "target_ref": "campaign--b9ee53d9-8311-49f0-bb55-fb3e33b5dde2"
        },
        {
            "type": "relationship",
            "spec_version": "2.1",
            "id": "relationship--8d42c2ed-89ef-4edc-916f-e9aeb7324650",
            "created": "2023-04-20T11:24:46.802247Z",
            "modified": "2023-04-20T11:24:46.802247Z",
            "relationship_type": "related-to",
            "source_ref": "campaign--b9ee53d9-8311-49f0-bb55-fb3e33b5dde2",
            "target_ref": "grouping--84e4d88f-44ea-4bcd-bbf3-b2c1c320bcb3"
        }
    ]
}
```

Bibliography

[1] European Commission and High Representative of the Union for Foreign Affairs and Security Policy. *Joint Communication to the European Parliament and the Council: The EU's Cybersecurity Strategy for the Digital Decade*. Accessed: 2025-01-04. 2020. url: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020JC0018 (cit. on p. 3).

[2] European Commission. *Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) (Text with EEA relevance)*. https://eur-lex.europa.eu/eli/dir/2022/2555. 2022 (cit. on pp. 3, 42, 102, 116).

[3] A. Furtună, V.-V. Patriciu, and I. Bica. "A structured approach for implementing cyber security exercises". *2010 8th International Conference on Communications*. IEEE. 2010, pp. 415–418 (cit. on pp. 7, 8).

[4] D. Schweitzer, D. Gibson, and M. Collins. "Active learning in the security classroom". *2009 42nd Hawaii International Conference on System Sciences*. IEEE. 2009, pp. 1–8 (cit. on p. 7).

[5] T. Augustine, R. C. Dodge, et al. "Cyber defense exercise: meeting learning objectives thru competition". *Proceedings of the 10th Colloquium for Information Systems Security Education*. 2006 (cit. on pp. 7, 8).

[6] R. S. Dewar. *Cybersecurity and cyberdefense exercises*. Tech. rep. ETH Zurich, 2018 (cit. on p. 7).

[7] R. Gurnani, K. Pandey, and S. K. Rai. "A scalable model for implementing Cyber Security Exercises". *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE. 2014, pp. 680–684 (cit. on p. 7).

[8] W. Schepens, D. Ragsdale, J. R. Surdu, J. Schafer, and R. N. Port. "The Cyber Defense Exercise: An evaluation of the effectiveness of information assurance education". *The Journal of Information Security*, 1 (2), (2002), pp. 1–14 (cit. on p. 7).

[9] R. Dodge and D. J. Ragsdale. "Organized cyber defense competitions". *IEEE International Conference on Advanced Learning Technologies, 2004. Proceedings.* IEEE. 2004, pp. 768–770 (cit. on p. 7).

[10] S. Karagiannis and E. Magkos. "Engaging students in basic cybersecurity concepts using digital game-based learning: Computer games as virtual learning environments". *Advances in Core Computer Science-Based Technologies*. Springer, 2021, pp. 55–81 (cit. on p. 7).

[11] W. J. Adams, E. Gavas, T. H. Lacey, and S. P. Leblanc. "Collective Views of the NSA/CSS Cyber Defense Exercise on Curricula and Learning Objectives." *CSET*. 2009 (cit. on p. 7).

[12] A. Conklin. "The use of a collegiate cyber defense competition in information security education". *Proceedings of the 2nd annual conference on Information security curriculum development*. 2005, pp. 16–18 (cit. on p. 7).

[13] A. Conklin. "Cyber defense competitions and information security education: An active learning solution for a capstone course". *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*. Vol. 9. IEEE. 2006, 220b–220b (cit. on pp. 7, 41).

[14] B. E. Mullins, T. H. Lacey, R. F. Mills, J. E. Trechter, and S. D. Bass. "How the cyber defense exercise shaped an information-assurance curriculum". *IEEE Security & Privacy*, 5 (5), (2007), pp. 40–49 (cit. on p. 7).

[15] B. E. Mullins, T. H. Lacey, R. F. Mills, J. M. Trechter, and S. D. Bass. "The impact of the nsa cyber defense exercise on the curriculum at the air force institute of technology". *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE. 2007, 271b–271b (cit. on p. 7).

[16] G. Vigna. "Teaching network security through live exercises". *IFIP World Conference on Information Security Education*. Springer. 2003, pp. 3–18 (cit. on p. 7).

[17] M. Mink and F. C. Freiling. "Is attack better than defense? Teaching information security the right way". *Proceedings of the 3rd annual conference on Information security curriculum development*. 2006, pp. 44–48 (cit. on p. 7).

[18] J. A. Rursch, A. Luse, and D. Jacobson. "IT-adventures: A program to spark IT interest in high school students using inquiry-based learning with cyber defense, game design, and robotics". *IEEE Transactions on Education*, 53 (1), (2009), pp. 71–79 (cit. on p. 7).

[19] G. B. White, D. Williams, and K. Harrison. "The CyberPatriot national high school cyber defense competition". *IEEE Security & Privacy*, 8 (5), (2010), pp. 59–61 (cit. on p. 7).

[20] W. J. Schepens and J. R. James. "Architecture of a cyber defense competition". *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*. Vol. 5. IEEE. 2003, pp. 4300–4305 (cit. on pp. 7, 8).

[21] A. Green and H. Zafar. "Addressing Emerging Information Security Personnel Needs. A Look at Competitions in Academia: Do Cyber Defense Competitions Work?" *AMCIS 2013 Proceedings*, 1, (2013), p. 257 (cit. on pp. 7, 8).

[22] V.-V. Patriciu and A. C. Furtuna. "Guide for designing cyber security exercises". *Proceedings of the 8th WSEAS International Conference on E-Activities and information security and privacy*. World Scientific, Engineering Academy, and Society (WSEAS). 2009, pp. 172–177 (cit. on p. 7).

[23] G. B. White, G. Dietrich, and T. Goles. "Cyber security exercises: testing an organization's ability to prevent, detect, and respond to cyber security events". *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. IEEE. 2004, 10–pp (cit. on p. 7).

[24] J. A. Mattson. "Cyber defense exercise: A service provider model". *IFIP World Conference on Information Security Education*. Springer. 2007, pp. 81–86 (cit. on p. 7).

[25] B. Sangster, T. O'Connor, T. Cook, R. Fanelli, E. Dean, C. Morrell, and G. J. Conti. "Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets." *CSET*. 2009 (cit. on p. 7).

[26] T. Sommestad and J. Hallberg. "Cyber security exercises and competitions as a platform for cyber security experiments". *Nordic conference on secure IT systems*. Springer. 2012, pp. 47–60 (cit. on p. 7).

[27] M. Samejima and H. Yajima. "IT risk management framework for business continuity by change analysis of information system". *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2012, pp. 1670–1674 (cit. on p. 7).

[28] R. Dodge, B. Hay, and K. Nance. "Standards-based cyber exercises". *2009 International Conference on Availability, Reliability and Security*. IEEE. 2009, pp. 738–743 (cit. on p. 7).

[29] M. Granåsen and D. Andersson. "Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study". *Cognition, Technology & Work*, 18 (1), (2016), pp. 121–143 (cit. on pp. 7, 8).

[30] M. Liljenstam, J. Liu, D. M. Nicol, Y. Yuan, G. Yan, and C. Grier. "Rinse: The real-time immersive network simulation environment for network security exercises (extended version)". *Simulation*, 82 (1), (2006), pp. 43–59 (cit. on p. 7).

[31] Y. Li, M. Liljenstam, and J. Liu. "Real-time security exercises on a realistic interdomain routing experiment platform". *2009 ACM/IEEE/SCS 23rd Workshop on Principles of Advanced and Distributed Simulation*. IEEE. 2009, pp. 54–63 (cit. on p. 7).

[32] D. H. Tobey. "A vignette-based method for improving cybersecurity talent management through cyber defense competition design". *Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research*. 2015, pp. 31–39 (cit. on p. 7).

[33] J. Kick. *Cyber exercise playbook*. Tech. rep. MITRE CORP BEDFORD MA, 2014 (cit. on p. 8).

[34] M. E. Planning. "Directors's Guideline for Civil Defence Emergency Management Groups, Wyd". *Ministry of Civil Defence & Emergency Management, Wellington*, (2008) (cit. on p. 8).

[35] K. A. Scarfone, T. Grance, and K. Masone. *Sp 800-61 rev. 1. computer security incident handling guide*. 2008 (cit. on p. 8).

[36] U. D. of Homeland Security. *DHS Cyber TTX for the Healthcare Industry*. https://www.hsdl.org/?abstract&did=789781. 2013 (cit. on p. 8).

[37] ISO Central Secretary. *Societal security — Guidelines for exercises*. en. Standard ISO22398:2013. Geneva, CH: International Organization for Standardization, 2013. url: https://www.iso.org/standard/50294.html (cit. on pp. 8, 41–43, 48).

[38] MITRE. *MITRE ATT&CK*. https://attack.mitre.org/. 2022 (cit. on pp. 8, 14, 48, 63, 139).

[39] Lockheed Martin. *The Cyber Kill Chain*. https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html. 2011 (cit. on pp. 8, 48).

[40] MITRE. *CVE*. https://cve.mitre.org/. 1999 (cit. on pp. 9, 48).

[41] ENISA. *National Exercise - Good Practice Guide*. https://www.enisa.europa.eu/publications/national-exercise-good-practice-guide. 2009 (cit. on p. 9).

[42] ANSII. *ORGANISING A CYBER CRISIS MANAGEMENT EXERCISE*. https://www.ssi.gouv.fr/guide/organising-a-cyber-crisis-management-exercise/. 2021 (cit. on p. 9).

[43] A. Salih, S. T. Zeebaree, S. Ameen, A. Alkhyyat, and H. M. Shukur. "A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection". *2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC)*. IEEE. 2021, pp. 61–66 (cit. on p. 9).

[44] K. Ren, Y. Zeng, Z. Cao, and Y. Zhang. "ID-RDRL: a deep reinforcement learning-based feature selection intrusion detection model". *Scientific Reports*, 12 (1), (2022), p. 15370 (cit. on p. 9).

[45] X. Liu and J. Liu. "Malicious traffic detection combined deep neural network with hierarchical attention mechanism". *Scientific Reports*, 11 (1), (2021), p. 12363 (cit. on p. 9).

[46] J. Kim, J. Kim, H. L. T. Thu, and H. Kim. "Long short term memory recurrent neural network classifier for intrusion detection". *2016 international conference on platform technology and service (PlatCon)*. IEEE. 2016, pp. 1–5 (cit. on p. 9).

[47] J. Z. Bakdash, S. Hutchinson, E. G. Zaroukian, L. R. Marusich, S. Thirumuruganathan, C. Sample, B. Hoffman, and G. Das. "Malware in the future? Forecasting of analyst detection of cyber events". *Journal of Cybersecurity*, 4 (1), (2018), tyy007 (cit. on p. 9).

[48] S. Fan, S. Wu, Z. Wang, Z. Li, J. Yang, H. Liu, and X. Liu. "Aleap: Attention-based LSTM with event embedding for attack projection". *2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)*. IEEE. 2019, pp. 1–8 (cit. on p. 9).

[49] C. Acarturk, M. Sirlanci, P. G. Balikcioglu, D. Demirci, N. Sahin, and O. A. Kucuk. "Malicious code detection: Run trace output analysis by LSTM". *IEEE Access*, 9, (2021), pp. 9625–9635 (cit. on p. 9).

[50] K. Gogineni, P. Derasari, and G. Venkataramani. "Foreseer: Efficiently Forecasting Malware Event Series with Long Short-Term Memory". *2022 IEEE International Symposium on Secure and Private Execution Environment Design (SEED)*. IEEE. 2022, pp. 97–108 (cit. on p. 9).

[51] S. J. Taylor and B. Letham. "Prophet: Forecasting at scale". *Journal of Open Source Software*, 3 (22), (2017), p. 651 (cit. on pp. 9, 11).

[52] M. N. Md Azam and N. A. Ramli. "Reported Malicious Codes Incident within Malaysia's Landscape: Time Series Modelling and a Timeline Analysis". *International Journal of Advanced Data Science and Intelligence Analytics*, 2 (2), (2022). url: https://amcs-press.com/index.php/ijadsia/article/view/65 (cit. on p. 9).

[53] I. H. Sarker, A. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng. "Cybersecurity data science: an overview from machine learning perspective". *Journal of Big data*, 7, (2020), pp. 1–29 (cit. on p. 9).

[54] M. Husák, J. Komárková, E. Bou-Harb, and P. Čeleda. "Survey of attack projection, prediction, and forecasting in cyber security". *IEEE Communications Surveys & Tutorials*, 21 (1), (2018), pp. 640–660 (cit. on p. 10).

[55] M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu. "Modeling and predicting cyber hacking breaches". *IEEE Transactions on Information Forensics and Security*, 13 (11), (2018), pp. 2856–2871 (cit. on p. 10).

[56] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey, and M. Liu. "Cloudy with a chance of breach: Forecasting cyber security incidents". *24th USENIX Security Symposium (USENIX Security 15)*. 2015, pp. 1009–1024 (cit. on p. 10).

[57] N. Sun, J. Zhang, P. Rimba, S. Gao, L. Y. Zhang, and Y. Xiang. "Data-driven cybersecurity incident prediction: A survey". *IEEE communications surveys & tutorials*, 21 (2), (2018), pp. 1744–1772 (cit. on p. 10).

[58] G. Werner, S. Yang, and K. McConky. "Time Series Forecasting of Cyber Attack Intensity". *Proceedings of the 12th Annual Conference on Cyber and Information Security Research*. CISRC '17. Oak Ridge, Tennessee, USA: Association for Computing Machinery, 2017. isbn: 9781450348553. doi: 10.1145/3064814.3064831. url: https://doi.org/10.1145/3064814.3064831 (cit. on p. 10).

[59] M. V. Carriegos and R. Á. Fernández-Díaz. "Towards Forecasting Time-Series of Cyber-Security Data Aggregates". *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*. Ed. by Á. Herrero, C. Cambra, D. Urda, J. Sedano, H. Quintián, and E. Corchado. Cham: Springer International Publishing, 2021, pp. 273–281. isbn: 978-3-030-57805-3 (cit. on p. 10).

[60] A. Okutan, G. Werner, S. J. Yang, and K. McConky. "Forecasting cyberattacks with incomplete, imbalanced, and insignificant data". *Cybersecurity*, 1, (2018), pp. 1–16 (cit. on p. 10).

[61] Z. Almahmoud, P. D. Yoo, O. Alhussein, I. Farhat, and E. Damiani. "A holistic and proactive approach to forecasting cyber threats". *Scientific Reports*, 13 (1), (2023), p. 8049 (cit. on pp. 10, 111).

[62] A. Couce-Vieira, D. R. Insua, and A. Kosgodagan. "Assessing and forecasting cybersecurity impacts". *Decision Analysis*, 17 (4), (2020), pp. 356–374 (cit. on p. 10).

[63] G. Werner, S. Yang, and K. McConky. "Time series forecasting of cyber attack intensity". *Proceedings of the 12th Annual Conference on cyber and information security research*. 2017, pp. 1–3 (cit. on p. 10).

[64] G. Werner, S. Yang, and K. McConky. "Leveraging intra-day temporal variations to predict daily cyberattack activity". *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE. 2018, pp. 58–63 (cit. on p. 10).

[65] A. Okutan, S. J. Yang, K. McConky, and G. Werner. "Capture: cyberattack forecasting using non-stationary features with time lags". *2019 IEEE Conference on Communications and Network Security (CNS)*. IEEE. 2019, pp. 205–213 (cit. on p. 10).

[66] B. Munkhdorj and S. Yuji. "Cyber attack prediction using social data analysis". *Journal of High Speed Networks*, 23 (2), (2017), pp. 109–135 (cit. on p. 10).

[67] P. Goyal, K. Hossain, A. Deb, N. Tavabi, N. Bartley, A. Abeliuk, E. Ferrara, and K. Lerman. "Discovering signals from web sources to predict cyber attacks". *arXiv preprint arXiv:1806.03342*, (2018) (cit. on p. 10).

[68] X. Qin and W. Lee. "Attack plan recognition and prediction using causal networks". *20th Annual Computer Security Applications Conference*. IEEE. 2004, pp. 370–379 (cit. on p. 10).

[69] M. Husák and J. Kašpar. "AIDA framework: real-time correlation and prediction of intrusion detection alerts". *Proceedings of the 14th international conference on availability, reliability and security*. 2019, pp. 1–8 (cit. on p. 10).

[70] J. Malik, A. Akhunzada, I. Bibi, M. Imran, A. Musaddiq, and S. W. Kim. "Hybrid deep learning: An efficient reconnaissance and surveillance detection mechanism in SDN". *IEEE Access*, 8, (2020), pp. 134695–134706 (cit. on p. 10).

[71] L. Bilge, Y. Han, and M. Dell'Amico. "Riskteller: Predicting the risk of cyber incidents". *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 2017, pp. 1299–1311 (cit. on pp. 10, 16).

[72] M. Husák, V. Bartoš, P. Sokol, and A. Gajdoš. "Predictive methods in cyber defense: Current experience and research challenges". *Future Generation Computer Systems*, 115, (2021), pp. 517–530 (cit. on p. 10).

[73] Y. Shen, E. Mariconti, P. A. Vervier, and G. Stringhini. "Tiresias: Predicting security events through deep learning". *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 592–605 (cit. on p. 10).

[74] N. Tavabi, P. Goyal, M. Almukaynizi, P. Shakarian, and K. Lerman. "Darkembed: Exploit prediction with neural language models". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018 (cit. on p. 10).

[75] O. Suciu, C. Nelson, Z. Lyu, T. Bao, and T. Dumitraș. "Expected exploitability: Predicting the development of functional vulnerability exploits". *31st USENIX Security Symposium (USENIX Security 22)*. 2022, pp. 377–394 (cit. on p. 10).

[76] N. Sun, M. Ding, J. Jiang, W. Xu, X. Mo, Y. Tai, and J. Zhang. "Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives". *IEEE Communications Surveys & Tutorials*, (2023) (cit. on p. 10).

[77] G. Bhardwaj, R. Gupta, A. P. Srivastava, and S. V. Singh. "Cyber Threat Landscape of G4 Nations: Analysis of Threat Incidents & Response Strategies". *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*. IEEE. 2021, pp. 75–79 (cit. on p. 10).

[78] M. Eto, D. Inoue, J. Song, J. Nakazato, K. Ohtaka, and K. Nakao. "Nicter: A large-scale network incident analysis system: Case studies for understanding threat landscape". *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. 2011, pp. 37–45 (cit. on p. 10).

[79] T. Kovanen, V. Nuojua, and M. Lehto. "Cyber threat landscape in energy sector". *ICCWS 2018 13th International Conference on Cyber Warfare and Security*. Academic Conferences and publishing limited. 2018, p. 353 (cit. on p. 10).

[80] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015 (cit. on p. 10).

[81] R. J. Hyndman, A. B. Koehler, and J. K. Ord. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Science & Business Media, 2008 (cit. on p. 11).

[82] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018 (cit. on p. 11).

[83] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose. "A state space framework for automatic forecasting using exponential smoothing methods". *International Journal of Forecasting*, 18 (3), (2002), pp. 439–454 (cit. on p. 11).

[84] S. Hochreiter and J. Schmidhuber. "Long short-term memory". *Neural Computation*, 9 (8), (1997), pp. 1735–1780 (cit. on p. 11).

[85] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016 (cit. on p. 11).

[86] T. Satyapanich, F. Ferraro, and T. Finin. "CASIE: Extracting Cybersecurity Event Information from Text". *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 8749–8757. doi: 10.1609/AAAI.V34I05.6401. url: https://doi.org/10.1609/aaai.v34i05.6401 (cit. on pp. 12, 15).

[87] K. Satvat, R. Gjomemo, and V. Venkatakrishnan. "Extractor: Extracting attack behavior from threat reports". *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2021, pp. 598–615 (cit. on p. 12).

[88] I. Sarhan and M. Spruit. "Open-cykg: An open cyber threat intelligence knowledge graph". *Knowledge-Based Systems*, 233, (2021), p. 107524 (cit. on p. 12).

[89] Y. Park and T. Lee. "Full-stack information extraction system for cybersecurity intelligence". *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 2022, pp. 531–539 (cit. on p. 12).

[90] K. Ahmed, S. K. Khurshid, and S. Hina. "CyberEntRel: Joint extraction of cyber entities and relations using deep learning". *Computers & Security*, 136, (2024), p. 103579 (cit. on p. 12).

[91] H. Zhang, G. Shen, C. Guo, Y. Cui, and C. Jiang. "EX-Action: Automatically Extracting Threat Actions from Cyber Threat Intelligence Report Based on Multimodal Learning". *Security and Communication Networks*, 2021 (1), (2021), p. 5586335 (cit. on p. 12).

[92] T. Sun, P. Yang, M. Li, and S. Liao. "An automatic generation approach of the cyber threat intelligence records based on multi-source information fusion". *Future Internet*, 13 (2), (2021), p. 40 (cit. on p. 12).

[93] P. Gao, F. Shao, X. Liu, X. Xiao, H. Liu, Z. Qin, F. Xu, P. Mittal, S. R. Kulkarni, and D. Song. "A system for efficiently hunting for cyber threats in computer systems using threat intelligence". *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE. 2021, pp. 2705–2708 (cit. on p. 12).

[94] C. Gao, X. Zhang, and H. Liu. "Data and knowledge-driven named entity recognition for cyber security". *Cybersecurity*, 4 (1), (2021), p. 9 (cit. on p. 12).

[95] H. Suryotrisongko, H. Ginardi, H. T. Ciptaningtyas, S. Dehqan, and Y. Musashi. "Topic modeling for cyber threat intelligence (cti)". *2022 Seventh International Conference on Informatics and Computing (ICIC)*. IEEE. 2022, pp. 1–7 (cit. on p. 12).

[96] H. Abdi, S. R. Bagley, S. Furnell, and J. Twycross. "Automatically Labeling Cyber Threat Intelligence reports using Natural Language Processing". *Proceedings of the ACM Symposium on Document Engineering 2023*. 2023, pp. 1–4 (cit. on p. 12).

[97] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. C. Cordeiro, M. Debbah, and T. Lestable. "Revolutionizing cyber threat detection with large language models". *arXiv preprint arXiv:2306.14263*, (2023) (cit. on p. 12).

[98] N. Lykousas and C. Patsakis. "Decoding developer password patterns: A comparative analysis of password extraction and selection practices". *Comput. Secur.*, 145, (2024), p. 103974. doi: 10.1016/J.COSE.2024.103974. url: https://doi.org/10.1016/j.cose.2024.103974 (cit. on p. 12).

[99] T. Koide, N. Fukushi, H. Nakano, and D. Chiba. "Detecting phishing sites using chatgpt". *arXiv preprint arXiv:2306.05816*, (2023) (cit. on p. 12).

[100] A. Chrysanthou, Y. Pantis, and C. Patsakis. "The anatomy of deception: Measuring technical and human factors of a large-scale phishing campaign". *Computers & Security*, 140, (2024), p. 103780 (cit. on p. 12).

[101] S. S. Roy, P. Thota, K. V. Naragam, and S. Nilizadeh. "From Chatbots to PhishBots?–Preventing Phishing scams created using ChatGPT, Google Bard and Claude". *arXiv preprint arXiv:2310.19181*, (2023) (cit. on p. 12).

[102] C. Patsakis, F. Casino, and N. Lykousas. "Assessing LLMs in malicious code deobfuscation of real-world malware campaigns". *Expert Syst. Appl.*, 256, (2024), p. 124912. doi: 10.1016/J.ESWA.2024.124912. url: https://doi.org/10.1016/j.eswa.2024.124912 (cit. on p. 12).

[103] Z. Song, Y. Tian, J. Zhang, and Y. Hao. "Generating fake cyber threat intelligence using the gpt-neo model". *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*. IEEE. 2023, pp. 920–924 (cit. on p. 12).

[104] M. D. Purba and B. Chu. "Extracting Actionable Cyber Threat Intelligence from Twitter Stream". *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE. 2023, pp. 1–6 (cit. on pp. 12, 14).

[105] F. Sufi. "An innovative GPT-based open-source intelligence using historical cyber incident reports". *Natural Language Processing Journal*, 7, (2024), p. 100074 (cit. on p. 12).

[106] ENISA. *ENISA Cybersecurity Threat Landscape Methodology*. https://www.enisa.europa.eu/publications/enisa-threat-landscape-methodology/. 2022 (cit. on pp. 13, 102, 134, 135).

[107] J. Liu and J. Zhan. "Constructing knowledge graph from cyber threat intelligence using large language model". *2023 IEEE International Conference on Big Data (BigData)*. IEEE. 2023, pp. 516–521 (cit. on pp. 13, 14).

[108] F. N. Motlagh, M. Hajizadeh, M. Majd, P. Najafi, F. Cheng, and C. Meinel. "Large language models in cybersecurity: State-of-the-art". *arXiv preprint arXiv:2402.00891*, (2024) (cit. on p. 13).

[109] M. Arazzi, D. R. Arikkat, S. Nicolazzo, A. Nocera, M. Conti, et al. "NLP-Based Techniques for Cyber Threat Intelligence". *arXiv preprint arXiv:2311.08807*, (2023) (cit. on p. 13).

[110] G. Siracusano, D. Sanvito, R. Gonzalez, M. Srinivasan, S. Kamatchi, W. Takahashi, M. Kawakita, T. Kakumaru, and R. Bifulco. "Time for action: Automated analysis of cyber threat intelligence in the wild". *arXiv preprint arXiv:2307.10214*, (2023) (cit. on pp. 13, 14).

[111] K. Park, J. Hong, W. Su, and H. Lee. "Machine Learning based Post Event Analysis for Cybersecurity of Cyber-Physical System". *arXiv preprint arXiv:2311.13488*, (2023) (cit. on p. 13).

[112] A. Zacharis, R. Gavrila, C. Patsakis, and D. Ikonomou. "AI-assisted Cyber Security Exercise Content Generation: Modeling a Cyber Conflict". *2023 15th International Conference on Cyber Conflict: Meeting Reality (CyCon)*. 2023, pp. 217–238. doi: 10.23919/CyCon58705.2023.10181930 (cit. on p. 13).

[113] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah. "Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence". *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 755–766 (cit. on p. 13).

[114] Z. Zhu and T. Dumitras. "Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports". *2018 IEEE European symposium on security and privacy (EuroS&P)*. IEEE. 2018, pp. 458–472 (cit. on p. 13).

[115] J. Zhao, Q. Yan, J. Li, M. Shao, Z. He, and B. Li. "TIMiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data". *Computers & Security*, 95, (2020), p. 101867 (cit. on p. 13).

[116] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu. "Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources". *Proceedings of the 33rd annual computer security applications conference*. 2017, pp. 103–115 (cit. on pp. 14, 15).

[117] S. E. Robertson and S. Walker. "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval". *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer. 1994, pp. 232–241 (cit. on p. 14).

[118] P. Gao, F. Shao, X. Liu, X. Xiao, Z. Qin, F. Xu, P. Mittal, S. R. Kulkarni, and D. Song. "Enabling efficient cyber threat hunting with cyber threat intelligence". *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE. 2021, pp. 193–204 (cit. on pp. 14, 15).

[119] V. Legoy, M. Caselli, C. Seifert, and A. Peter. "Automated retrieval of att&ck tactics and techniques for cyber threat reports". *arXiv preprint arXiv:2004.14322*, (2020) (cit. on pp. 14, 15).

[120] Z. Li, J. Zeng, Y. Chen, and Z. Liang. "AttacKG: Constructing technique knowledge graph from cyber threat intelligence reports". *European Symposium on Research in Computer Security*. Springer. 2022, pp. 589–609 (cit. on pp. 14, 15).

[121] M. T. Alam, D. Bhusal, Y. Park, and N. Rastogi. "Looking beyond IoCs: Automatically extracting attack patterns from external CTI". *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*. 2023, pp. 92–108 (cit. on pp. 14, 15).

[122] N. Rani, B. Saha, V. Maurya, and S. K. Shukla. "TTPHunter: Automated Extraction of Actionable Intelligence as TTPs from Narrative Threat Reports". *Proceedings of the 2023 Australasian Computer Science Week, ACSW 2023, Melbourne, VIC, Australia, 30 January 2023- 3 February 2023*. ACM, 2023, pp. 126–134. doi: 10.1145/3579375.3579391. url: https://doi.org/10.1145/3579375.3579391 (cit. on pp. 14, 15, 21, 139).

[123] MITRE. *Threat Report ATT&CK Mapper (TRAM)*. https://github.com/center-for-threat-informed-defense/tram/. 2023 (cit. on pp. 14, 15, 21, 139).

[124] N. Rani, B. Saha, V. Maurya, and S. K. Shukla. "TTPXHunter: Actionable Threat Intelligence Extraction as TTPs form Finished Cyber Threat Reports". *arXiv preprint arXiv:2403.03267*, (2024) (cit. on pp. 14, 15).

[125] Y. Hu, F. Zou, J. Han, X. Sun, and Y. Wang. "Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model". *Computers & Security*, 145, (2024), p. 103999 (cit. on p. 14).

[126] Y. Fengrui and Y. Du. "Few-Shot Learning of TTPs Classification Using Large Language Models". *Preprints.org*, (2024). doi: 10.20944/preprints202401.0372.v1. url: https://doi.org/10.20944/preprints202401.0372.v1 (cit. on pp. 14, 15).

[127] Y. Zhang, J. Chen, Z. Cheng, X. Shen, J. Qin, Y. Han, and Y. Lu. "Edge propagation for link prediction in requirement-cyber threat intelligence knowledge graph". *Information Sciences*, 653, (2024), p. 119770 (cit. on p. 14).

[128] S. Fujii, N. Kawaguchi, T. Shigemoto, and T. Yamauchi. "Cyner: Information extraction from unstructured text of cti sources with noncontextual iocs". *International Workshop on Security*. Springer. 2022, pp. 85–104 (cit. on p. 14).

[129] J. Liu, J. Yan, J. Jiang, Y. He, X. Wang, Z. Jiang, P. Yang, and N. Li. "TriCTI: an actionable cyber threat intelligence discovery system via trigger-enhanced neural network". *Cybersecurity*, 5 (1), (2022), p. 8 (cit. on pp. 14, 15).

[130] Y. Hu, F. Zou, J. Han, X. Sun, and Y. Wang. "Llm-Tikg: Threat Intelligence Knowledge Graph Construction Utilizing Large Language Model". *SSRN*, (2023). Available at SSRN: https://ssrn.com/abstract=4671345 or http://dx.doi.org/10.2139/ssrn.4671345 (cit. on pp. 14, 15).

[131] Y. Schwartz, L. Benshimol, D. Mimran, Y. Elovici, and A. Shabtai. "LLMCloudHunter: Harnessing LLMs for Automated Extraction of Detection Rules from Cloud-Based CTI". *arXiv preprint arXiv:2407.05194*, (2024) (cit. on pp. 14, 15, 139).

[132] N. Tsinganos and I. Mavridis. "Building and Evaluating an Annotated Corpus for Automated Recognition of Chat-Based Social Engineering Attacks". *Applied Sciences*, 11 (22), (2021), p. 10871 (cit. on p. 16).

[133] Z. Fang, M. Xu, S. Xu, and T. Hu. "A framework for predicting data breach risk: Leveraging dependence to cope with sparsity". *IEEE Transactions on Information Forensics and Security*, 16, (2021), pp. 2186–2201 (cit. on p. 16).

[134] Google. *Google Gemini*. https://gemini.google.com/]. 2023 (cit. on p. 21).

[135] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". *arXiv preprint arXiv:2403.05530*, (2024) (cit. on pp. 21, 139).

[136] OPENAI. *GPT-4o*. https://platform.openai.com/docs/models/gpt-4o]. 2023 (cit. on pp. 21, 69, 139).

[137] Stanford Center for Biomedical Informatics Research. *Protégé*. Version 5.6.1. 2023. url: https://protege.stanford.edu/ (cit. on p. 24).

[138] M. Honnibal, I. Montani, and Explosion AI. *spaCy: Industrial-strength Natural Language Processing*. Version 3.7.2. 2023. url: https://spacy.io/ (cit. on p. 24).

[139] I. Montani and M. Honnibal. *Prodigy: An annotation tool for AI, machine learning and NLP*. Berlin, Germany, 2018. url: https://prodi.gy/ (cit. on pp. 24, 55).

[140] R. Řehůřek and P. Sojka. "Software Framework for Topic Modelling with Large Corpora". *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA. Valletta, Malta, 2010, pp. 45–50. url: https://radimrehurek.com/gensim/ (cit. on p. 24).

[141] A. Zacharis and C. Patsakis. "AiCEF: an AI-assisted cyber exercise content generation framework using named entity recognition". *International Journal of Information Security*, 22 (5), (2023), pp. 1333–1354 (cit. on pp. 27, 31).

[142] A. Zacharis, V. Katos, and C. Patsakis. "Integrating AI-driven threat intelligence and forecasting in the cyber security exercise content generation lifecycle". *International Journal of Information Security*, (2024), pp. 1–20 (cit. on p. 28).

[143] A. Zacharis, R. Gavrila, C. Patsakis, and C. Douligeris. "Optimising AI Models for Intelligence Extraction in the Life-cycle of Cybersecurity Threat Landscape Generation". *Journal of Information Security and Applications*, (2025), X–Y (cit. on p. 29).

[144] A. Zacharis, R. Gavrila, C. Patsakis, and D. Ikonomou. "Ai-assisted cyber security exercise content generation: Modeling a cyber conflict". *2023 15th International Conference on Cyber Conflict: Meeting Reality (CyCon)*. IEEE. 2023, pp. 217–238 (cit. on p. 30).

[145] A. Zacharis, R. Gavrila, and C. Patsakis. *AICEF: An AI-Powered Cyber Exercise Content Generation Framework, BLACKHAT ASIA*. 2023. url: https://www.blackhat.com/asia-23/arsenal/schedule/index.html#aicef-an-ai-powered-cyber-exercise-content-generation-framework-31299 (cit. on p. 31).

[146] R. Gavrila and A. Zacharis. "Advancements in Malware Evasion: Analysis Detection and the Future Role of AI". *Malware: Handbook of Prevention and Detection*. Ed. by D. Gritzalis, K.-K. R. Choo, and C. Patsakis. Cham: Springer Nature Switzerland, 2025, pp. 275–297. isbn: 978-3-031-66245-4. doi: 10.1007/978-3-031-66245-4_12. url: https://doi.org/10.1007/978-3-031-66245-4_12 (cit. on p. 32).

[147] M. Karjalainen, T. Kokkonen, and S. Puuska. "Pedagogical aspects of cyber security exercises". *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE. 2019, pp. 103–108 (cit. on p. 41).

[148] OASIS OPEN. *STIX version 2.1*. https://www.oasis-open.org/standard/stix-version-2-1/. 2021 (cit. on p. 48).

[149] R. MacIntyre. *Penn Treebank tokenizer (sed script source code)*. 1995 (cit. on pp. 55, 104, 137).

[150] R. T. CRAIG. "Generalization of Scott's Index of Intercoder Agreement". *Public Opinion Quarterly*, 45 (2), (Jan. 1981), pp. 260–264. issn: 0033-362X. doi: 10.1086/268657. eprint: https://academic.oup.com/poq/article-pdf/45/2/260/5432504/45-2-260.pdf. url: https://doi.org/10.1086/268657 (cit. on pp. 57, 137).

[151] N. Wilhelmson and T. Svensson. *Handbook for planning, running and evaluating information technology and cyber security exercises*. Försvarshögskolan (FHS), 2011 (cit. on p. 59).

[152] J. Hakala and J. Melnychuk. *Russia's Strategy in Cyberspace*. Tech. rep. NATO STRATCOM COE, 2021. url: https://stratcomcoe.org/cuploads/pfiles/Nato-Cyber-Report_15-06-2021.pdf (cit. on pp. 89, 92).

[153] M. T. I. Center. *Destructive malware targeting Ukrainian organizations*. Tech. rep. 2022. url: https://www.microsoft.com/en-us/security/blog/2022/01/15/destructive-malware-targeting-ukrainian-organizations/ (cit. on p. 92).

[154] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied linear statistical models*. McGraw-hill, 2005 (cit. on p. 103).

[155] P. F. Velleman and D. C. Hoaglin. *Applications, basics, and computing of exploratory data analysis*. Duxbury Press, 1981 (cit. on p. 104).

[156] J. S. Armstrong and F. Collopy. "Error measures for generalizing about forecasting methods: Empirical comparisons". *International journal of forecasting*, 8 (1), (1992), pp. 69–80 (cit. on p. 104).

[157] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation". *Journal of machine Learning research*, 3 (Jan), (2003), pp. 993–1022 (cit. on p. 108).

[158] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora". *Proceedings of the 2009 conference on empirical methods in natural language processing*. 2009, pp. 248–256 (cit. on p. 108).

[159] C. Sievert and K. Shirley. "LDAvis: A method for visualizing and interpreting topics". *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 2014, pp. 63–70 (cit. on p. 108).

[160] M. Röder, A. Both, and A. Hinneburg. "Exploring the space of topic coherence measures". *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015, pp. 399–408 (cit. on p. 108).

[161] J. Cohen. "A coefficient of agreement for nominal scales". *Educational and psychological measurement*, 20 (1), (1960), pp. 37–46 (cit. on p. 137).

[162] ENISA. *ENISA Cybersecurity Threat Taxonomy*. https://www.enisa.europa.eu/topics/cyber-threats/threats-and-trends/enisa-threat-landscape/threat-taxonomy. 2016 (cit. on p. 137).

[163] L. S. Penrose. "The elementary statistics of majority voting". *Journal of the Royal Statistical Society*, 109 (1), (1946), pp. 53–57 (cit. on p. 148).

[164] C. Sammut and G. I. Webb. *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated, 2017 (cit. on p. 148).

Part V

Publications

International Journal of Information Security (2023) 22:1333–1354
https://doi.org/10.1007/s10207-023-00693-z

REGULAR CONTRIBUTION

AiCEF: an AI-assisted cyber exercise content generation framework using named entity recognition

Alexandros Zacharis[1] · **Constantinos Patsakis**[2,3]

Published online: 19 April 2023
© The Author(s) 2023

Abstract
Content generation that is both relevant and up to date with the current threats of the target audience is a critical element in the success of any cyber security exercise (CSE). Through this work, we explore the results of applying machine learning techniques to unstructured information sources to generate structured CSE content. The corpus of our work is a large dataset of publicly available cyber security articles that have been used to predict future threats and to form the skeleton for new exercise scenarios. Machine learning techniques, like named entity recognition and topic extraction, have been utilised to structure the information based on a novel ontology we developed, named Cyber Exercise Scenario Ontology (CESO). Moreover, we used clustering with outliers to classify the generated extracted data into objects of our ontology. Graph comparison methodologies were used to match generated scenario fragments to known threat actors' tactics and help enrich the proposed scenario accordingly with the help of synthetic text generators. CESO has also been chosen as the prominent way to express both fragments and the final proposed scenario content by our AI-assisted Cyber Exercise Framework. Our methodology was assessed by providing a set of generated scenarios for evaluation to a group of experts to be used as part of a real-world awareness tabletop exercise.

Keywords Cyber security exercise scenario · Artificial intelligence · Cyber security exercise ontology

1 Introduction

Cyber security exercises (CSEs) are increasingly becoming an integral part of the cybersecurity training landscape [20], providing a hands-on experience to personnel of both public and private organisations worldwide. A CSE, as described in the ISO Guidelines for Exercises [18], is "*a process to train for, assess, practice, and improve performance in an organisation*". ENISA defines a CSE as "*a planned event during which an organisation simulates cyber-attacks or information security incidents or other types of disruptions to test the organisation's cyber capabilities, from being able to detect a security incident to the ability to respond appropriately and minimise any related impact.*" [7].

Constantinos Patsakis
kpatsak@unipi.gr

Alexandros Zacharis
alexandros.zacharis@enisa.europa.eu

[1] European Union Agency for Cybersecurity (ENISA), Athens, Greece

[2] Department of Informatics, University of Piraeus, Karaoli & Dimitriou 80, 18534 Piraeus, Greece

[3] Athena Research Center, Marousi, Greece

1.1 Problem setting and objectives

The creation of CSE content is a painstaking process that requires a deep understanding of the current threat landscape and the historical threats and incidents faced by an entity and the corresponding sector. Furthermore, training employees with simulated incidents is the closest method to testing the preparedness and effectiveness of measures and procedures set in place. Creating a relevant and dynamic content for developing CSE scenarios requires expertise and resources often lacking among most organisations.

The main objective of our work is automating the generation of structured CSE scenarios based on a pool of unstructured information with little experience in scenario building expected from the Exercise Planner (EP).

The standard method for preparing an exercise scenario [18] lays down three layers, namely events, incidents, and injects. After developing a scenario, an organisation must ensure that it contains only necessary information. More-

over, it must be designed to test participants' capabilities in a stressful environment. Events, at the first level, provide the general description of an exercise scenario. Depending on previously decided objectives and aims, the number of events can differ from one exercise to another. Each event would have a specific set of consequences at the second level. These consequences are called incidents. An event can have multiple consequences, which can affect each other. On the third level, injects facilitate the communication of events and incidents to the exercise participants. An ideal inject would provide exercise information and problems to be solved. At the same time, it would indirectly force participants to act on those consequences and make decisions.

The proposed scenarios should satisfy the specifications provided by the EP. Such specifications can be the training topics and objectives, the sector to focus on or specific threats of interest that are currently or will be trending in the future. For simplicity, in what follows, when referring to sectors, we will refer to the ones of Directive (EU) 2022/2555 of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) [12]; however, any other such classification can be used. More specifically, the objectives can be summarised as follows:

1. Create an ML-powered Exercise Generation Framework that would:
   (a) Generate structured exercise scenarios that reflect a sector's current or future threat landscape, including potential threat actors and the corresponding techniques, tactics, and procedures (TTPs).
   (b) Generate scripted events and incidents that could materialise in the context of a real attack against an organisation belonging to any NIS 2 defined Sector
   (c) Identify and describe artefacts that could accompany the exercise scenarios as potential injects
2. The generated scenarios should be expressed in a structured way or format, following an Ontology. The generated outputs should be both machine and human-readable.
3. The proposed methodology and tools created should provide qualitative and quantitative added value in CSE development and cyber-awareness by measuring the following Key Performance Indicators (KPIs):
   (a) Improve the speed in CSE generation (quantitative)
   (b) Improve quality in CSE generation (qualitative) for inexperienced EPs
   (c) Improve the relevance of proposed CSE scenarios to the current threat landscape (qualitative)

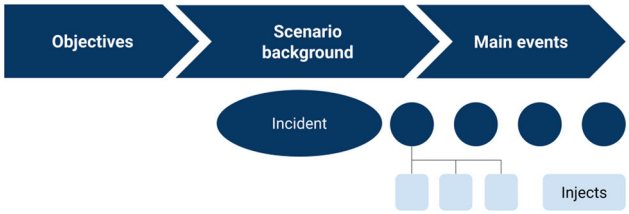


Fig. 1 Cyber exercise structure according to ISO 22398:2013 [18]

The use of case studies will help measure the results of the KPIs set by comparing the traditional exercise generation methods and tools versus the proposed ones through an evaluation provided by an Ad-hoc Cyber Awareness Expert Group [1] [2] that will peer review the outputs of the aforementioned methodology.

1.2 Main contributions

The contribution of this work is twofold. Initially, we can identify future cyber-attack trends in a specific sector and propose customised awareness training topics by clustering them accordingly. Then, we automate the process of generating the corresponding content for cyber awareness exercises with machine learning (ML).

Our proposed methodology, which a set of tools will accompany, allows an inexperienced EP to fully structure CSE scenarios from free text following our proposed Cyber Exercise Scenario Ontology (CESO). The exercise structure will follow the traditional Scenario—Events -Incidents—Injects tree structure ISO 22398:2013 [18] as depicted in Fig. 1. Additional cyber exercise content will be generated to complement the scenario and proposals for the fittest of a set of given training topics to better prepare an organisation for an imminent cyber crisis.

Through our work, we fill the gap in the lack of expertise by the average cyber security expert that acts as an Exercise Planner and provides the tools and the methodology to design CSE scenarios in an easy, automated, and structured way. To achieve this, we combine the power of machine learning (ML) and, more specifically, named entity recognition (NER) with a set of novel Cyber Exercise Scenario Ontology (CESO) and CSE scenario generation framework dubbed AiCEF. Finally, an evaluation methodology and its results are presented along with ideas for future work.

[1] https://www.enisa.europa.eu/topics/cybersecurity-education/ad-hoc-working-group-awareness-raising.

[2] "The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union Agency for Cybersecurity (ENISA). Neither the European Union institutions nor any person acting on their behalf may be held responsible for any use that may be made of the information contained therein.

2 Related work

CSEs, also known as Cyber Defense Exercises (CDX), have been considered an effective way to implement an engaging security awareness training [13, 42] experience. CSEs have been characterised as a highly effective method to provide an ultimate learning experience [3], helping individuals or teams of varying expertise improve a range of skills related to information security. Furthermore, via exercising, organisations can uncover gaps in security policies, procedures and resources [9, 16] leading to awareness training, tools and policy improvements.

Previous work in the CSE domain [40] has highlighted the use of cyber defence competitions or live-attack exercises as a very effective way of teaching information security [10, 19], helping teams design, implement, manage and defend a network of computers [1, 6, 7, 30, 31]. Vigna [46] and Mink [27] further support these findings.

Further research was conducted on cyber defence competitions [36, 49] and the most suited architecture [41] and tools and techniques to be used to create an active learning experience were described by Green et al. [15]. Patriciu and Furtuna [34] presented several steps and guidelines to be followed when designing a CSE. White [48] introduced a different approach to such live CSEs, presenting lessons learned and providing suggestions to help organisations run their own exercises. Other works in the literature examined how to run CSEs, using a service provider model [26].

CSEs can be used as a tool to generate scientifically valuable datasets for future security research [38, 43] and help uncover hidden risk from weak Security policies and/or procedures [37]. CSEs can even be used to measure performance against specific standards [11] or team effectiveness based on behavioural assessment techniques [14]. Moreover, experiments using various platforms like the RINSE simulator [23] or using realistic inter-domain routing experiment platform [22], for the rendering of network behaviour.

Focusing further on the human aspect, Job Performance Modelling (JPM), using vignettes for improving cybersecurity talent management through cyber defence competition design, was described by Tobey [44].

A successful CSE counts heavily on the use of a robust scenario. Exercise scenarios must describe worst-case scenarios that participants can relate to and are realistic enough to trigger seamless engagement. Intuitive scenarios can be a powerful tool that can predict future states or situations [3], [13]; incorporating issues to be resolved, interactions and consequences [15], [14] leading to a constructive training experience.

An exercise's scenario is a sequential, narrative account of a hypothetical incident that provides the catalyst for the exercise and is intended to introduce situations that will inspire responses and thus allow demonstration of the exercise objectives [41]. In the context of CSEs, a scenario defines the training environment that will lead participants towards fulfilling the exercise objectives [21] set. The cyber security problem described in a scenario itself portrays a structured representation named Master Scenario Events List (MSEL), which serves as the script for the execution of an exercise [41]. CSE scenarios formats can vary [35], but two are the most prevalent:

- Outlined scenarios: Provide a general summary of the impact of an event on assets. [39]
- Detailed scenarios: Contain exhaustive information sequentially describing the event's impact on specific services or sections of an organisation, along with a timeline for restoring key functions. [17]

Recent trends in attack recognition utilise AI, ML, and NLP tools and techniques to empower their efficiency. However, there needs to be a more dedicated methodology focusing on CSE scenario generation. There is a need for a methodologically built and annotated CE corpus that could train multiple algorithms for Cyber Exercise elements. Such a corpus should focus on the syntactic and semantic characteristics of the cyber exercise components and broaden our understanding of the malicious patterns used in cyber incidents that can be reused for CSE material. A similar approach to the one used in building and evaluating an annotated Corpus for automated Recognition attacks has been utilised [45], only this time to extract CSE relevant objects.

Following Cyber Security related ontology creation examples [33], ontology -based scenario modelling for CSEs have already been proposed [47]. Still, an ontology that is truly compatible with Machine Learning algorithms is missing and will be the focus of our work.

3 Cyber exercise scenario ontology (CESO)

Our work so far highlighted the need for a common CSE scenario ontology for translating the various parts of an exercise while keeping a close link to popular already used ontologies for cyber incident representations. The analysis of the domain revealed many taxonomies for different areas of the cybersecurity domain (types of attacks, vulnerabilities, sectors, harm) but those needed to be linked together in a model that allows for an EP to represent a CSE accurately.

To build our ontology, the following questions were raised:

1. What is the scope of the ontology?
2. Should we consider reusing existing ontologies or taxonomies?
3. What are the important terms in the ontology?

The scope of the ontology was determined by asking competency questions to experienced EPs that helped us identify the most important terms. A key priority was interoperability to ensure that the proposed ontology could be integrated with existing tools and frameworks. Moreover, the proposed ontology should describe a cyber security incident using popular cyber security frameworks. Finally, the ontology should be easily implementable using extraction via named entity recognition (NER) to allow the easy ingestion of online content.

We also used the domain expert's knowledge to identify prominent existing ontologies and ways to reuse them. The steps followed were:

1. Define the scope of our ontology,
2. Identify other ontologies or taxonomies that can be used/reused,
3. Define the main concepts and the relationships between them,
4. Define the properties of the concepts,
5. Implement the ontology.

3.1 Scope

The scope of the defined model was to target an efficient and robust way of representing cyber incidents in the context of a CSE. After all, a CSE is a collection of simulated incidents provided to players in an orchestrated way to achieve the exercise's objectives.

The exercise ontology presented is incident-centric, focusing on using a bottom-up approach that allows us to identify and describe incidents first so we can group them into Events and then cover the full generation of CSE scenarios that fit the high-level objectives set.

The first building blocks, incidents, are assigned injects and mitigation actions that match the expected scope of the scenario. Injection timing is configured on the attribute level of each object. As we build toward the higher level of the exercise, the scenario is formed. The selected format should allow for the scenario's portability to various existing tools (ex. MISP[3]) and support a decentralised type of CSE execution.

3.2 Ontologies/taxonomies to be (re)used

A set of existing ontologies, taxonomies, frameworks, standards, and formats have been explored with relevance to cyber security and a focus on the representations of the key element of CSEs from the point of their very building blocks being the incidents to be simulated. Our research concluded that a combination of the following would provide the necessary means: ISO 22398 [18], MITRE ATT&CK [29] and Cyber Kill Chain [24], MITRE CVE [28], and STIX 2.1 [32].

We chose STIX 2.1 as the basis for our ontology, which defines a taxonomy of cyber threat intelligence to be extended to cover our need to describe CSE scenarios. ISO 22398 best describes the structure of the cyber exercise components and was used to help us repurpose STIX 2.1 to cover our scope. The STIX 2.1 model describes an adversary and adversary activities in appropriate data structures by default. STIX Domain Objects cover: Threat Actor; Malware; Tools; Campaign; Intrusion Sets, and Attack Patterns (referencing the Common Attack Pattern Enumeration and Classification taxonomy, CAPEC), perfectly covering what is called incident and injects in the CSE nomenclature. STIX 2.1 supports by default the MITRE ATT&CK, MITRE CVE and Cyber Kill Chain frameworks, helping us achieve our goal for maximum interoperability. Moreover, intelligence (CTI) in a consistent and machine-readable manner, allowing security communities to understand better what computer-based attacks they are most likely to face and anticipate and/or respond to those attacks faster and more effectively.

This helps us build on top of these communities to reuse existing tools and share CSE scenarios represented in the very same format.

3.3 Scenario augmented model

Based on the bottom-up approach, a scenario augmented model (SAM) is proposed in two layers that cover both the informational and operational aspects with the same objects but utilise different attributes.

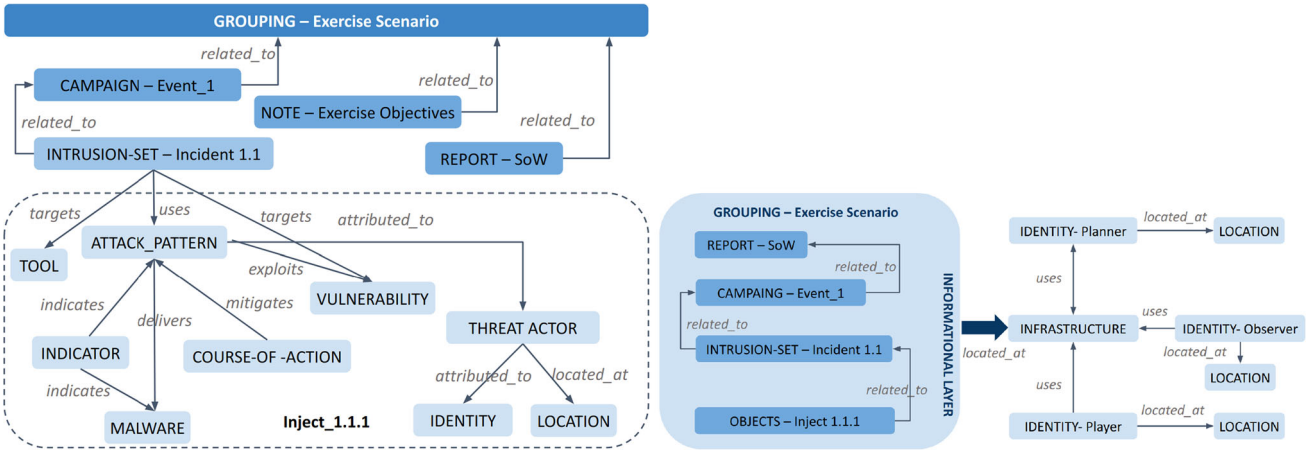
The informational layer covers the context and main attributes of scenarios. Figure 2a describes the key relationships in the informational layer.

The whole exercise is grouped using the grouping objects. The object holds information related to the exercise's name, description and scenario. All Events, Objectives and their matching objects (Campaign, Note, Report) are related to the Exercise Scenario along with the matching "State of the World (SoW)". The SoW includes details such as the status of various simulated systems and networks, the simulated geopolitical landscape, and any simulated incidents or events that have taken place.

One or more Incidents (Intrusion Set) can be related to Events. From there, various objects with interlinked dependencies form the Inject in a Course of Actions Instance that refers to all related objects of an Attack Pattern.

An Inject can contain the following objects: Attack Pattern, Tool, Vulnerability, Indicator, Malware Threat Actor (who is attributed and Identity and is located at a Location) and a Course Of Action. Injects do not have to be related to an Event or Incident. Such examples are the STARTEX or

[3] https://www.misp-project.org/.



(a) Informational Layer of CESO. (b) Operational Layer of CESO.

Fig. 2 The informational (left) and operational (right) layers of CESO

Table 1 CSE components to STIX 2.1 objects mapping

CSE Component	STIX 2.1 Object
Exercise details and scenario background	Grouping
Objectives	Note
Events	Campaign
Incidents	Intrusion Set
State of the World (SoW)	Report
Injects	Tool, vulnerability, threat-actor, identity, location, attack pattern, malware, indicator, course-of-action, observed data, malware analysis, report
Exercise Platform	Infrastructure
Exercise Participant	Identity, location

ENDEX,[4] which can be represented only with a Course of Action object but are directly related to the Scenario.

The Scenario Operational Layer describes an exercise scenario's execution flow, mainly dealing with injects delivery to the intended recipients. There are two major interrelated parts: (1) the events/injects, which describe the detailed activities of the scenario and expected actions from the participants, and (2) the Participants.

The whole scenario, including Events, Incidents, and Injects, is stored in an Infrastructure object, representing the Exercise Platform. This platform is used by EPs (Identity) to design and conduct the exercise, Observers, and Players to interact with the Scenario. All Participants are located in the same or different Locations. The Operational Layer is illustrated in Fig. 2b.

3.4 Implementing the ontology

Keeping the structure of CSE intact, the following STIX 2.1 Objects have been repurposed to fulfil our goal to represent the main CSE components successfully covering SAM along with matching relationships (Table 1).

Objects STIX 2.1 defined objects as per specifications.

Relationships All relationships are implemented as per STIX 2.1 relationship object specifications. The relationships in Table 2 (representing the edges of the graph) have been identified between key objects, but more can be used.

Object Extension STIX 2.1 objects, extended with additional attributes/properties to cover the needs of CESO, as shown in Table 3.

4 Automated generation of cybersecurity exercise scenarios

To create the envisioned ML-powered Exercise Generation Framework, we opted to use Python and develop a set of tools that would perform a set of individual tasks in the form of steps, which would help an EP, regardless of his/her experience, to create timely and targeted CSEs. Conceptually, we split the process into six steps. Namely, data collection, data

[4] The [START]ing and [END]ing [EX]ercise injects.

Table 2 Relationships matrix

Source Object	Destination object	Relationship
Campaign	Grouping	related_to
Note	Grouping	
Report	Grouping	
Intrusion-Set	Campaign	
Course-Of-Action	Grouping	
Intrusion-Set	Tool	targets
Intrusion-Set	Vulnerability	
Intrusion-Set	Attack-Pattern	uses
Identity	Infrastructure	
Attack-Pattern	Threat-Actor	attributed_to
Threat-Actor	Identity	
Identity	Location	located_at
Attack-Pattern	Malware	delivers
Attack-Pattern	Indicator	indicates
Indicator	Malware	
Attack-Pattern	Vulnerability	exploits
Course-Of-Action	Attack-Pattern	mitigates
Course-Of-Action	Vulnerability	

processing and mapping, trend prediction, incident generation, enhancement, and storyline generation. The proof of concept framework we developed is AiCEF, and its general outline is illustrated in Fig. 4.

Its main components that are relevant to the work presented in this paper are the following:

- CESO: The Cyber Exercise Scenario Ontology used to describe the various components of a CSE
- AiCEF: The Cyber Exercise Framework used to model CSEs based on CESO with the use of Machine Learning
- MLCESO: The ML models trained to parse text and extract objects based on CESO
- IncGen: The incident generation module that models a CSE incident from the MLCESO extracted objects based on CESO
- CEGen: The cyber exercise generation module that models a CSE from the MLCESO extracted objects based on CESO
- KDb: A knowledge pool of incidents stored in a database. Extracted objects and other characteristics, including the STIX 2.1 blob, are stored in the database

To facilitate the reader, we map these components in a timeline diagram, see Fig. 3. This way, one can get a quick grasp of the role of each component in the flow and navigate through the rest of the sections understanding how these pieces fit in the greater picture.

The modular approach of AiCEF allows for customisation and local refinements and enables more interoperability. In the following paragraphs, we detail each component and then present the main steps to generate a concrete CSE scenario using AiCEF modules, providing some examples.

4.1 Machine learning to CESO (MLCESO)

The most important step in our methodology is the creation of the ML pipeline that will parse free text and extract objects in CESO, as defined in the previous section. To do so, we need to train our ML following a well-structured methodology consisting of three phases: Corpus Building, Corpus Annotation, and Corpus Evaluation using NER, which we detail below.

4.1.1 Corpus building

As shown in Table 4, four Incident Sources have been identified as the initial input to our corpus. All these websites cover a wide variety of cyber security incidents in article format that date many years. For simplicity, in this work, we collected incidents from 2020–01 till 2022–03, which accounts for 2000 articles. All relevant articles were collected through automated web scraping.

Then, the raw text was processed using Natural Language Processing (NLP) techniques to form a reduced Incidents Corpus (IC). Initially, all text was converted to the UTF-

Table 3 Objects extension matrix

Object	Attribute added	Type	Description
Course-of-action	Difficulty (optional)	Integer	An integer from 1 to 5 (1 being easy and five being hard) declaring how difficult a course of action is evaluated to be executed by the player to resolve an incident
Grouping	Scenario (mandatory)	String	A description that provides more details and context about the exercise scenario
Identity	Recipient_Group (optional)	String	The name of the recipient group in which players are split to receive different injects

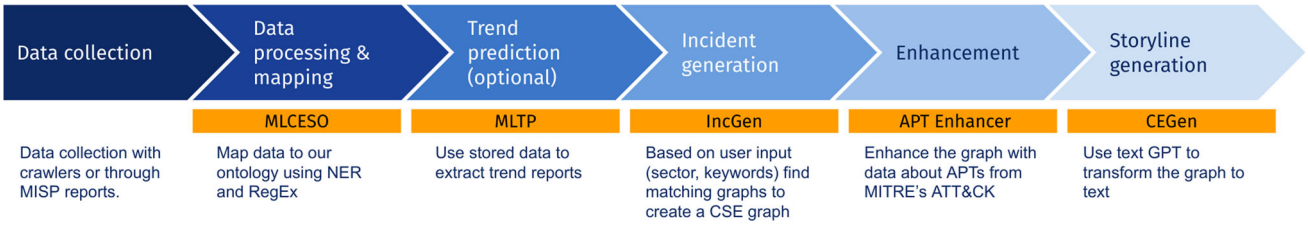


Fig. 3 Process flow and the corresponding modules of AiCEF

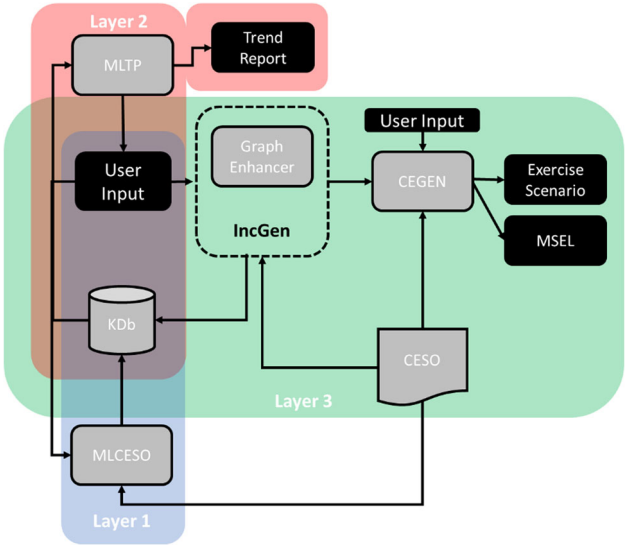


Fig. 4 High-level overview of AiCEF

Table 4 Corpus collection count

Webpage	Articles
`bleepingcomputer.com`	1000
`securityaffairs.co`	150
`zdnet.com`	350
`databreaches.net`	500
Total	2000

8 encoding scheme. Using dictionaries and the Textblob library,[5] we performed spelling corrections and removed special characters. Empty lines, specific stopwords and specific punctuation marks were removed using traditional NLP libraries like NLTK[6] and spaCy. [7] Moreover, all HTML or other programming codes, URLs, and paths were removed. Any illegal characters were also stripped, and all text was transformed to lowercase.

The standard Penn Treebank [25] tokenisation rules were utilised for sentence tokenisation, and finally, standardisation processes were applied to tune the Incidents Text to facilitate annotation. At the end of this step, a corpus composed of Incidents was formed. As discussed, the corpus, from now on referred to as IC, contains 2000 cyber security articles. This accounts for 35.745 sentences containing 819.690 words leading to a vocabulary of 24,594 terms. An example of a corpus line ready for annotation is the following:

```
{"text":"revil sodinokibi
   ransomware targets chinese
   users with dhl spam"}
```

4.1.2 Corpus annotation

Following the CESO ontology, a simple model was developed comprising six steps to represent the annotation task. Entities and interconnections were formally described to align the efforts of converting words to tags in an Annotators Reference Document (ARD). This file, along with the corpus guidelines and CESO ontology, was given to the annotators to perform the annotation task using Prodigy.[8] After completing the annotation, an inter-annotator agreement assessment took place using Cohen's Kappa metric, and the gold standard version of the IC was finally produced.

Our annotation methodology consists of the following steps.

Step 1: Setting the Annotation Objectives The main annotation objective was to create the appropriate semantic target to facilitate IC recognition by assigning the correct tag to in-context words in a sentence. Labelling all related words or sequences of words or text spans in the Cyber Incident context was crucial to perform efficient NER or text classification later. Each word or text span was labelled with a type identifier (tag) drawn from a vocabulary created based on the CESO ontology. It indicated what various terms denote in the context of a Cyber Incident and how they interconnect between them.

Our objective is to identify keywords, syntax, and semantic characteristics to detect i) threat actors, ii) cyber security

[5] https://github.com/sloria/TextBlob.

[6] https://www.nltk.org/.

[7] https://spacy.io/.

[8] https://prodi.gy.

Table 5 Annotation tags per category

Category	Tag	Link to CESO and STIX 2.1
Attacker	ATTACKER_TYPE	Threat Actor Attribute
	ATTACKER_NAME	Threat Actor Attribute, Identity
	ATTACKER_ORIGIN	Location
Attack	MALWARE_TYPE	Malware Attribute
	MALWARE_NAME	Malware Attribute
	ATTACK_TYPE (TECHNIQUE)	Attack Pattern
	VULNERABILITY	Vulnerability
Victim	SECTOR	Identity Attribute, Scenario
	ASSETS	Threat Actor Attribute
	TECHNOLOGY	Tool

Table 6 Annotation tags per category example

Category	Tag	Annotator A-Tags	Annotator B: Tags
TEXT	qbot malware dropped via context aware phishing campaign infects the energy sector.		
	russian hacking group claims 1000 windows machines compromised.		
Attacker	ATTACKER_TYPE	Hacking group	
	ATTACKER_NAME		
	ATTACKER_ORIGIN	Russian	Russian
Attack	MALWARE_TYPE	malware	malware
	MALWARE_NAME	qbot	qbot
	ATTACK_TYPE (TECHNIQUE)	phishing campaign	phishing
	VULNERABILITY		
Victim	SECTOR	energy	energy
	ASSETS	windows machines	windows machines
	TECHNOLOGY	windows	windows

incidents, and iii) victim characteristics, to tag them accordingly.

Step 2: Specifications Definition A concrete representation of the Annotation model to be used is created based on CESO.

An abstract model that practically represented the annotation objectives was defined. A three-category classification (Attacker, Attack, Victim) was introduced as the basis of this abstract model for identifying cyber-incident related terms in the text analysed. The category *other* represents all remaining words out of context.

Our model M consists of a vocabulary of terms T, the relations between these terms R, and their interpretation I. Thus, our model can be represented as $M =< T, R, I >$ where:

- T={CESO, Attacker, Attack, Victim, Other}
- R={CESO:: = Attacker|Attack|Victim|Other}
- I={Attacker= "list of attacker related terms in vocabulary", Attack ="list of Cyber Security Incident or Attack terms in vocabulary",
  Victim = {"list of victim-related terms in vocabulary"}
  Other = {"Other terms not related to the attacks"}}

Step 3: Annotator Reference To help annotators in element identification and element association with the appropriate tags, we provided them with documentation containing the tags in Table 5, which have been identified and mapped accordingly.

Step 4: Annotation Task the annotation process is performed

The annotation task aimed to label the words of the IC corpus based on their semantic and syntactic characteristics. Two cybersecurity experts were assigned to label the words based on their semantic characteristics. By annotating the semantic characteristics of the words with Prodigy, the context of each sentence was translated into CESO. Table 6 presents the annotation in action through some examples.

Table 7 Consistency matrix

Annotator		B				Total
	Category	Attacker	Attack	Victim	Other	
A	Attacker	397	10	4	24	435
	Attack	13	1722	8	9	1752
	Victim	10	2	926	15	953
	Other	16	10	12	21, 416	21,454
Total		436	1744	950	21,464	24,594

Table 8 AI models' scores

Category	Tag	Presicion	Recall	F1
Attacker	ATTACKER_TYPE	100.00	83.33	90.11
	ATTACKER_NAME	95.29	87.10	91.01
	ATTACKER_ORIGIN	Used Native Spacy LOC tag (no training)		
Attacker	MALWARE_TYPE	80.56	76.32	78.38
	MALWARE_NAME	95.29	87.10	91.01
	ATTACK_TYPE (TECHNIQUE)	88.60	87.07	87.83
	VULNERABILITY	87.50	84.00	85.71
Victim	SECTOR	85.84	84.07	84.95
	ASSETS	87.02	89.06	88.03
	TECHNOLOGY	87.60	89.93	88.70

Step 5-Golden Standard Creation: the final version of the annotated Incident corpus is generated.

The inter-annotator agreement (IAA) was validated using Cohen's Kappa [8]. The formula used is defined as follows:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

where p_0 expresses the relative observed agreement, and p_e is the hypothetical probability of chance agreement.

The produced IC corpus has $N = 24594$ terms and $m = 4$ categories, and both annotators (A and B) agreed for the Attacker category 397 times, for the Attack category 1722 times, for the Victim 932 times and for the Irrelevant 21416.

Table 7 shows the contingency matrix where each x_{ij} represents the multitude of terms that annotator A classified in category i, but Annotator B is classified in category j, with $i, j \in \{1, 2, 3, 4\}$. The proportions on the diagonal (x_{ii}) represent the proportion of terms in each category for which the two annotators agreed on the assignment.

The observed agreement p_o is:

$$p_o = \frac{397 + 1722 + 926 + 21416}{24594} = 0, 996$$

and the expected change agreement; thus, the proportion of terms which would be expected to agree by chance is:

$$p_e = \frac{\frac{436\times435}{24594} + \frac{1744\times1752}{24594} + \frac{950\times953}{24594} + \frac{21464\times21454}{24594}}{24594}$$

$$= 0, 768(76, 8\%)$$

so, according to Eq. 1 the Cohen's Kappa is $k = \frac{p_0 - p_e}{1 - p_e} = \frac{0,228}{0,232} = 0, 98$. Thus, based the Cohen's kappa value of 0.98, we can safely conclude [50] that the level of agreement for the corpus annotation task was almost perfect.

4.1.3 Training and evaluation using NER

The following methodology has been used to train and evaluate our Named Entity Recognition (NER) agent.

1. **Preprocessing** The corpus has already been annotated, with each line of the corpus stored as a list of token-tag pairs. Each token was represented by a word embedding using the pre-trained English language model of the spaCy NLP library.
2. **Build a model using spaCy**
3. **Training** Training was conducted in spaCy by specifying a loss function to measure the prediction error and a batch-wise gradient descent algorithm for optimisation. One NER model was created per object. To improve accuracy, several iterations were conducted by expanding the annotation and retraining the model until an F1 Score of $\approx$ %80% was reached.
   One NER model was trained per object as presented in Table 8.

Table 9 AI models' scores vs reviewers evaluation. H: Hit, P: Partial, M: Miss

Category	TAG	Reviewer 1			Reviewer 2			Average			F1
	SCORE TYPE	H	P	M	H	P	M	H	P	M	
Attacker	ATTACKER_TYPE	90	5	5	90	5	5	90	5	5	90.11
	ATTACKER_NAME	63	20	17	59	24	17	61	22	17	91.01
	ATTACKER_ORIGIN	70	25	15	65	30	15	67.5	17.5	15	98
	MALWARE_TYPE	82	11	7	80	12	8	81	11.5	7.5	78.38
Attack	MALWARE_NAME	72	14	24	71	14	25	71.5	14	24.5	91.01
	ATTACK_TYPE (TECHNIQUE)	84	11	5	86	9	5	85	10	5	87.83
	VULNERABILITY	75	10	15	79	10	15	77	10	15	85.71
Victim	SECTOR	84	16	0	86	13	1	85	14.5	0.5	84.95
	ASSETS	90	7	3	90	8	2	90	7.5	2.5	88.03
	TECHNOLOGY	90	8	2	86	12	2	88	10	2	88.70

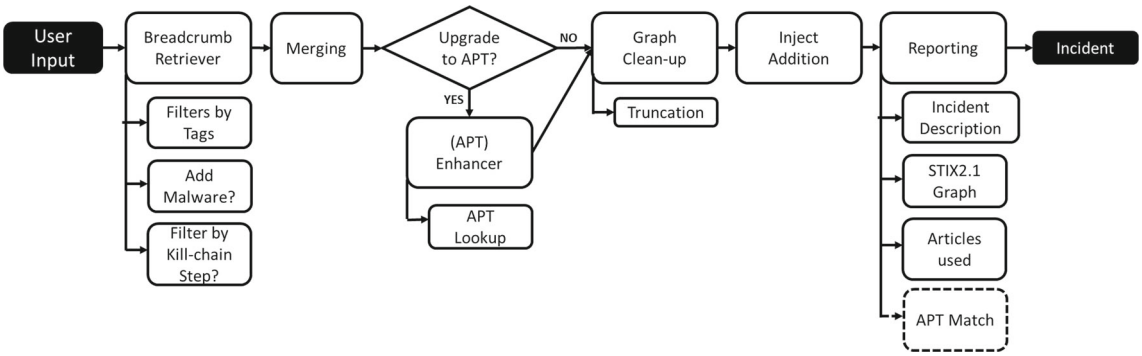


Fig. 5 The workflow of IncGen

4. **Evaluation** The performance assessment of the model was conducted by applying the model to the preprocessed validation data.

While the results seem satisfactory, one can achieve further performance improvements in some tags.

We made an extra evaluation step with two experts against a set of 100 articles not used before in the training or evaluation steps. The aim was to evaluate the models against the selected tags empirically. The two reviewers have scored the NER accuracy per tag as presented in Table 9:

- **HIT** The tag was correctly assigned or not.
- **PARTIAL** The tag was correctly assigned or not, but not for all values
- **MISS** The tag was either assigned wrongly or was not assigned at all when it should

The following findings should be highlighted:

1. The hit rate of four (4) NER models has been identified as very weak, with an abnormal difference from the F1 score identified in the previous step.
2. Names of Attackers or Malware can be a very vague topic to tackle using NER.
3. The Attacker's Origin cannot be properly identified with the use of the out-of-the-box SpaCy LOC NER model. Locations are identified but can be related to the victim or are irrelevant to the attacker's origin.
4. The vulnerability NER model misses the correct formatting of CVE. This issue can be solved using a regex that accurately detects CVE in the text in combination with the model generated.

4.2 Incident generation and enhancement (IncGen)

Incident creation is the most important step of the scenario generation procedure and consists of several steps to achieve maximum customisation (Fig. 5). All of the steps can be automated, generating a variety of Incidents from which a Planner can choose to fit most.

The EP can choose to provide specific text or articles for conversion to Incidents or rely on a dynamic generation based on filtering parameters and a search of the existing database. Incidents can be enhanced with activity simulating TTPs of known APT actors.

Table 10 Knowledge DB content per source

Source	Count
bleepingcomputer.com	1368
securityaffairs.co	169
zdnet.com	495
databreaches.net	938
Total	2970

To generate scenarios, a set of texts was used as a baseline and parsed to map with CESO for processing. The sources in Table 10 were utilised to create the knowledge database (KDb). To ensure relevance, a threshold system *maturity* was introduced to evaluate the maturity of the parsed articles and NER extracted tags. The scoring system, ranging from 0 to 185, is shown in Algorithm 1. In the implementation, a threshold of 50 was set to consider a text relevant for representing a standalone incident in AiCEF.

Algorithm 1 Compute the maturity of a parsed text.

Require: Set of Tags T
$maturity \leftarrow 0$
if $Attacker_Type \in T$ OR $Attack_Type \in T$ **then**
  $maturity \leftarrow 50$
  if $Vulnerability \in T$ **then** $maturity \leftarrow maturity + 15$
  else $maturity \leftarrow maturity - 10$
  end if
  if $Malware \in T$ **then** $maturity \leftarrow maturity + 15$
  else $maturity \leftarrow maturity - 10$
  end if
  if $Attack_Type \in T$ **then**
    $maturity \leftarrow maturity + 15$
    if $Attacker_Type \in T$ **then**
      $maturity \leftarrow maturity + 50$
      if $Technology \in T$ **then** $maturity \leftarrow maturity + 10$
      end if
      if $Sector \in T$ **then** $maturity \leftarrow maturity + 10$
      end if
      if $Assets \in T$ **then** $maturity \leftarrow maturity + 10$
      end if
      if $Attackers_Origin \in T$ **then** $maturity \leftarrow maturity +$ 10
      end if
    end if
  end if
end if
return $maturity$

Two types of enhancements were applied to improve the automatically NER exported tags, namely Regular Expression (REGEX), which is a sequence of characters that defines a search pattern, and Hard-coded groups of Strings. Thus, the following tags have been further enhanced:

- Attackers Name: NER + Hardcoded Groups of Strings from MITRE APT list [29]
- Attackers Origin: No NER, Hardcoded Groups of Strings,
- Malware Name: NER + Hardcoded Groups of Strings from MITRE APT list,
- Technique: NER + Hardcoded Groups of Strings from MITRE APT list,
- Vulnerability: NER + CVE REGEX.

The above enhancements greatly improved the tag detection rates, achieving almost 99% in the Vulnerability tag. Moreover, based on the analysis of the most prominent extracted tags, the tag groups of Table 11 were assigned to the training topics meta tag to help categorise text for later use in an exercise scenario-building process. An output report and visualisation (using stixview[9] library) of IncGen utilising the improved MLCESO tag detection can be seen in Fig. 6.

4.3 APT enhancer

To simulate the activity of APT groups, a STIX 2.1 structure was created for each actor using the Groups from MITRE. Attributes and TTPs were automatically extracted to populate the database, generating a STIX 2.1 graph for comparison and enhancement purposes. During incident enhancement, the extracted graph is compared to known APT actors and the most similar is proposed for enhancement. The similarity score, based on a set of weighted properties and ranging from 0 to 100, is calculated using the STIX 2.1 Python API. In AiCEF, the EP can completely or partially merge the draft incident graph with that of known APT actors.

4.4 Storyline text generation

The Storyline Text Generator (STG) creates synthetic text based on predefined input. Using a Python text generator and Generative Pre-trained Transformer 2 (GPT-2),[10] AI large-scale unsupervised language model, which can create coherent paragraphs of text from small pieces of text input.

4.5 Trend prediction module (MLTP)

The trend prediction module provides EP with valuable information by analysing the KDb and extracting trends based on predetermined training objectives to generate a trend report. The MLTP process consists of three steps:

1. Receiving input such as Filter Tags
2. Extracting incident statistics based on specified sector and Training Objective

[9] https://github.com/traut/stixview.

[10] https://openai.com/blog/better-language-models/.

Table 11 Training topics

Training topic	Tags
INCIDENT HANDLING	MALWARE, RANSOMWARE, APT, CYBER, ATTACK, WEBSITE, HACKER, EXPLOIT, ZERO-DAY
GDPR	PRIVACY, DATA LEAKAGE, PERSONAL DATA, EXFILTRATION, CLOUD, SENSITIVE DATA, DATA, GOOGLE DRIVE, AWS, MEDICAL DATA, PASSPORT
CYBER HYGIENE	PASSWORD, ACCOUNT, USERNAME, LOGIN, ACCOUNTS, FILES, CREDENTIALS
PHISHING and SOCIAL ENGINEERING	PHISHING, SCAM, FRAUD, VISHING, IMPERSONATION, BEC, EMAIL, GMAIL
SOCIAL MEDIA	FACEBOOK, TWITTER, LINKEDIN, META, INSTAGRAM
BYOD	MOBILE, ANDROID, IOS, LAPTOP, IOT, GOOGLE PLAY

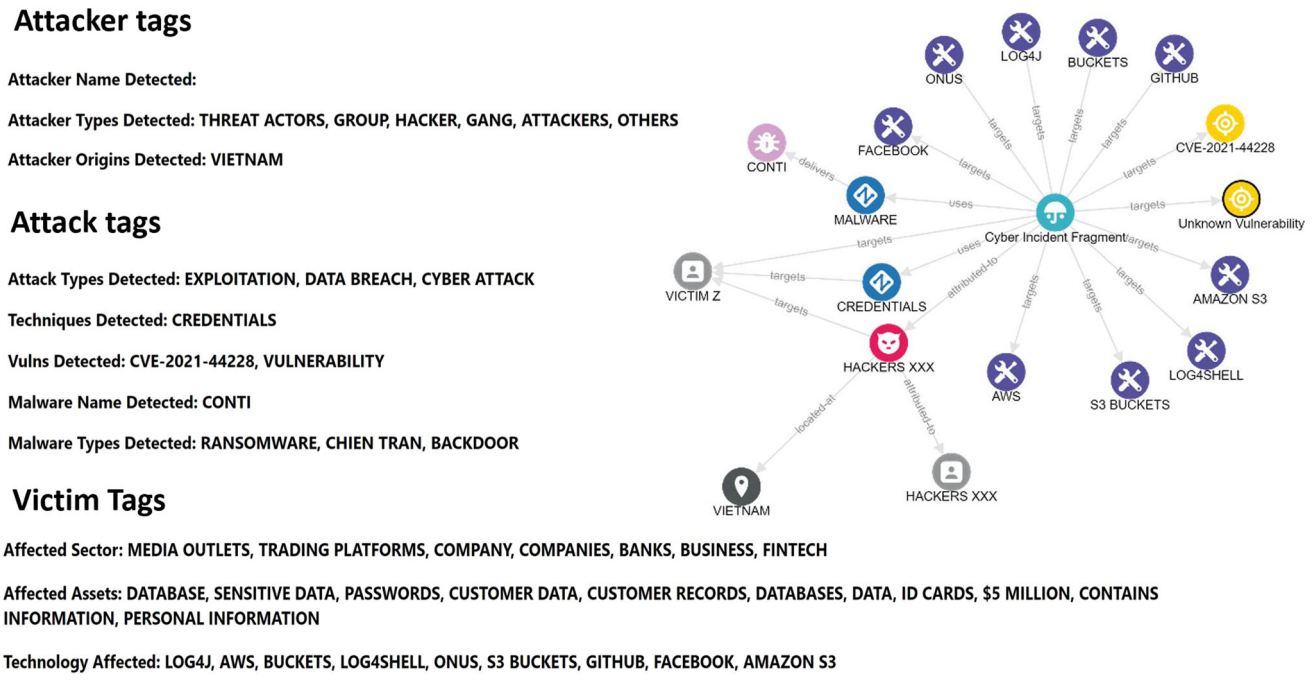


Fig. 6 IncGen output report and visualisation

3. Performing time-series analysis to plot and calculate future trends for a specific Attack Type and/or Training Objective.

In our implementation, we chose the SARIMA[11] equation to represent the trends on the existing KDb of 2970 articles as represented in Table 10. However, in future work, we intend to investigate further methods to boost the capabilities of MLTP, including the identification of micro-trends as the existing results are very promising [2].

[11] SARIMA is Seasonal ARIMA, or simply put, ARIMA with a seasonal component. ARIMA is a statistical analysis model that uses time-series data to predict future trends.

4.6 Putting everything together

Let us summarise the use of AiCEF and its modules with an example. An EP populates the Knowledge database (KDb) with incidents of interest, which are then converted into graphs based on the CESO ontology. When the EP wishes to create a new scenario for a cyber security exercise, they provide AiCEF with a set of keywords. To assist the planning process, AiCEF can generate a trend report that identifies trends relevant to the objectives at the time of the exercise execution. Based on the keywords, AiCEF crawls its database for the most relevant articles and returns a corresponding graph. The EP can then enhance the graph by merging it with that of known threat groups and filtering the graph according to the intended Cyber Kill Chain phases to be simulated. The resulting incident graph representation is then ready to be

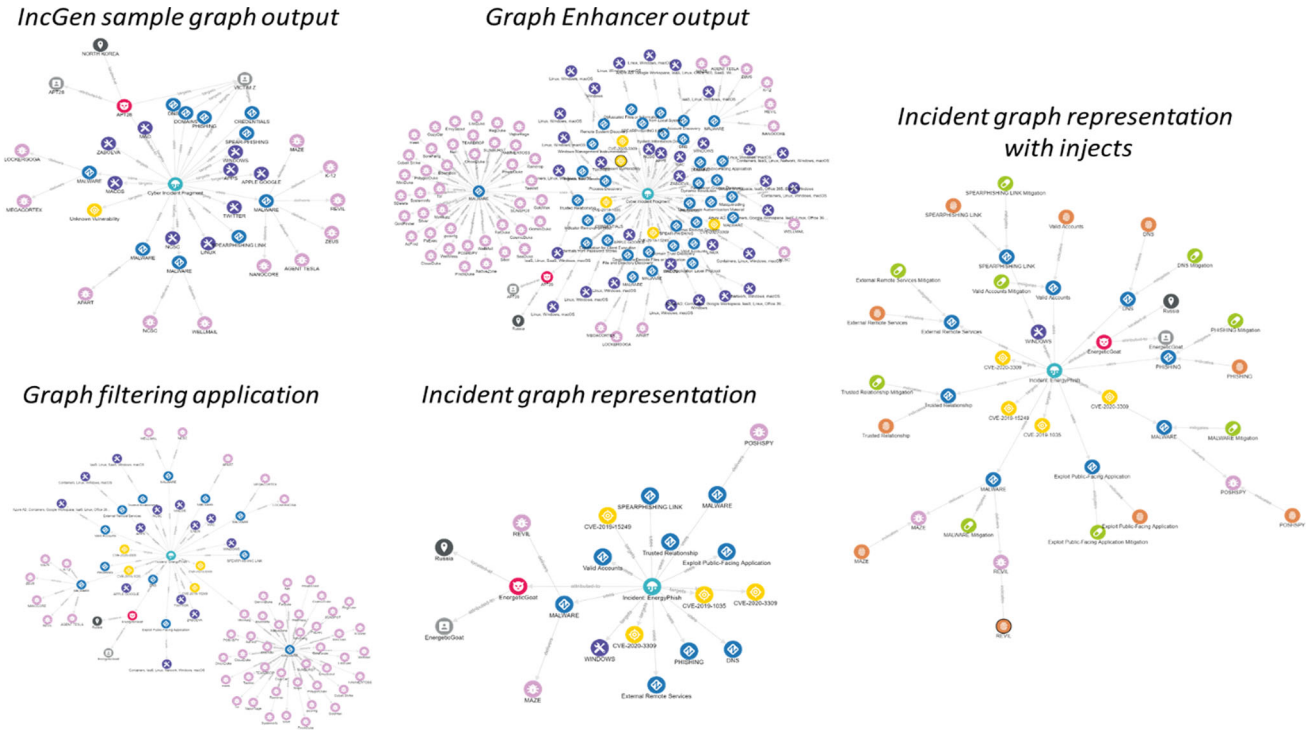


Fig. 7 IncGen execution flow with intermediate representation steps

populated with injects. A representation of the progress of an incident graph generation can be visualised in Fig. 7.

This process is repeated multiple times to generate the number of wanted incidents for a specific CSE. The EP follows the CEGen flow to compile a full exercise and generate a scenario (Fig. 8) and Exercise graph (Fig. 9).

5 Evaluation methodology and results

We developed a case study to help measure the effectiveness of our proposed framework and underlying methodology. To this end, the steps below were followed.

1. **Scenario Content Generation** A group of exercise planners of varying expertise have been used to generate the same exercise scenario using traditional exercise means and the AiCEF methodology and tools while being monitored on timeliness, effectiveness, creativity and methodology used.
2. **Content Evaluation** The reports were anonymised and given to a group of evaluators to grade the complexity, technical depth and richness of lessons learnt on the generated subset of exercise scenarios as per Objectives and KPIs set through a questionnaire.
3. **Results collection and Analysis** The results of this process were evaluated against the previously set KPIs to estimate:
   (a) Improved speed in Cyber Exercise Scenario generation (quantitative) using AiCEF.
   (b) Improvement in quality in Cyber Exercise Scenario generation (qualitative) for inexperienced Planners using AiCEF.
   (c) Improved relevance of proposed Cyber Exercise Scenarios to the current threat landscape (qualitative) using AiCEF.

5.1 Scenario content generation

Four EPs were selected to individually generate a CSE scenario according to the provided high-level exercise requirements and specifications, see Fig. 10. The EPs were split into two groups based on their previous experience with the task. All EPs have deep knowledge of cyber security, and their skill sets resemble that of a CISO.

Both groups consisted of one experienced and one inexperienced planner. The first group was briefly introduced to the basics of developing CSE scenarios, while the second one was provided with a course on using AiCEF and the accompanying tools. Both groups were provided with the same Scenario Template (ST) to fill in as an output of their task.

Then, we created a third group, consisting of Scripted Exercise Planner (SEP), using different parameters and flows of the AiCEF methodology and toolset.

The provided ST had the following generic structure:

State of the World (SoW)

Description: **STG Input**:: '*State sponsored Threat actor EnergeticGoat has historically targeted government organizations, non-government organizations (NGOs), think tanks, military, IT service providers, health technology and research, and telecommunications providers. With this latest attack, EnergeticGoat attempted to target the Energy Sector*'
STG Output: "State sponsored Threat actor EnergeticGoat has historically targeted government organizations, non-government organizations (NGOs), think tanks, military, IT service providers, health technology and research, and telecommunications providers. With this latest attack, EnergeticGoat attempted to target the Energy Sector in South Europe. More specifically the following targets have been identified: 1) EU energy firms that have received investments from the Energysubsidiary, 2) SA government units and departments who receive funding from the Energysubsidiary, 3) the SA energy regulator, 4) the Energy Sector's main lobbying and public relations organization, 5) the Energy Sector's largest customer, the Energysubsidiary, and 6) the Energy Sector's corporate regulator, the Energy Regulator of SA (ERA). Energy Sector entities and their customers and regulators are usually among the least likely organizations to be attacked by EnergeticGoat.
For example, the Energy Regulator of SA (ERA) is not a high-profile organization. It is a largely unregulated regulatory body that provides general oversight of the energy sector and serves as a key regulator and enforcement body of relevant energy laws in EU. Another example would be the Energy Corporation (Zesa). Zesa is a state-funded enterprise that provides essential infrastructure and related services to its citizens, including the electricity, gas, and broadband services that power large swaths in EU.
The attackers may have targeted other Energy Sector actors, such as Eskom or the SA Energy Institute (SAEI). However, it is thought that EnergeticGoat is specifically targeting Zesa and other organizations as a result of the threat actor's perceived inability to target the SA Energy Sector as a whole. EnergeticGoat's choice to target a South African energy firm with its latest attack is indicative of this perception."

Scenario Summary

EXERCISE NAME: Energy Test
EXERCISE OBJECTIVES: Test detection of Attacks and Incident Handling Procedures
EVENT 1: EnergyWipeOut ### Description: **STG Input**:: '*State sponsored Threat Actor EnergeticGoat is believed to be behind cyber-attacks against Energy Corp. The attackers used PHISHING & DATA DESTRUCTION techniques to*'
STG Output: "State sponsored Threat Actor EnergeticGoat is believed to be behind cyber-attacks against Energy Corp. The attackers used PHISHING & DATA DESTRUCTION techniques to infect the system with malware. In addition to the malware used, the attackers also used other tools to get around the security countermeasures of ENERGYCORP. It is understood that it will take at least a few days for the authorities to identify the attackers as they have only just informed the company about the attack. A spokesperson for the company said the company did not have any details as to the motive behind the attack as it was still investigating. "We are working with law enforcement to bring this to the most serious level. The situation is under investigation and we cannot comment any further than that"
*** Incident 1.1: EnergyPhish *** Description: **STG Input**:: '*EnergeticGoat deployed Phishing attacks against Energy Corp. Malicious word documents were sent*'
STG Output:"EnergeticGoat deployed Phishing attacks against Energy Corp. Malicious word documents were sent to Energy Corp. employees in an attempt to obtain personal information. On October 26, 2022, a new employee of Energy Corp., received a suspicious email containing a malicious word document. The email identified as 'malware' was in fact a phishing attack with the subject 'Vendors of Energy Corp. under Federal Investigation"

Fig. 8 Sample text of an AI-generated exercise

- Section 1: Storyline (SoW)
- Section 2: Scenario and MSEL
- Section 3: Scenario Analysis
- Section 4: Resources Used

We provided detailed instructions on the expected content per paragraph to all involved planners to streamline the information of the generated reports and create homogeneous outputs to be evaluated in the later step.

As a result, five complete exercise scenarios were generated, as shown in Table 12.

5.2 Scenario content evaluation

To evaluate the scenarios above, we conducted an anonymous online survey from 01/09/2022 to 30/09/2022. To avoid bias, we invited a number of evaluators from different cyber awareness and cyber exercise groups with varying expertise, ethnicity, and focus sectors to participate in the evaluation process. More precisely, we invited the Ad-hoc Cyber Awareness Expert Group of ENISA. In total, 16 experts responded, whose demographic statistics are illustrated in Table 13. Given that we have a representation of 66% of the group, we believe that the sample is significant, as they are experts. Moreover, we highlight that their allocation has been made through independent criteria, not from us, but from an individual international organisation on cyber security such as ENISA, which avoids possible biases.

The survey was in the form of an online questionnaire consisting of 11 questions. Eight questions were used to evaluate the generated Scenarios, two to be used as Turing test to determine whether the AI used could be identified by humans and a set of complementary questions for demographic and future improvement purposes. All five scenarios were provided using only the "Eval_Tag" parameter for tracking purposes without providing additional information on the authors of the scenarios.

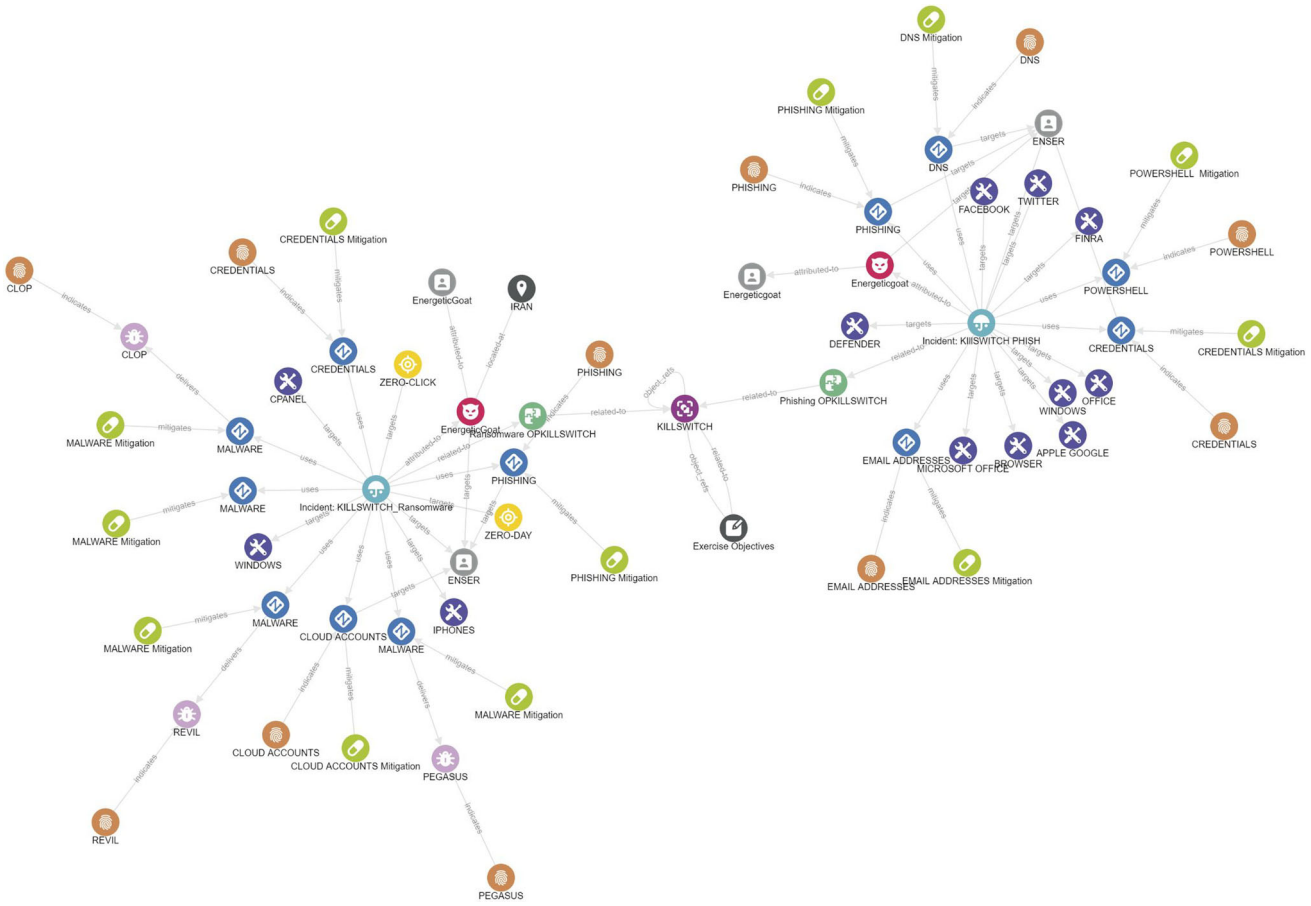


Fig. 9 Sample exercise graph visualisation

> Generate a CSE scenario for a cyber awareness exercise by filling in a provided Scenario Template. The CSE should include 2 events consisting of 1 incident each. All incidents should be accompanied by a short description of indicative injects to be sent to players. At least 3 inject descriptions per incident should be provided.
> The company which will use the scenario is an **Energy Service Provider** and all its Employees can be potential Players.
> The exercise should last between 2-4 hours and can include technical artefacts for analysis.
> The two main objectives of the exercise are:
>
> 1. Provide awareness to employees regarding Phishing Attacks
> 2. Provide awareness to employees regarding Ransomware Attacks
>
> These objectives can be updated and more can be added. The Scenario Development task will be timed and should not last more than 4 hours

Fig. 10 Task definition

The eight scenario evaluation questions and their corresponding scores in parenthesis were the following:

1. How do you evaluate the relevance of the State of the World text to the Objectives of the Exercise? (0–4)
2. How do you evaluate the relevance of the selected Events to the Objectives of the exercise? (0–4)
3. How do you evaluate the relevance of the selected Incidents to the Objectives of the exercise? (0–4)
4. How do you evaluate the Complexity of the Scenario? (0–1)
5. How do you evaluate the Technical Depth of the Scenario? (0–2)
6. How do you evaluate the Threat Actor's description? (1–3)
7. How do you evaluate the used resources? (0–2)
8. Would you be willing to use this Scenario based on the task description? (0–4)

To evaluate the use of AI for exercise content generation, we asked the expert the following questions:

1. How was the scenario generated?
2. How skilled was the planner?

Other questions revolved around the overall scenario development process:

Table 12 Details of the generated scenarios

Eval_Tag	Explanatory Name Tag	Exercise Expertise	AiCEF	Duration	Other Tools
ExSc1	Sc1:Exp(erienced)Hum(an)	YES	NO	2 h 00min	Google, MITRE
ExSc2	Sc2:Nov(ice)Hum(an)	NO	NO	2 h 35 min	Google
ExSc3	Sc3:Exp(erienced)Hum(an) and AI	YES	YES	1 h 20min	Online Resources
ExSc4	Sc4:Nov(ice)Hum(an) and AI	NO	YES	2 h 10 min	Google
ExSc5	Sc5:AiCEF	N/A	YES	20 min	–

Table 13 Demographics of the experts

Countries	#
Greece	2
Austria	1
Italy	1
Belgium	2
Poland	1
Spain	1
Romania	1
Portugal	1
Czechia	1
Netherlands	2
France	2
Finland	1

(a) Countries of origin of the experts

Sector	#
Governmental	6
Energy	2
ICT	4
Critical Infra	1
aw enforcement	1
Education	1
Other	1

(b) Sector that experts are working in

Seniority	#
Novice (small exercises)	5
Medium (medium scale exercises for a few years)	3
Expert (EU level, cross country exercises)	4
Senior (large-sized exercises)	1
None	3

(c) Seniority self-assessment

1. How much time did you invest in the Scenario Content Development?
2. How do you define the scope/objectives of the exercise?
3. How do you define the scenario content?
4. What tools did you use to create the scenario or define the objectives if any?

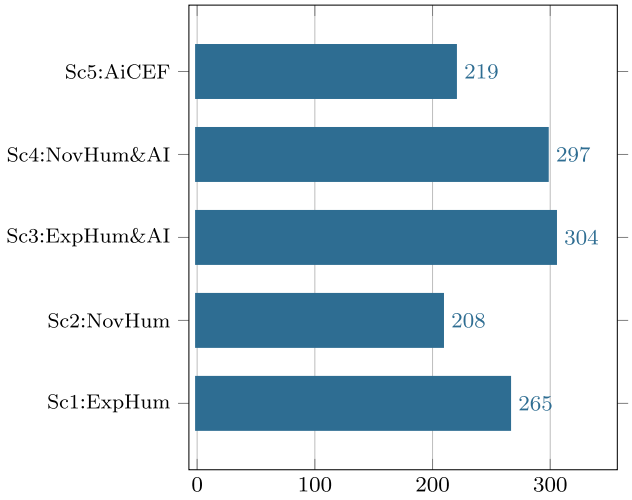


Fig. 11 Overall performance of evaluated scenarios based on total score

Finally, evaluators were asked to rank AI-powered tools as follows:

- Rank the following AI-powered tools that could be created to support the design and implementation of future cyber exercises:
  - Automated extraction of Exercise Objects (Incidents, Injects) from unstructured information and DB storage
  - Lead generation for trend prediction of Training Topics
  - Automated Enrichment's of content to match realistic patterns and relationships of known Attackers
  - Automated Cyber Exercise Script/Scenario Generation

5.3 Results analysis

The analysis of the input provided a good understanding of the strengths and potential areas for improvement of AiCEF. It also provided better insight into the exercise Scenario creation process, with good inputs for future improvement based on the experience of real EPs (Fig. 11).

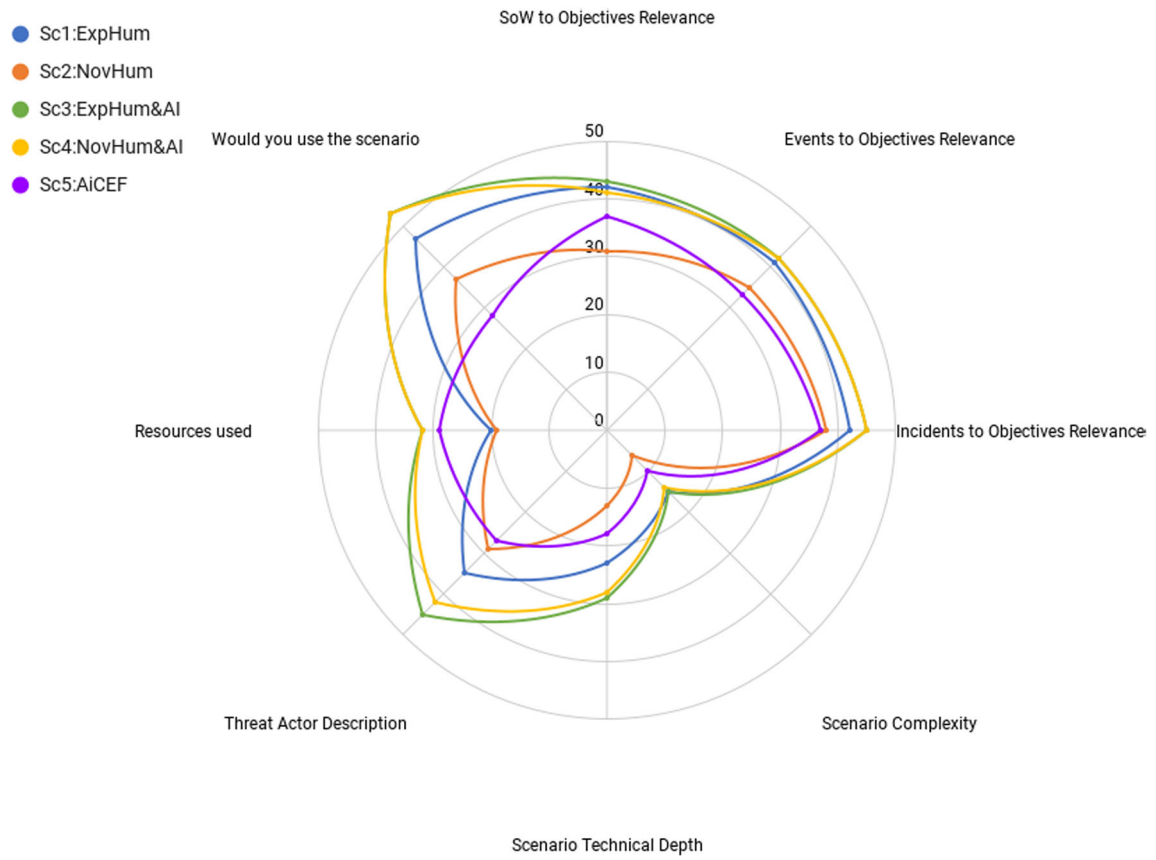


Fig. 12 Scenario evaluation parameters

Based on analysis of the provided input, we can safely conclude that both scenarios Sc3:ExpHum and AI and Sc4:NovHum and AI have scored higher than any other scenario with the help of AiCEF, see Fig. 14. Currently, the hybrid scenario generation approach of a human exercise planner using AiCEF outperforms a seasoned exercise planner, even when a planner is a novice. Furthermore, the Scripted Exercise Planner generated a relatively good Scenario (Sc5:AiCEF) that can be evaluated as equal, if not better, than that of a novice planner (SC2:NovHum) (Fig. 12).

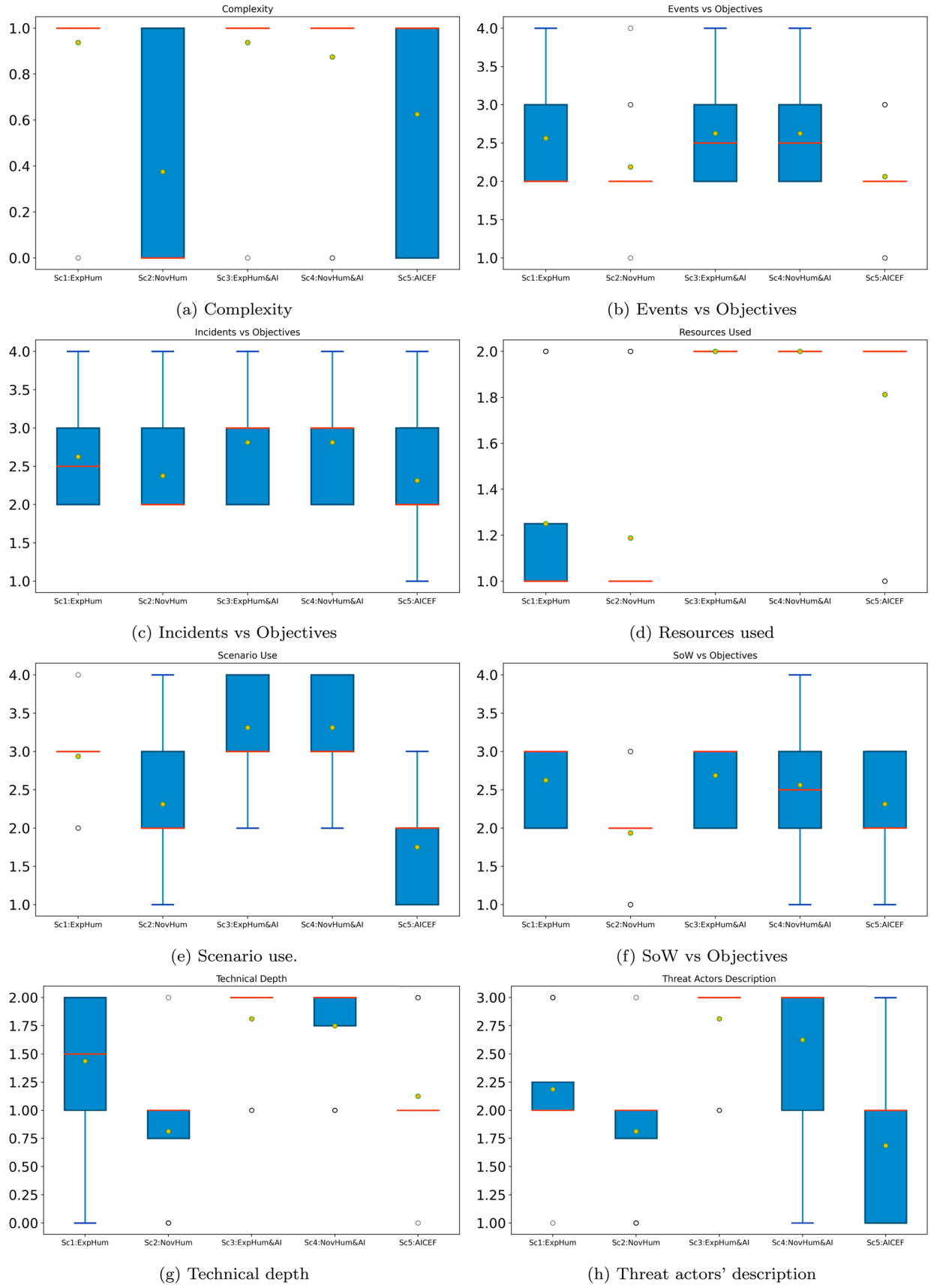
In what follows, we provide a breakdown of the parameters evaluated to highlight the strengths and weaknesses of using AiCEF based on the experts' input.

The use of AiCEF by a Scripted Exercise Planner performed well (top 3, outperforming humans) in *Relevant Resources*, *Events Relevance*, and *Scenario Technical Depth*. On the other hand, AiCEF did not perform as well in the following aspects: *Threat Actor Description*, *Scenario Complexity*, and *Incidents to Objectives Relevance*. The above can be justified by the fact that the raw generated content can include conflicting information or content that might not match the high-level context requested. After human curation, the content can be easily improved to compete with a seasoned exercise planner. In fact, AiCEF used by humans helped them excel in *Scenario Creation*, dominating all categories versus their human counterparts. The human expert using AiCEF (Sc3:ExpHum&AI) managed to create a better scenario 33,33% faster than his expert peer using regular tools (Sc1:ExpHum) (Fig. 13).

Nevertheless, the most impressive finding was that novice planners using AiCEF (Sc4:NovHum&AI) outperform a seasoned exercise planner (Sc1:ExpHum), as seen in Fig. 12, providing a good indication of the capabilities of the proposed framework. Note that the scenario performance developed by the novice planner with the help of AiCEF matches, among others, that of a Seasoned Planner in the question: *Would you use the scenario?*". Even more, evaluators could not distinguish the pure AI-generated content (ExSC5) based on Table 14, categorising the scenario as either hybrid or human-made. Indeed, the results were like those of a novice human planner.

On the question: "*How do you define the scope/objectives of the exercise?"* most evaluators replied with two or more of the following options, with known incidents and lessons learnt along with risk assessment as the most prevalent replies (Fig. fig:score).

On the question: "*How do you define the scenario content?*" most evaluators replied with two or more options, with



(a) Complexity

(b) Events vs Objectives

(c) Incidents vs Objectives

(d) Resources used

(e) Scenario use.

(f) SoW vs Objectives

(g) Technical depth

(h) Threat actors' description

Fig. 13 Score range for the Q1–8 of 16 evaluators

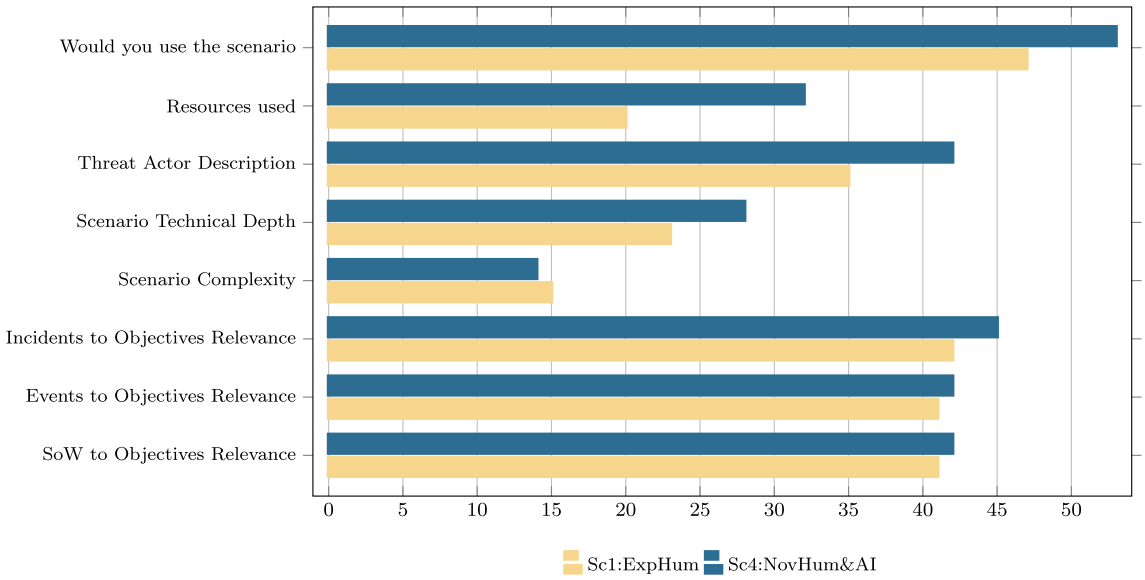


Fig. 14 Novice planner with AiCEF (Sc4:NovHum&AI) versus senior exercise planner (Sc1:ExpHum)

Table 14 Turing test to evaluate the performance of AI

Scenario	Human	AI and human	AI
Sc1:ExpHum	1	13	2
Sc2:NovHum	9	2	5
Sc3:ExpHum&AI	0	9	7
Sc4:NovHum&AI	1	10	5
Sc5:AiCEF	6	5	5

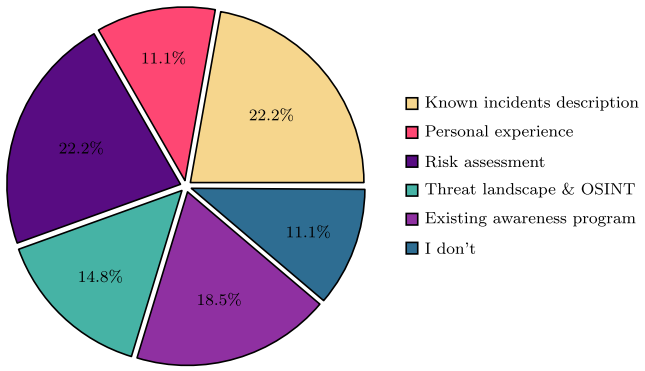
news and articles being the most important source followed by the known incident option (Fig. 15).

The evaluators replied to the question "*How much time do you invest in the Scenario Content Development?*" with an average of 53 h. This means that tools which can improve the CSE scenario content development process by reducing time without compromising the quality could be of great use.

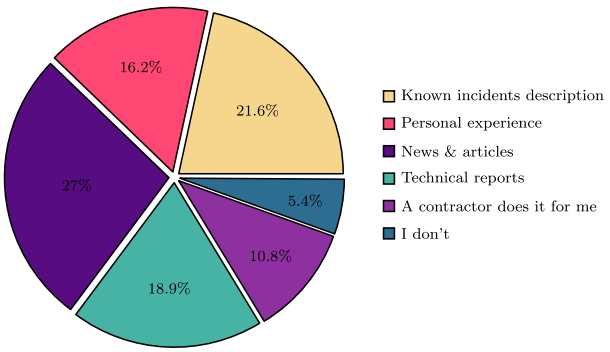
Finally, for the question "*What tools did you use to create the scenario or define the objectives if any?*", the responses varied between Google Search, Cyber Security (News) websites, MS Office, and Internet/Table Top Research.

6 Conclusions and future work

The shortage of cybersecurity experts and awareness is a well-known and big worldwide challenge. CSEs can address some of the aspects of this problem; however, the shortage of experts to develop new CSEs coupled with the timeliness and relevance of the developed CSEs requires novel solutions. In this work, we try to fill in this gap by facilitating the



(a) Experts' responses to "How do you define the scope/objectives of the exercise?" question.



(b) Experts' responses to "How do you define the scenario content?" question.

Fig. 15 Experts' responses

work of EPs with the use of AI. To this end, we developed a novel AI-powered exercise generation framework called AiCEF, which generates structured exercise scenarios that reflect the current or future threat level that an organisation faces, including potential threat actors and TTPs. Moreover, it generates scripted events that could happen in the context of a real attack against a specific organisation belonging to one of the NIS2 critical infrastructure sectors. AiCEF also identifies and describes artefacts that could accompany the exercise scenarios. To this end, AiCEF uses a new ontology that we built, named CESO, and with which we were able to generate structured exercise scenarios that can be both machine and human-readable.

Our proposed methodology and developed tools can provide tangible qualitative and quantitative added value in CSE development and Cyber Awareness in various ways. For instance, in our experiment, the total time for the CSE scenario generation is decreased by 33.33% without impacting the quality. In fact, AiCEF improves the quality of CSE scenario generation for an inexperienced/novice EP by elevating the generated scenario quality to the same level as an experienced EP. Finally, the relevance of proposed CSE scenarios is aligned with that of the current threat landscape, as indicated by evaluating all the generated scenarios using AiCEF.

While AiCEF might be rather efficient, there is room for various improvements. For instance, for operational usage, more sources must be parsed (ex., threat reports and alerts) to generate more diverse scenarios. While Generative Pre-trained Transformers (GTP-2 and GTP-3) [4] [5] might create a textual output of very good quality, it would be even better if the text synthesiser were based only on Cyber Security related resources so that the generated text is even more relevant and uses, e.g. better technical terms. As indicated in the evaluation, AiCEF could be benefited from further improvements to enhance the threat actor description section. Finally, we plan to enhance AiCEF to detect the Cyber Kill Chain phases automatically using NER and create relevant CSE injects for a number of popular categories like phishing while also automating the inject description and content generation using AI-powered text synthesis.

Acknowledgements This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the project CyberSec4Europe (https://www.cybersec4europe.eu) (Grant Agreement no. 830929), and Horizon Europe Programme, as part of the project LAZARUS (https://lazarus-he.eu/) (Grant Agreement no. 101070303). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Funding Open access funding provided by HEAL-Link Greece.

Data availability The used data are publicly available.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval The authors declare full compliance with ethical standards. This article does not contain any studies involving humans or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Adams, W.J., Gavas, E., Lacey, T.H., Leblanc, S.P.: Collective views of the NSA/CSS cyber defense exercise on curricula and learning objectives. In: CSET (2009)
2. Zacharis, A., Gavrila, C.P.R.: AI-assisted cyber crisis management exercise content generation: Modelling a cyber conflict. In: 15th International Conference on Cyber Conflict (CyCon 2023). IEEE (2023)
3. Augustine, T., Dodge, R.C., et al.: Cyber defense exercise: meeting learning objectives thru competition. In: Proceedings of the 10th Colloquium for Information Systems Security Education (2006)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, S., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020a)
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, S., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020b)
6. Conklin, A.: The use of a collegiate cyber defense competition in information security education. In: Proceedings of the 2nd Annual Conference on Information Security Curriculum Development, pp. 16–18 (2005)
7. Conklin, A.: Cyber defense competitions and information security education: An active learning solution for a capstone course. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), vol. 9, pp. 220b–220b. IEEE (2006)
8. Craig, R.T.: Generalization of Scott's index of intercoder agreement. Publ. Opin. Q. **45**(2), 260–264 (1981). https://doi.org/10.1086/268657
9. Dewar, R.S.: Cybersecurity and Cyberdefense Exercises. Tech. rep, ETH Zurich (2018)
10. Dodge, R., Ragsdale, D.J.: Organized cyber defense competitions. In: Proceedings of IEEE International Conference on Advanced Learning Technologies, pp. 768–770. IEEE (2004)

11. Dodge, R., Hay, B., Nance, K.: Standards-based cyber exercises. In: 2009 International Conference on Availability, Reliability and Security, pp. 738–743. IEEE (2009)
12. European Commission: Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) (Text with EEA relevance). https://eur-lex.europa.eu/eli/dir/2022/2555 (2022)
13. Furtună, A., Patriciu, V.V., Bica, I.: A structured approach for implementing cyber security exercises. In: 2010 8th International Conference on Communications, pp. 415–418. IEEE (2010)
14. Granåsen, M., Andersson, D.: Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. Cognit. Technol. Work **18**(1), 121–143 (2016)
15. Green, A., Zafar, H.: Addressing emerging information security personnel needs. a look at competitions in academia: Do cyber defense competitions work. In: AMCIS 2013 Proceedings, vol. 1, p. 257 (2013)
16. Gurnani, R., Pandey, K., Rai, S.K.: A scalable model for implementing cyber security exercises. In: 2014 International Conference on Computing for Sustainable Global Development (INDIACom), pp. 680–684. IEEE (2014)
17. of Homeland Security UD: DHS Cyber TTX for the healthcare industry. https://www.hsdl.org/?abstract&did=789781 (2013)
18. ISO Central Secretary: Societal security - guidelines for exercises. Standard ISO22398:2013, International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/50294.html (2013)
19. Karagiannis, S., Magkos, E.: Engaging students in basic cybersecurity concepts using digital game-based learning: computer games as virtual learning environments. In: Advances in Core Computer Science-Based Technologies, pp 55–81. Springer (2021)
20. Karjalainen, M., Kokkonen, T., Puuska, S.: Pedagogical aspects of cyber security exercises. In: 2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 103–108. IEEE (2019)
21. Kick, J.: Cyber exercise playbook. Tech. rep, MITRE CORP BEDFORD MA (2014)
22. Li, Y., Liljenstam, M., Liu, J.: Real-time security exercises on a realistic interdomain routing experiment platform. In: 2009 ACM/IEEE/SCS 23rd Workshop on Principles of Advanced and Distributed Simulation, pp. 54–63. IEEE (2009)
23. Liljenstam, M., Liu, J., Nicol, D.M., Yuan, Y., Yan, G., Grier, C.: Rinse: the real-time immersive network simulation environment for network security exercises (extended version). Simulation **82**(1), 43–59 (2006)
24. Lockheed Martin: The Cyber Kill Chain. https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html (2011)
25. MacIntyre, R.: Penn treebank tokenizer (sed script source code) (1995)
26. Mattson, J.A.: Cyber defense exercise: A service provider model. In: IFIP World Conference on Information Security Education, pp. 81–86. Springer (2007)
27. Mink, M., Freiling, F.C.: Is attack better than defense? teaching information security the right way. In: Proceedings of the 3rd annual conference on Information security curriculum development, pp. 44–48 (2006)
28. MITRE: CVE. https://cve.mitre.org/ (1999)
29. MITRE: MITRE ATT&CK. https://attack.mitre.org/ (2022)
30. Mullins, B.E., Lacey, T.H., Mills, R.F., Trechter, J.E., Bass, S.D.: How the cyber defense exercise shaped an information-assurance curriculum. IEEE Secur. Privacy **5**(5), 40–49 (2007)
31. Mullins, B.E., Lacey, T.H., Mills, R.F., Trechter, J.M., Bass, S.D.: The impact of the nsa cyber defense exercise on the curriculum at the air force institute of technology. In: 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07), pp. 271b–271b. IEEE (2007b)
32. OASIS OPEN: STIX version 2.1. https://www.oasis-open.org/standard/stix-version-2-1/ (2021)
33. Pastuszuk, J., Burek, P., Ksieopolski, B.: Cybersecurity ontology for dynamic analysis of it systems. Procedia Comput. Sci. **192**, 1011–1020 (2021)
34. Patriciu, V.V., Furtuna, A.C.: Guide for designing cyber security exercises. In: Proceedings of the 8th WSEAS International Conference on E-Activities and information security and privacy, World Scientific and Engineering Academy and Society (WSEAS), pp. 172–177 (2009)
35. Planning, M.E.: Directors's Guideline for Civil Defence Emergency Management Groups, wyd. Ministry of Civil Defence & Emergency Management, Wellington (2008)
36. Rursch, J.A., Luse, A., Jacobson, D.: It-adventures: A program to spark it interest in high school students using inquiry-based learning with cyber defense, game design, and robotics. IEEE Trans. Educ. **53**(1), 71–79 (2009)
37. Samejima, M., Yajima, H.: It risk management framework for business continuity by change analysis of information system. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1670–1674. IEEE (2012)
38. Sangster, B., O'Connor, T., Cook, T., Fanelli, R., Dean, E., Morrell, C., Conti, G.J.: Toward instrumenting network warfare competitions to generate labeled datasets. In: CSET (2009)
39. Scarfone, K.A., Grance, T., Masone, K.: Sp 800-61 rev. 1. computer security incident handling guide (2008)
40. Schepens, W., Ragsdale, D., Surdu, J.R., Schafer, J., Port, R.N.: The cyber defense exercise: an evaluation of the effectiveness of information assurance education. J. Inf. Secur. **1**(2), 1–14 (2002)
41. Schepens, W.J., James, J.R.: Architecture of a cyber defense competition. In: SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483), vol. 5, pp. 4300–4305. IEEE (2003)
42. Schweitzer, D., Gibson, D., Collins, M.: Active learning in the security classroom. In: 2009 42nd Hawaii International Conference on System Sciences, pp. 1–8. IEEE (2009)
43. Sommestad, T., Hallberg, J.: Cyber security exercises and competitions as a platform for cyber security experiments. In: Nordic conference on secure IT systems, pp. 47–60. Springer (2012)
44. Tobey, D.H.: A vignette-based method for improving cybersecurity talent management through cyber defense competition design. In: Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research, pp. 31–39 (2015)
45. Tsinganos, N., Mavridis, I.: Building and evaluating an annotated corpus for automated recognition of chat-based social engineering attacks. Appl. Sci. **11**(22), 10871 (2021)
46. Vigna, G.: Teaching network security through live exercises. In: IFIP World Conference on Information Security Education, pp. 3–18. Springer (2003)
47. Wen, S.F., Yamin, M.M., Katt, B.: Ontology-based scenario modeling for cyber security exercise. In: 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 249–258. IEEE (2021)
48. White, G.B., Dietrich, G., Goles, T.: Cyber security exercises: testing an organization's ability to prevent, detect, and respond to cyber security events. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences (2004), p. 10. IEEE (2004)

49. White, G.B., Williams, D., Harrison, K.: The cyberpatriot national high school cyber defense competition. IEEE Secur. Privacy **8**(5), 59–61 (2010)
50. Wilhelmson, N., Svensson, T.: Handbook for planning, running and evaluating information technology and cyber security exercises. Försvarshögskolan (FHS) (2011)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2023 15th International Conference on Cyber Conflict
Meeting Reality
T. Jančárková, D. Giovannelli, K. Podiņš, I. Winther
2023 © NATO CCDCOE Publications, Tallinn

Permission to make digital or hard copies of this publication for internal use within NATO and for personal or educational use when for non-profit or non-commercial purposes is granted provided that copies bear this notice and a full citation on the first page. Any other reproduction or transmission requires prior written permission by NATO CCDCOE.

AI-assisted Cyber Security Exercise Content Generation: Modeling a Cyber Conflict

Alexandros Zacharis
European Union Agency
for Cyber Security
alexandros.zacharis@enisa.europa.eu

Razvan Gavrila
European Union Agency
for Cyber Security
razvan.gavrila@enisa.europa.eu

Constantinos Patsakis
Department of Informatics
University of Piraeus
kpatsak@unipi.gr

Demosthenes Ikonomou
European Union Agency
for Cyber Security
demosthenes.ikonomou
@enisa.europa.eu

Abstract: A cyber conflict can be defined as a cyberattack or a series of attacks that target the critical functions of a country. Such attacks can potentially wreak havoc on government and civilian infrastructure and disrupt critical systems, resulting in damage to the state and even loss of life. National bodies are usually expected to run cyber crisis exercises to prevent such attacks and prepare for their impact. Developing risk scenarios that are both relevant and up to date with the current threat landscape is a critical element in the success of any cyber exercise, especially a cyber conflict scenario.

Our work explores the results of applying machine learning to unstructured information sources to generate structured cyber exercise content in preparation for or during a destructive cyber conflict. We collected a dataset of publicly available cyber security articles and used them to assess future threats and as a skeleton for new exercise scenarios. We utilize named-entity recognition to structure the information based on a novel ontology. With the help of graph comparison methodologies, we match the generated scenarios to known threat actors' tactics, techniques, and procedures and enrich the final scenario accordingly, with the help of synthetic text generators following our novel artificial-intelligence-assisted cyber exercise framework (AiCEF). Our framework has been evaluated on its efficiency and speed and can produce structured cyber exercise scenarios in real time, provided with incident descriptions

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

in raw text format or a set of keywords. By deep diving into a pool of pre-tagged incidents, AiCEF can build exercise content from scratch, assisting inexperienced exercise planners in generating a scenario quicker and achieving a level of quality similar to an experienced planner or subject matter expert.

We have assessed our methodology for relevance and preparedness by applying it to a real cyber conflict use case to model two categories of crisis management exercise scenarios: pre-conflict and post-conflict initiation. Thus, we assess whether the generated scenarios match the attack trends and the news feeds that were not used in training the AiCEF and prove that we can provide targeted and customized awareness of upcoming incidents.

Keywords: *cyber conflict, cyber awareness, cyber exercises scenario, artificial intelligence, machine learning, named-entity recognition*

1. INTRODUCTION

Cyber security exercises (CSE) are increasingly becoming an integral part of the cybersecurity training landscape [1], providing hands-on experience to personnel of both public and private organizations worldwide. The ISO Guidelines for Exercises [2] define a CSE as "a process to train for, assess, practice, and improve performance in an organisation." Similarly, ENISA defines a CSE as "a planned event during which an organisation simulates cyber-attacks or information security incidents or other types of disruptions to test the organisation's cyber capabilities, from being able to detect a security incident to the ability to respond appropriately and minimise any related impact." [3]

The success of such exercises can only be measured by their outcomes and ability to address real-world needs. As a result, each exercise must prioritize the prompt identification of training objectives, topic-specific preparation, and mirroring of real-world scenarios. To achieve these aims, exercise planners (EPs) must invest substantial effort, particularly when creating material miming cyber operations.

The Current Problem

Cyber operations are best used in combination with electronic warfare, disinformation campaigns, anti-satellite attacks, and precision-guided munitions. The objective is to degrade informational advantage and intangible assets (e.g., data) for operational advantage. Cyber operations can also be used for political effect by disrupting finance,

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

energy, transportation, or government services [4]. This inherited complexity makes creating realistic scenarios difficult, especially for inexperienced cyber exercise planners or those lacking subject matter expertise. The planner must make a plethora of choices when creating a scenario to not only meet the learning objectives but also make it timely and appealing to the target audience. The research question is whether the above process can be automated. While previous work [5] has shown that the planning process can be modeled [6], [7] and automated to a great extent, the relevance and timeliness aspects have not yet been explored. Therefore, the core problem that this work studies is the relevance of the generated exercises and their usefulness in a real-world scenario based on public information sources.

Proposed Solution and Contribution

Based on the above, this paper illustrates how an EP can generate structured and realistic CSE scenarios based on a pool of pre-tagged incident information despite having little prior experience in scenario development. We also demonstrate how using machine learning (ML) and our artificial-intelligence-assisted cyber exercise framework (AiCEF) [5], such scenarios can be generated from the ground up, in particular the threats, attacks, assets, vulnerabilities, and attack vectors used by sophisticated threat actors, as well as the potential impact and severity of the resulting cyber incidents.

Thus, the main contributions reported in this article include but are not limited to:

1) mapping real-world threat information using our cyber exercise scenario ontology (CESO);
2) using this mapping to generate AI-enriched scenarios to improve their realism;
3) identifying the best strategy for using known, relevant incidents to predict and train on incidents that will materialize in the future.

Overview of This Paper

The rest of the paper is divided into four sections. First, we summarize previous literature on cybersecurity exercises and cyber drills. Then, we provide an overview of the AiCEF framework and how it uses ML to generate cybersecurity exercise scenarios. We focus on the cyber exercise scenario ontology (CESO) and the scenario augmented model (SAM) used, and on how EPs can use advanced persistent threat (APT) enhancers to simulate the activity of sophisticated threat actors more precisely. Next, we present the methodology to generate the test scenarios step by step. We have created simulated incident datasets for the public administration, energy, and information and communications technology (ICT) sectors for this work. In Section 5, we compare our datasets to observed cybersecurity attacks to provide

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

an overview of our main use cases and findings by assessing whether the synthetic scenarios match observed real-world cyber trends. Our main focus has been the cyber incidents that public sources linked to Ukraine before and after the early phases of the 2022 aggression. We made this choice based on the duration of the conflict and the abundance of resources covering it, providing the perfect testbed for the development of future exercise scenarios in an attempt to cover future events.

2. RELATED WORK

Previous research in the field of cybersecurity exercises [8]–[10] has highlighted the importance of cyber drills [11] in aiding teams in designing, implementing, managing, and defending a computer network [3], [12]–[17]. Patriciu and Furtuna [18], [19] propose various processes and criteria for developing a cybersecurity exercise [20].

Green et al. [21] and Rursch et al. [22] have conducted additional research on cyber defense competitions [23]; [24] examines the best-suited architecture, tools, and strategies to create an active learning experience [25]–[27], whereas [28] takes a different approach to live cyber drills, offering lessons learned and making recommendations to assist organizations in conducting their own exercises. Other studies have examined how to run a cybersecurity exercise using service providers [29], [30].

In the literature, most researchers discuss serious games and how they can be used to train future practitioners, focusing on various aspects of capture-the-flag (CTF) challenges. The core concept is that the gamification of a thought process can significantly improve the learning process [31], [32] and seems rather effective in cybersecurity [33]–[36].

Nevertheless, despite the issues that CTF challenges and other gamification approaches might have, the crucial issues are the identification of the learning objectives, the positioning, and the timeliness of the content [37], [38]. Undoubtedly, for core cybersecurity skills, these aspects are not that thorny. Yet they become a dominant problem when the target training group is not a novice or when multiple skills have to be trained across diverse groups. As a result, these aspects require a lot of manual effort, good knowledge of the field, and an overview of how different sectors operate [39].

Moreover, cybersecurity exercises can be used to generate scientifically useful datasets for future security research [40], [41] and to identify hidden risks from ineffective security policies and/or processes [42], [43].

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

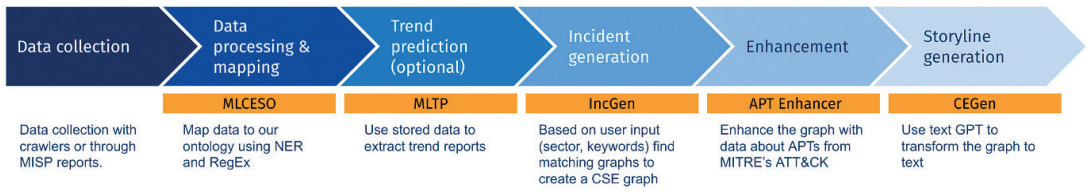
3. CESO AND THE AiCEF FRAMEWORK

AiCEF [5] is an ML-powered cyber exercise generation framework developed in Python consisting of various tools to help an EP create a timely and targeted CSE scenario regardless of their experience level. Its main components relevant to the work presented in this paper are the following:

1) CESO: the cyber exercise scenario ontology used to describe the various components of a CSE;
2) AiCEF: the cyber exercise framework used to model CSEs based on CESO using ML;
3) MLCESO: the ML models trained to parse text and extract objects based on CESO;
4) IncGen: the incident generation module that models a CSE incident from the MLCESO-extracted objects based on CESO;
5) CEGen: the cyber exercise generation module that models a CSE from the MLCESO-extracted objects based on CESO;
6) KDb: a knowledge pool of incidents stored in a database. Extracted objects and other characteristics, including the STIX 2.1 blob, are stored in the database;
7) APT Enhancer: a module that helps enhance any STIX 2.1 incident graph with objects of the modeled activity of known APTs [44].

Figure 1 shows these components in a timeline diagram to help the reader get a quick grasp of the role of each component in the flow and navigate through the rest of the sections understanding how these pieces fit in the bigger picture.

FIGURE 1: PROCESS FLOW AND THE CORRESPONDING MODULES OF AICEF [5]



We briefly describe each of these components in the following paragraphs, but the interested reader can refer to [5] for more detail.

Cyber Exercise Scenario Ontology (CESO)

A CSE is a collection of simulated incidents provided to players in an orchestrated way to achieve the exercise's objectives. CESO [5], the ontology developed to support AiCEF, is incident-centric, focusing on using a bottom-up approach that allows us to

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

identify and describe incidents first so we can group them into events and then cover the full generation of CSE scenarios that fit the high-level objectives set.

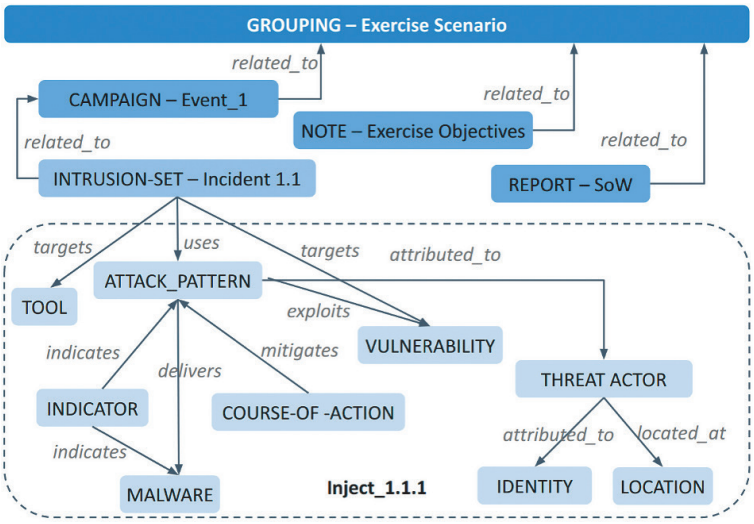
We have explored a set of existing ontologies, taxonomies, frameworks, standards, and formats relevant to cyber security and focus on the representations of the critical elements of CSEs, considering that their building blocks are the very incidents to be simulated. Our research concluded that a combination of ISO 22398 [2], MITRE ATT&CK [45] and Cyber Kill Chain [29], MITRE CVE [17], and STIX 2.1 [22] would provide the necessary means.

We chose STIX 2.1 as the basis for our ontology, which defines a taxonomy of cyber threat intelligence to be extended to cover our need to describe a CSE scenario.

Scenario Augmented Model

Based on the bottom-up approach, we propose a scenario augmented model (SAM) in two layers. First is the informational layer (see Figure 2) that covers the scenario's context and main attributes.

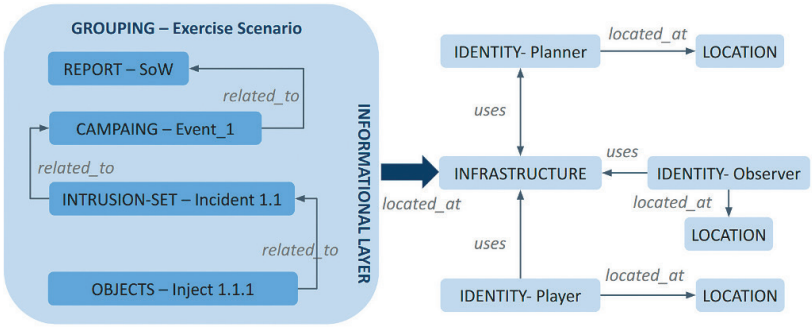
FIGURE 2: INFORMATIONAL LAYER OF CESO [5]



Second is the operational layer (see Figure 3) that describes an exercise scenario's execution flow, mainly dealing with injects delivery to the intended recipients. The operational layer has two major interrelated parts: 1) the events/injects, which describe the detailed activities of the scenario and the expected actions from the participants, and 2) the participants.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

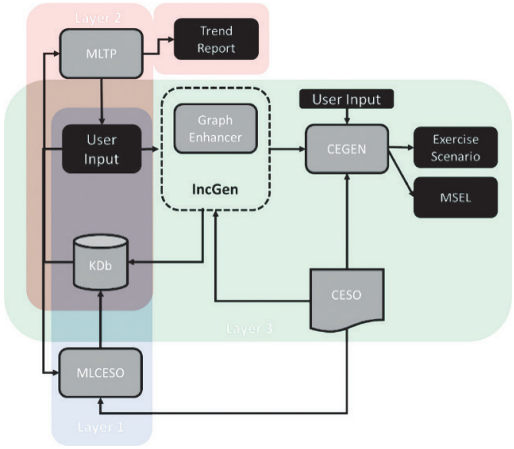
FIGURE 3: OPERATIONAL LAYER OF CESO [5]



Automated Generation of Cybersecurity Exercise Scenarios

The proof-of-concept ML-powered exercise generation framework developed (AiCEF) is illustrated in Figure 4.

FIGURE 4: THE AICEF MODULES [5]



More concretely, to generate a CSE scenario using AiCEF, the EP must perform the following steps:

i. The MLCESO module converts relevant articles or free text into incident breadcrumbs, which are mapped to the CESO ontology and stored in the KDb database.
ii. IncGen generates several incidents based on provided meta tags by merging relevant incident breadcrumbs.
iii. The APT Enhancer module can be used to enhance the generated incidents by simulating known APT activity and filling in the missing information.
iv. CEGen allows the EP to create a CSE scenario by defining various attributes, including the CSE name, number of events and incidents.
v. The scenario is generated in STIX 2.1 format along with a state-of-the-world storyline.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

In the following paragraphs, we detail these steps and modules of interest for this work.

Machine Learning to CESO (MLCESO)

The most important step in our methodology is the creation of the ML pipeline that parses free text and extracts objects [46], [47] using named-entity recognition (NER), mapping them to CESO, as defined above. Separate models have been trained to cover the categories of tags as listed in Table I, which are later interlinked to form the STIX 2.1 enhanced incident graph.

TABLE I: ANNOTATION TAGS PER CATEGORY

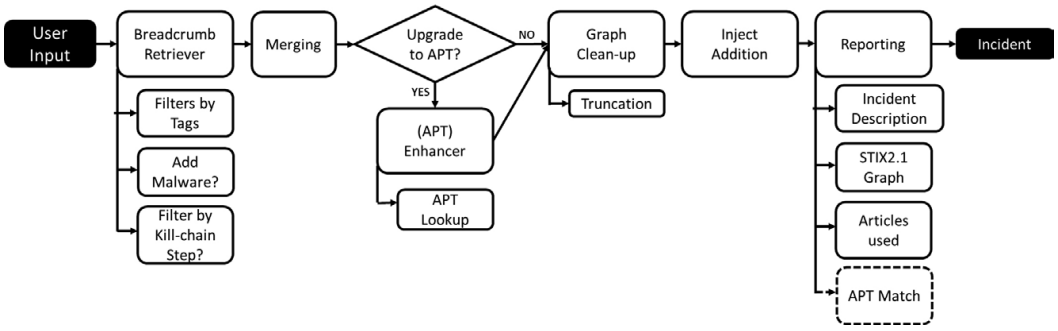
Category	(NER) Tag	Link to CESO and STIX 2.1
Attacker	ATTACKER_TYPE	Threat actor (attribute)
	ATTACKER_NAME	Threat actor (attribute), identity
	ATTACKER_ORIGIN	Location
Attack	MALWARE_TYPE	Malware (attribute)
	MALWARE_NAME	Malware (attribute)
	ATTACK_TYPE	Attack pattern
	VULNERABILITY	Vulnerability
Victim	SECTOR	Identity (attribute), scenario
	ASSETS	Threat actor (attribute)
	TECHNOLOGY	Tool

Incident Generation and Enhancement (IncGen) and Knowledge Database (KDb)

Incident creation is the most important step of the scenario generation procedure and consists of several steps to achieve maximum customization (see Figure 5). All the steps can be automated, generating a variety of incidents from which an EP can choose the fittest.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

FIGURE 5: INCIDENT GENERATION FLOW [5]



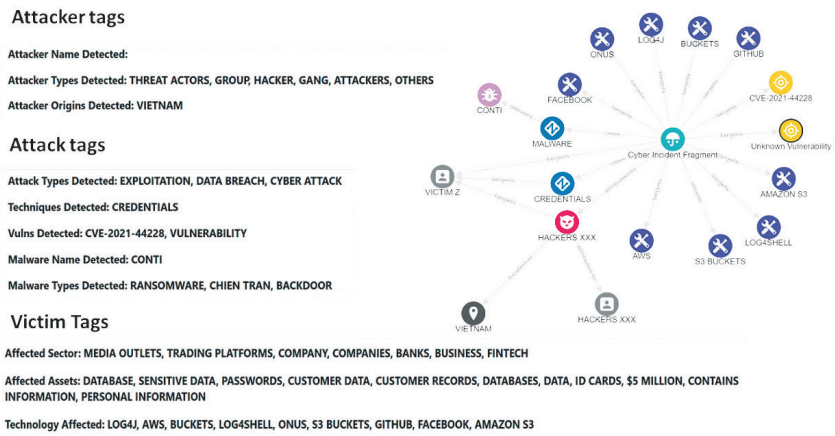
An output report and visualization of IncGen utilizing improved MLCESO tag detection can be seen in Figure 6.

The EP can provide specific text or articles for conversion to incidents or rely on a dynamic generation based on filtering parameters and a search of the existing database. The current knowledge database (KDb) consists of pre-parsed and modeled articles as described in Table II, spanning from 1 January 2020 to 1 March 2022.

TABLE II: KNOWLEDGE DATABASE CONTENT PER SOURCE

Source	Count
bleepingcomputer.com	1,368
securityaffairs.co	169
zdnet.com	495
databreaches.net	938
Total	**2,970**

FIGURE 6: INCIDENT GENERATION OUTPUT AND VISUALIZATION EXAMPLE

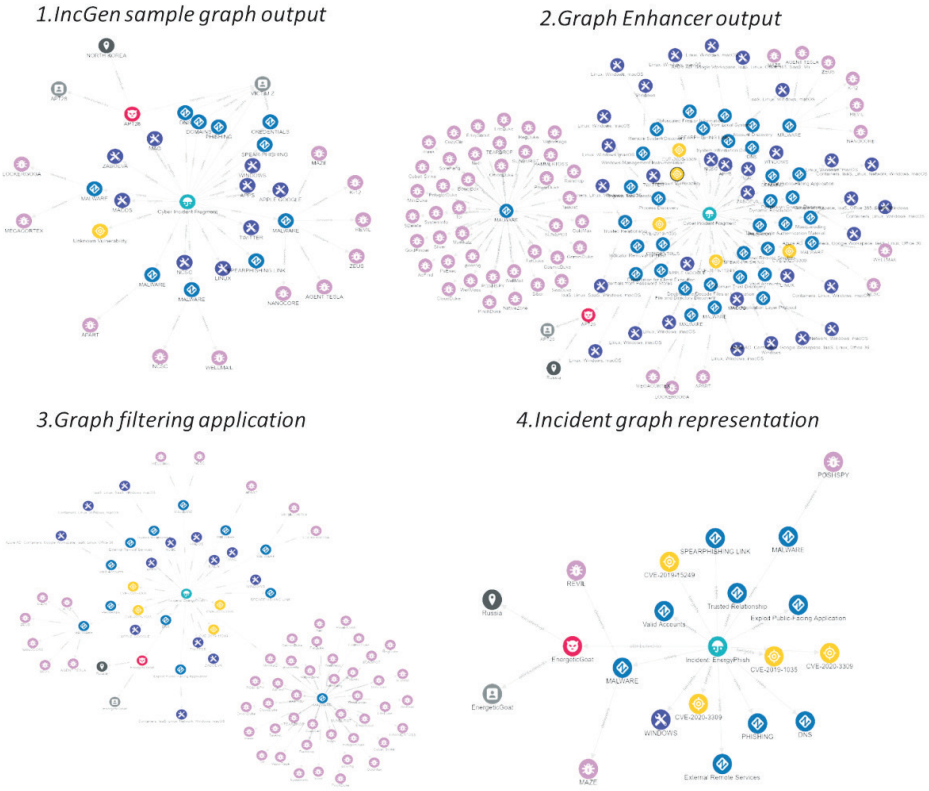


Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

APT Enhancer

Incidents can be enhanced with activity simulating the tactics, techniques and procedures (TTPs) of known APT actors. To simulate the activity of known APT groups, a basic STIX 2.1 structure was created per actor using the groups from MITRE, from which various attributes and TTPs were automatically extracted to populate our database. Thus, we generated a STIX 2.1 graph that can be used to compare and enhance other graphs at will. Currently, 125 APT actors can be simulated. A preview of the intermediate steps is shown in Figure 7.

FIGURE 7: APT ENHANCING AN INCIDENT IN FOUR STEPS (NO INJECTS ADDED)



4. PROPOSED METHODOLOGY

We propose a methodology to measure the effectiveness of AI-generated cyber security incidents for use in CSE scenarios. To do this, sets of incidents that occurred before and during a cyber conflict (pre-conflict and post-conflict sets) are collected and divided into sector-specific pools. The Pre-Conflict Pools are then compared to the actual incidents that took place (covered in the Post-Conflict Pool) to evaluate the methodology and determine the best strategy for generating CSE incidents with AI.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

Here is a brief description of the content of the pools generated:

Pre-Conflict Pools (per sector)

Pool 0 contains a random selection of incidents generated based on reports that have taken place prior to the conflict.
Pool 1 contains a relevant and targeted selection of incidents generated based on reports published prior to the conflict of interest.
Pool 2 contains relevant and targeted reports published prior to the conflict of interest, merged to formulate richer incidents.
Pool 3 contains relevant and targeted reports published prior to the conflict of interest, merged to formulate richer incidents enhanced based on known threat actors' TTPs.

Post-Conflict Pool (per sector)

Pool 0 contains three incidents: two relevant and targeted incidents generated based on reports published during the conflict of interest, and a merge of a relevant and targeted selection of incidents combined into a single richer incident.

A presentation of the implementation of our methodology using the AiCEF toolset can be found below.

PHASE 1: Incident Modeling (Pre-Conflict)

Step 1.1: Identify relevant pre-conflict incidents described in various resources and extract text

Step 1.2: Parse text using our MLCESO and store extracted objects in KDb using a PRE-CONFLICT meta tag

Step 1.3: Perform a trend analysis of the parsed incidents in Step 1.2 to discover the most prominent Attack Types & Techniques and Threat Actors and filter by sectors impacted

Step 1.4: Extract from KDb or generate synthetic incidents and form pools per impacted sector:

Pool 0: Extract related incidents over a threshold based on queries of the original KDb

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

Pool 1: Extract only rich incidents over a threshold with the PRE-CONFLICT meta tag.
Pool 2: Merge up to three[1] articles describing the same incident, using Incident Generator (IncGen). Incidents must have the PRE-CONFLICT meta tag.
Pool 3: Merge articles using Incident Generator (IncGen) and APT Enhancement based on known Threat Actors extracted in Step 1.3

PHASE 2: Incident Modeling (Post-Conflict)

Step 2.1: Identify relevant post-conflict incidents described in various resources and extract text

Step 2.2: Parse text using our MLCESO and store extracted objects in KDb using a POST-CONFLICT meta tag

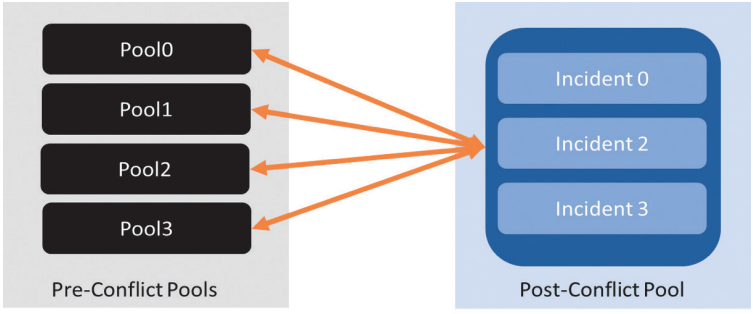
Step 2.3: Generate incidents and form a pool per affected sector:

i. Extract two rich incidents (Incident1-2) over a threshold
ii. Merging up to three relevant articles under a single Incident (Incident0)

PHASE 3: Comparison

Calculate the average coverage percentage of the Post-Conflict Pool versus the Pre-Conflict Pool sets to identify the best strategy in CSE creation using AiCEF.

FIGURE 8: COMPARISON PHASE



The comparison function used (Figure 8) to calculate coverage between modeled incidents can be summarized as follows:

[1] The number of similar articles of the same incident to be merged is a convention based on the average size of the pre-conflict incidents generated so as to be comparable to the one of Incident0 modelled post conflict.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

Step 3.1: We are about to compare PreConflictPool(i)= {Inc0, Inc1…} of varying size *i=n* to PostConflictPool = (Inc0,Inc1,Inc2) with *size=m* to calculate its coverage.

Step 3.2: For each IncX in PreConflictPool0 and PostConflictPool, we extract and store in a list the name attributes of the following objects: *Attack_Pattern, Tool, Vulnerability, Malware, Location* objects. We end up with the following lists:

PreConflictPool(i) = {Inc0[Attack_Pattern.name, Tool.name, Vulnerability. name, Malware.name, Location.name], …}

PostConflictPool(j) = {Inc0[Attack_Pattern.name, Tool.name, Vulnerability. name, Malware.name, Location.name], …}

Step 3.3: We perform a one-to-one comparison between the two lists and calculate the total coverage of the PostConflictPool Incident per PostConflictPool incident duplets as follows:

$$TotalCoverage(i) = \frac{AttackPattern_{\ Coverage} + Tool_{\ Coverage} + Vunerability_{\ Coverage} + Malware_{\ Coverage} + Location_{\ Coverage}}{5}$$

where an example coverage of a *Vulnerability* Object between two incidents can be calculated as the (Number of PreConflictPool(i).Inc0[*Vulnerability*] list items that exist in PostConflictPool(i).Inc0[*Vulnerability*]) list / (PostConflictPool(i). Inc0[*Vulnerability*] list Size) * 100.

Then, we calculate the AverageTotalCoverage of all PreConflictPool incidents per PostConflictPool Incident:

$$AvarageTotalCoverage = \frac{\sum_{i=0}^{n} TotalCoverage(i)}{n}$$

5. USE CASE AND RESULTS

Based on the methodology described above, we covered the following use case to help us assess whether the generated scenarios match the attack trends and prove that AiCEF can provide targeted and customized awareness in real life.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

We modeled incidents in the recent Russian-Ukrainian cyber conflict along with historical data into incident sets. The sets (pre-conflict and post-conflict) were parsed automatically and consisted of known instances of cyber incidents linked to attacks against Ukraine that have been described in publications and other reliable specialized sources. For this research, we analyzed stories reported by credible publications and reputable security researchers. Before including it in the dataset, each entry was reviewed manually to verify its relevance. We selected a total of 46 articles for the pre-conflict set and 14 for the post-conflict set and classified these into three sectors of interest: public administration, energy and ICT.

Finally, we modeled two exercises, one based on a selected pre-conflict set and the other based on the post-conflict set of incidents or relevant sectors to compare the overall training coverage achieved against future incidents.

PHASE 1: Incident Modeling (Pre-Conflict) in Practice

After implementing all the steps of our methodology, we derived the following outputs per step:

Step 1.1: We identified 46 articles as the basis for our research describing incidents that took place between January 2020 and January 2022. Although 24 February 2022 is considered the start of the conflict in our specific example, no articles or incidents covering February 2022 were used for our analysis.

Step 1.2 and Step 1.3: Performing a trend analysis of the parsed incidents in Step 1.2, we discovered a set of interesting Attack Types & Techniques and Threat Actors per sectors impacted, as illustrated in Table III.

TABLE III: TOP OBJECTS EXTRACTED

Sector	Attack Types & Techniques	Threat Actor
Public Administration	DOS, DDOS, MALWARE, DATA EXFILTRATION, PHISHING, DEFACEMENT	APT28, SANDWORM, TURLA, APT29
Energy	WIPER, RANSOMWARE, ESPIONAGE	STRONTIUM, KRYPTON, GAMAREDON, BLACKENERGY, TURLA, FANCY BEAR, SOFACY, NOBELIUM, UNC2452, PAWN STORM, SEDNIT
ICT	DOS, DDOS, WIPER, MALWARE	ACTINIUM, NOBELIUM, STRONTIUM

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

For actors such as APT28, Sandworm, Turla, and APT29, the findings were verified against the CCDCOE study of 2021 [4], and for ACTINIUM, against Microsoft's January 2022 report [48].

Step 1.4: Pools with the following size were then formatted per sector ([Pool0(40), Pool1(6), Pool2(40), Pool3(40)]), with 40 articles generated per Pool2-3 using AiCEF.

PHASE 2: Incident Modeling (Post-Conflict) in Practice
After implementing all the steps of our methodology, we derived the following outputs per step:

Step 2.1: We identified 14 articles covering the three sectors of interest for the period of March 2022 – June 2022.

Step 2.2 and Step 2.3:

i. Two incidents (Incident1-2) rich in objects were generated per sector.
ii. Both were merged into a single incident (Incident0).

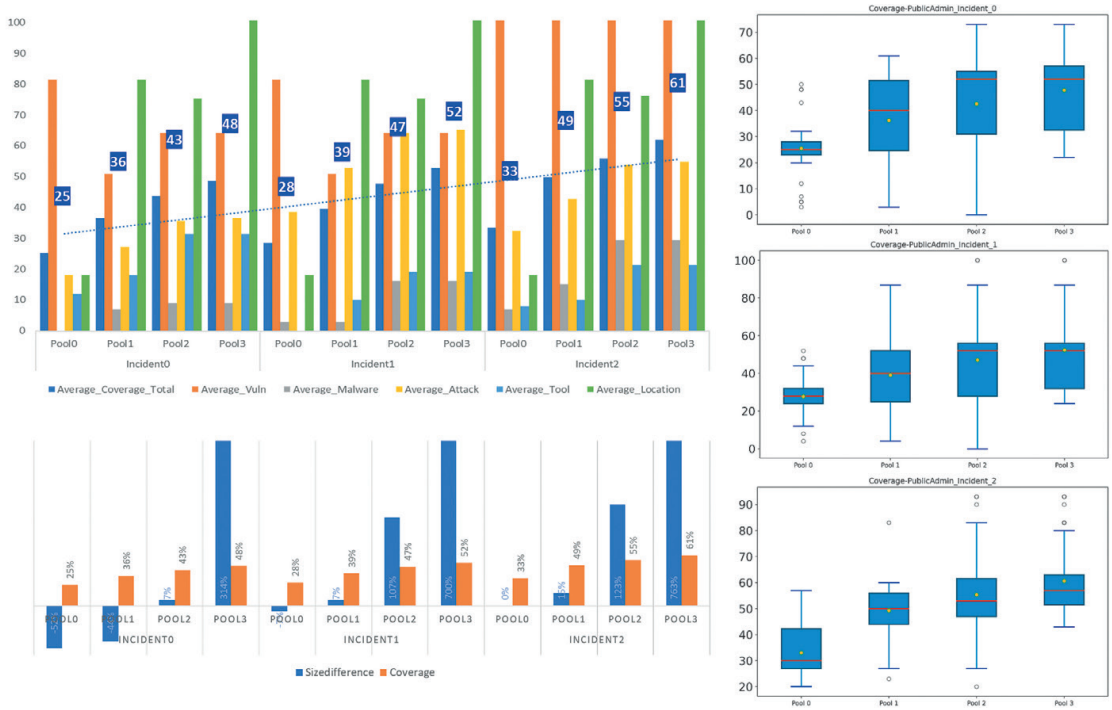
PHASE 3: Comparison and Results
After performing all possible comparisons, both at the incident and exercise level, the following coverage results were retrieved.

Public Administration
For the public administration sector, which was the most impacted in the conflict examined in terms of attack volume, we were able to determine the following, based on Figure 9:

i. Pre-Conflict Pool 2 Incidents covered the three post-conflict incidents selected with an improved rate of an average of approximately 20%. Overall, the coverage achieved was close to 50%, meaning that attacks used during the conflict were not similar to the ones performed before the conflict.
ii. APT enhancement considerably improved the coverage, but the complexity of the incidents and the overall exercise generated rose drastically, providing little training benefit.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

FIGURE 9: PUBLIC ADMINISTRATION INCIDENTS COVERAGE RESULTS



Energy Sector

The energy sector is considered the most critical to target during conflict. Taking its importance into account, we concluded that sufficient training coverage was achieved. More specifically as seen in Figure 10:

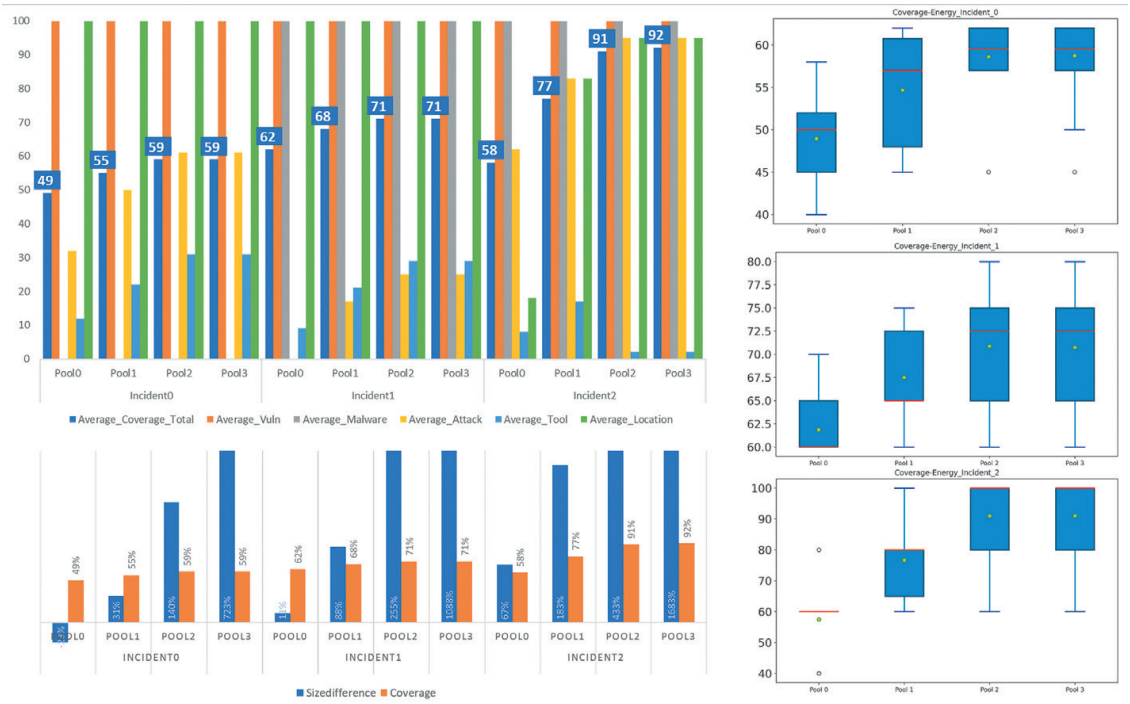
i. Pre-Conflict Pool 2 Incidents covered the three post-conflict incidents with an improved rate of an average of approximately 17%. Overall, the coverage achieved for Pool 2 was close to 74%, meaning that attacks used during the conflict were similar to the ones performed before the conflict, providing a good space for training and exercises to improve awareness.

ii. Although APT enhancement did not improve the coverage generated, incidents in both Pools 2 and 3 managed to cover 100% for the Post-Conflict Incident 2 with a constant rate of over 60% for all 40 incidents generated.

iii. Malware prediction, although low in the automatic statistics extracted seen below, can be considered satisfactory upon manual analysis. As an example, AiCEF proposed INDUSTROYER as the selected malware to exercise with, when in reality post-conflict incidents of the energy sector were populated by the INDUSTROYER v.2.0, a malware of the same family.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

FIGURE 10: ENERGY SECTOR INCIDENTS COVERAGE RESULTS



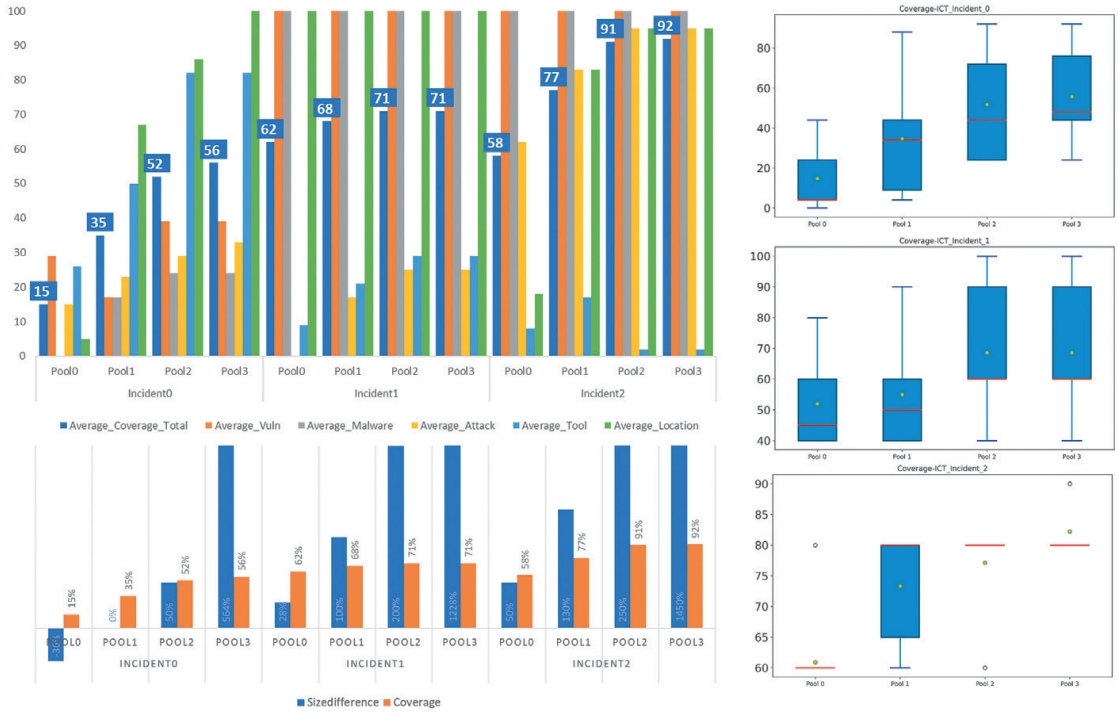
ICT Sector

For the ICT sector, based on our analysis represented in Figure 11:

i. Pre-Conflict Pool 2 incidents covered the three post-conflict incidents with an improved rate of an average of approximately 26%.
ii. The coverage achieved for Pool 2 was close to 72%, meaning that attacks used during the conflict were similar to the ones performed before the conflict.
iii. Although APT enhancement significantly improved the coverage, the generated incidents in both Pools 2 and 3 managed to cover 100% of the Post-Conflict Incident 1 with an average of 70% for all 40 generated incidents.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

FIGURE 11: ICT SECTOR INCIDENTS COVERAGE RESULTS



Targeted Exercise Coverage

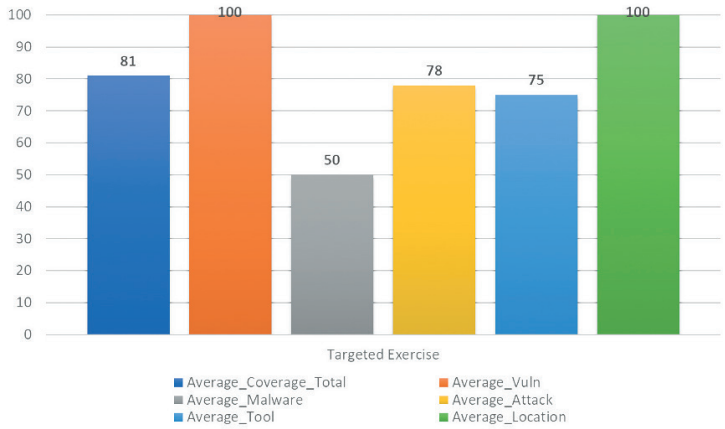
When comparing pre-conflict incidents with post-conflict incidents in the three sectors, observations made it clear that Pre-Conflict Pool 2 (relevant merged incidents) can achieve the highest coverage with a justifiable graph size difference against Post Conflict incidents. We decided to perform a final comparison, but now at the exercise level, by creating a CSE (named Targeted Exercise) consisting of one event with three incidents, one for each sector. The incident in each sector was a merge of all Incidents of the Pre-Conflict Pool 1 (removing duplicate objects).

We then created a post-conflict exercise by simply modeling a CSE consisting of one event with three incidents, which were the Incident 0's of the Post-Conflict Pool of each sector.

The exercise bundles were compared, and the overall average total coverage achieved was 81% (Figure 12) with a 304% size graph difference, meaning that our proposed exercise graph was three times greater than that of the post-conflict exercise based on real incidents.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

FIGURE 12: EXERCISE COVERAGE RESULTS



The specific vulnerabilities used by the attackers, along with the fact that the attackers used unknown vulnerabilities, were predicted by AiCEF at 100%. The framework was also good at detecting the attack techniques and tools used by the attackers in the incidents, paving the way for well-scoped exercise generation.

6. CONCLUSIONS

Based on analysis of the data collected in the above-described use case, we have reached some valuable conclusions on how to create insightful cyber security exercises that can prepare a country or a sector-specific organization ahead of a cyber conflict. Furthermore, with AiCEF we can now generate more realistic incidents and, as a consequence, better-scoped CSEs, following a statistically proven strategy that will offer the best coverage of training objectives in the expectance of future conflicts. The results of this study show that the Pre-Conflict Pool 1 incident generation strategy performs better than a random generation of sector-specific incidents. Using past incidents as a model for exercises can help achieve better coverage of exercise objectives up to an average of 13%.

The Pre-Conflict Pool 2 incident generation strategy, which involves merging multiple resources to describe the same incident, is the most effective in creating incidents using AiCEF, achieving an average of +21% overall coverage of the exercise objective set. The APT enhancement strategy (Pre-Conflict Pool 3) can achieve better overall coverage results but with a high complexity impact. In the future, more innovative methods should be used to enhance a graph based on known actors in a more effective manner.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

Additionally, exercises that cover multiple sectors impacted by the same actors can achieve higher coverage (81%) of training objectives than those covering a single sector (public administration: 50%; energy: 74%; ICT: 72%).

As a final remark, we can confidently state that AiCEF can automate the generation of well-scoped CSE scenarios to prepare a specific sector or a country ahead of a cyber conflict. Scenarios can be adapted in real time as the conflict progresses, with little cyber exercise planning expertise, through the power of AI, providing accurate training material.

ACKNOWLEDGMENTS

This work was supported by the European Commission under the Horizon Europe Programme, as part of the project LAZARUS (Grant Agreement no. 101070303).

The information and views set out in this article are those of the author (s) and do not necessarily reflect the official opinion of the European Union Agency for Cybersecurity (ENISA). Neither the European Union institutions nor any person acting on their behalf may be held responsible for any use that may be made of the information contained therein.

REFERENCES

[1] M. Karjalainen, T. Kokkonen, and S. Puuska, "Pedagogical aspects of cyber security exercises," in *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, IEEE, 2019, pp. 103–108.

[2] *Societal security—guidelines for exercises. Standard ISO22398:2013*, ISO Central Secretary, International Organization for Standardization, Geneva, CH, 2013. [Online]. Available: https://www.iso.org/standard/50294.html

[3] A. Conklin, "Cyber defense competitions and information security education: An active learning solution for a capstone course," in Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), IEEE, vol. 9, 2006, pp. 220b–220b.

[4] J. Hakala and J. Melnychuk, "Russia's Strategy in Cyberspace," NATO STRATCOM COE, 2021. [Online]. Available: https://stratcomcoe.org/cuploads/pfiles/Nato-Cyber-Report_15-06-2021.pdf

[5] A. Zacharis and C. Patsakis, "AiCEF: An AI-assisted cyber exercise content generation framework using named entity recognition," *International Journal of Information Security*, 2023.

[6] R. Gurnani, K. Pandey, and S. K. Rai "A scalable model for implementing cyber security exercises," in *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, 2014, pp. 680–684.

[7] S. F. Wen, M. M. Yamin and B. Katt, "Ontology based scenario modeling for cyber security exercise," in *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, IEEE, 2021, pp. 249–258.

[8] M. Mink and F. C. Freiling, "Is attack better than defense? Teaching information security the right way," in *Proceedings of the 3rd Annual Conference on Information Security Curriculum Development*, 2006, pp. 44–48.

[9] W. Schepens, D. Ragsdale, J. R. Surdu, J. Schafer, and R. N. Port, "The cyber defense exercise: An evaluation of the effectiveness of information assurance education," *Journal of Information Security*, vol. 1(2), pp. 1–14, 2002.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

[10] G. Vigna, "Teaching network security through live exercises," in *IFIP World Conference on Information Security Education*, Monterey, CA, USA, 2003, pp. 3–18.
[11] R. Dodge and D. J. Ragsdale, "Organized cyber defense competitions," in *Proceedings. IEEE International Conference on Advanced Learning Technologies*, 2004, pp. 768–770.
[12] W. J. Adams, E. Gavas, T. H. Lacey, and S. P. Leblanc, "Collective views of the NSA/CSS cyber defense exercise on curricula and learning objectives," presented at the 2nd Workshop on Cyber Security Experimentation and Test (CSET 09), Montreal, Canada, 2009.
[13] A. Conklin, "The use of a collegiate cyber defense competition in information security education," in *Proceedings of the 2nd Annual Conference on Information Security Curriculum Development*, 2005, pp. 16–18.
[14] Y. Li, M. Liljenstam, and J. Liu, "Real-time security exercises on a realistic interdomain routing experiment platform," in *2009 ACM/IEEE/SCS 23rd Workshop on Principles of Advanced and Distributed Simulation*, IEEE, 2009, pp. 54–63.
[15] M. Liljenstam, J. Liu, D. M. Nicol, Y. Yuan, G. Yan, and C. Grier, "Rinse: The real-time immersive network simulation environment for network security exercises (extended version)," *Simulation*, vol. 82(1), pp. 43–59.
[16] B. E. Mullins, T. H. Lacey, R. F. Mills, J. E. Trechter, and S. D. Bass, "How the cyber defense exercise shaped an information-assurance curriculum," *IEEE Security & Privacy*, vol. 5(5), pp. 40–49, 2007.
[17] B. E. Mullins, T. H. Lacey, R. F. Mills, J. M. Trechter, and S. D. Bass, "The impact of the NSA cyber defense exercise on the curriculum at the air force institute of technology," in *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, IEEE, 2007, pp. 271b–271b.
[18] A. Furtună, V. V. Patriciu, and I. Bica, "A structured approach for implementing cyber security exercises," in *2010 8th International Conference on Communications*, IEEE, 2010, pp. 415–418.
[19] V. V: Patriciu and A. C. Furtuna, "Guide for designing cyber security exercises," in *Proceedings of the 8th WSEAS International Conference on E-Activities and Information Security and Privacy*, World Scientific and Engineering Academy and Society (WSEAS), 2009, pp. 172–177.
[20] J. Kick, "Cyber exercise playbook," MITRE Corporation, Bedford, MA, USA, Tech. Rep., 2014.
[21] A. Green and H. Zafar, "Addressing emerging information security personnel needs. a look at competitions in academia: Do cyber defense competitions work?" *AMCIS 2013 Proceedings*, 1:257, 2013.
[22] J. A. Rursch, A. Luse, and D. Jacobson, "IT-Adventures: A program to spark it interest in high school students using inquiry-based learning with cyber defense, game design, and robotics," *IEEE Transactions on Education*, vol. 53(1), pp. 71–79, 2009.
[23] J. W. Schepens and J. R. James, "Architecture of a cyber defense competition," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance* (Cat. No. 03CH37483), IEEE, 2003, vol. 5, pp. 4300–4305.
[24] G. B. White, D. Williams, and K. Harrison, "The cyberpatriot national high school cyber defense competition," *IEEE Security & Privacy*, vol. 8(5), pp. 59–61, 2010.
[25] T. Augustine, R. C. Dodge *et al*., "Cyber defense exercise: meeting learning objectives thru competition," in *Proceedings of the 10th Colloquium for Information Systems Security Education*, 2006.
[26] M. Granåsen and D. Andersson, "Measuring team effectiveness in cyber-defense exercises: a crossdisciplinary case study," *Cognition, Technology & Work*, vol. 18(1), pp. 121–143, 2016.
[27] D. Schweitzer, D. Gibson, and M. Collins, "Active learning in the security classroom," in *2009 42nd Hawaii International Conference on System Sciences*, IEEE, 2009, pp. 1–8.
[28] G. B. White, G. Dietrich, and T. Goles, "Cyber security exercises: testing an organization's ability to prevent, detect, and respond to cyber security events," in *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, Big Island, HI, USA, 2004.
[29] J. A. Mattson, "Cyber defense exercise: A service provider model," in *IFIP World Conference on Information Security Education*, Springer, 2007, pp. 81–86.
[30] M. E. Planning, *Directors's guideline for civil defence emergency management groups, wyd*. (2008). Ministry of Civil Defence & Emergency Management, Wellington.
[31] F. Laamarti, M. Eid, and A. El Saddik, "An overview of serious games," *International Journal of Computer Games Technology*, 2014, doi: 10.1155/2014/358152.
[32] Y. Chou, *Actionable Gamification: Beyond points, badges, and leaderboards*. Packt Publishing Ltd, 2019.
[33] S. Trepte *et al*., "Do people know about privacy and data protection strategies? Towards the 'Online Privacy Literacy Scale' (OPLIS)," in *Reforming European data protection law*, Dordrecht: Springer, 2015, pp. 333–365.
[34] B. D. Cone *et al*., "A video game for cyber security training and awareness," *Computers & Security*, vol. 26.1, pp. 63–72.52, 2007.

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

[35] H. Berger and A. Jones, "Cyber security & ethical hacking for SMEs," in *Proceedings of the 11th International Knowledge Management in Organizations Conference on The Changing Face of Knowledge Management Impacting Society*, 2016, pp. 1–6.
[36] M. Hendrix, A. Al-Sherbaz, and V. Bloom, "Game based cyber security training: are serious games suitable for cyber security training?" *International Journal of Serious Games*, vol. 3(1), pp. 53–61, 2016.
[37] D. H. Tobey, "A vignette-based method for improving cybersecurity talent management through cyber defense competition design," in *Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research*, 2015, pp. 31–39.
[38] N. Wilhelmson and T. Svensson, *Handbook for Planning, Running and Evaluating Information Technology and Cyber Security Exercises*. Försvarshögskolan (FHS), 2011.
[39] U.S. Department of Homeland Security, 2013, "Cyber Tabletop Exercise (TTX)," Homeland Security Digital Library. [Online]. Available: https://www.hsdl.org/?abstract&did=789781
[40] B. Sangster, T. O'Connor, T. Cook, R. Fanelli, E. Dean, C. Morrell, and G. J. Conti, "Toward instrumenting network warfare competitions to generate labeled datasets," in CSET, 2009.
[41] T. Sommestad and J. Hallberg, "Cyber security exercises and competitions as a platform for cyber security experiments," in *Nordic Conference on Secure IT Systems*, Springer, 2012, pp. 47–60.
[42] M. Samejima and H. Yajima, "IT risk management framework for business continuity by change analysis of information system," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2012, pp. 1670–1674.
[43] K. A. Scarfone, T. Grance, and K. Masone, *Sp 800-61 Rev. 1. Computer Security Incident Handling Guide*, National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, USA, 2008.
[44] MITRE, 2022. [Online]. Available: https://attack.mitre.org/
[45] J. Pastuszuk, P. Burek, and B. Ksieopolski, "Cybersecurity ontology for dynamic analysis of it systems," *Procedia Computer Science*, vol. 192, pp. 1011–1020, 2021.
[46] R. MacIntyre, University of Pennsylvania, PA. 1995. *Penn Treebank Tokenizer*. [Online]. Available: https://www.cis.upenn.edu/~treebank/tokenizer.sed
[47] N. Tsinganos and I. Mavridis, "Building and evaluating an annotated corpus for automated recognition of chat-based social engineering attacks," *Applied Sciences*, vol. 11(22), 10871, 2021, doi: 10.3390/app112210871.
[48] Microsoft Threat Intelligence Center, 2022, "Destructive malware targeting Ukrainian organizations," Microsoft. [Online]. Available: https://www.microsoft.com/en-us/security/blog/2022/01/15/destructive-malware-targeting-ukrainian-organizations/

Authorized licensed use limited to: University of Piraeus. Downloaded on September 04,2024 at 15:06:52 UTC from IEEE Xplore. Restrictions apply.

International Journal of Information Security (2024) 23:2691–2710
https://doi.org/10.1007/s10207-024-00860-w

REGULAR CONTRIBUTION



Integrating AI-driven threat intelligence and forecasting in the cyber security exercise content generation lifecycle

Alexandros Zacharis[1] · **Vasilios Katos**[2] · **Constantinos Patsakis**[3,4]

Published online: 10 May 2024
© The Author(s) 2024

Abstract
The escalating complexity and impact of cyber threats require organisations to rehearse responses to cyber-attacks by routinely conducting cyber security exercises. However, the effectiveness of these exercises is limited by the exercise planners' ability to replicate real-world scenarios in a timely manner that is, most importantly, tailored to the training audience and sector impacted. To address this issue, we propose the integration of AI-driven sectorial threat intelligence and forecasting to identify emerging and relevant threats and anticipate their impact in different industries. By incorporating such automated analysis and forecasting into the design of cyber security exercises, organisations can simulate real-world scenarios more accurately and assess their ability to respond to emerging threats. Fundamentally, our approach enhances the effectiveness of cyber security exercises by tailoring the scenarios to reflect the threats that are more relevant and imminent to the sector of the targeted organisation, thereby enhancing its preparedness for cyber attacks. To assess the efficacy of our forecasting methodology, we conducted a survey with domain experts and report their feedback and evaluation of the proposed methodology.

Keywords Cyber security exercise scenarios · Machine learning · Threat intelligence · Threat forecasting

1 Introduction

Cyber Security Exercises (CSE) have gained prominence in the training landscape, providing hands-on experience to personnel across various industries. A CSE, as described in the ISO Guidelines for Exercises [31], is "*a process to train for, assess, practice, and improve performance in an organisation*". However, the effectiveness of these exercises is often hindered by the inability to replicate real-world scenarios in a timely manner relevant to the sector methodology. Our research aims to identify and bridge the gap in context-relevant CSE scenarios generated by sectorial cyber security experts and seasoned exercise planners. To address this challenge, this work proposes integrating AI-driven sectorial tag coverage techniques into the CSE scenario generation process. Although threat intelligence and forecasting have been used for various cyber security tasks, including CSE scenario design, our fully automated AI-driven methodology is the first of its kind to incorporate forecasting and scenario generation in a single flow. Using our approach, organisations can generate sector-specific threat intelligence that improves the realism and relevance of future CSE scenarios, especially when integrated into exercise generation frameworks such as the AI-assisted Cyber Exercise Framework (AiCEF) [66].

The research questions we aim to address it this work are:

- Can we produce reliable sectorial threat landscapes using AI models and methodologies?
- How much can experts and nonexperts developing CSE scenarios benefit from our methodology?

To address these questions, the following objectives have been set for this work and can be summarised as follows:

Constantinos Patsakis
kpatsak@unipi.gr

Alexandros Zacharis
alexandros.zacharis@enisa.europa.eu

Vasilios Katos
vkatos@bournemouth.ac.uk

[1] European Union Agency for Cybersecurity (ENISA), Chalandri, Greece

[2] Bournemouth University, Poole, UK

[3] Department of Informatics, University of Piraeus, Karaoli & Dimitriou 80, 18534 Piraeus, Greece

[4] Athena Research Center, Marousi, Greece

1. Develop an AI-driven threat intelligence and forecasting methodology that accurately identifies current and emerging trends with their potential impact in different sectors from any given dataset and incorporates them into a CSE scenario. To achieve this objective, we evaluate several existing forecasting algorithms using quantifiable criteria to identify those that provide accurate predictions for a given dataset.
2. Evaluate the effectiveness of integrating AI-driven sectorial threat intelligence and forecasting into the exercise generation process by comparing the realism and relevance of CSEs created using our methodology versus the traditional human experts' approach. Additional effort was devoted to implementing our methodology into a fully automated prompt generation methodology that non-expert exercise planners can easily use. To achieve this objective, we have conducted surveys based on a case study to measure the perceived relevance of CSE with AI-driven threat intelligence and forecasting compared to traditional exercise generation. No other AI-driven exercise generation framework exists against which we can compare our methodology. For this reason, we compared our methodology against human-based approaches by tasking experts in this field. We have then collected quantitative results that help us compare our methodology against the human expert's methods used in identifying current and emerging threats and compare the results against those of renounced threat landscapes used as a baseline.

Our analysis, coupled with valuable insights from field experts, has yielded tangible results affirming that the integration of our proposed methodology into the Cyber Security Exercise generation process significantly improves the realism and relevance of CSE scenarios. This enhancement can drastically improve organisations' ability to confront real-world cyber threats effectively.

The structure of the remainder of this work is as follows. In Sect. 2, we delineate the current state-of-the-art about on AI-driven techniques to enhance cyber security exercises and improve organisations' preparedness for emerging threats, focusing on machine learning techniques used for threat forecasting. Our methodology is presented in Sect. 3, where we deep dive into its implementation, focusing on data collection, data and trend analysis, topic modelling and extraction. We then compare different forecasting and trend prediction models and evaluate them for use in the following steps. All previous steps are incorporated into our AI-driven prompt generation (AiDPG) tool, which can be used to create tailored CSE scenarios. We present our evaluation results in Sect. 4, using a case study focusing on the energy sector and involving both sectorial and exercise experts. We conclude this article by presenting the lessons learned and future work in this area of interest in Sect. 5.

2 Related work

The following paragraphs provide a brief overview of the related work. First, we discuss the relevant forecasting models, focusing only on the machine learning models we used in our study. Then, we discuss the trends in generating CSEs and the use of forecasting methods.

2.1 Forecasting models

The main forecasting models that we consider in the scope of this work, along with their limitations, are the following.
SARIMAX (Seasonal Auto Regressive Integrated Moving Average with Exogenous Variables) [9] is a popular statistical model for time series forecasting. It extends the ARIMA (AutoRegressive Integrated Moving Average) [9] model by incorporating seasonal patterns and exogenous variables. Unlike ARIMA, which effectively analyses and predicts time series data with stationary or non-stationary properties, SARIMAX captures the relationships between past observations, forecast errors, and external factors to provide more accurate forecasts. However, SARIMAX models assume linearity and stationarity in the data, which may not always be true in real-world scenarios. SARIMAX models do not handle missing data or outliers well, which requires additional preprocessing steps. Interpreting the coefficients and understanding the impact of exogenous variables can be complex.
Exponential Smoothing (ES) [28, 30] is a widely used time series forecasting technique that assigns exponentially decreasing weights to past observations. It is based on the assumption that recent observations have greater significance in predicting future values. Yet, ES may struggle to capture complex non-linear patterns in the data.
Simple Exponential Smoothing (SES) [29] is a specific form of exponential smoothing that only considers the current observation and a single smoothing parameter (alpha). It provides a straightforward method for forecasting by exponentially weighting past observations and updating the forecast based on the weighted average.
Nonetheless, SES assumes a constant level and does not capture trends or seasonality, making it less suitable for data with such patterns. Moreover, it may struggle with data that have irregular or changing patterns over time and is more suitable for short-term forecasting rather than long-term predictions.
Prophet [58] is a forecasting framework developed by Facebook's Core Data Science team. It combines the components of additive regression models with flexible seasonality estimation. Prophet incorporates trend components, seasonal

effects, and additional regressors to generate forecasts. It also effectively handles outliers and missing data points. However, Prophet assumes that trends and seasonality are additive and follow specific patterns, which may not always hold in diverse datasets. The automatic change-point detection in Prophet might not always accurately identify complex or abrupt changes.

Long Short-Term Memory (LSTM) [20, 24] is a type of **recurrent neural network (RNN)** architecture commonly used to model and forecast time series data. LSTMs are designed to overcome the limitations of traditional RNNs in capturing long-term dependencies and handling vanishing or exploding gradients. In LSTM models, the network learns to selectively retain or forget information from past observations, allowing it to capture and retain important patterns over longer sequences.

On the other hand, training LSTM models requires more data to capture long-term dependencies and avoid over-fitting effectively. LSTM models can be computationally intensive. Hyper-parameter tuning for LSTM models can be challenging, as the optimal architecture and parameter settings may vary depending on the specific dataset and problem.

2.2 Cyber security exercise scenarios and forecasting

The integration of AI-driven techniques into CSEs has gained significant traction in recent years as organisations seek to enhance their preparedness for emerging threats, enhancing traditional CSE scenario-building methodologies. AI-assisted cyber exercise content generation framework (AiCEF) [66] harnesses the power of Named Entity Recognition (NER) to extract tags based on a Cyber Exercise Scenario Ontology (CESO) through its Machine Learning to Cyber Exercise Scenario Ontology (MLCESO) conversion module. Successful CSEs hinge on the effectiveness of their scenarios, which must simulate real-world situations that resonate with participants while maintaining a level of realism that fosters engagement and learning. Realistic scenarios, as conceptualised in traditional exercise scenario-building methodologies [5, 17], serve as powerful tools for predicting future events, incorporating relevant issues, interactions, and potential consequences. Examples of realistic scenarios modelled from past cyber conflicts can be found in [67] partially foreseeing future cyber events. These scenarios, according to Green and Zafar [23] and Granåsen et al. [22], lead to constructive training experiences that enhance participant understanding and preparedness.

More concretely, an exercise scenario is a sequential narrative account of a hypothetical incident that catalyses the exercise and intends to introduce situations that will inspire responses and thus allow demonstration of the exercise objectives [51]. ENISA acknowledges the benefits of sectorial exercises scenario in their National Exercise - Good Practice Guide [12], an approach recognised by national cyber security authorities like ANSII in their guide on organising cyber crisis management exercises [3]. In the context of CSEs, a scenario defines the training environment that will lead participants towards fulfilling the exercise objectives set [33]. The cyber security problem described in the scenario itself portrays a structured representation named Master Scenario Events List (MSEL), which serves as the script to execute an exercise [51].

To generate a useful MSEL, we need to successfully detect and classify cyber attacks relevant to the needs of the trainee, as emphasised in [36, 47, 49]. Interestingly, applying advanced machine learning techniques to predict future cyberattack trends has gained significant momentum in cyber security. Long Short-Term Memory (LSTM) networks constitute a specific type of recurrent neural network that has been particularly noted for their ability to learn long-term dependencies and patterns in time-series data, which is characteristic of cyber attack vectors. For instance, [34] demonstrated the efficacy of LSTMs in classifying intrusion detection patterns, emphasising their potential to identify sophisticated cyber threats by analysing historical data patterns. Similar studies to ours can be considered in the cases of [1, 6, 15, 19] where LSTM networks are proposed as a method to forecast cyber incidents with a focus on malware infections.

Parallel to the application of LSTMs, the Neural Prophet model has emerged as a novel approach for time-series forecasting [58]. Although specific studies focusing on the use of this model in cyber attack predictions are scarce, its robustness in detecting complex patterns is essential to predict cyber incidents, as reported by [40] with promising results. In the realm of more traditional statistical methods, ARIMA models have been a mainstay for time-series forecasting and have found applications in cyber-security. Sarker et al. [50] discuss integrating machine learning approaches like ARIMA in cyber-security, highlighting its effectiveness, especially in scenarios with limited computational resources or where model interpretability is vital. Following the same path, data mining and machine learning techniques can be applied to VSE scenario development to help develop a mitigation plan for future incident trends and cyber security incident timelines as proposed in [26].

Further to analysing existing threats, modelling [64] predicting upcoming cyber incidents [37, 55] and the overall threat trends are equally important. Several studies focus on the short-term prediction of the number or source of attacks to be expected in the following hours or days [2, 7, 10, 11, 21, 25, 27, 37, 39, 41–43, 45, 61–63]. Most of these works make predictions in restricted settings; e.g., against a specific entity or organisation for which historical data are available [39, 43, 63]. Forecasting attack occurrences has been attempted with statistical methods, especially when parametric data distribu-

tions could be assumed [62, 63], as well as by using machine learning models [21] and deep learning [52]. Other methods adopt a Bayesian setting and build event graphs suitable for estimating the conditional probability of an attack following a given chain of events [45]. Such techniques rely on libraries of predefined attack graphs to identify the known attacks that are most likely to occur. Research has also covered zero-day attacks [54, 57] with less impressive results.

Almahmoud et al. [2] introduced a machine learning based proactive approach for long-term cyber-attack prediction, allowing early detection of potential threats. This approach could facilitate cyber security experts in prioritising resources and mitigation measures, including training through CSEs, a type of proactive cyber defence [56].

The existing literature showcases the potential for AI-driven topic extraction and forecasting techniques to be applied in the identification and prioritisation of both current and future threats in the context of CSEs, minimising their impact upon materialisation [7, 16]. Our work is the first to infuse the outputs of fully automated AI-driven threat intelligence and forecasting into the CSE content generation life cycle. Comparative results are provided against the current, traditional CSE scenario development methodologies used, which are a combination of manual and human-driven effort, underlining the potential of our methodology.

3 Proposed methodology

3.1 Generic overview

While in this section, we outline the proposed methodology for analysing a dataset of cyber security incidents using the AiCEF framework, specifically the MLCESO module; it is generic enough to allow its individual use in other domains and applications. Additionally, we propose prompts based on these trends, prioritising the needs per sector through our AI-driven Prompt Generation (AiDPG) methodology. In terms of the threat intelligence collection methodology used, we decided to follow some of the steps proposed by ENISA [13] to create our own sectorial threat intelligence (STI) report with the difference of using AI exclusively as a means of data processing instead of human analysts. In what follows, when referring to sectors, we will refer to those considered in the NIS 2 Directive [14]; however, any other such classification can be used. The outline of our methodology is illustrated in Fig. 1, consisting of five distinct steps.

In what follows, we assume access to a comprehensive and representative dataset of cyber security incidents, in text format, that spans over several years. Moreover, we assume that all incidents are unique to prevent biases in our predictions. Nevertheless, to exemplify, as we discuss in the following paragraphs, we point the reader to a manually curated dataset,

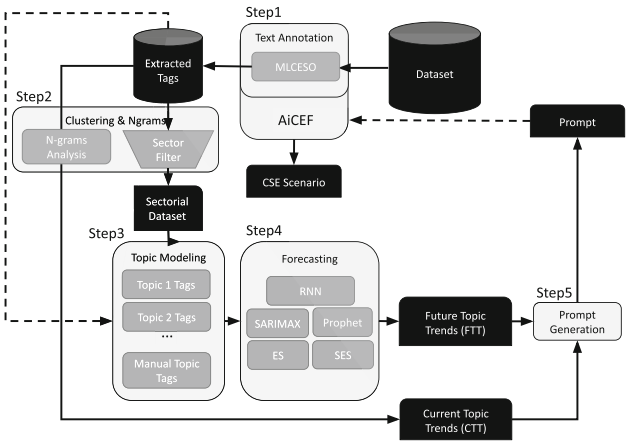


Fig. 1 Proposed methodology overview

and we use it as a reference for showcasing and validating our methodology.

1. **Text annotation using AiCEF** The AiCEF framework is employed, being the only available CSE-specific framework, and we use the MLCESO module for preprocessing (Paragraph 3.2) and text annotation, extracting relevant exercise-related objects, such as attacker type, attack type, and exploited vulnerabilities (Paragraph 3.3).
2. **Clustering and N-Gram Analysis** Incidents are clustered into sectors based on the extracted word tags, to identify similarities and group similar incidents for a more detailed N-gram analysis (Paragraph 3.4).
3. **Topic Modelling** We apply the Latent Dirichlet Allocation (LDA) algorithm for topic modelling, extracting topics within the context of elements that might be involved in a CSE scenario (Paragraph 3.5). Given the increasing use of large language models (LLMs) and their exceptional results in multiple domains [18, 32, 44, 59], we anticipate that soon LLMs will be able to automate these steps of our methodology relatively efficiently.
4. **Forecasting** Occurrence analysis is performed to capture the incident frequency, patterns, and trends over time, extracting insights about the temporal aspects of the data. The latter allows us to identify current and future trends, patterns, anomalies, and emerging trends in cyber security incidents within each sector. Although any single time series analysis algorithm can be used for this step, we decided to put SARIMAX, ES, SES, Prophet, and LSTM to the test and evaluate them using R-squared (R^2) [35], mean absolute error (MAE) [35], mean absolute deviation (MAD) [60] and mean absolute percentage error(MAPE) [4] metrics to identify the best-performing ones for this task (Paragraph 3.6).
5. **Prompt Generation** An exercise prompt is considered a detailed representation of a CSE scenario, containing

all the important cyber security elements that will be later taken into consideration to draft a CSE. Actionable prompts are proposed back to AiCEF via our AiDPG methodology, using the identified trends to address cyber security challenges and mitigate risks but also prioritising training topics based on sector-specific requirements and characteristics (Paragraph 3.7).

3.2 Dataset and data preprocessing

As shown in Table 1a, four incident sources have been identified as input to generate our corpus. These websites contain unique articles previously selected and classified in the Hackmageddon dataset,[1] a reputable webpage that describes cyber security incidents since 2011. For simplicity, in this work, we collected unique incidents from 2019–01 to 2022–12 (Table 1b), which accounts for 4.957 articles. We spanned our collection over four years to adhere to the limitations paused by the time series analysis algorithms to be used in later steps. Note that while there were undoubtedly attacks before 2019, neither the amount of information nor the depth can be comparable to today's figures. The issuance of the Presidential Policy Directive 41 (PPD-41) in 2016 in the USA, titled "United States Cyber Incident Coordination" and in parallel with the introduction of the General Data Protection Regulation (GDPR) in the same year, were the main drivers in pushing organisations both private and public to report cyber security incidents. Therefore, even if significant cyber security incidents existed before 2016, they were not reported, let alone published. As a result, if one simply compares the number of reports before 2019 that GDPR was enforced, one would notice a massive surge in the numbers, as the bulk of incidents were hidden under the carpet. Of course, that would introduce severe biases in our dataset. Therefore, we opt to use a smaller yet more representative dataset.

All relevant articles were collected through automated web scraping. Then, the plain text of each article was processed using Natural Language Processing (NLP) techniques to form the reduced Incidents Corpus (IC). Initially, all texts were converted to the UTF-8 encoding scheme. Using dictionaries and the Textblob library,[2] we corrected spelling and removed special characters. Empty lines, specific stopwords, and punctuation marks were removed using traditional NLP libraries such as NLTK[3] and spaCy. [4] In addition, all HTML and other programming codes, URLs, and paths were removed.

The standard Penn Treebank [38] tokenisation rules were used for sentence tokenisation to tune the incidents' text and facilitate annotation. At the end of this step, a corpus composed of timestamped incidents was formed.

Table 1 Composition of our dataset

(a) dataset quota per source

Source	# Incidents
`bleepingcomputer.com`	2431
`securityaffairs.co`	339
`zdnet.com`	794
`databreaches.net`	1393
Total	**4957**

(b) Dataset yearly distribution

Year	# Incidents
2019	1088
2020	1348
2021	1248
2022	1353
Total	**4957**

3.3 Text annotation

To annotate our dataset, pre-trained named entity recognition models (NER) (Table 2) were used. These models are part of AiCEF's MLSECO module and can annotate text with predefined tags, as described in Table 3. The output of the annotation step was stored in a Knowledge Database (KDb). An example of the information stored in KDb can be seen in Table 4.

3.4 N-gram analysis

By applying NER and extracting important keywords related to the above categories, we have already taken a crucial step in preparing the data for N-gram analysis, a computational technique to extract meaningful patterns and insights from text data. It involves identifying and analysing sequences of N consecutive words (N-grams) that occur frequently within a given corpus. In the context of processing cyber security incidents, N-gram analysis can provide valuable information about the prevalent attack types, techniques, attacker names, malware types, assets, vulnerabilities, and sectors affected by these incidents. Thus, we can leverage this technique to gain a deeper understanding of the relationships and trends in the cyber security incidents that we have in our database.

By performing a top-ten terms analysis (for Attack Types, Techniques, Attackers, Malware Types, Assets, Vulnerabilities, and Sectors) over four years' worth of data in KDb, we were able to visualise the results after clustering similar notion phrases, where deemed necessary. Furthermore, we can uncover relationships between different categories. For

[1] https://www.hackmageddon.com

[2] https://github.com/sloria/TextBlob

[3] https://www.nltk.org/

[4] https://spacy.io/

Table 2 AI models' scores

Category	Tag	Presicion	Recall	F1
Attacker	ATTACKER_TYPE	100.00	83.33	90.11
	ATTACKER_NAME	95.29	87.10	91.01
	ATTACKER_ORIGIN	Used Native Spacy LOC tag (no training)		
Attack	MALWARE_TYPE	80.56	76.32	78.38
	MALWARE_NAME	95.29	87.10	91.01
	ATTACK_TYPE (TECHNIQUE)	88.60	87.07	87.83
	VULNERABILITY	87.50	84.00	85.71
Victim	SECTOR	85.84	84.07	84.95
	ASSETS	87.02	89.06	88.03
	TECHNOLOGY	87.60	89.93	88.70

Table 3 Annotation tags per category

Category	Tag	Link to CESO & STIX 2.1
Attacker	ATTACKER_TYPE	Threat Actor Attribute
	ATTACKER_NAME	Threat Actor Attribute, Identity
	ATTACKER_ORIGIN	Location
Attack	MALWARE_TYPE	Malware Attribute
	MALWARE_NAME	Malware Attribute
	ATTACK_TYPE (TECHNIQUE)	Attack Pattern
	VULNERABILITY	Vulnerability
Victim	SECTOR	Identity Attribute, Scenario
	ASSETS	Threat Actor Attribute
	TECHNOLOGY	Tool

instance, frequent co-occurrence of N-grams like "APT3" and "PlugX malware" may indicate the involvement of specific advanced persistent threat groups and the utilisation of sophisticated malware in cyber attacks.

To showcase an example of threat intelligence extracted for the given dataset, we provide the output of our methodology for the sector tag. After analysing the 4957 cyber security incidents, the following distinct clusters or words representing at least seven heavily impacted sectors were formulated:

- **Commerce:** COMMERCE, COMPANY, COMPANIES, BUSINESS, BUSINESSES, RETAIL SECTOR, ENTERPRISE, ENTERPRISES, ORGANIZATIONS, ORGS, RETAIL, FIRMS, FIRM
- **ICT&Technology:** E-COMMERCE, GAMING, INFRASTRUCTURE, ITECH, ELECTRONICS, IT SERVICES, INTERNET SERVICE PROVIDERS, MARKETPLACE, ONLINE SERVICES, INFORMATION TECHNOLOGY, CLOUD HOSTING, SERVICE PROVIDER, HOSTING, TECHNOLOGY, TECH, CLOUD STORAGE, CLOUD STORAGE SERVICES, NETWORK OPERATOR, IT, TELECOMS, TELCO, TELECOMMUNICATIONS PROVIDER, TELECOMMUNICATIONS, TELECOM, CLOUD SERVICES
- **Finance:** BANKS, STOCK EXCHANGE, BANK, FINANCE, FINANCIAL, BANKING SECTOR
- **Healthcare:** HEALTHCARE, CLINICS, CLINIC, HOSPITAL, MEDICAL PROVIDERS, HEALTHCARE SYSTEM, HEALTH, MEDICAL CENTER, MEDICAL, PHARMACEUTICAL
- **Education:** EDUCATION, SCHOOLS, SCHOOL, UNIVERSITY, UNIVERSITIES, COLLEGE, EDUCATION SECTOR, EDUCATIONAL, HIGHER EDUCATION
- **Government:** GOVERNMENT, PARLIAMENT, MINISTRY, MINISTRIES, GOVT, GOVTS, GOVERNMENTS, GOVERNMENT AGENCIES, POLITICAL, MUNICIPALITIES, ELECTION, CITY, CITIES, FEDERAL AGENCIES

Further to this, smaller clusters of words representing less impacted sectors also emerged, as seen below:

- **Energy:** ENERGY, ENERGY SECTOR, NUCLEAR, OIL, GAS, POWER, PETROL, POWER SUPPLY
- **Defence:** DEFENCE, DEFENSE, LAW ENFORCEMENT, LAW ENFORCEMENT AGENCIES, LAW ENFORCEMENT AGENCY, POLICE, MILITARY

Table 4 Annotation example using AiCEF (MLCESO)

Field	Value
id	2246
Name	202203-2021-0153-58be06f8-f18c-4e1b-a243-9ecb13442636
URL	https://www.bleepingcomputer.com/news/security/bluenoroff-hackers-steal-crypto-using-fake-metamask-extension/
Text	BlueNoroff hackers steal crypto using fake MetaMask extension..
Month	1
Year	2022
Maturity	150
MetaTags	LINUX, SCRIPTS, HACKS, POWERSHELL, CRYPTOCURRENCY, THREAT ACTOR, PASSWORD, MALWARE, EXCEL, WORD, USER CREDENTIALS, CREDENTIALS, POWERSHELL, CHROME EXTENSION, WINDOWS, HACKERS, FILES, ATTACKERS, NORTH KOREA, BROWSER, BROWSER EXTENSIONS, CVE-2017-0199, DATA BREACH, TEMPLATE INJECTION, WALLET, FINANCIAL, GROUP, VULNERABILITY, BLUENOROFF, THREAT ACTORS, DATA
Attacker_Name	BLUENOROFF
Attacker_Type	HACKERS, GROUP, THREAT ACTOR
Threat_Actor_Type	crime-syndicate
Attack_Type	DATA BREACH
Technique	CREDENTIALS, POWERSHELL
Vuln	TEMPLATE INJECTION, VULNERABILITY, CVE-2017-0199
Malware_Name	
Malware_Type	SCRIPTS, MALWARE
Attacker_Origin	NORTH KOREA
Sector	FINANCIAL
Assets	FILES, WALLET, USER CREDENTIALS, CRYPTOCURRENCY, PASSWORD, DATA
Technology	WINDOWS, LINUX, EXCEL, WORD, BROWSER, POWERSHELL, BROWSER EXTENSIONS, CHROME EXTENSION

- **Aerospace:** SPACE, AEROSPACE, AVIATION
- **Media:** MEDIA, MEDIA OUTLETS, NEWS AGENCY, NEWSPAPER, RADIO STATION
- **Maritime:** MARITIME, PORT, SHIP, SHIPS
- **Food:** FOOD, FOOD CHAIN

According to the analysis dataset from 2019 to 2022, the ten largest sector clusters affected are represented per yearly occurrence in total volume in Fig. 2. Since some incidents could have impacted more than one sector, they appear in multiple sectors affected.

Following the same process for all other labels: Assets, Vulnerabilities, Attack Types, Attacker Names, Malware Types, and Malware Names, we can derive the respected threat intelligence. This information is then stored in our KDb per sector, providing us with the Sectorial Threat Intelligence (STI) needed to enrich CSE scenarios.

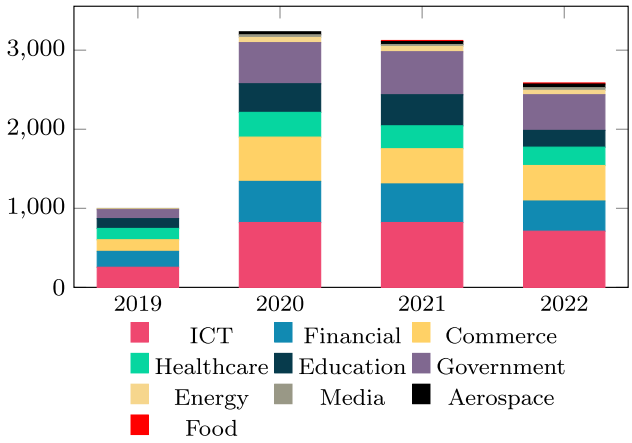


Fig. 2 Top 10 sectors impacted between 2019 and 2022 in volume

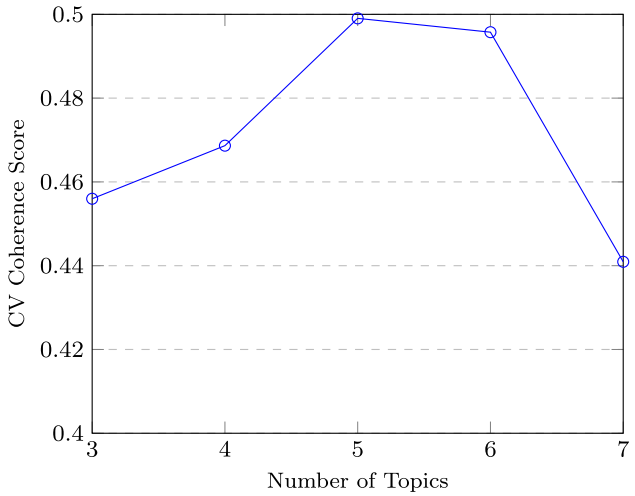


Fig. 3 CV coherence score for different numbers of topics

3.5 Clustering through topic modelling

Topic modelling, specifically key phrase cluster analysis, plays a significant role in understanding and extracting meaningful insights from text corpora. Although any topic modelling algorithm like BERTopic and non-negative matrix factorization (NMF) can be used for this step, we decided to incorporate LDA in our methodology. [8, 46, 53]. LDA is a generative statistical model used to discover topics within a collection of documents. It assumes that each document is a mixture of various topics, and each topic is a distribution over words. LDA scales well with large datasets, making it suitable for analysing extensive collections of documents. LDA is often preferred when dealing with a large text corpus, as NMF can be computationally more expensive. BERTTopic, like NMF requires more computational resources, however, the interpretability of the extracted topics can be challenging, especially compared to methods like LDA [65].

To properly implement the LDA algorithm, the optimal number of topics must be first identified based on a coherence score usually calculated using coherence models such as the CV coherence [48]. The coherence score measures the degree of semantic similarity between words within each topic, and its value can range from zero to one. Comparing the coherence scores obtained for different numbers of topics, the number of topics that yields the highest coherence score is considered the optimal number of topics for the given corpus.

Following the above steps, we conclude that the optimal number of topics to be extracted for the given dataset is five (5), as illustrated in Fig. 3.

Once we trained our LDA model and settled for five topics, we extracted a set of unique words per topic, allowing us to categorise them with specific tags, reported in Table 5

3.6 Forecasting, trend prediction and model evaluation

As already discussed, we tested various time series analysis methodologies, such as SARIMAX, Exponential Smoothing (ES), Simple Exponential Smoothing (SES), Prophet, and Long Short-Term Memory (LSTM) modelling using Recurrent Neural Networks (RNN), following recommendations of previous research findings [2]. Nonetheless, our goal was not to perform an exhaustive comparison of all possible time series analysis methodologies but to assess whether time-series analysis can support our effort of generating relevant cyber security scenarios with a limited incident dataset (of up to four years in this case) of cyber security incidents in hand. To this end, we performed trend prediction using data for three years (2019–2021) of the available dataset to predict the values for the fourth year (2022). Then, to evaluate our models, we focused on four metrics, namely: R-squared (R^2), mean absolute error (MAE), mean absolute deviation (MAD) and mean absolute percentage error(MAPE).

1. R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s) and was used to assess the goodness of fit of our regression models.
2. MAE represents the average absolute difference between the predicted and actual values and provided use with a straightforward measure of prediction accuracy. (Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics were also considered for the same task but were dismissed due to the tasks sensitivity to outliers, where MAE is preferred)
3. MAPE expresses the average percentage difference between predicted and actual values, providing a relative measure of accuracy. MAPE is the most used measure for forecasting error since the variable's units are scaled to percentage units, which makes it easier to understand.
4. MAD is similar to MAE and is used in forecasting to quantify the accuracy of a forecasting model. It represents the average magnitude of forecast errors and helps assess how well the model predicts future values.

As a general guideline, having at least two yearly seasonal data cycles is often recommended to capture seasonal patterns effectively. Assuming that a single seasonal cycle corresponds to one year for monthly data, we would ideally want to have a minimum of 24 observations (two years) to perform a meaningful time series analysis. Theoretically, this allows the models to capture seasonal variations and assess their significance. Clearly, having more observations improves the accuracy and reliability of the analysis and forecasting results. Additionally, the complexity of the time series patterns and the specific goals of the analysis may also impact

Table 5 Topics and Content

Unique words	Tag
HEALTH, RECORD, RECORDS, PHI, PI, PERSONAL, DATA, BIOMETRIC, SSNS, BREACH, DATABASE, ACCOUNT, DETAILS, IDENTIFICATION, FILES	Personal Data Breaches
PHISHING, SPEAR-PHISHING, SPEAR, WALLET, ETHEREUM, SCAM, SIM, SWAPPING, SPOOF, IMPERSONATION, LOGINS, EMAIL, MEDIA, ACCOUNT, LINK	Phishing & social engineering
DDOS, DOS, EXPLOITATION, INFRASTRUCTURE, WINDOWS, LINUX, WEB, INTERNET, DENIAL, SSH, NETWORK, IP, SERVER, ROUTER, CHAIN, SUPPLY, SHELL, INTERNET, PROXY, DNS	Network and System Exploitation
CREDIT, CARD, CARDS, CREDENTIALS, DATA, BANKING, PASSWORD, PAYMENT, DATABASE, SKIMMING, SKIMMER, LOGIN, KEYLOGGING, LINK, COOKIE, DEBIT	Payment Card Fraud & Online Theft
ACCESS, REMOTE, RAT, TROJAN, EXTORTION, RANSOMWARE, TROJANS, VIRUS, SCREEN, VOICE, VIDEO, CLIENT, SCREENSHOTS, ZERO-DAY, O-DAY, MACOS, IOS, ANDROID, SPYING, PHOTO, STEGANOGRAPHY	RATs, Ransomware, Spyware

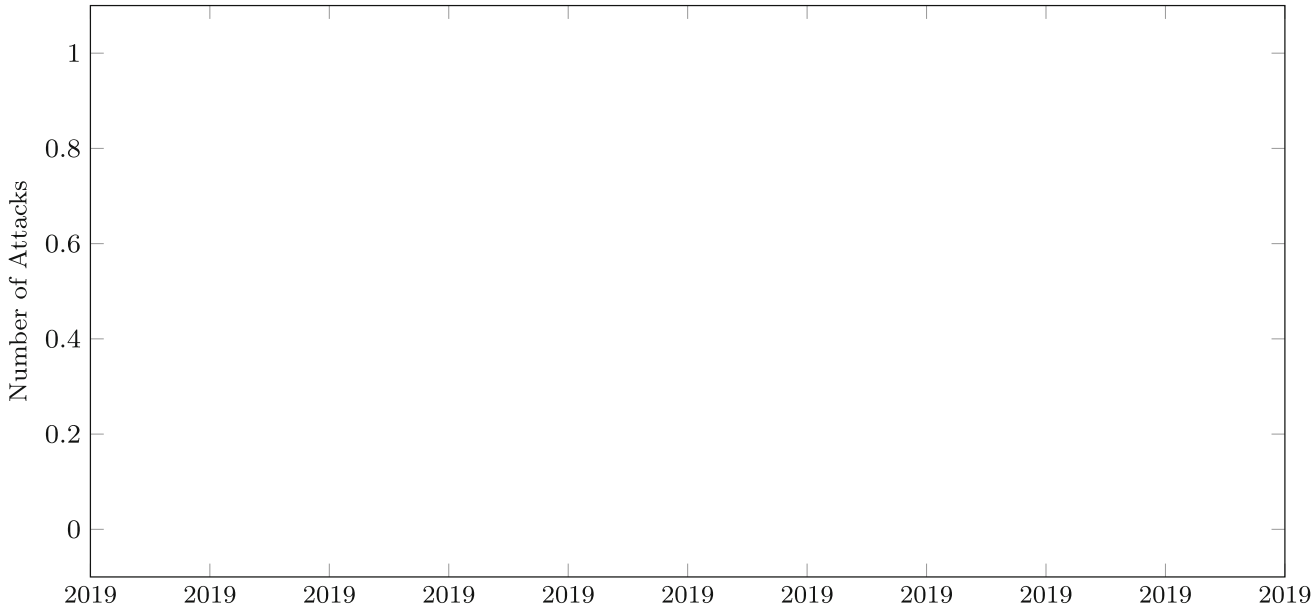


Fig. 4 Cyber security incidents 2019–2022

the minimum requirement. Nevertheless, this was impossible due to the lack of accurate reporting of cyber security events before 2019.

Given the dataset of incidents covering 2019 to 2022, described in Table 1b, we extracted the monthly statistics per topic and calculated the volume fluctuation per month, visualised in Fig. 4.

By providing the time series of cyber attack volumes to a set of forecasting models, we evaluated the most effective ones for trend prediction purposes, given the limitations mentioned above. We used 90% of the dataset to train the models and 10% for testing. We aimed to use the models trained with statistics of 2019 till the first semester of 2022 to predict incidents of the second semester of 2022 and validate them with the actual incidents that took place around the same period. Rather than aiming towards exact prediction matching for each month, we focus our effort on predicting the incident volume trends per period as follows: *upward* (↑), *slightly upward* (↗), *stable (-)*, *slightly downward* (↘), and *downward* (↓). The trend is calculated as the difference between two consecutive periods' average volume of incidents (Algorithm 1). In our case, the period used was six months.

Figure 5 maps the predicted values per month for the second semester of 2022 against the real volume of incidents, while Figs. 6 and 7 present the various model evaluation metrics.

Then, using this dataset, we decided to forecast the trends for 2022 for all topics (Table 6) and then performed a dedicated forecast by filtering our dataset per sector. We focused

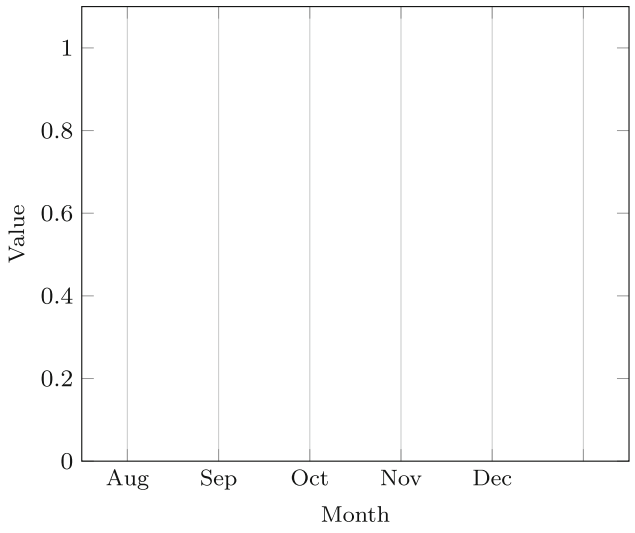


Fig. 5 Cyber security incidents prediction results (2nd Semester 2022)

Algorithm 1 Trend Value Labelling

```
Given TrendValue, ranging from negative to positive.
if TrendValue ≥ 6 then
    Label ← "Upward" {Trend is upward}
else if 3 < TrendValue < 6 then
    Label ← "Slightly upward" {Trend is slightly upward}
else if -3 ≥ TrendValue ≤ 3 then
    Label ← "Stable" {Trend is stable}
else if -6 < TrendValue < -3 then
    Label ← "Slightly downward" {Trend is slightly downward}
else
    Label ← "Downward" {Trend is downward}
end if
```

our efforts on four sectors, namely energy, healthcare, government, and finance, as seen in Table 7, using the forecasting algorithms side by side to ease the comparison of the results. This approach can help us generate tailored leads and forecasts and, as a result, more focused CSE scenarios per sector.

Four forecasting methods, namely LSTM, Prophet, ES, and SES, returned positive R^2 values, with LSTM being the fittest algorithm, closely followed by Prophet. The R^2 values of ES and SES, while positive, show that these models do not explain the variability of the response data around the mean; therefore, the predictions cannot be considered accurate. On the contrary, while not ideal, the range of R^2 values of both LSTM and Prophet implies a good fitness of the models. Finally, the negative values of SARIMAX clearly illustrate that this method is unsuitable for our dataset.

Based on the fact that LSTM had low MAE, MAPE, and MAD scores, it can be considered the most suitable algorithm for this task. This is validated by the fact that LSTM was one of the best in trend prediction when forecasting future volumes of the five topics per sector, as presented in Table 7. Similarly, Prophet had low MAE, MAPE, and MAD scores and satisfactory trend predictions when forecasting future

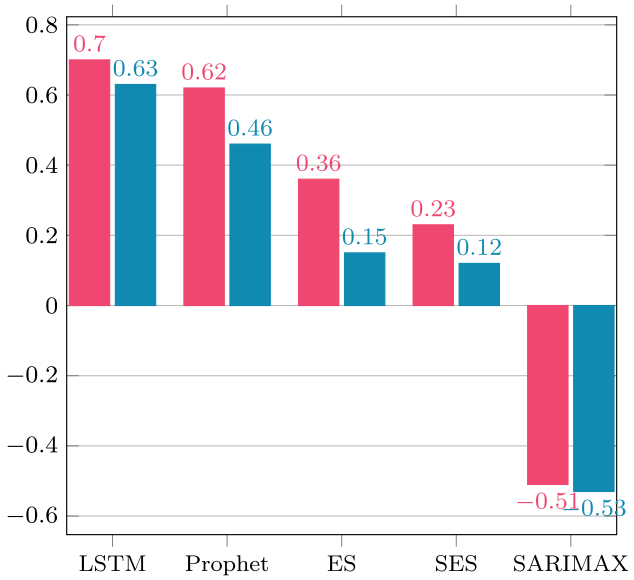


Fig. 6 R^2 Scores (Training, Testing Dataset) per Forecasting Model

volumes for the five topics per sector. Therefore, it can also be considered suitable for the task. SARIMAX, beyond the negative R^2 values, had high MAE, MAPE, and MAD values, validating that it cannot be used for accurate forecasting or trend prediction in our dataset.

From the other two methods with low R^2 values, ES had high MAE, MAPE and MAD scores, which was mapped in the poor performance when forecasting future incident volumes for the five topics per sector. Finally, even if SES did not fit well based on the R^2 values, it had the lowest MAE, MAPE and MAD scores and its performance was very satisfying for forecasting future incident volumes for the five topics per sector. We interpret the above as an indication that, while the predictions were not very accurate, the model was able to fit enough to identify the trends. For the rest of the steps in our proposed methodology, we will use the top three performing algorithms (Prophet, LSTM, SES) for further evaluation and comparison while dismissing ES and SARIMAX due to low performance in the selected metrics. When implementing our methodology in real life, the use of a single forecasting algorithm is adequate to provide the wanted predictions.

3.7 AI-driven prompt generation (text synthesis)

Our prompt generation process focuses on creating textual prompts based on a set of specific requirements by mixing user input with Current Topic Trends (CTTs) and Future Topic Trend Predictions (FTTPs) extracted from previous steps. The user provides a dataset of incidents and the sector of interest (optional) as input to our methodology. The user can also define the forecasting model to be used. Our AI-

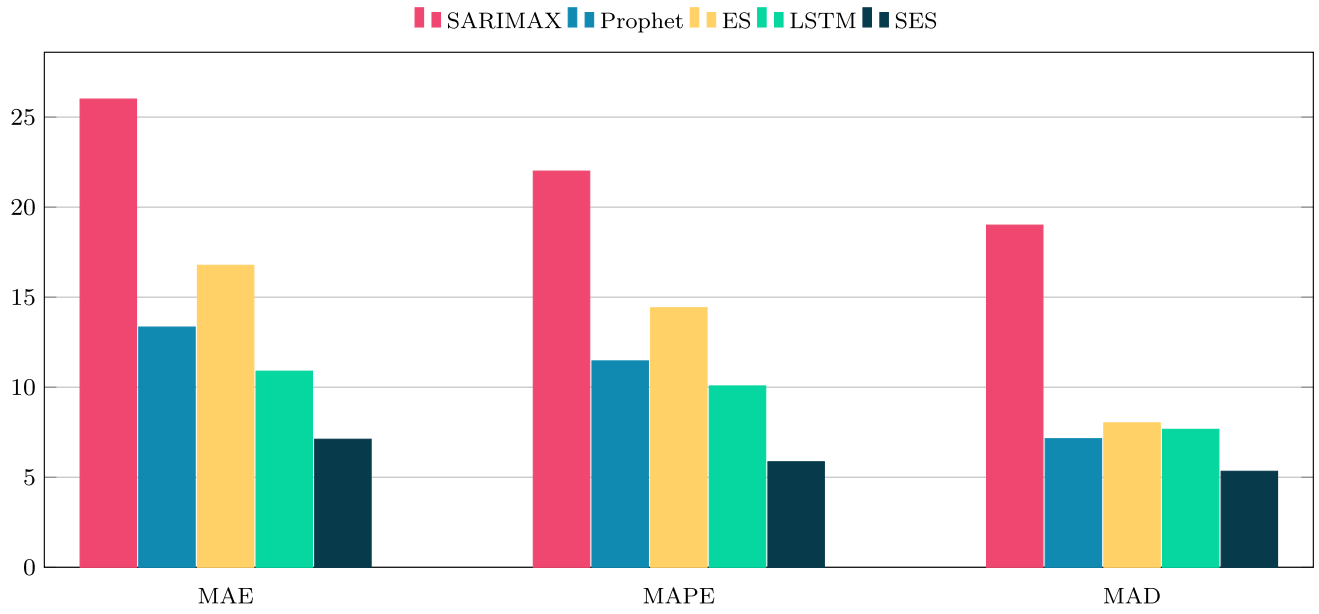


Fig. 7 MAE, MAPE, and MAD score per forecasting model

Table 6 Trend forecasting results for the 2nd Semester of 2022

	Actual Trend 2022	SARIMAX	Prophet	ES	SES	LSTM
All data-topic agnostic	–	↓	↓	↓	↓	–
RATS, ransomware, trojans	↘	↓	–	–	↘	–
Network_System_Exploit	↓	–	↓	↓	↓	↓
Phishing, Social_Eng	↘	↑	↑	↑	↘	↘
Personal_Data_Breach	↘	↘	↓	↑	↘	↘
Card_Fraud_Online_Theft	↘	–	↓	↑	–	–

driven methodology extracts the following information from the dataset based on the user input:

- **Current Topic Trends (CTTs)**: The top ten keywords in volume for the following tag categories: [ATTACK TYPE], [ATTACKER NAME], [VULNERABILITY], [ASSETS], [TECHNOLOGY], [MALWARE NAME], [MALWARE TYPE] for a given [SECTOR] over the course of a year. This is achieved by executing steps 1 and 2 of our methodology.
- **Future Topic Trend Predictions (FTTPs)**: A set of keywords that stem from previously defined topics (Table 5) and that exhibit a clear upward trend based on our trend prediction methodology by using the user selected forecasting algorithm. This is achieved by executing steps 1, 2, 3 and 4 of our methodology.

A textual prompt template consisting of four sections (with fixed structure along with dynamic sections) is then used to generate prompts that will form the basis of a CSE scenario consisting of the fifth and final step of our methodology.

An example of a prompt template can be seen in Fig. 8

The resulting output can then be edited or used as input to AiCEF, and a CSE incident will be generated, ready to be incorporated in an exercise scenario, as visualised in Fig. 9. Our methodology can produce meaningful, sectorial-focused prompts and scenarios with trend prediction incorporated seamlessly into the prompt without the need for cyber security expertise from the user.

4 Evaluation methodology and results

4.1 Evaluation methodology

To evaluate and measure the effectiveness of our proposed framework and the underlying methodology, we developed a case study focusing on the energy sector. The energy sector, present in both NIS and NIS2 [14] Directives, was selected due to its high critical impact on other industries if affected, but also due to its unique threat landscape that attracts a specific set of threat actors with unique motivation, Tactics, Techniques, and Procedures (TTPs). To this end, we followed a three-step methodology.

Table 7 Sectorial trend forecasting results for 2022

Sector	Actual Trend 2022	LSTM	Prophet	SARIMAX	ES	SES
Energy						
Sectorial Incidents	–	–	–	↗	↗	–
Personal_Data_Breach	–	–	–	↗	↗	–
Phishing_Social_Eng	–	–	–	–	–	–
Netwok_System_Exploit	–	–	↗	↗	↗	–
Card_Fraud_Online_Theft	–	–	↗	↗	↗	–
Rats_Ransomware_Trojans	–	–	–	–	↗	–
Healthcare						
Sectorial Incidents	–	–	–	↗	–	–
Personal_Data_Breach	–	–	–	↗	–	–
Phishing_Social_Eng	–	–	–	–	–	–
Network_System_Exploit	–	–	–	↗	↗	–
Card_Fraud_Online_Theft	–	–	–	↗	↗	–
Rats_Ransomware_Trojans	–	–	–	–	–	–
Government						
Sectorial Incidents	↓	–	–	–	↓	–
Personal_Data_Breach	↓	–	–	–	↓	–
Phishing_Social_Eng	↓	–	–	–	↓	↓
Network_System_Exploit	–	–	–	↑	↗	–
Card_Fraud_Online_Theft	↓	–	–	–	↓	↓
Rats_Ransomware_Trojans	↘	–	–	↗	↓	↓
Financial						
Sectorial Incidents	↓	–	↘	↓	↓	↓
Personal_Data_Breach	↓	–	↘	↓	↓	↓
Phishing_Social_Eng	↓	↓	↘	↓	↓	↓
Network_System_Exploit	–	–	–	–	–	–
Card_Fraud_Online_Theft	↓	–	↘	↓	↓	↓
Rats_Ransomware_Trojans	↓	–	↓	↓	↓	↓

1. **Prompt generation** We assigned exercise planners (EPs) with varying expertise to generate a set of textual prompts as input to our AiDPG methodology. Two groups of experts were used: experts in the field of cyber exercise development and experts with knowledge of the specific sector but with little or no experience in the exercise development role. The choice of creating a group of human experts consisting of the two subgroups mentioned above was made so that the two could combine their expertise and generate an exercise with both sectorial focus and technical value for the participant. Our approach of creating mixed groups of exercise planners with sectorial expertise and exercise planning background is not new; in fact, it is promoted by national cyber security agencies such as ANSII.[5] Consequently, three sets of prompts were generated by humans and a fourth one using our AiDPG methodology:
   - Prompt Set 1: The Exercise experts prompt set, created by the experts in CSE development.
   - Prompt Set 2: The Sector experts prompt set, created by experts in the sector examined (ENERGY).
   - Prompt Set 3: The Human experts prompt set, which is the combination of sets 1 & 2.
   - Prompt Set 4: The AiDPG prompt set was generated purely using our methodology focusing on the sector of interest (ENERGY) with forecasting enabled.

   AiDPG (no forecasting) prompt set was also generated as a subset of Prompt Set 4 and will be used only for limited comparisons against other prompt sets.
2. **Prompt Evaluation** All four sets of prompts were lexicologically analysed to identify keyword patterns and biases, using AiCEF for annotation, leading to the following tag sets:

[5] https://www.ssi.gouv.fr/uploads/2021/09/anssi-guide-organising_a_cyber_crisis_management_exercise-v1.0.pdf

Prompt Template:

- **Section 1: Attack Type and Sector**
  - Example: ATTACK TYPE in the SECTOR.
- **Section 2: Trending TTPs and Impact based on STI**
  - Example: Attackers ATTACKER NAME used VULNERABILITY to affect ASSETS.
  - The malware employed is MALWARE NAME MALWARE TYPE.
  - Technologies used or abused: TECHNOLOGY.
- **Section 3: Future Trending Topics** (*This section can be omitted if not enough data are available to perform forecasting)
  - Use NUM of the following trending elements: SECTOR TOPIC[0] ITEMS
- **Section 4: Styling and Formatting**
  - Start the prompt with: ["*Generate a cyber security incident describing an* "]. Prompt size: [200] words max

Example Prompt:

Generate a cyber security incident describing a BREACH in the ENERGY SECTOR. Attackers CONTI used CVE-2022-24002 to affect PASSWORDS, DATA. The malware employed is CONTI RANSOMWARE. Technologies used or abused: WINDOWS. The following trending elements were used in the incident: PHISHING, SPEAR-PHISHING, MALWARE.

Fig. 8 The form of our template above and an example prompt, below

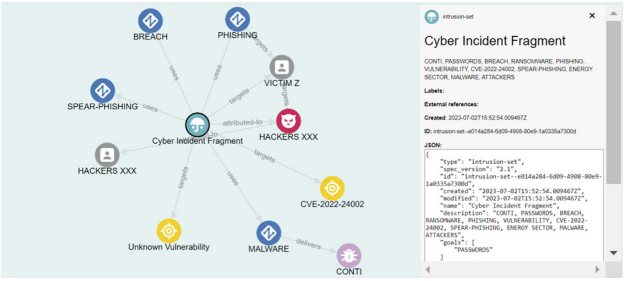


Fig. 9 STIX2.1 output generated by AiCEF

- "Exercise experts" Tag set: Prompt Set 1 (The Exercise experts prompt set) was annotated, forming this tag set.
- "Sector experts" Tag set: Prompt Set 2 (The Exercise experts prompt set) was annotated, forming this tag set.
- "Human Experts" Tag set: Prompt set 3 (All Human experts prompt set) was annotated, forming this tag set.

Prompt Specs: Generate a CSE prompt for a cyber awareness exercise. The prompt should not be longer than 600 words and should include as many technical details as possible related to a cyber security incident. The company which will use the scenario is an **Energy Service Provider** and all its employees can be potential Players. The incident generated should focus on the existing threat landscape and should train employees to prepare for current or future threat trends relevant to the designated sector in an exercise that will take place in 2023.

Fig. 10 Task definition

- "AiDPG Energy" Tag set: The Prompt set 4 (AiDPG prompt set) was annotated, forming this tag set.
- "Energy Threat Landscapes" Tag set: A collection of Sectorial Threat Landscapes (STLs) were annotated, forming this baseline Tag set.

Using these five tag sets, we performed a number of comparisons in an attempt to identify the prompt maturity and coverage of sector-oriented threat landscapes by comparing prompts generated using our methodology to those generated by human experts.

3. **Trend Forecasting Evaluation**
   Finally, to further evaluate and verify our trend prediction and forecasting findings, we surveyed the two groups of experts to test our forecasts against human expertise and intuition.

In the following paragraphs, we drill down to each step and present our evaluation results.

4.2 Prompt generation

Twenty-six cyber security experts were selected to generate five (5) individual CSE scenarios prompts each, according to the high-level requirements and specifications, as illustrated in Fig. 10.

All experts have cyber security backgrounds, and their skill sets resemble those of a Chief Information Security Officer (CISO) or similar. For the cyber exercise expert group, we invited several planners with high expertise in cyber security exercise management, mixed ethnicity, and focus sectors. More precisely, we invited members of a Cyber Awareness Expert Group, [6] from which 12 exercise planners responded. The sectorial experts group had a cybersecurity background for the energy sector (Members of the European Network of Transmission System Operators for Gas or European Network of Transmission System Operators for Electricity) and acted as CISOs for their organisations. In total, 14 sectorial experts responded. The demographic distribution of all experts is illustrated in Table 8.

[6] https://www.enisa.europa.eu/topics/cybersecurity-education/ad-hoc-working-group-awareness-raising

Table 8 Demographics of the experts

(a) Countries of origin of the experts

Country	#	Country	#
France	3	Luxembourg	1
Netherlands	3	Lithuania	1
Czechia	2	Austria	1
Spain	2	Italy	1
Belgium	2	Finland	1
Greece	2	Ireland	1
Denmark	2	Portugal	1
Croatia	1	Poland	1
Germany	1	Total	**26**

(b) Sector that experts are working in

Sector	#
Energy	14
Public Administration	5
ICT	2
Transport	2
Law Enforcement	2
Education	1

Not all of the experts provided the maximum prompt number of five (5) as requested, so we collected only fifty-five (55) prompts per group, 110 human prompts in total. All experts provided at least three prompts.

The prompt specs were inputted in our methodology, and an AiDPG prompt set was generated, focusing on the energy sector concluding the prompt generation task.

4.3 Prompt evaluation

To evaluate the sets of generated prompts, we applied keyword extraction using AiCEF.

Two types of evaluation metrics were used:

- Prompt maturity, using AiCEF's maturity calculation algorithm, to evaluate the richness of the prompt in terms of exercise-related objects.
- Tag coverage of the human and AI-assisted tag sets against that of a sectorial baseline. This metric helps evaluate the relevance of the prompts generated.

4.3.1 Prompt maturity

Based on AiCEF's maturity calculation algorithm, the maximum maturity score for any text can be 150 points, indicating that the text analysed has the maximum richness of information, covering all MLCESO tags in Table 3.

Table 9 Maturity scores

Label Name	Prompts	Maturity Score
Sector Experts	55	50,41
Exercise Experts	55	73,94
AiDPG (no forecasting)	55	135
AiDPG	55	150

Table 9 illustrates the average maturity score per group of prompts. The average maturity score of AiDPG (150) (even with no forecasting (135)) clearly exceeds the scores of any of the two human expert groups.

Indeed, these scores provide sufficient proof that our sector-focused AiDPG methodology can help human CSE planners improve their prompts and effectively exercise scenarios. More precisely, a 50.4% improvement for the exercise experts and an improvement of 66.67% for sector experts can be achieved in terms of prompt maturity.

4.3.2 Tag coverage

After analysing and comparing the accuracy of the lexicological elements of the prompt sets generated by exercise experts, sector experts (energy experts) and their combination along with our AiDPG methodology, we defined a baseline set of tags. The baseline tag set was generated by extracting tags (see Table 3) from acknowledged STL reports. More precisely, we used three STLs to evaluate our methodology: the Deloitte Oil and Gas (Energy) Threat Landscape, [7] the Energy Sector 2022 Threat Landscape by CITALID and Sekoia.io [8] and the Electric Avenue report by BlueViv & Neurosoft.[9] These threat landscapes cover the broader energy sector, using cyber security incidents from as early as 2017 up to early 2023. The unique tags per report are visualised in Fig. 11.

The analysis of the following tag sets can be visualised in Fig. 12.

- "Exercise experts" Tag set
- "Sector experts" Tag set
- "Human Experts" Tag set
- "AiDPG Energy" Tag set
- "Energy Threat Landscapes" Tag set

[7] https://www2.deloitte.com/us/en/pages/risk/articles/oil-and-gas-sectors-threat-landscape.html

[8] https://blog.sekoia.io/the-energy-sector-2022-cyber-threat-landscape/

[9] https://neurosoft.gr/the-electric-avenue-an-overview-of-the-energy-sectors-threat-landscape/

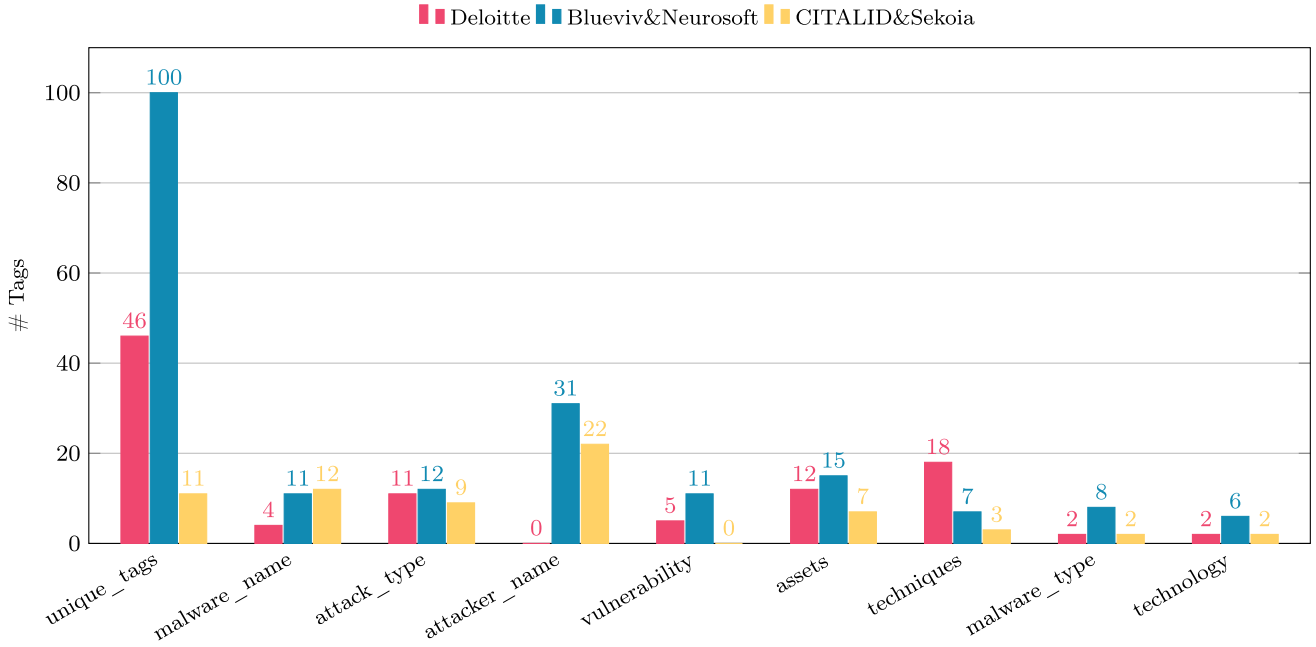


Fig. 11 Unique annotation tags values per energy landscape report

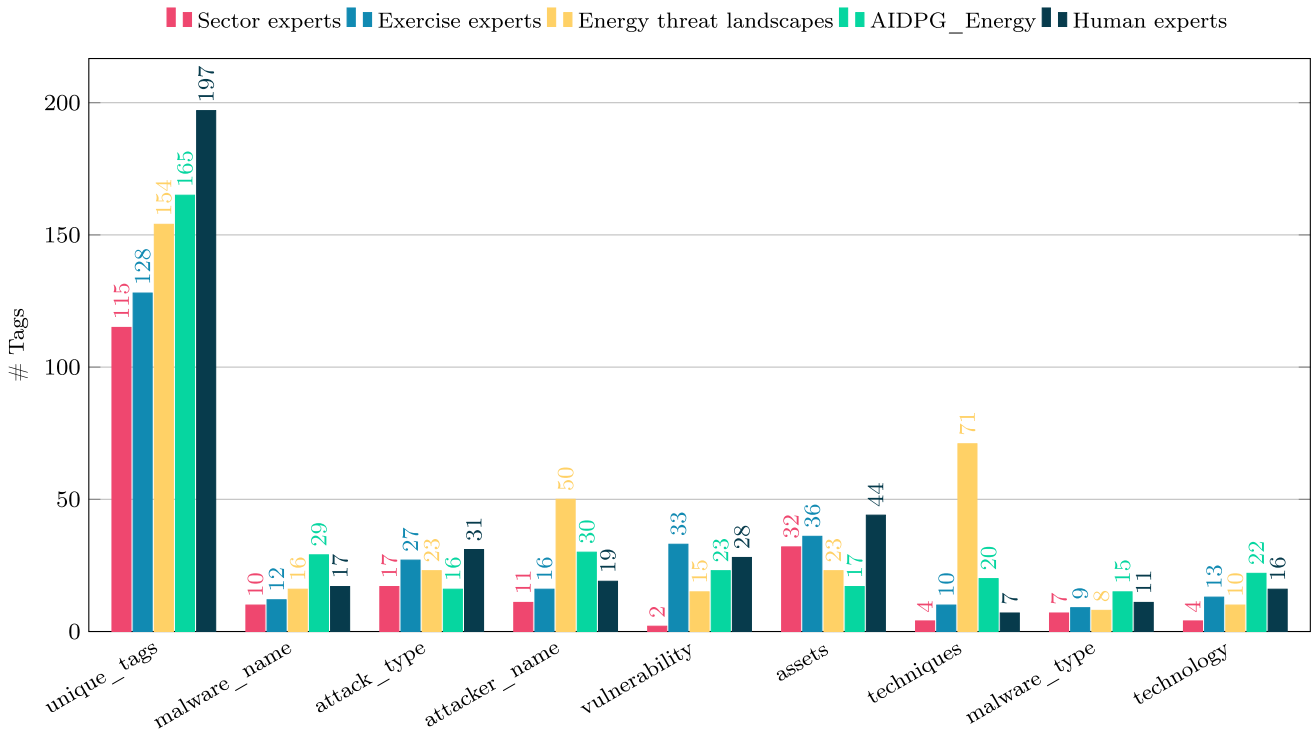


Fig. 12 Unique annotation tag volume comparison

Finally, the overlapping unique tags for "ENERGY THREAT LANDSCAPES", "HUMAN EXPERTS", and "AiDPG ENERGY" are represented in Fig. 13 providing a better visual representation of the coverage between Prompt and eventually Tag sets.

To verify the superiority of AiDPG methodology over human-expert-generated prompts, we use Jaccard similarity to calculate the similarity between two sets of unique phrases based on the number of common phrases and the total number of distinct phrases in the sets.

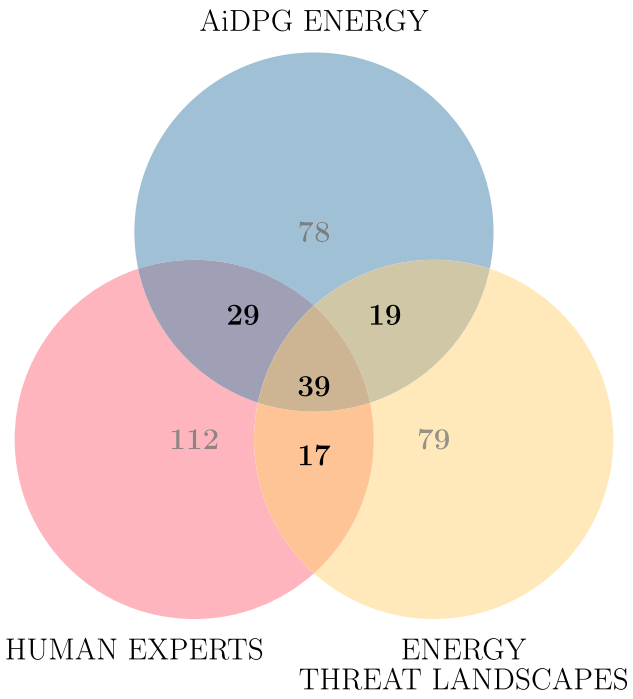


Fig. 13 Unique tags of STLs versus Human Experts and AiDPG

The Jaccard similarity coefficient (JSC) is defined as the size of the intersection of the sets divided by the size of the union of the sets.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where:

J(A, B) represents the Jaccard similarity coefficient (JSC) between sets A and B.

$|A \cap B|$ represents the size (cardinality) of the intersection of sets A and B.

$|A \cup B|$ represents the size (cardinality) of the union of sets A and B.

After comparing the Jaccard similarity, it is clear that the prompts generated with our AiDPG methodology are more relevant to the set generated when combining all energy threat landscapes:

- JSC for "AiDPG Energy" and "Energy Threat Landscapes" with value: 0.22
- JSC for "Human Experts" and "Energy Threat Landscapes" with value: 0.19
- JSC for "Sector Experts" and "Energy Threat Landscapes" with value: 0.16
- JSC for "Exercise Experts" and "Energy Threat Landscapes" with value: 0.17

Focusing on which group of experts can benefit more from our AiDPG methodology, we calculate the unique annotation tag categories intersection per group and measure the number of tags that are present in both "Energy Threat Landscapes" and "AiDPG Energy" but not in the individual human generated sets. By doing this, we assume that, at a minimum, the content of the "Energy Threat Landscapes" that is also part of the "AiDPG Energy" set can improve the prompt and be relevant.

Let A represent the set "Energy Threat Landscapes", B represent the set "AiDPG Energy", and C represent a human-generated set. The number of tags in the intersection of A and B but not in C can be calculated as follows:

$$\text{Benefit} = \|A \cap B\| - \|A \cap B \cap C\|$$

$$\text{BenefitPercentage} = \frac{\text{IntersectionCount}}{\|A\|} \times 100\%$$

Using the formula above for the annotation tags in Table 3, we can calculate the individual benefit percentages that the two human groups of experts could achieve using our methodology, Fig. 14.

The results underline the effectiveness of our methodology since both expert groups can benefit from at least 18.83% more relevant tags proposed by AiDPG, not selected by them but present in the "Energy Threat Landscapes" set.

4.4 Trend forecasting evaluation

To compare our sectorial trend forecasting results, a survey was run against our pool of specialised planners. The survey aimed to collect insights on the prioritisation of cyber threats and assets/resources based on both established threat landscape data and personal research, with a focus on preparing for incidents that are anticipated to occur in the coming months in the Energy sector. The ENISA Threat Landscape 2022[10] was used to extract the following eight prime cyber threats and assets affected and link them directly to the five types of incidents.

Cyber Threats: Ransomware, Malware, Disinformation/ misinformation, Threats against availability—(DOS, DDOS), Phishing & Social engineering threats, Threats against data (data breaches & leaks), Supply-chain attacks, Physical threats with security impact.

Assets Affected: Company Data, Email or other Credentials, Personal Information, Website, Network or other Infrastructure, Sensitive Personal Information, Monetary Funds, Reputation.

The planners were asked to provide feedback on the following questions:

[10] https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

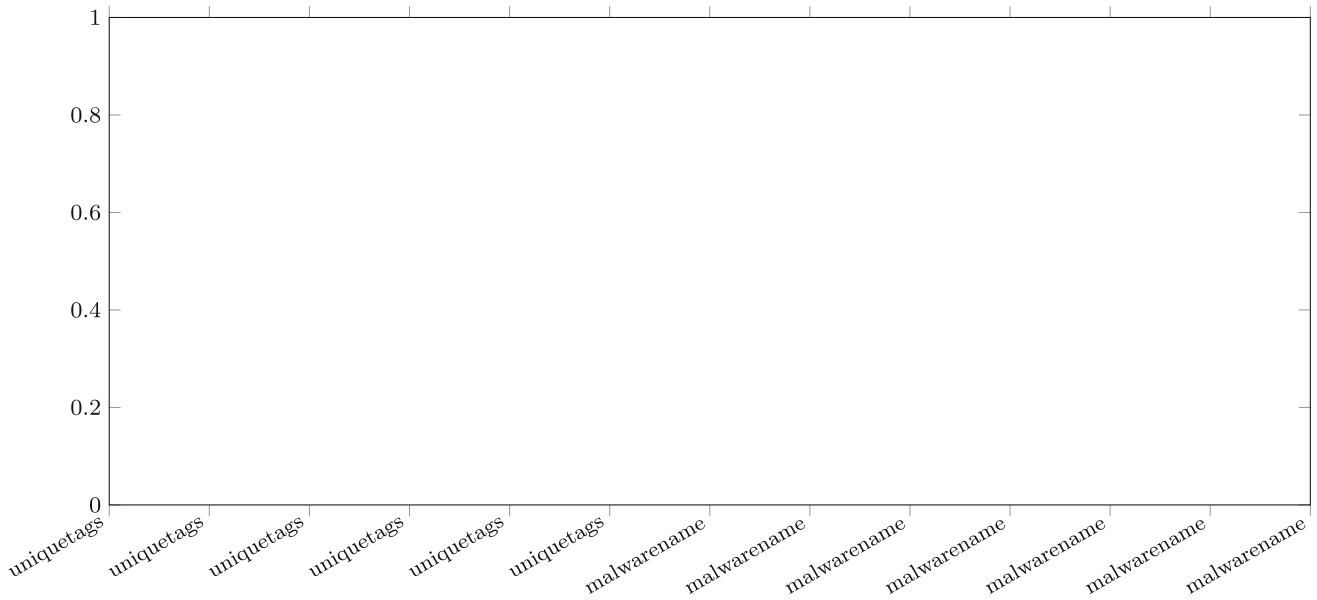


Fig. 14 Benefit percentage per tag and expert group

1. What is your trend prediction on the overall volume of cyber incidents in 2023 compared to 2022 affecting the energy sector?
2. What is your trend prediction for the following types of incidents in 2023 compared to 2022 affecting the energy sector?
   - (Personal) Data Breaches
   - Phishing & Social Engineering
   - Network & System Exploitation
   - Payment Card Fraud & Online Theft
   - Remote Access Trojans (RATs), Ransomware, Spyware

The possible answers were: *upward* (↑), *slightly upward* (↗), *stable (-)*, *slightly downward* (↘), and *downward* (↓)

We then implemented our proposed trend forecasting methodology for energy-related incidents for 2023 using the top three scoring models (Prophet, SES, and LSTM). A comparison of our results against the expectations of the two expert groups can be found in Tables 10 and 11, where both the normalised forecasted score (as per Algorithm 1) and trend is presented.

Regarding forecasting the overall sectorial incident trends, the two groups of humans agreed that there would be a "Slightly Upward" trend in 2023, a result also captured by our LSTM model. When comparing human foresight on sectorial incident trends per topic, we notice that the two expert groups do not align. Focusing more on the foresight of sectorial experts, we see that both Prophet and LSTM are making trend predictions that are close to those of humans. The results are satisfying in that, in most cases, our methodology did not produce the opposite results to the humans' foresight. Yet, our methodology is based on incident volumes and cannot replicate human foresight consisting of broader knowledge aspects that can contribute to a change trends like geopolitics. Only the actual data of cyber incidents against the energy sector in 2023 can prove whether our trend forecast or human insight was more accurate in the predictions made, a task to be performed in future work.

Table 10 Sectorial trend forecasting results for 2023

Forecast Type	Overall Energy Attacks trend
Prophet	7, Upward Trend (↑)
SES	0, Stable (–)
LSTM	3.17, Slightly Upward (↗)
SECTOR EXPERTS	3.9, Slightly Upward (↗)
EXERCISE EXPERTS	3.75, Slightly Upward (↗)

5 Conclusions and future work

CSEs can be a powerful tool in improving the organisation's preparedness against the ever-changing threat landscape. However, experts' shortages, timeliness, and relevance of the developed CSEs require novel solutions. To address this challenge, our proposal revolves around incorporating AI-driven sectorial threat landscaping and forecasting into existing exercise generation frameworks like AiCEF. This integration aims to identify emerging and significant threats and predict their potential impact in various sectors. By infusing these forecasts into creating cyber security exercise prompts using AiDPG and, eventually, scenarios, organisations can

Table 11 Sectorial forecast results per topic for the year 2023

Forecast Type	Card Fraud & Online Theft	Network System Exploit	Personal Data Breach	Phishing & Social Eng	Rats, Ransomware, Trojans
Prophet	4, ↗	9, ↑	7, ↑	5, ↑	4, ↗
SES	1, –	3, –	0, –	0, –	0, –
LSTM	2.1, –	3.5, ↗	3.1, ↗	1.9, –	2.4, –
SECTOR EXPERTS	2.71, –	4.29, ↑	3, –	4.64, ↑	4.15, ↑
EXERCISE EXPERTS	2.08, ↘	3.58, ↗	2.41, ↘	4.75, ↑	4.58, ↑

achieve more precise simulations of real-world cyber security incidents and better evaluate their preparedness against emerging threats.

Our proposed methodology identified the gap in context relevant CSE scenarios generated by sectorial cyber security experts and seasoned exercise planners. By introducing a novel methodology, we can provide tangible quantitative added value in CSE scenario development by tackling the problems of relevance and timeliness that are independent of the EP's experience. More precisely, through our AI-driven topic extraction model and by clustering sector-specific annotation tags, we can compare various forecasting models and accurately identify emerging (sectorial) cyber threats for a given set of cyber security incidents. To this end, we compared SARIMAX, ES, Simple Exponential Smoothing (SES), Prophet, and Long Short-Term Memory (LSTM) modelling using Recurrent Neural Networks (RNN). Our experiments illustrated that for this dataset, LSTM and Prophet were the two fittest algorithms for this task.

Regarding prompt maturity, the sectorial-focused AiDPG methodology can indeed assist human CSE planners in improving their prompts and effectively the exercise scenarios. More specifically, a 50.4% improvement for the exercise experts and a 66.67% improvement for the sector experts can be achieved in terms of prompt maturity, generating not only richer but also more sector-focused scenarios.

After calculating the Jaccard similarity, it is clear that the prompts generated by our AIDPG methodology are more relevant to the set generated when combining all energy threat landscapes than the set generated by all human experts combined:

- "AiDPG Energy" and "Energy Threat Landscapes" with value: 0.22
- "HUMAN Experts" and "Energy Threat Landscapes" with value: 0.19.

Even more, by leveraging our user surveys to measure the perceived relevance of CSE with AI-driven threat intelligence compared to traditional exercise generation, we managed to quantify the value of our methodology. The results underline the effectiveness of our methodology since both expert groups surveyed, Exercise Experts and Sector Experts, can improve by at least 18.83% the prompt and effective scenario relevance to a specific sector.

Limitations to our methodology have been identified regarding the volume and time distribution of past incidents needed in order to form accurate forecasting. For Prophet specifically, 3 years of observations is the minimum data to predict a fourth year in sequence. We consider the amount of observation of three years to be the minimum for satisfactory predictions to incorporate forecasting results in our proposed methodology. Further to the above, forecasting algorithms cannot replace human foresight based on the broader knowledge of a domain that leads human experts to expect trend changes. We believe that the use of Large Language Models will contribute to the improvement of our methodology towards bridging this gap.

In the future evolution of our work, we will aim to fully automate threat landscape generation procedures by processing any given data set. In the short term, we plan to collect cyber incident statistics for 2023. We will then properly evaluate the results of our presented case study on the energy sector to identify if our trend forecast methodology was more accurate in predicting trends against the human experts' insight. While our methodology proves to be relatively efficient in focused prompt and scenario generation, more tests on non-expert humans should be conducted to identify further benefits in the CSE creation and customisation process. To verify our hypothesis, we plan to embed the AiDPG methodology in AiCEF to help automatically produce CSE scenarios and measure the adoption and satisfaction of the final output by experts in the domain.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10207-024-00860-w.

Acknowledgements This work was supported by the European Commission under the Horizon Europe Programme, as part of the projects CyberSecPro (https://www.cybersecpro-project.eu) (Grant Agreement no.101083594) and LAZARUS (https://lazarus-he.eu/) (Grant Agreement no. 101070303). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Funding Open access funding provided by HEAL-Link Greece.

Data Availability The data used in this work is available at https://github.com/alexzacharis/AI-driven-Threat-Intelligence-and-Forecasting.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Acarturk, C., Sirlanci, M., Balikcioglu, P.G., Demirci, D., Sahin, N., Kucuk, O.A.: Malicious code detection: run trace output analysis by lstm. IEEE Access **9**, 9625–9635 (2021)
2. Almahmoud, Z., Yoo, P.D., Alhussein, O., Farhat, I., Damiani, E.: A holistic and proactive approach to forecasting cyber threats. Sci. Rep. **13**(1), 8049 (2023)
3. ANSII: Organising a cyber crisis management exercise (2021). https://www.ssi.gouv.fr/guide/organising-a-cyber-crisis-management-exercise/
4. Armstrong, J.S., Collopy, F.: Error measures for generalizing about forecasting methods: empirical comparisons. Int. J. Forecast. **8**(1), 69–80 (1992)
5. Augustine, T., Dodge, R.C., et al.: Cyber defense exercise: meeting learning objectives thru competition. In: Proceedings of the 10th Colloquium for Information Systems Security Education (2006)
6. Bakdash, J.Z., Hutchinson, S., Zaroukian, E.G., Marusich, L.R., Thirumuruganathan, S., Sample, C., Hoffman, B., Das, G.: Malware in the future? Forecasting of analyst detection of cyber events. J. Cybersecur. **4**(1), tyy007 (2018)
7. Bilge, L., Han, Y., Dell'Amico, M.: Riskteller: predicting the risk of cyber incidents. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1299–1311 (2017)
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
9. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. John Wiley & Sons (2015)
10. Carriegos, M.V., Fernández-Díaz, R.Á.: Towards forecasting time-series of cyber-security data aggregates. In: Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., Corchado, E. (eds.) 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020), pp. 273–281. Springer International Publishing, Cham (2021)
11. Couce-Vieira, A., Insua, D.R., Kosgodagan, A.: Assessing and forecasting cybersecurity impacts. Decis. Anal. **17**(4), 356–374 (2020)
12. ENISA: National Exercise—Good Practice Guide (2009). https://www.enisa.europa.eu/publications/national-exercise-good-practice-guide
13. ENISA: ENISA CYBERSECURITY THREAT LANDSCAPE METHODOLOGY (2022). https://www.enisa.europa.eu/publications/enisa-threat-landscape-methodology/
14. European Commission: Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive) (Text with EEA relevance) (2022). https://eur-lex.europa.eu/eli/dir/2022/2555
15. Fan, S., Wu, S., Wang, Z., Li, Z., Yang, J., Liu, H., Liu, X.: Aleap: attention-based lstm with event embedding for attack projection. In: 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC), pp. 1–8. IEEE (2019)
16. Fang, Z., Xu, M., Xu, S., Hu, T.: A framework for predicting data breach risk: leveraging dependence to cope with sparsity. IEEE Trans. Inf. Forensics Secur. **16**, 2186–2201 (2021)
17. Furtună, A., Patriciu, V.V., Bica, I.: A structured approach for implementing cyber security exercises. In: 2010 8th International Conference on Communications. IEEE, pp. 415–418 (2010)
18. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint (2023) arXiv:2303.15056
19. Gogineni, K., Derasari, P., Venkataramani, G.: Foreseer: efficiently forecasting malware event series with long short-term memory. In: 2022 IEEE International Symposium on Secure and Private Execution Environment Design (SEED). IEEE, pp. 97–108 (2022)
20. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
21. Goyal, P., Hossain, K., Deb, A., Tavabi, N., Bartley, N., Abeliuk, A., Ferrara, E., Lerman, K.: Discovering signals from web sources to predict cyber attacks (2018). arXiv preprint arXiv:1806.03342
22. Granåsen, M., Andersson, D.: Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. Cogn. Technol. Work **18**(1), 121–143 (2016)
23. Green, A., Zafar, H.: Addressing emerging information security personnel needs. A look at competitions in academia: Do cyber defense competitions work? AMCIS 2013 Proceedings 1, 257 (2013)
24. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
25. Husák, M., Kašpar, J.: Aida framework: real-time correlation and prediction of intrusion detection alerts. In: Proceedings of the 14th International Conference on Availability, Reliability and Security, pp. 1–8 (2019)
26. Husák, M., Komárková, J., Bou-Harb, E., Čeleda, P.: Survey of attack projection, prediction, and forecasting in cyber security. IEEE Commun. Surv. Tutor. **21**(1), 640–660 (2018)
27. Husák, M., Bartoš, V., Sokol, P., Gajdoš, A.: Predictive methods in cyber defense: current experience and research challenges. Future Gener. Comput. Syst. **115**, 517–530 (2021)
28. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts (2018)
29. Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S.: A state space framework for automatic forecasting using exponential smoothing methods. Int. J. Forecast. **18**(3), 439–454 (2002)
30. Hyndman, R.J., Koehler, A.B., Ord, J.K.: Forecasting with Exponential Smoothing: The State Space Approach. Springer Science & Business Media (2008)
31. ISO Central Secretary: Societal security—guidelines for exercises. Standard ISO22398:2013, International Organization for Standardization, Geneva, CH (2013). https://www.iso.org/standard/50294.html

32. Jansen, B.J., Sg, Jung, Salminen, J.: Employing large language models in survey research. Nat. Lang. Process. J. **4**(100), 020 (2023)
33. Kick, J.: Cyber Exercise Playbook. Tech. rep, MITRE CORP BEDFORD MA (2014)
34. Kim, J., Kim, J., Thu, H.L.T, Kim, H.: Long short term memory recurrent neural network classifier for intrusion detection. In: 2016 International Conference on Platform Technology and Service (PlatCon). IEEE, pp. 1–5 (2016)
35. Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models. McGraw-hill (2005)
36. Liu, X., Liu, J.: Malicious traffic detection combined deep neural network with hierarchical attention mechanism. Sci. Rep. **11**(1), 12,363 (2021)
37. Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., Liu, M.: Cloudy with a chance of breach: Forecasting cyber security incidents. In: 24th USENIX Security Symposium (USENIX Security 15), pp. 1009–1024 (2015)
38. MacIntyre, R.: Penn treebank tokenizer (sed script source code) (1995)
39. Malik, J., Akhunzada, A., Bibi, I., Imran, M., Musaddiq, A., Kim, S.W.: Hybrid deep learning: an efficient reconnaissance and surveillance detection mechanism in sdn. IEEE Access **8**, 134,695-134,706 (2020)
40. Md Azam, M.N., Ramli, N.A.: Reported malicious codes incident within malaysia's landscape: time series modelling and a timeline analysis. Int. J. Adv. Data Sci. Intell. Anal. **2**(2) (2022). https://amcs-press.com/index.php/ijadsia/article/view/65
41. Munkhdorj, B., Yuji, S.: Cyber attack prediction using social data analysis. J. High Speed Netw. **23**(2), 109–135 (2017)
42. Okutan, A., Werner, G., Yang, S.J., McConky, K.: Forecasting cyberattacks with incomplete, imbalanced, and insignificant data. Cybersecurity **1**, 1–16 (2018)
43. Okutan, A., Yang, S.J., McConky, K., Werner, G.: Capture: cyber-attack forecasting using non-stationary features with time lags. In: 2019 IEEE Conference on Communications and Network Security (CNS). IEEE, pp. 205–213 (2019)
44. Patsakis, C., Lykousas, N.: Man vs the machine in the struggle for effective text anonymisation in the age of large language models. Sci. Rep. (2023)
45. Qin, X., Lee, W.: Attack plan recognition and prediction using causal networks. In: 20th Annual Computer Security Applications Conference. IEEE, pp. 370–379 (2004)
46. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 248–256 (2009)
47. Ren, K., Zeng, Y., Cao, Z., Zhang, Y.: Id-rdrl: a deep reinforcement learning-based feature selection intrusion detection model. Sci. Rep. **12**(1), 15,370 (2022)
48. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
49. Salih, A., Zeebaree, S.T., Ameen, S., Alkhyyat, A., Shukur, H.M.: A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection. In: 2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC). IEEE, pp. 61–66 (2021)
50. Sarker, I.H., Kayes, A., Badsha, S., Alqahtani, H., Watters, P., Ng, A.: Cybersecurity data science: an overview from machine learning perspective. J. Big Data **7**, 1–29 (2020)
51. Schepens, W.J., James, J.R.: Architecture of a cyber defense competition. In: SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483). IEEE, vol. 5, pp. 4300–4305 (2003)
52. Shen, Y., Mariconti, E., Vervier, P.A., Stringhini, G.: Tiresias: predicting security events through deep learning. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, pp. 592–605 (2018)
53. Sievert, C., Shirley, K.: Ldavis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 63–70 (2014)
54. Suciu, O., Nelson, C., Lyu, Z., Bao, T., Dumitraş, T.: Expected exploitability: predicting the development of functional vulnerability exploits. In: 31st USENIX Security Symposium (USENIX Security 22), pp. 377–394 (2022)
55. Sun, N., Zhang, J., Rimba, P., Gao, S., Zhang, L.Y., Xiang, Y.: Data-driven cybersecurity incident prediction: a survey. IEEE Commun. Surv. Tutor. **21**(2), 1744–1772 (2018)
56. Sun, N., Ding,. M., Jiang, J., Xu, W., Mo, X., Tai, Y., Zhang, J.: Cyber threat intelligence mining for proactive cybersecurity defense: a survey and new perspectives. IEEE Commun. Surv. Tutor. (2023)
57. Tavabi, N., Goyal, P., Almukaynizi, M., Shakarian, P., Lerman, K.: Darkembed: exploit prediction with neural language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
58. Taylor, S.J., Letham, B.: Prophet: forecasting at scale. J. Open Source Softw. **3**(22), 651 (2017)
59. Törnberg, P.: Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning (2023). arXiv preprint arXiv:2304.06588
60. Velleman, P.F., Hoaglin, D.C.: Applications, Basics, and Computing of Exploratory Data Analysis. Duxbury Press (1981)
61. Werner, G., Yang, S., McConky, K.: Time series forecasting of cyber attack intensity. In: Proceedings of the 12th Annual Conference on Cyber and Information Security Research, Association for Computing Machinery, New York, NY, USA, CISRC '17 (2017a). https://doi.org/10.1145/3064814.3064831. https://doi.org/10.1145/3064814.3064831
62. Werner, G., Yang, S., McConky, K.: Time series forecasting of cyber attack intensity. In: Proceedings of the 12th Annual Conference on cyber and information security research, pp. 1–3 (2017)
63. Werner, G., Yang, S., McConky, K.: Leveraging intra-day temporal variations to predict daily cyberattack activity. In: 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, pp. 58–63 (2018)
64. Xu, M., Schweitzer, K.M., Bateman, R.M., Xu, S.: Modeling and predicting cyber hacking breaches. IEEE Trans. Inf. Forensics Secur. **13**(11), 2856–2871 (2018)
65. YILDIZ, K., BÜYÜKTANIR, B.: Comparison of lda, nmf and bertopic topic modeling techniques on amazon product review dataset: a case study. International Conference on Computing, Intelligence and Data Analytics (2023)
66. Zacharis, A., Patsakis, C.: AiCEF: an ai-assisted cyber exercise content generation framework using named entity recognition. Int. J. Inf. Secur. (2023). https://doi.org/10.1007/s10207-023-00693-z
67. Zacharis, A., Gavrila, R., Patsakis, C., Ikonomou, D.: Ai-assisted cyber security exercise content generation: Modeling a cyber conflict. In: 2023 15th International Conference on Cyber Conflict: Meeting Reality (CyCon), pp. 217–238 (2023). https://doi.org/10.23919/CyCon58705.2023.10181930

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Journal of Information Security and Applications 90 (2025) 104037

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

Optimising AI models for intelligence extraction in the life cycle of Cybersecurity Threat Landscape generation

Alexandros Zacharis [a], Razvan Gavrila [a], Constantinos Patsakis [a,b,*], Christos Douligeris [a]

[a] Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece
[b] Information Management Systems Institute, Athena Research Centre, Artemidos 6, Marousi 15125, Greece

ARTICLE INFO

Keywords:
Machine learning
Cybersecurity Threat Landscape
Threat intelligence
Named entity recognition
Large language model

ABSTRACT

The increasing complexity and frequency of cyber attacks in the modern digital environment demand continuous vigilance and proactive strategies to manage risks effectively. Conventional approaches to generating intelligence for Cybersecurity Threat Landscape (CTL) reports are often resource-intensive and time-consuming, as they depend on manual identification, collection, and analysis of relevant electronically stored information (ESI). This study investigates the potential of artificial intelligence (AI) to transform CTL generation, reducing manual classification and tagging while improving efficiency and accuracy.

We focus on evaluating the classification performance of several Large Language Models (LLMs), including Gemini 1.5 Pro, GPT-4o, but also Bidirectional Encoder Representations from Transformers (BERT) based models like TRAM and TTPHunter along with custom Named Entity Recognition (NER) models, using a dataset previously annotated by human experts. Our findings demonstrate the promising results of AI-driven intelligence extraction for CTL report generation, streamlining cybersecurity operations by automating routine tasks and providing precise and timely threat intelligence. However, the variability in model performance suggests the importance of hybrid approaches needed to achieve the accuracy of human annotation. Therefore, we propose a novel voting agreement-based methodology, harvesting the most from the combined AI model capabilities to effectively address the complexities of cybersecurity threat intelligence extraction.

1. Introduction

The contemporary cybersecurity landscape is characterised by a constantly shifting spectrum of threats and vulnerabilities which require robust strategies for effective threat detection, analysis, and mitigation [1–3], especially when dealing with Advanced Persistent Threat (APT) actors. To navigate these challenges, organisations adopt various methodologies to construct Cybersecurity Threat Landscapes (CTLs), which provide comprehensive insights into current cyber risks and support informed decision-making. A prominent methodology established by the European Union Agency for Cybersecurity (ENISA) offers a structured approach to CTL generation [4], involving phases such as identification, collection, preservation, processing, review, analysis, and production of electronically stored information (ESI).

Although existing methodologies are effective, the manual processes involved in CTL generation pose significant limitations, including time inefficiencies, intensive use of resources, high costs, and the risk of human error, such as incorrect annotation. Addressing these challenges requires innovative approaches that leverage emerging technologies to automate and streamline CTL report generation. Artificial Intelligence (AI) and Machine Learning (ML) are at the forefront of these advancements, offering robust data analysis, pattern recognition, and decision-making capabilities.

In this work, we evaluate the feasibility of an AI-driven approach to CTL generation designed to extract key intelligence while limiting manual intervention, aiming to maximise efficiency, accuracy, and scalability. To validate the effectiveness of our approach, we compare the latest AI/ML models against a team of human cybersecurity experts. By comparing the results obtained through AI automation with those derived from human expertise, our objective is to assess our proposed methodology's efficacy, reliability, and practical utility for the given task.

Problem Definition, Motivation, and Contributions: The primary challenge addressed in this work is the lack of automated and systematically evaluated approaches for generating Cyber Threat Landscape (CTL) reports using ML/AI models. Traditional methods rely heavily on

* Corresponding author.
E-mail addresses: Alexandros.zacharis@enisa.europa.eu (A. Zacharis), razvan.gavrila@unipi.gr (R. Gavrila), kpatsak@unipi.gr (C. Patsakis), cdoulig@unipi.gr (C. Douligeris).

https://doi.org/10.1016/j.jisa.2025.104037

Available online 1 April 2025
2214-2126/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

manual analysis, which is resource-intensive and prone to inconsistencies. Moreover, while various AI/ML models for cyber security threat intelligence feature extraction exist, it is still unclear how closely they compete with the human analyst capabilities regarding CTL generation. The challenge here lies in the nature of the attributes to be extracted. The reason is that these attributes are complex to contextualise and are even difficult for human cyber security analysts to extract. These challenges significantly hinder the timely and cost-effective generation of actionable cybersecurity intelligence, a critical need in today's rapidly evolving threat landscape.

Our research is motivated by the following practical questions:

- How effectively can AI-driven methods automate and enhance the generation of CTL reports compared to manual methods?
- What types of AI models are the most effective for extracting attributes that are usually found in a CTL report?
- What are the trade-offs between accuracy, efficiency, and scalability when using AI models versus manual approaches for CTL report generation?

To this end, we are the first to explore the potential of modern AI/ML models in automating CTL report generation and systematically evaluating their performance. Therefore, the key contributions of this work are as follows:

- **AI/ML powered CTL generation pipeline:** We are the first to propose the extension of an existing CTL generation methodology to utilise AI/ML for automated analysis and extraction of key attributes. More specifically we propose a hybrid approach to the analysis step combining human effort with AI/ML models.
- **Comparative Analysis:** We evaluate the performance of state-of-the-art Large Language Models (LLMs), including Gemini 1.5 Pro [5,6] and GPT-4o [7], in conjunction with high-performing BERT-based models like TRAM [8], TTPHunter [9] and custom Named Entity Recognition (NER) models, applying them to the same corpus of cybersecurity reports.
- **Quantitative Assessment:** We measure and compare the quality, cost, and speed of AI-generated CTL reports with those produced by human experts, providing a rigorous evaluation of the trade-offs involved. For clarity, a comparative analysis is provided per extracted feature.
- **Optimisation in attribute extraction:** We introduce a novel Voting-Based Agreement technique to improve the precision of AI models, particularly to extract challenging attributes such as MITRE TTPs and to test their effectiveness.

As a result, this work represents a significant advance in the automation of CTL generation, addressing a critical gap in the cybersecurity domain by providing actionable insights for both researchers and practitioners.

The structure of the remainder of this work can be summarised as follows. The related work is presented in Section 2, focusing on threat landscapes and notable methodologies, along with AI models used for various cybersecurity-related tasks, specifically threat intelligence extraction. Section 3 provides more information on the methodology followed in deciding which parts of a CTL AI-powered intelligence extraction can help and how to evaluate which models fit the task. The selection of the suitable dataset is also described, along with the limitations faced. Section 4 describes the human annotation procedure undertaken by cybersecurity experts, while Section 5 presents the model selection and AI annotation process. In Section 6, the evaluation metrics, comparative results, and findings are presented, while Section 7 introduces the proposed Voting-Based Agreement Methodology along with improved accuracy results. Section 8 concludes our work by highlighting the most important results and presenting our next steps.

2. Background and related work

The field of CTL generation has seen significant advancements over the years as a response to the increasing complexity of cyber threats and the need for more efficient and accurate threat analysis [10–12]. The current threat landscape generation methodologies focus on manual effort in both collection and analysis, striving to adopt advanced technologies like AI and ML to become more efficient. In what follows, we first discuss the developments in attribute extraction for Cyber Threat Intelligence (CTI), focusing on potential uses in CTL generation. Then, we present popular CTL drafting methodologies and examples that formulated the basis of our proposed work and target output.

2.1. AI and ML in cyber threat intelligence

Natural Language Processing (NLP) techniques have been extensively applied in cyber threat intelligence analysis, as demonstrated by methods such as Casie [13], Extractor [14], Open-CyKG [15], SecIE [16], and CyberEntRel [17]. These methods leverage advanced NLP models (e.g., BiLSTM, BERT, RoBERTa) to extract actionable insights from text. However, adapting these models to the cyber threat domain requires substantial preprocessing and fine-tuning efforts.

Previous studies have achieved extraction of cyber-related attributes from cyber threat intelligence reports [18–23] using NLP to formulate actionable outputs with utilisation in a variety of cyber security domains, including Cyber Security Exercises [24].

LLMs have been proven to successfully identify emerging cybersecurity threats, vulnerabilities, and attack trends and enhance existing threat intelligence pipelines [25] as well as detect secrets in source code [26]. Phishing detection and analysis have also been studied in the context of LLMs [27–29] including deobfuscation of malicious payloads [30].

The application of NLP and LLMs in cyber incident analysis has shown significant promise in enhancing the efficiency and effectiveness of the cyber threat detection process [31,32]. It enabled analysts to extract, understand, and communicate incident data more comprehensively from both network threat intelligence [31] and social media sources [32]. An innovative GPT-based open-source intelligence pipeline has been developed by Sufi [33], using historical cyber incident reports to extract a seven-dimension intelligence. The GPT-driven analyses extracted attributes such as Actor Type, Target, Attack Source (i.e., the country from which the attack originated), Attack Destination (i.e., the targeted country), Attack Level, Attack Type, and Attack Timeline.

2.2. MITRE tactics, techniques, and procedures extraction

There are various approaches to facilitate the extraction of actionable threat intelligence from reports. For instance, a priority is the accurate extraction of IOCs. To this end, iACE [34] focuses on extracting IOCs, and to achieve this, it uses regular expressions and a custom NER model. Going a step further, ChainSmith [35] classifies the IOCs to the corresponding attack phase, while TIMiner [36] uses word embeddings and syntactic dependencies to extract them and prioritise them according to sector impact.

Moving upper to Bianco's pyramid of pain, the goal is to pass from simple IOCs, e.g., hashes, URLs, and domains, to the tooling and the methods that threat actors use. Given that MITRE's ATT&CK framework has become a common ground for most practitioners and researchers, the goal is to extract the Tactics, Techniques, and Procedures (TTPs) from threat intelligence reports. Indeed, there is a constant improvement in identifying new methods such as ontology, graphs, and keyword matching powered by AI models of various sizes, including LLMs. TTPDrill [37] uses an ontology-based approach with BM25 [38] weighting method to classify threat actions to known attack patterns. Both TTPDrill and THREATRAPTOR [39] adopt an unsupervised NLP

Table 1
State of the art on MITRE TTP extraction. Notation: A: Accuracy, P: Precision, and R: Recall.

Article	Year	Extraction Technique	TTP Extraction	Metrics[a]	Dataset
TTPDrill [37]	2017	TF–IDF, Ontology-based, and improved BM25 similarity rank. Unsupervised NLP	MITRE TTPs	P:0.84 R:0.82	Symantec
Casie [13]	2020	BiLSTM	MITRE TTPs	P:0.77 R:0.77 F1:0.77	CyberWire
rcATT [40]	2020	TF–IDF and ML Models (KNN, Decision Tree, Random Forest, and Extra Tree)	MITRE TTPs	P:0.30 R:0.44 F1:0.30	MITRE Enterprise attack
ThreatRaptor [39]	2021	Unsupervised NLP	Other TTPs	P:0.96 R:0.97 F1:0.96	DARPA TC
TriCTI [50]	2022	Unsupervised NLP using web crawler	Other TTPs	P:0.87 R:0.87 F1:0.87	Labelled IoCs
AttacKG [41]	2022	Entity-based dependency graph and Graph-alignment algorithm	Other TTPs	P:0.782 R:0.860 F1:0.819	DARPA TC
TTPHunter [9]	2023	BERT/RoBERTa followed by Linear Classifier	MITRE TTPs (TOP 50)	P:0.89 R:0.88 F1:0.88	Custom
LADDER [42]	2023	Extract attack phases using a sequence tagging model and map these patterns to TTP using cosine similarity	MITRE TTPs	P:0.65 R:0.63 F1:0.64	Custom
TRAM [8]	2023	SciBERT followed by Linear Classifier	MITRE TTPs (TOP 50)	P:0.94 R:0.94 F1:0.93	Custom
LLM-TIKG [51]	2023	Fine-tuned LLaMA-2-7B	MITRE TTPs	A:0.97 P:0.96 R:0.99 F1:0.98	Custom
TTPXHunter [43]	2024	Domain-specific SecureBERT with Linear Classifier	MITRE TTPs	P:0.97 R:0.96 F1:0.97	Custom
Fengrui and Du [47]	2024	Fine-tuned LLaMA-2-7B	MITRE TTPs	P:0.86 R:0.89 F1:0.87	MITRE ATT&CK STIX
LLMCloudHunter [52]	2024	GPT-4.o	MITRE TTPs	P:0.66 R:0.74 F1:0.69	Custom

[a] Metrics: A: Accuracy P:Precession, R:Recall, F1-Score.

pipeline based on rule-based and information retrieval techniques. RcATT [40], a machine learning tool, applies Term Frequency–Inverse Document Frequency (TF–IDF) for classification but struggles with word context and synonyms. AttacKG [41] uses graph template matching but faces issues with adjectives describing threats. LADDER [42] enhances threat intelligence by extracting attack patterns with a three-step process. TTPHunter [9] utilises BERT and RoBERTa for contextual understanding but is limited by small datasets. MITRE's TRAM [8] uses SciBERT to improve TTP extraction with TTPXHunter further [43] being one of the latest state-of-the-art models using Domain-specific SecureBERT with Linear Classifier and showcasing improved extraction results.

Purba and Chu [32] and Siracusano et al. [44] utilised GPT-3.5 for tasks ranging from extracting Indicators of Compromise (IoCs) to generating structured CTI formats (e.g., STIX). Liu and Zhan [45] employed ChatGPT to create graphical representations of CTI data. Further advancements were made by Hu et al. [46] and Fengrui and Du [47], who explored the use of pretrained and fine-tuned LLM models. They leveraged GPT-3.5 and ChatGPT for data annotation and augmentation, respectively, to prepare datasets for fine-tuning the LLaMA2-7B model. Specifically, Hu et al. [48] utilised the fine-tuned LLaMA2-7B to construct knowledge graphs, while Fengrui and Du [47] focused on Tactics, Techniques, and Procedures (TTP) classification. CyNer [49] and TriCTI [50] developed a custom web crawler to retrieve unstructured text across selected trusted websites and manually annotated a subset for evaluation purposes. LLM-TIKG [51] also developed a custom web crawler to collect intelligence from selected platforms, but this study differs from the study presenting CyNer in that it utilises an LLM (GPT) for annotation. Finally, LLMCloudHunter [52] harnesses the power of LLMs to perform automated extraction of sigma rules from Cloud-Based CTI, incorporating a TTP extraction module. Table 1 provides an overview of current state of the art in extracting MITRE TTPs with the corresponding metrics and reference datasets used for their assessment.

2.3. Cybersecurity threat landscapes

Various commercial threat landscapes are annually published over the last years, focusing on different topics. Threat intelligence and nation-state activity reports such as CrowdStrike's Global Threat Report,[1] and Microsoft's Digital Defense Report[2] provide deep insights into APTs, cyber espionage, and nation-state-sponsored attacks. Reports that focus on breach analysis and incident patterns, such as the Verizon Data Breach Investigations Report (DBIR)[3] IBM X-Force Threat Intelligence Index,[4] and Cisco's Annual Cybersecurity Report,[5] offer detailed breakdowns of data breaches, attack vectors, and industry-specific insights.

ENISA publishes an annual Threat Landscape report[6] that identifies and assesses emerging and evolving cybersecurity threats. This report provides an analysis of trends based on observed incidents and developments. In what follows, we use ENISA's methodology [4] since it provides structured steps for CTL generation, emphasising the systematic identification, collection, and analysis of electronically stored information (ESI) to assess cyber risks. While robust, this methodology relies heavily on manual processes, which can be rather time-consuming and prone to human errors when applied to real-world scenarios. Cyber threat intelligence has emerged as a pivotal aspect of cybersecurity practices, aimed at understanding and anticipating cyber threats for proactive defence [44,45,53,54]. This field encompasses collecting, analysing, and disseminating information about current or potential attacks threatening information systems security [55]. By leveraging CTI, organisations can identify specific threats and vulnerabilities, enabling them to effectively tailor their security measures. This intelligence-gathering process draws from a wide array of sources, including historical cyber attack data [56], hacker forums, malware samples, and more, to provide a comprehensive view of the cyber threat landscape, which helps in enhancing the security posture against known threats and aids in predicting and mitigating future cyber attacks.

The value of our work lies in our unique approach, which leverages the power of cyber threat intelligence (CTI) extraction using AI/ML models to address the gap in automating cyber threat landscape generation reports. Although CTL generation methodologies exist, none has been extended to incorporate AI/ML models in the generation life cycle. Also, no prior research has quantified the benefits of such an approach in terms of quality, time or cost. Although a wide variety of methods and tools exist, primarily focusing on the extraction of a set of attributes (ex., Indicators of Compromise (IOCs), Threat Name, Attack Type, Actor

[1] https://www.crowdstrike.com/global-threat-report/
[2] https://www.microsoft.com/en-us/security/business/security-intelligence-report
[3] https://www.verizon.com/business/resources/reports/dbir/
[4] https://www.ibm.com/reports/threat-intelligence
[5] https://www.cisco.com/c/en/us/products/security/cybersecurity-reports.html
[6] The latest is published here https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024

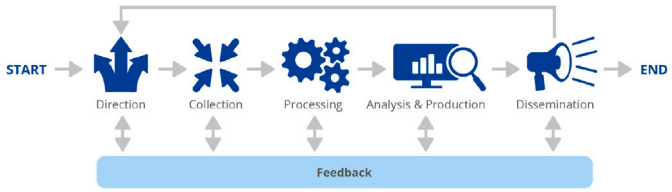


Fig. 1. Overview of ENISA CTL methodology, source [4].

Type, Target, Attack Source or Attack Destination), our research focuses on a subset of key attributes that comprise a cyber threat landscape report, with some being difficult to extract (attackers motivation, the affected sector, MITRE TTPs), even for human experts.

3. Proposed AI model comparison methodology

To better scope the flow of our work, we decided to strictly follow the steps of the well-defined ENISA CTL methodology (see Fig. 1) and pinpoint the part of the methodology in which AI models can contribute the most, potentially replacing human effort.

The output of the ENISA methodology is a set of actionable, timely, and accurate insights based on a collection of timestamped resources. It focuses on at least one sector and tries to answer questions like:

- Who are the potential adversaries and their motivations?
- What are the most common attack vectors and threats?
- What techniques are adversaries using to achieve their goals?

Using as a reference the content of the ENISA methodology templates ([4], B ANNEX: Reference Templates, Page 36), we decided to focus our effort on recreating insights relevant to five key attributes: (i) Attack Type, (ii) Motivation, (iii) Threat Name, (iv) Sector Affected, and (v) Tactics, Techniques, and Procedures (TTPs). Other attributes usually referenced in CTLs are Indicators of Compromise (IoCs) and weaponised vulnerabilities, but we decided to limit our scope to the five attributes mentioned above, as these cannot be easily extracted using pattern-based methods. For instance, a weaponised vulnerability could be referred to as a CVE, which is easy to extract with, e.g., regular expressions; nevertheless, a source report may describe the vulnerability, especially when there is no CVE for it. Similar approaches can be applied to hashes and URLs; nonetheless, one needs to be aware of the context of the URL, as it can have multiple interpretations in a report.

Our work starts with *collecting* the right dataset, then *processing* it using human analysts and different types of AI models to identify which of them better fits the assigned task. The results of our *analysis* can then determine which model or combination of models can be used to *produce* core parts of a CTL (Fig. 2). This work does not examine the automation of text generation based on the provided analysis results, a task that generative AI can also automate, but we consider it beyond the scope of this work.

3.1. Dataset selection

The core dataset used for this study was meticulously curated from open-source cybersecurity outlets over 12 months, ensuring a comprehensive representation of the evolving threat landscape. The dataset comprises 1000[7] cybersecurity-related texts, averaging 170 words per text and ranging from threat intelligence reports, security blog posts, to social media posts, comprising the Incident Corpus (IC) in English. These texts were selected to cover a diverse range of cybersecurity

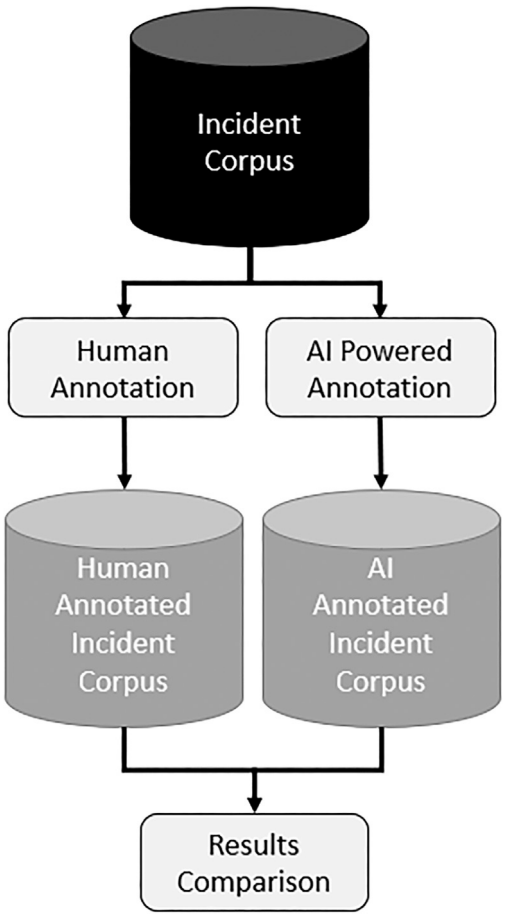


Fig. 2. Overview of our comparison methodology.

threats, including various types of cyberattacks, targeted sectors, underlying motivations, and specific threat actors' behaviour or malware used.

A complementary dataset of 100 texts, averaging 718 words per text, was synthetically generated only to cover the particular case of MITRE's TTP extraction, comprising the TTPs Corpus (TTPC) in English. Due to the fact that two of the models selected, namely TRAM and TTPHunter, for this task had a limitation on detecting only the top 50 MITRE techniques, we decided to generate an additional dataset focusing on these techniques to make evaluation against LLMs fair. The texts were generated using either descriptions or examples provided by the MITRE ATT&CK dataset for enterprise techniques v15.1[8] when characterising a technique. Using a text generator, we merged two or more techniques simultaneously, resulting in a synthesised report. Our experiments indicated that all techniques included in our TTP Corpus (TTPC) were correctly classified at least once by one of the tested models.

3.2. Dataset pre-processing

All text found in both corpuses were processed using Natural Language Processing (NLP) techniques. Initially, all text was converted to the UTF-8 encoding scheme. Using dictionaries and the Textblob library,[9] we corrected spelling errors and removed special characters.

[7] Note for reviewers: These articles represent a subset of the original corpus used to create the 2023 ENISA Threat Landscape [**Footnote will be deleted before publication]

[8] https://github.com/alexzacharis/AICTL/blob/main/MITRE/enterprise-attack-v15.1-techniques.xlsx

[9] https://github.com/sloria/TextBlob

Table 2
Cohen's Kappa consistency matrix.

Annotator		B						Total
	Category	Destruction	Disruption	Espionage	Financial Gain	Ideology	Unknown	
A	Destruction	165	1	0	0	0	0	**166**
	Disruption	1	315	0	1	0	0	**317**
	Espionage	0	0	94	1	0	1	**96**
	Financial Gain	0	1	1	362	1	0	**365**
	Ideology	0	1	1	1	25	0	**28**
	Unknown	0	0	0	0	0	28	**28**
Total		166	318	96	365	26	29	**1000**

Empty lines, specific stopwords, and specific punctuation marks were removed using traditional NLP libraries such as NLTK.[10] and spaCy[11] Moreover, all HTML or other programming codes were removed.

The standard Penn Treebank [57] tokenisation rules were used for sentence tokenisation, and finally, standardisation processes were applied to adjust the text to facilitate annotation. An example of a corpus text ready for annotation is the following: "text": "The National Institute of Cybersecurity Incibe has detected a phishing campaign impersonating WiZink bank".

```
{
  "text": "The National Institute of
     Cybersecurity Incibe has detected a
     phishing campaign impersonating
     WiZink bank."
}
```

All datasets used for this work can be found in the Research Data Policy and Data Availability Statements section.

4. Human annotation

To establish a reliable benchmark, two cybersecurity experts with the role of a Senior Cyber Threat Analyst[12] manually reviewed, categorised, and annotated each text. The common human experts' classifications served as the ground truth against which the AI models' performance was evaluated. The experts' deep domain knowledge ensured that the classifications accounted for the nuanced contexts that AI models might overlook. After completing the classification, an inter-annotator agreement assessment took place using Cohen's Kappa metric [58], and finally, the gold standard version of the IC was produced.

Each incident in IC was classified according to the ENISA taxonomy [59] using three categories of five attributes/tags describing the Attacker, the Attack and the Victim, which were scoped for extraction. The agreed values are in Table 4. The inter-annotator agreement (IAA) was validated using Cohen's Kappa [60]. The formula used is defined as follows:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

where p_0 expresses the relative observed agreement, and p_e is the hypothetical probability of chance agreement.

Cohen's Kappa was calculated separately for all five attributes. Table 2 shows an example of the contingency matrix for the motivation attribute with its Cohen Kappa calculated to 0.98. A summary of the calculation of Cohen's kappa values for the respective five attributes is depicted in Table 3.

Minor differences in classification were identified for *Motivation* and *Attack Type* attributes but were mediated after a second examination, and agreement was reached between the annotators. A higher number of differences were also identified in the human classification of the attributes of TTPs and *Sector*. Most of these cases were mediated upon the second examination, but in some cases, more than one tag could have been correctly attributed. To resolve this issue, a merging solution was proposed for cases where both categorisations were considered valid. This solution allows incidents to have more than one value for these attributes. The completion of this step produced the *Human Annotated Incident Corpus*.

Table 3
Cohen's Kappa values.

Attribute	Cohen's Kappa Value
Threat Name	100
Motivation	0.98
Attack Type	0.92
TTPs	0.81
Sector	0.83

Table 4
Classification values per tag.

Category	Tag-Attribute	Values
Attacker	Threat_Name	Any ATTACKER NAME or MALWARE NAME or NONE
	Motivation	DESTRUCTION, DISRUPTION, ESPIONAGE, FINANCIAL GAIN, IDEOLOGY, UNKNOWN
Attack	Attack_type	DATA, DOS/DDOS/RDOS, MALWARE, MISINFORMATION/DISINFORMATION, RANSOMWARE, SUPPLY CHAIN ATTACK, WEB THREATS,bZERO DAY, SOCIAL ENGINEERING THREATS, UNKNOWN
	Tactics, Techniques and Procedures	TTP ID [61]
Victim	Sector	CHEMICALS, DEFENSE, DRINKING WATER, EDUCATION, ENERGY, FINANCIAL, FOOD, GOVERNMENTAL, HEALTHCARE, ICT, LAW ENFORCEMENT & MILITARY, MANUFACTURING, MEDIA AND COMMUNICATION, POSTAL/COURIER, RETAIL, SERVICES, SPACE, TARGETED INDIVIDUAL, TRANSPORT, WASTE MANAGEMENT, WASTE WATER, UNKNOWN

5. Model selection and AI annotation

We have selected multiple LLMs for evaluation. These LLMs represent advanced state-of-the-art natural language processing models. More specifically, Gemini 1.5 Pro [6] (paid version) and GPT-4o [7] (paid version) have been evaluated along with BERT-based models, like TRAM [8] and TTPHunter [9] which were trained on the specific

[10] https://www.nltk.org/
[11] https://spacy.io/
[12] Two of the authors have this professional role

Table 5
Model usage per attribute extracted.

	Gemini 1.5 Pro	GPT-4o	TTPHunter	TRAM	AiCEF NER
Threat Name	✓	✓	–	–	✓
Motivation	✓	✓	–	–	–
Attack Type	✓	✓	–	–	✓
MITRE TTPs (*Top 50)	✓	✓	✓	✓	–
Sector	✓	✓	–	–	✓

extraction task.[13]. GPT-4o has already been benchmarked in Threat Intelligence Extraction tasks [52] while no references in the current literature have been identified for Gemini 1.5 Pro. The two BERT-based models (TRAM [8], TTPHunter) are selected since they are considered two of the best on the given task and are publicly available for testing. Models like DarkBert [62] were initially tested to determine whether they could be used for these tasks; however, they could not extract the needed information efficiently. Alongside the LLMs, we employed custom NER models tailored for cybersecurity text analysis, developed in our previous work [24]. These models were trained in domain-specific data and focused on extracting relevant cybersecurity incident-related information that was compliant with the taxonomy of the CTL methodology we chose to implement. The BERT and NER models are the baseline of the simplest form of trained-on-the-task AI models that can be used offline, utilising significantly fewer resources than LLMs while maintaining privacy. The latter is rather significant, as organisations may want to use sensitive reports as input for CTLs, and using them with a proprietary online LLM may leak critical information.

Each model was tasked with categorising the cybersecurity texts according to the five predefined attributes. The LLMs and NER models processed the texts independently, generating their respective categorisations. Table 5 summarises the AI models usage based on the extracted CTL attributes. The completion of this step produced the *AI Annotated Incident Corpus*.

6. Evaluation metrics & comparative analysis

To assess the performance of AI models, we utilised a combination of quantitative metrics focusing on the balance between correctly identifying categories (precision), the coverage of all relevant instances (recall), and the harmonic mean of precision and recall (F1-score). We performed a qualitative analysis of the errors to understand each model's misclassifications. This involved a review of cases where models deviated from the human-expert categorisation.

6.1. Comparative analysis and findings

Based on the F1 scores presented in Fig. 3, we can observe critical strengths and weaknesses in each model:

Gemini 1.5 Pro, an LLM not explicitly trained for the task, performs well on attributes such as Threat Name (0.9549) and Motivation (0.7450). However, it struggles with MITRE TTPs (0.5484), which may require more domain-specific knowledge for accurate detection.

GPT-4o (another LLM) has the highest score in Threat Name (0.9666), but its performance drops in Motivation (0.6598) and MITRE TTPs (0.4828). Like Gemini, it shows good generalisation overall but weaker detection for domain-specific attributes.

TTPHunter, a BERT model fine-tuned for TTP detection, performs poorly on MITRE TTPs (0.2213), which is its core focus. This finding, in particular, underlines the importance of testing pre-trained models in a given corpus subset to control the results. Even if a model is trained on a given task, it does not mean that results will be similar to the ones reported in its official release. The corpus on which the model was trained could differ significantly in the lexicological structure from that provided for our task, leading to significant deviations from the expected F1 score to the one achieved.

TRAM, another fine-tuned model for TTP detection, achieves a moderate F1 score on MITRE TTPs detection (0.4465), but is significantly better than TTPHunter.

AiCEF's NER, a named entity recognition model fine-tuned for cybersecurity tag extraction tasks, excelled in Attack Type (0.8929) and Sector (0.6411) detection but underperformed in other tasks, showing its strength in targeted information extraction but limited general capability compared to LLMs. The high performance of the NER model is attributed to the fact that it was trained using the exact same taxonomy used for the CTL experiment.

Overall, the LLMs; Gemini 1.5 Pro and GPT-4o, offered more generalisation across attributes but faltered in specialised areas like MITRE TTPs extraction. In contrast, task-specific models like AiCEF NER and TRAM excelled in narrowly defined attributes but lacked versatility across broader detection tasks.

6.2. Attribute-specific insights

Diving deeper into the statistics collected during the model annotation phase, we identified a series of valuable insights into the performance of each type of model used.

6.2.1. Threat name identification

Table 6 presents the performance of three models on Threat Name Detection using True Positives (TP), False Positives (FP), False Negatives (FN), Precision, Recall, and F1 Score metrics.

Gemini 1.5 Pro achieved perfect precision (1.0) and strong recall (0.914). Its F1 score of 0.9549 reflects its solid overall performance, although it misses a small number of cases (86 FNs).

GPT-4o outperformed Gemini, detecting 935 TPs with 0 FPs, giving it 1.0 precision and better recall (0.9349). This results in the highest F1 score (0.9666), making GPT-4o the best at balancing precision and recall.

AiCEF NER, with 713 TPs and 465 FPs, suffers from low precision (0.6111) and moderate recall (0.7309), leading to a much lower F1 score (0.6662). Its high false positives and lower detection rate highlight its limitations compared to large LLMs. This problem can be partially attributed to the training dataset, which was limited to incidents before 2023, indicating the importance of model updates.

In summary, Gemini 1.5 Pro and GPT-4o (large LLMs) excel in generalisation, while AiCEF NER, although a task-specific model, struggles with precision and recall, making it less reliable for this detection task.

Our research reveals a notable similarity between sources; see Table 7, regarding several notorious threat groups, such as NoName057(16), which all models constantly identified.

The LockBit ransomware group exhibits considerable discrepancies among the sources. The human count indicates 66 cases, although AiCEF NER reports 63, GPT-4o reports 64, and Gemini categorises the data into two groups: 50 for the "LockBit ransomware group" and 18 for "LockBit" (malware). This suggests a disparity in the models' recognition and classification of this group, which is also found in the literature. Similarly, RansomHouse and LeakBase are two entities that are only visible in the classification that humans provide and not in the output of Gemini 1.5 pro, GPT-4o, or AiCEF NER.

It is also worth noting that human experts could not annotate a threat name in 352 of the articles analysed. In many instances, the text did not clearly indicate the potential attacker.

[13] TTPXHunter model [43] is not yet available for evaluation, so we fall back to the publicly available models

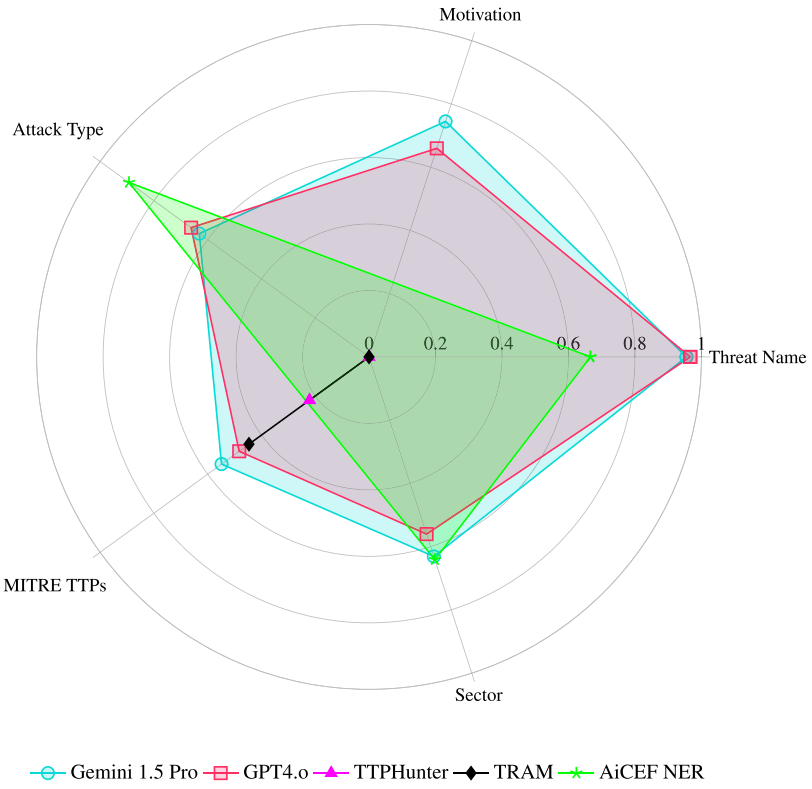


Fig. 3. F1 Score Comparison on Attribute Detection per Model (Spider Diagram).

Table 6
Performance metrics for threat name detection.

Threat Name (1000)	TP	FP	FN	Precision	Recall	F1 Score
Gemini 1.5 Pro	914	0	86	1	0.914	0.9549
GPT-4o	935	0	65	1	0.9349	0.9666
AiCEF NER	713	465	269	0.6111	0.7309	0.6662

6.2.2. Motivation identification

Table 8 compares the performance of two LLMs, Gemini 1.5 Pro and GPT-4o, in Motivation detection, given that no BERT or NER model was available to evaluate this task.

Gemini 1.5 Pro achieved a precision of 0.7631 and recall of 0.7276, leading to a solid F1 score of 0.7450. It balances precision and recall well, although some false positives and missed cases slightly impact performance.

GPT-4o detection resulted in lower precision (0.6074) but comparable recall (0.7201). The F1 score of 0.6598 is lower than that of Gemini Pro, mainly due to the model's higher false positive rate.

In summary, both models performed relatively well in the task given, but the results are not good enough to conclude that this task can be assigned exclusively to AI models.

Our dataset reveals substantial discrepancies in model performance among various categories (see Fig. 4). Gemini 1.5 Pro outperforms GPT-4o in most categories, but GPT-4o demonstrates strong performance in detecting 'Disruption' related events, although exhibits a significant decline in accuracy for 'Destruction', achieving a **0%**hit rate. This indicates that GPT-4o may have constraints in managing narrower or more specific threat classifications. Moreover, our analysis shows that Gemini 1.5 Pro performs subpar for the used dataset.

6.2.3. Attack type classification

Table 9 presents the performance metrics for Attack Type Detection across three models—Gemini 1.5 Pro, GPT-4o, and AiCEF NER.

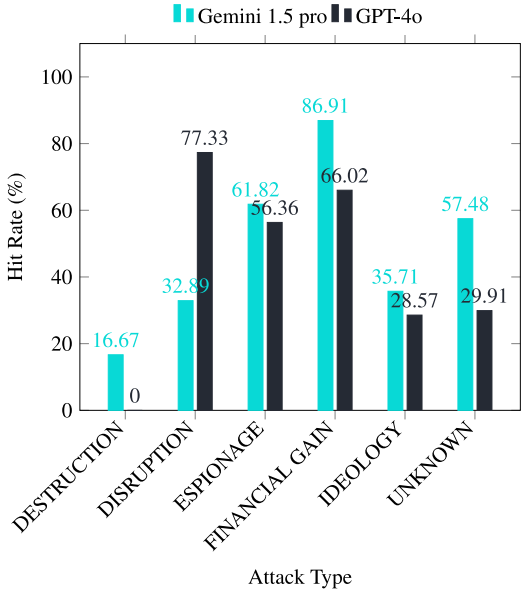


Fig. 4. Motivation (%) Model Hit Rate Comparison.

Gemini 1.5 Pro achieves a precision of 0.6632 and a recall of 0.6017, leading to an F1 score of 0.6311. Its relatively low recall indicates it misses many attack types, while a moderate false positive rate impacts precision.

GPT-4o has consistent precision and recall (0.6616). This balance results in a slightly better F1 score (0.6616) than Gemini, though both models struggle with false positives and miss detections.

AiCEF NER excels in this task, with a high precision of 0.9499 and a recall of 0.8414, leading to the best F1 score (0.8929). This model

Table 7
Attacker name True Positive Hit count excluding "UNKNOWN".

Attacker Name	Human Hits	Gemini 1.5 pro Hits	GPT-4o Hits	AiCEF NER Hits
NoName057(16)	110	111	107	110
LockBit	66	50 (LockBit ransomware group)/18 (LockBit)	64	63
Anonymous Sudan	35	34	35	–
Play ransomware group	18	24	16	–
BianLian	11	11	11	–
Royal ransomware group	8	–	8	–
RansomHouse	8	–	–	–
LeakBase	8	–	–	–
BlackCat ALPHV	7	6 (BlackCat/ALPHV ransomware group)	7	–

Table 8
Performance metrics for motivation detection.

Motivation (1000)	TP	FP	FN	Precision	Recall	F1 Score
Gemini 1.5 Pro	622	193	233	0.7631	0.7276	0.7450
GPT-4o	548	354	213	0.6074	0.7201	0.6598

Table 9
Performance metrics for attack type detection.

Attack Type (990)	TP	FP	FN	Precision	Recall	F1 Score
Gemini 1.5 Pro	595	302	395	0.6632	0.6017	0.6311
GPT-4o	655	335	335	0.6616	0.6616	0.6616
AiCEF NER	833	44	157	0.9499	0.8414	0.8929

performs significantly better, especially in precision, because of its low false positive rate. This can be attributed to its training using the same taxonomy used for the CTL construction.

Our findings are illustrated in Fig. 5. It is worth noting that all models performed well in identifying "Malware" and "Ransomware". For example, both Gemini 1.5 pro and GPT-4o achieved hit rates above **90%** in the "Ransomware" category. This consistency highlights the models' proficiency in handling common and well-defined cyber threats. On the contrary, the "Intrusion" category exposed significant weaknesses, particularly in Gemini 1.5 pro, where the hit rate was only **25%**, coupled with a false positive count of **45**. This indicates that the model may be overrelying on specific indicators that do not generalise well across different intrusion scenarios.

6.2.4. Sector identification

Table 10 summarises the performance metrics for Sector detection across three models—Gemini 1.5 Pro, GPT-4o, and AiCEF NER.

Gemini 1.5 Pro achieves a precision of 0.6659 and a recall of 0.6000. Its F1 score of 0.6312 indicates moderate performance, with a noticeable number of missed detections and false positives affecting overall effectiveness.

GPT-4o achieves lower precision (0.5477) and similar recall (0.5737). Its F1 score (0.5604) is the lowest among the three models, reflecting challenges in accurately identifying sectors exacerbated by a high false positive rate.

AiCEF NER yields a higher precision of 0.7382 and a lower recall (0.5670). The resulting F1 score (0.6411) indicates decent performance, although it still misses a significant number of actual sectors.

In Sector Detection, Gemini 1.5 Pro leads with the highest F1 score, followed closely by AiCEF NER, which demonstrates better precision but suffers from higher false negatives. Overall, while all models face difficulties in recall, Gemini 1.5 Pro shows the most balanced performance on this task.

Going more in-depth on the TP rates per sector, an overview can be seen in Fig. 6, where a pattern is clearly visible. All models struggle to categorise the Retail industry, while GPT-4o and AiCEf NER also score low in sectors like Healthcare, Finance, and Energy.

Table 10
Performance metrics for sector detection.

Sector (990)	TP	FP	FN	Precision	Recall	F1 Score
Gemini 1.5 Pro	578	290	384	0.6659	0.6000	0.6312
GPT-4o	533	440	396	0.5477	0.5737	0.5604
AiCEF NER	556	197	424	0.7382	0.5670	0.6411

Table 11
Performance Metrics for MITRE TTPs.

MITRE TTPs (345)	TP	FP	FN	Precision	Recall	F1 Score
Gemini 1.5 Pro	220	239	125	0.4792	0.6384	0.5484
GPT-4o	238	386	107	0.3806	0.6899	0.4828
TTPHunter	111	550	234	0.1675	0.3217	0.2213
TRAM	146	164	199	0.4710	0.4232	0.4465

6.2.5. MITRE TTP identification

Table 11 presents the performance metrics for detecting MITRE TTPs across four models: Gemini 1.5 Pro, GPT-4o, TTPHunter and TRAM. All models focused only on detecting the Top 50 Mitre TTPs using the TTP Corpus (TTPC). This limitation was enforced for the training of the BERT models (TTPHunter and TRAM) and was applied (at prompt level) to the LLMs (Gemini 1.5 Pro and GPT-4o) in order to achieve comparable results. We consider this a multi-labelling problem, since our dataset was generated to simulate the reality of reports containing more than one TTPs, and so the scoring has been done based on this assumption.

Gemini 1.5 Pro leads in precision (0.324) and recall (0.6389). Its F1 score of 0.43 indicates moderate performance, with a notable number of false positives affecting its overall effectiveness.

GPT-4o has lower precision (0.276) despite a higher recall (0.69). The F1 score of 0.394 reflects its struggle with false positives, indicating challenges in effectively identifying true positives.

TTPHunter yielded very low precision (0.1426) and recall (0.3214). The F1 score of 0.1982 signifies poor performance, primarily due to high false positive and false negative rates.

TRAM achieved a precision of 0.3204 and a recall of 0.4232. Its F1 score of 0.3667 indicates subpar performance, similar to Gemini but with a slightly lower recall.

In MITRE TTPs detection, Gemini 1.5 Pro performs best, particularly in the recall, but it suffers from a high false positive rate, leading to a moderate F1 score. GPT-4o shows slightly better detection capability with more true positives but also struggles with precision. TTPHunter exhibits significantly poorer performance than TRAM, with TTPHunter having the lowest metrics overall.

Although TRAM has a lower detection rate than GPT-4o, it produces fewer false positives, leading to very similar F1 scores. All models face challenges in balancing precision and recall, with high false positive rates being a critical issue. This highlights the difficulty of accurately detecting MITRE TTPs as a task overall.

6.3. Key performance findings

On *Threat Name detection* detection, both LLMs (Gemini 1.5 Pro and GPT-4o) demonstrated high precision and high recall, indicating that

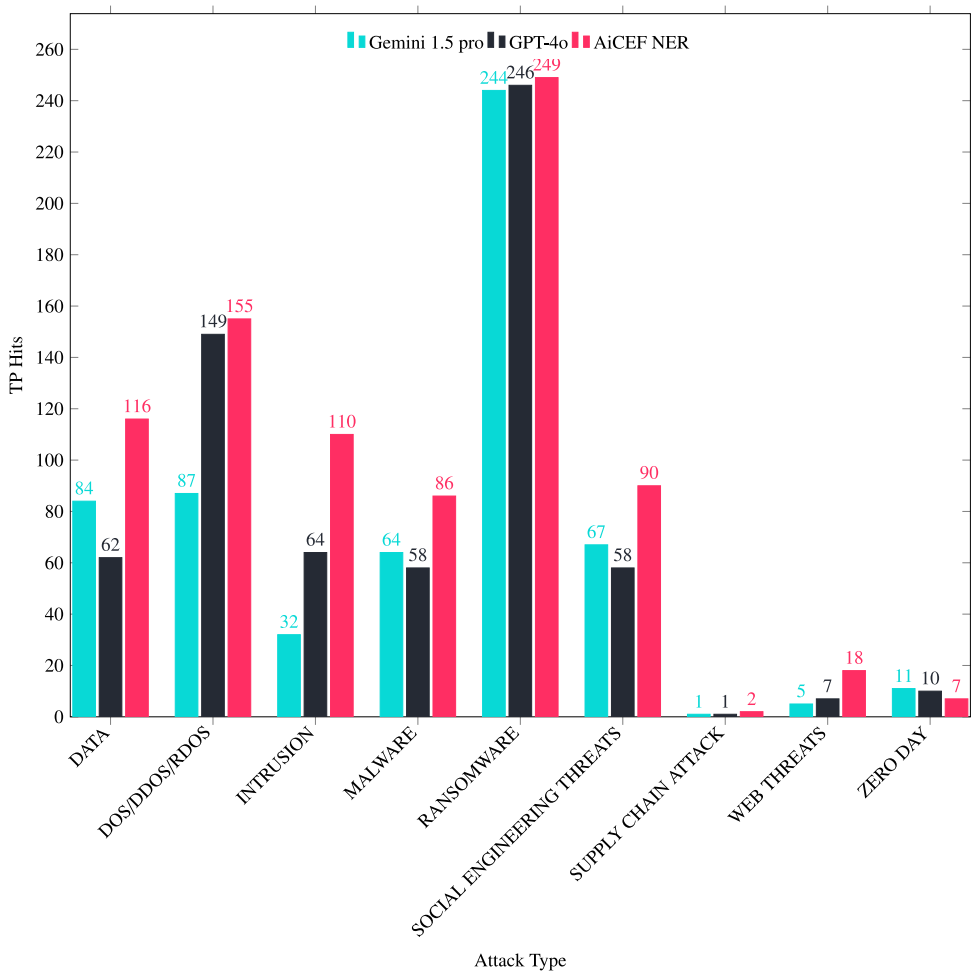


Fig. 5. True Positive (TP) Comparison for Attack Type.

LLMs can effectively identify threat names with minimal false positives, making them suitable for real-time applications, including intelligence extraction for CTL creation.

On *Motivation* detection, moderate performance was achieved overall. Gemini 1.5 Pro outperformed GPT-4o in precision and F1 score, indicating that LLMs can generally effectively identify the motivation behind threat actor actions. The high false positive rate is still an issue. For this reason, LLMs can be used as a first step to a hybrid classification scheme where humans can verify the annotations proposed. Custom models trained on the task may enhance precision further for specific contexts, but they should be rigorously evaluated to avoid high false positives.

On *Attack Type* detection, AiCEF's NER outperformed both LLMs, showcasing the effectiveness of task-specific models for precise attribute extraction and can be used in intelligence extraction for CTL creation.

On *Sector* detection, both LLMs performed poorly in precision and recall, indicating a significant gap in their effectiveness for sector detection. AiCEF NER showed better precision but faced challenges in recall, highlighting that while custom models can be more trusted than LLMs, there is plenty of room for improvement. AI models would benefit from access to more diverse and representative training data to improve their performance in identifying specific sectors as defined in the threat landscape to be developed.

Based on these findings, both LLMs and NER models can be used as the first step in a hybrid classification scheme in which humans can verify the annotations (pre-tagging).

In MITRE TTP detection, all models struggled with low detection and high false positive rates. TRAM model performed relatively well compared to LLMs. This approach can ensure a relatively robust and time-friendly solution for analysts while also limiting the risk of exposing threat-related information to public LLM providers.

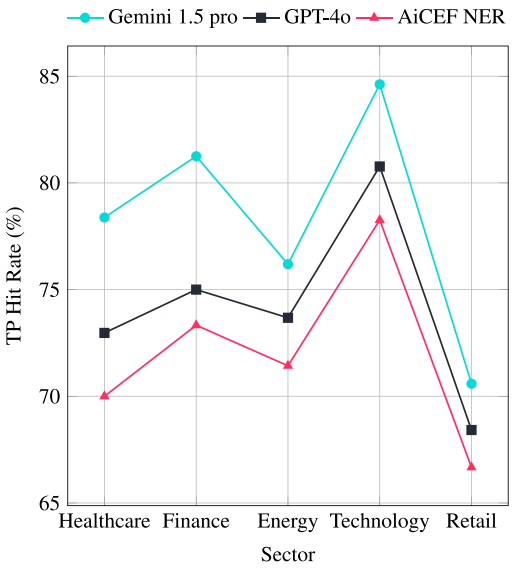


Fig. 6. True Positive Hit Rates (%) breakdown by Model per Sector.

Table 12
Annotation resources per task breakdown.

	Human Expert	Gemini 1.5 Pro	GPT-4o	TRAM	ThreatHunter	AiCEF NER
Average Time (min) per Text						
Threat Name - Motivation - Attack Type - Sector	10	0.02	0.02	n/a	n/a	0.01
MITRE TTPs	15	0.3	0.3	1	0.1	n/a
Average Cost (€) per Text						
Threat Name - Motivation - Attack Type - Sector	n/a	0,023	0,065	n/a	n/a	0,003
MITRE TTPs	n/a	0,023	0,14	0,003	0,003	n/a

More generally, locally trained MITRE TTP detection remains challenging and requires further exploration and enhancement. To address this observation, we propose a novel methodology that does not rely on model retraining but on model crowdsourcing.

6.4. Resource comparison findings

During our research, we attempted to estimate the potential savings for an organisation that aims to build its own CTL by automating the intelligence extraction process using AI. While these findings, presented in Table 12, are preliminary and should be treated with caution (e.g., we did not take into account the time needed to correct the FP and FN), these estimates can provide a good indication of what are the potential financial and time savings for an organisation if such approaches will be improved.

We also recorded the average time needed to annotate a text, and we discovered that, on average, it takes approximately 10 min to read the contents, assign the tags and reach an agreement between two annotators for the four attributes (Threat Name – Motivation – Attack Type – Sector). TTP extraction requires more time, as the text is usually richer in threat information and typically contains more technical details. On average, we recorded that it takes 15 more minutes per text. For the AI models, while the time needed to extract the basic tags is negligible, extracting TTPs is compute-intensive and may take up to 1 min for longer texts.

In terms of costs, calculating the exact price varies heavily on the experts' seniority and the rate per country. We estimate[14] that for a dataset of 1000 articles, the average costs for annotation of the five attributes examined will be above €10,000. We will refrain from making an exact prediction as part of this work, as we believe it to be beyond the scope of this work, but the number above can indicate the range of improvement achieved.

Using GPT-4o, the cost to extract the tags for the same dataset was approximately €65, roughly €0065 per text. The extraction of TTPs is estimated to cost approximately €140, roughly €0,14 per text. In total, extracting all attributes from 1000 articles using only GPT-4.o would cost €205.

Gemini 1.5 Pro analysis was covered by the free tier of a standard Premium subscription for one month, for a total of €23 (flat rate), which covered the extraction of all attributes from the 1000 texts. For easy comparison, we divided the flat cost by 1000 articles analyses, ending with a price of €0023. Still, many more articles could have been analysed under the same flat rate charge, leading to lower cost per text.

Similarly, a flat rate could be applied to the local models. We have calculated the costs of spinning off and installing the needed packages on a virtual machine in a public cloud provider. Analysing the entire dataset would cost approximately €3, or €0003 per article. Obviously, more texts could have been analysed under the same flat rate charge, leading to a lower cost per entry.

Finally, specific aspects such as the expert time needed for training, verifying the tags detected by the model and opportunities for fine-tuning would need to be accounted for to accurately estimate the potential savings for a given task, such as collecting statistical threat data related to a particular sector. Nevertheless, for such a task, based on the limited tests we performed, we saw a decrease from 10 to 2 min needed per text to extract non-TTP related attributes (time spent to check the pre-annotated tags).

6.5. Proposed annotation workflow

Considering the above, the use of AI models for intelligence extraction for CTL generation can be represented in a hybrid workflow, as seen in Fig. 7. Specific types of AI models perform better than others based on the attribute annotated and should be preferred, as indicated by the results collected. Attributes like Threat Name and Attack Type can be annotated purely by AI models, while others like Motivation, Sector and MITRE TTPs still require human curating. AI can be utilised as an initial annotation prior to or in parallel with human annotation and play a supportive role in the decision-making of the categorisation of these attributes.

7. Optimisation through voting-based agreement methodology

Based on the findings presented above, we decided to test whether a classic systematic voting mechanism [63] could ensure consensus between multiple AI models, denoted as $M_1, M_2, \ldots, M_n$, to resolve the problem of high false positives observed in the MITRE TTP extraction task. The proposed method uses voting to determine the final consensus output. For this experiment, we decided to test two algorithms, one based on simple majority voting (Algorithm 1) and one based on weighted voting (Algorithm 2) where each model is assigned a weight based on its historical performance, accuracy, or confidence in its predictions. Finally, we also test Ensemble learning, a machine learning technique aggregating two or more models to produce better predictions. In other words, an ensemble model combines several individual models to make more accurate predictions than a single model alone [64].

For both methods, we assume that:

- All models are pre-trained and exhibit competence in the specified domain.
- Each model produces outputs represented as one or more MITRE TTP labels.
- For the algorithm to work, at least three models should vote.

7.1. Simple majority voting

In the simple majority method, the goal is to aggregate the outputs of the models to find a consensus decision. The label is accepted if N models agree on a label based on a user-defined majority *Threshold*. We set the threshold to **Threshold=2**, where at least two models agree out of three models available and **Threshold=3**, where three models agree for four models available.

[14] https://www.salaryexplorer.com/average-salary-wage-comparison-germany-cybersecurity-analyst-c81j19087

Algorithm 1 The simple majority voting method.

Input: Set of AI models $\{M_1, M_2, \ldots, M_n\}$
Input: Test instance X
Input: User-defined threshold T
Output: Aggregated consensus output C
for each model M_i in $\{M_1, M_2, \ldots, M_n\}$ **do**
  Get output O_i from $M_i(X)$
end for
Initialise a count dictionary count for TTP labels
for each output O_i **do**
  Increment count of O_i in count
end for
Initialise an empty list accepted_labels
for each TTP label C_j in count **do**
  if count of $C_j \geq T$ **then**
    Add C_j to accepted_labels
  end if
end for
if accepted_labels is empty **then**
  Return: No consensus found
end if
Find the TTP label C_{max} with the highest count in count from accepted_labels
Set $C \leftarrow C_{max}$
return C

7.2. Weighted voting

For the weighted voting method, we assume that each model has its strengths in terms of coverage and precision so that we would assign weights to their predictions based on these factors. A trial and error approach was followed to verify our intuition, thoroughly testing all possible voting weighted combinations with weights between 0.5 and 5 for each model, for groups of three or more models. Defining the right threshold was also important, so we decided to test values of five and above (>=5) to identify the best F1 score achieved.

For the three models weighted voting system (Gemini 1.5 Pro - GPT-4o - TRAM), after testing all possible weight combinations, we identified that Threshold=5 yields, 305 combinations scored above the threshold of 0.5484. The maximum F1 score achieved was 0.5924. As far as the best-performing weight ranges per model, we concluded that ideal ranges of values are: Gemini 1.5 Pro: [5], GPT-4o:[0.5-4.5], TRAM:[0.5-4.5] where GPT-4o and TRAM combined score should be at least five.

For the four model weighted voting system (Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter), after testing all possible weight combinations for a threshold set to T=5, 1881 combinations scored above the threshold of 0.5484. The maximum F1 score achieved was 0.5958. As far as the best-performing weight ranges per model, we concluded that the ideal ranges of values are: Gemini 1.5 Pro: [5], GPT-4o:[0.5-4], TRAM:[0.5-4], TTPHunter:[0.5-4], where GPT-4o, TRAM and TTPHunter combined scores should be five or above.

Based on the previous F1 performance of these models, we adopt the following weighting strategy: A max weight of 5 to the Gemini 1.5 Pro model to leverage its superior F1 score and a balanced weight of 0.5-4.5 distributed between GPT-4o, TRAM, and TTPHunter to benefit of their contributions. A distribution analysis of the weight combinations that achieve the maximum F1 score using the weighted voting method can be found in Fig. 8. Finally, the optimal threshold is Threshold=5 for the method proposed.

7.3. Ensemble learning: Stacking model

In this attempt, we combined the predictions of individual models as features and trained a meta-model to make the final prediction. This

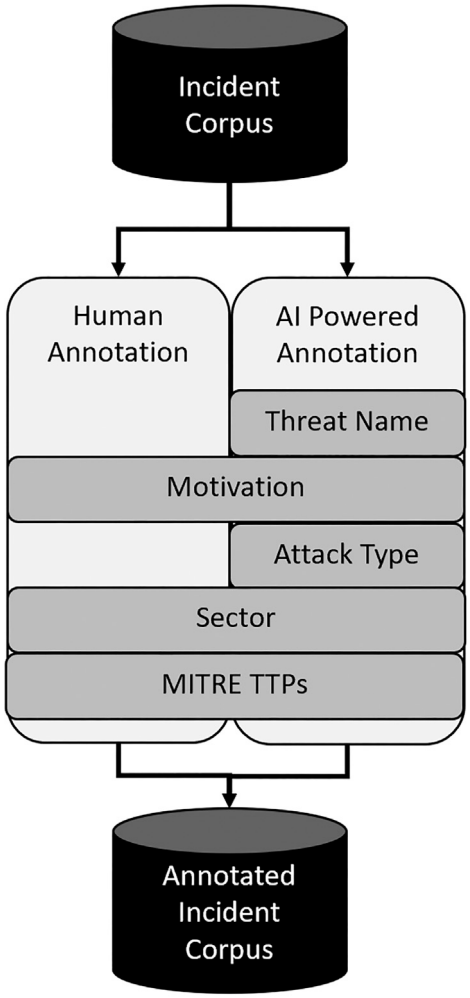


Fig. 7. Proposed hybrid annotation methodology.

Algorithm 2 The weighted voting method.

Input:
  M: Set of models $\{M_1, M_2, \ldots, M_n\}$
  W: Set of corresponding weights for each model $\{w_1, w_2, \ldots, w_n\}$
  $T_{predicted}$: Set of techniques predicted by each model M_i
  $Threshold$: Voting threshold (e.g., 5)
  T: Set of all possible techniques
Output:
  T_{final}: Set of final predicted techniques
Initialise:
  $Votes[t] \leftarrow 0, \forall t \in T$ {Initialise vote count for each technique}
for each model $M_i \in M$ **do**
  for each predicted technique $t \in T_{predicted}[M_i]$ **do**
    $Votes[t] \leftarrow Votes[t] + w_i$ {Add weight of model M_i to the technique t}
  end for
end for
$T_{final} \leftarrow \emptyset$ {Initialise the set of final predicted techniques}
for each technique $t \in T$ **do**
  if $Votes[t] \geq Threshold$ **then**
    $T_{final} \leftarrow T_{final} \cup \{t\}$ {Include t in the final prediction set if vote exceeds threshold}
  end if
end for
return T_{final}

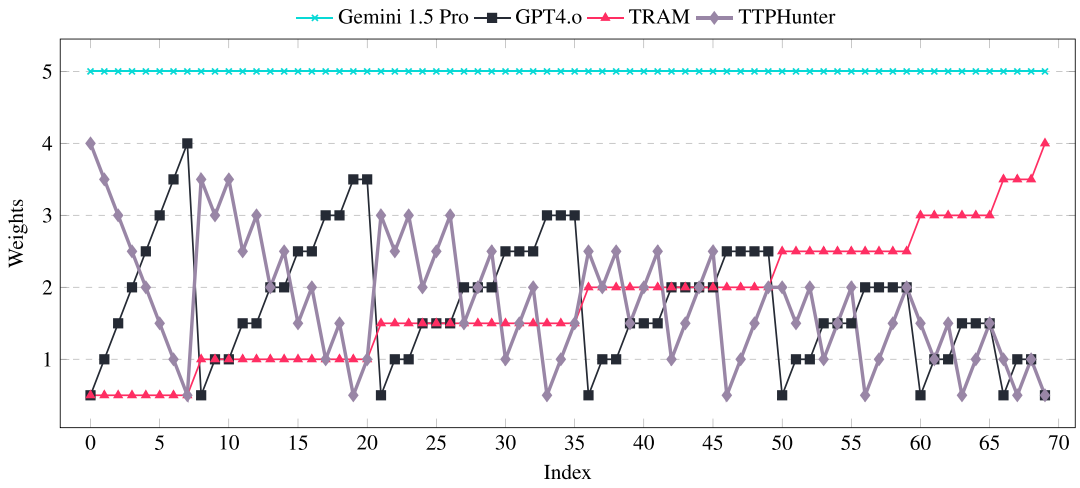


Fig. 8. Distribution of Weights achieving maximum F1 Score (0.5958) in MITRE TTP extraction.

meta-model learns how to weigh the outputs of each model and how to aggregate them in an optimal way.

Initially, we assume that each pre-trained model (Gemini, GPT, TRAM, and TTPHunter) generates its own predictions. Moreover, we assume that the outputs from these models form a new dataset (i.e., the ensemble's feature space). Thus, we train a simple classifier on this new dataset to predict the final TTP labels.

In our case, the implementation was adapted to train the stacking meta-model using multi-label classification. The Multi Output Classifier[15] was used to wrap around the Logistic Regression model to handle multi-label output. To this end, we collect the predictions from the four models for each sample in our dataset and use them as inputs (features) to a meta-model. We train the meta-model on the actual ground truth labels to learn the best way to combine the predictions from the base models.

7.4. Voting results comparison

Upon reviewing Table 11, it is clear that our proposed methodology surpasses the maximum F1 score of the fittest model used (Gemini 1.5 Pro. F1 score: 0.5484) and reduces the number of false positives below the average rate of the models combined (Average FP (3 models): 263, Average FP (4 models):334,75) to be considered as an optimisation to the Intelligence Extraction process of CTL generation.

We achieved our objective after running all possible combinations of the four available models (Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter) for both voting methods and training a stacking model. Table 13 showcases the successful combinations in which the previous best F1 score is improved without retraining existing models while the false positive rate is also reduced.

Weighted voting is the most effective of the two traditional voting methodologies. The four-model combination (Gemini 1.5 Pro—GPT-4o—TRAM—TTPHunter) achieved a 4.74

The simple majority voting method achieved an F1 score increase of 3.11% utilising only the 3-model combination of Gemini 1.5 Pro - GPT-4o - TRAM, reducing the false positive rate by 32 units. Simple majority voting for four models failed to improve the F1 score. This potentially indicates that the fourth model used (TTPHunter) degraded performance.

Ensemble learning and, more specifically, a newly trained stacking model (based on predictions of all four models Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter) using ninety (90) samples to predict 10 out of our available dataset yielded satisfactory results with an F1 score of 0.5714, translating to a 2.30% increase in the F1 score.

If we have a larger dataset of annotated text, we believe that a more robust model can be trained to extract tags for the MITRE's TTP attribute.

8. Conclusions and future work

Our evaluation of various AI models for intelligence extraction in CTL generation has demonstrated significant potential for automating and optimising this process. By comparing the performance of LLMs such as Gemini 1.5 Pro and GPT-4o, alongside BERT-based models such as TRAM and TTPHunter, as well as custom NER systems against human performance, we found that AI-driven approaches can streamline threat intelligence extraction, reducing the burden on human analysts and cost, while ensuring timely and precise data processing.

LLMs like Gemini 1.5 Pro and GPT-4o demonstrated strong generalisation and can be trusted for attribute extraction, such as Threat Name. In contrast, models like AiCEF NER excelled in targeted attribute extraction, such as Attack Type and Sector, but lacked the versatility of LLMs in handling broader detection tasks. In extracting motivation-related intelligence, LLMs produced average results. Without more specialised task models, the results are inconclusive. A topic extraction model could be trained to achieve better intelligence extraction for the motivation attribute, surpassing the current performance of LLMs. In detecting specialised attributes like MITRE TTPs, LLM models had an average performance with a high false positive rate. In contrast, specialised models for the task, like TRAM and TTPHunter, performed poorly. This task is clearly challenging for AI models and humans alike, and new methodologies must be proposed to address it.

Automating the intelligence extraction process using AI models presents significant potential for reducing costs and time spent compared to human annotation. While human annotation of a dataset of 1000 articles would cost some thousands of euros, AI models can drastically reduce these costs to as low as €200 under specific subscription models. Additionally, the time savings are substantial as AI models almost instantly complete the basic tag extraction. In contrast, TTP extraction, which is more compute-intensive, is still much faster than manual methods. With further refinement and integration of hybrid approaches, organisations could potentially reduce the costs of CTL generation by more than half while significantly accelerating the process. In the future, we intend to expand our proposed voting-based extraction scheme to all attributes to assess better the potential of our approach to more than one threat intelligence attribute.

Our proposed methodology focuses on MITRE TTP extraction, leverages existing models without retraining, and demonstrates promising

[15] https://scikit-learn.org/dev/modules/generated/sklearn.multioutput.MultiOutputClassifier.html

Table 13
Optimisation Methods Comparison: Simple Majority Voting vs. With Weights vs. Stacking Model.

Model Combination	Precision	Recall	F1 Score	Improvement
Simple Majority Voting				
Gemini 1.5 Pro - GPT-4o - TRAM	0.4979	0.6812	0.5795	+3.11%
Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter	0.4649	0.6957	0.5585	+1.01%
With Weights				
Gemini 1.5 Pro - GPT-4o - TRAM	0.5010	**0.7246**	0.5924	+ 4.40%
Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter	0.5194	0.6986	**0.5958**	**+4.74%**
Stacking Model				
Gemini 1.5 Pro - GPT-4o - TRAM - TTPHunter	**1.0000**	0.4000	0.5714	+2.30%

results and significant potential for future research. Using voting-based methods and ensemble learning, we outperformed the highest F1 score achieved by an individual model (Gemini 1.5 Pro with an F1 score of 0.5484) on this rather complex extraction topic. Specifically, a weighted voting combination of four models (Gemini 1.5 Pro, GPT-4.o, TRAM, TTPHunter) achieved a 4.74% improvement of the F1 score, dropping the FP rate by 16 points and improving TP by 21 points. The simple majority voting method also delivered a 3.11% F1 score improvement using the three-model combination, reducing the false positive rate by 32 units. Additionally, a stacking model trained on a small dataset showed a 2.30% increase in F1 score, with the potential for further optimisation as more data becomes available. Our proposed methodology can be extended to any combination of LLM and BERT-based models and easily extended to cover extraction of all MITRE TTPs, a path we intend to explore.

The presented approach can potentially automate and improve a cybersecurity expert's overall throughput when preparing organisational threat landscapes. In particular, we estimate that by leveraging our proposed methodology, a cybersecurity expert presented with an AI pre-annotated report could, depending on their objectives, boost their performance by up to 80% (see Table 12). As mentioned, this performance enhancement will depend on the error-rate tolerance level. If the analysis is performed to compile a threat landscape that would feed the broad risk management framework of the organisation, the focus could be on detecting high-confidence tags while processing the maximum amount of reports. Similarly, identifying controls that would be adequate to respond to the broadest range of TTPs relevant to the operating context of the organisation may require more sophisticated and specialised models (training time would need to be factored in). We intend to research and verify some of these hypotheses in our future research.

Research data policy and data availability statements

The datasets analysed as part of this work are available at https://github.com/alexzacharis/AICTL.

CRediT authorship contribution statement

Alexandros Zacharis: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Razvan Gavrila:** Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis, Data curation. **Constantinos Patsakis:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Formal analysis, Conceptualization. **Christos Douligeris:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Constantinos Patsakis reports financial support was provided by European Commission. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the European Commission under the Horizon Europe Programme, as part of the projects CyberSecPro (https://www.cybersecpro-project.eu) (Grant Agreement no.101083594) and NERO (https://nerocybersecurity.eu/) (Grant Agreement no. 101127411). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Data availability

Data has been published on GitHub.

References

[1] Tang B, Wang J, Yu Z, Chen B, Ge W, Yu J, et al. Advanced Persistent Threat intelligent profiling technique: A survey. Comput Electr Eng 2022;103:108261.
[2] Rahman MR, Williams L. From threat reports to continuous threat intelligence: a comparison of attack technique extraction methods from textual artifacts. 2022, arXiv preprint arXiv:2210.02601.
[3] Zhou Y, Tang Y, Yi M, Xi C, Lu H. CTI view: APT threat intelligence analysis system. Secur Commun Networks 2022;2022(1):9875199.
[4] ENISA. ENISA Cybersecurity Threat Landscape Methodology. 2022, https://www.enisa.europa.eu/publications/enisa-threat-landscape-methodology/.
[5] Google. Google Gemini. 2023, https://gemini.google.com/.
[6] Reid M, Savinov N, Teplyashin D, Lepikhin D, Lillicrap T, Alayrac J-b, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. 2024, arXiv preprint arXiv:2403.05530.
[7] OPENAI. GPT-4o. 2023, https://platform.openai.com/docs/models/gpt-4o.
[8] MITRE. Threat Report ATT&CK Mapper (TRAM). 2023, https://github.com/center-for-threat-informed-defense/tram/.
[9] Rani N, Saha B, Maurya V, Shukla SK. TTPHunter: Automated extraction of actionable intelligence as TTPs from narrative threat reports. In: Proceedings of the 2023 australasian computer science week. ACM; 2023, p. 126–34. http://dx.doi.org/10.1145/3579375.3579391.
[10] Bhardwaj G, Gupta R, Srivastava AP, Singh SV. Cyber threat landscape of G4 nations: Analysis of threat incidents & response strategies. In: 2021 2nd international conference on intelligent engineering and management. IEEE; 2021, p. 75–9.
[11] Eto M, Inoue D, Song J, Nakazato J, Ohtaka K, Nakao K. Nicter: A large-scale network incident analysis system: Case studies for understanding threat landscape. In: Proceedings of the first workshop on building analysis datasets and gathering experience returns for security. 2011, p. 37–45.
[12] Kovanen T, Nuojua V, Lehto M. Cyber threat landscape in energy sector. In: ICCWS 2018 13th international conference on cyber warfare and security. Academic Conferences and publishing limited; 2018, p. 353.
[13] Satyapanich T, Ferraro F, Finin T. CASIE: extracting cybersecurity event information from text. In: The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020. AAAI Press; 2020, p. 8749–57. http://dx.doi.org/10.1609/AAAI.V34I05.6401.
[14] Satvat K, Gjomemo R, Venkatakrishnan V. Extractor: Extracting attack behavior from threat reports. In: 2021 IEEE European symposium on security and privacy. IEEE; 2021, p. 598–615.
[15] Sarhan I, Spruit M. Open-cykg: An open cyber threat intelligence knowledge graph. Knowl-Based Syst 2021;233:107524.
[16] Park Y, Lee T. Full-stack information extraction system for cybersecurity intelligence. In: Proceedings of the 2022 conference on empirical methods in natural language processing: industry track. 2022, p. 531–9.

[17] Ahmed K, Khurshid SK, Hina S. CyberEntRel: Joint extraction of cyber entities and relations using deep learning. Comput Secur 2024;136:103579.

[18] Zhang H, Shen G, Guo C, Cui Y, Jiang C. EX-action: Automatically extracting threat actions from cyber threat intelligence report based on multimodal learning. Secur Commun Networks 2021;2021(1):5586335.

[19] Sun T, Yang P, Li M, Liao S. An automatic generation approach of the cyber threat intelligence records based on multi-source information fusion. Futur Internet 2021;13(2):40.

[20] Gao P, Shao F, Liu X, Xiao X, Liu H, Qin Z, et al. A system for efficiently hunting for cyber threats in computer systems using threat intelligence. In: 2021 IEEE 37th international conference on data engineering. ICDE, IEEE; 2021, p. 2705–8.

[21] Gao C, Zhang X, Liu H. Data and knowledge-driven named entity recognition for cyber security. Cybersecurity 2021;4(1):9.

[22] Suryotrisongko H, Ginardi H, Ciptaningtyas HT, Dehqan S, Musashi Y. Topic modeling for cyber threat intelligence (cti). In: 2022 seventh international conference on informatics and computing. ICIC, IEEE; 2022, p. 1–7.

[23] Abdi H, Bagley SR, Furnell S, Twycross J. Automatically labeling cyber threat intelligence reports using natural language processing. In: Proceedings of the ACM symposium on document engineering 2023. 2023, p. 1–4.

[24] Zacharis A, Patsakis C. AiCEF: an AI-assisted cyber exercise content generation framework using named entity recognition. Int J Inf Secur 2023. http://dx.doi.org/10.1007/s10207-023-00693-z.

[25] Ferrag MA, Ndhlovu M, Tihanyi N, Cordeiro LC, Debbah M, Lestable T. Revolutionizing cyber threat detection with large language models. 2023, arXiv preprint arXiv:2306.14263.

[26] Lykousas N, Patsakis C. Decoding developer password patterns: A comparative analysis of password extraction and selection practices. Comput Secur 2024;145:103974. http://dx.doi.org/10.1016/J.COSE.2024.103974.

[27] Koide T, Fukushi N, Nakano H, Chiba D. Detecting phishing sites using chatgpt. 2023, arXiv preprint arXiv:2306.05816.

[28] Chrysanthou A, Pantis Y, Patsakis C. The anatomy of deception: Measuring technical and human factors of a large-scale phishing campaign. Comput Secur 2024;140:103780.

[29] Roy SS, Thota P, Naragam KV, Nilizadeh S. From chatbots to PhishBots?–Preventing phishing scams created using ChatGPT, google bard and claude. 2023, arXiv preprint arXiv:2310.19181.

[30] Patsakis C, Casino F, Lykousas N. Assessing LLMs in malicious code deobfuscation of real-world malware campaigns. Expert Syst Appl 2024;256:124912. http://dx.doi.org/10.1016/J.ESWA.2024.124912.

[31] Song Z, Tian Y, Zhang J, Hao Y. Generating fake cyber threat intelligence using the gpt-neo model. In: 2023 8th international conference on intelligent computing and signal processing. IEEE; 2023, p. 920–4.

[32] Purba MD, Chu B. Extracting actionable cyber threat intelligence from Twitter stream. In: 2023 IEEE international conference on intelligence and security informatics. IEEE; 2023, p. 1–6.

[33] Sufi F. An innovative GPT-based open-source intelligence using historical cyber incident reports. Nat Lang Process J 2024;7:100074.

[34] Liao X, Yuan K, Wang X, Li Z, Xing L, Beyah R. Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016, p. 755–66.

[35] Zhu Z, Dumitras T. Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports. In: 2018 IEEE European symposium on security and privacy. IEEE; 2018, p. 458–72.

[36] Zhao J, Yan Q, Li J, Shao M, He Z, Li B. TIMiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data. Comput Secur 2020;95:101867.

[37] Husari G, Al-Shaer E, Ahmed M, Chu B, Niu X. Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources. In: Proceedings of the 33rd annual computer security applications conference. 2017, p. 103–15.

[38] Robertson SE, Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR'94: proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval, organised by Dublin city university. Springer; 1994, p. 232–41.

[39] Gao P, Shao F, Liu X, Xiao X, Qin Z, Xu F, et al. Enabling efficient cyber threat hunting with cyber threat intelligence. In: 2021 IEEE 37th international conference on data engineering. IEEE; 2021, p. 193–204.

[40] Legoy V, Caselli M, Seifert C, Peter A. Automated retrieval of att&ck tactics and techniques for cyber threat reports. 2020, arXiv preprint arXiv:2004.14322.

[41] Li Z, Zeng J, Chen Y, Liang Z. AttacKG: Constructing technique knowledge graph from cyber threat intelligence reports. In: European symposium on research in computer security. Springer; 2022, p. 589–609.

[42] Alam MT, Bhusal D, Park Y, Rastogi N. Looking beyond iocs: Automatically extracting attack patterns from external CTI. In: Proceedings of the 26th international symposium on research in attacks, intrusions and defenses. 2023, p. 92–108.

[43] Rani N, Saha B, Maurya V, Shukla SK. TTPXHunter: Actionable threat intelligence extraction as TTPs form finished cyber threat reports. 2024, arXiv preprint arXiv:2403.03267.

[44] Siracusano G, Sanvito D, Gonzalez R, Srinivasan M, Kamatchi S, Takahashi W, et al. Time for action: Automated analysis of cyber threat intelligence in the wild. 2023, arXiv preprint arXiv:2307.10214.

[45] Liu J, Zhan J. Constructing knowledge graph from cyber threat intelligence using large language model. In: 2023 IEEE international conference on big data. IEEE; 2023, p. 516–21.

[46] Hu Y, Zou F, Han J, Sun X, Wang Y. Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model. Comput Secur 2024;145:103999.

[47] Fengrui Y, Du Y. Few-shot learning of TTPs classification using large language models. 2024, http://dx.doi.org/10.20944/preprints202401.0372.v1, preprints.org.

[48] Zhang Y, Chen J, Cheng Z, Shen X, Qin J, Han Y, et al. Edge propagation for link prediction in requirement-cyber threat intelligence knowledge graph. Inform Sci 2024;653:119770.

[49] Fujii S, Kawaguchi N, Shigemoto T, Yamauchi T. Cyner: Information extraction from unstructured text of cti sources with noncontextual iocs. In: International workshop on security. Springer; 2022, p. 85–104.

[50] Liu J, Yan J, Jiang J, He Y, Wang X, Jiang Z, et al. TriCTI: an actionable cyber threat intelligence discovery system via trigger-enhanced neural network. Cybersecurity 2022;5(1):8.

[51] Hu Y, Zou F, Han J, Sun X, Wang Y. Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model. 2023, Available at SSRN: https://ssrn.com/abstract=4671345 or http://dx.doi.org/10.2139/ssrn.4671345.

[52] Schwartz Y, Benshimol L, Mimran D, Elovici Y, Shabtai A. LLMCloudHunter: Harnessing LLMs for automated extraction of detection rules from cloud-based CTI. 2024, arXiv preprint arXiv:2407.05194.

[53] Motlagh FN, Hajizadeh M, Majd M, Najafi P, Cheng F, Meinel C. Large language models in cybersecurity: State-of-the-art. 2024, arXiv preprint arXiv:2402.00891.

[54] Arazzi M, Arikkat DR, Nicolazzo S, Nocera A, Conti M, et al. NLP-based techniques for cyber threat intelligence. 2023, arXiv preprint arXiv:2311.08807.

[55] Park K, Hong J, Su W, Lee H. Machine learning based post event analysis for cybersecurity of cyber-physical system. 2023, arXiv preprint arXiv:2311.13488.

[56] Zacharis A, Gavrila R, Patsakis C, Ikonomou D. AI-assisted cyber security exercise content generation: Modeling a cyber conflict. In: 2023 15th international conference on cyber conflict: meeting reality. 2023, p. 217–38. http://dx.doi.org/10.23919/CyCon58705.2023.10181930.

[57] MacIntyre R. Penn treebank tokenizer (sed script source code). 1995.

[58] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46.

[59] ENISA. ENISA cybersecurity threat taxonomy. 2016, https://www.enisa.europa.eu/topics/cyber-threats/threats-and-trends/enisa-threat-landscape/threat-taxonomy.

[60] Craig RT. Generalization of Scott's index of intercoder agreement. Public Opin Q 1981;45(2):260–4. http://dx.doi.org/10.1086/268657, arXiv:https://academic.oup.com/poq/article-pdf/45/2/260/5432504/45-2-260.pdf.

[61] MITRE. MITRE ATT&CK. 2022, https://attack.mitre.org/.

[62] Jin Y, Jang E, Cui J, Chung J, Lee Y, Shin S. DarkBERT: A language model for the dark side of the internet. In: Rogers A, Boyd-Graber JL, Okazaki N, editors. Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics; 2023, p. 7515–33. http://dx.doi.org/10.18653/V1/2023.ACL-LONG.415.

[63] Penrose LS. The elementary statistics of majority voting. J R Stat Soc 1946;109(1):53–7.

[64] Sammut C, Webb GI. Encyclopedia of machine learning and data mining. Springer Publishing Company, Incorporated; 2017.

Chapter 12
Advancements in Malware Evasion: Analysis Detection and the Future Role of AI



Razvan Gavrila and Alexandros Zacharis

Abstract This chapter examines the evolution of malware evasion tactics and techniques, focusing on analysis detection. While the topic is well-researched and broad, we intend to provide an overview of malware dynamic analysis and evasion. The information we gather can be used as a starting point for understanding the current landscape but also as a resource for potential new research opportunities. The chapter also provides a historical backdrop of malware evolution, emphasising the tactics used to hide malicious activities from security tools and analysts. We highlight the arms race between malware developers and cybersecurity defenders, focusing on various anti-analysis techniques such as environment detection, reverse Turing tests, and the use of artificial intelligence. Our contribution also pinpoints the growing significance of AI in both improving and countering malware evasion strategies and what would be some potentially interesting areas for future research.

12.1 Introduction

Since the early days of Stoll's Cuckoo's Egg [49], security researchers and forensic investigators have been faced with the challenge of detecting and analysing deceptive malicious software. The cuckoo's egg metaphor has not only enriched the lexicon of cybersecurity but also inspired the naming of popular tools in the field, such as the widely used open-source dynamic analysis platform, Cuckoo Sandbox [47]. In this spirit, we provide an overview of the arms race between malware authors and those tasked to understand and analyse its behaviour. In particular, we focus on the evolution of dynamic analysis methods as well as on the increasingly sophisticated tactics and techniques developed to circumvent these efforts.

R. Gavrila · A. Zacharis (✉)
University of Piraeus, Chalandri, Greece
e-mail: Razvan.Gavrila@enisa.europa.eu; Alexandros.zacharis@enisa.europa.eu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

D. Gritzalis et al. (eds.), *Malware*, Advances in Information Security 91,
https://doi.org/10.1007/978-3-031-66245-4_12

12.1.1 1990s: The Dawn of the Sandbox

While there is no definitive early history of dynamic analysis environments used for malware study, we can assume that the first efforts were made by the early antivirus companies. These initiatives likely date back to the 1990s, with the emergence of polymorphic viruses, such as "1260" (aka "V2Px") [8], and obfuscation techniques (see the "Tequila" virus). These novel pieces of malicious code made signature detection less effective and led to the development of binary unpacking environments. Both malware employed algorithms to mutate their code or encryption, creating new, unique malware versions with each infection cycle. For example, the "1260" virus used two sliding keys to decrypt its body while injecting junk instructions into its decryptor.[1] This ability to dynamically change made it difficult to detect because it required more advanced, behaviour-based analysis techniques that could identify malicious activity patterns rather than relying on static signatures based on byte sequences. Thus, building environments that could enable analysis of such behaviours originates from a business necessity.

Operating at scale became more critical as the ownership of personal computers increased, and the adoption of Microsoft Windows and its Office ecosystem became prevalent. Thus, while analysis environments enabled antivirus companies to quickly determine potential malign behaviours by monitoring the interaction with the underlying system [51], a growing need emerged to ensure that these environments took into account market developments, e.g. the wide use of Windows. Another important opportunity at that time was the emergence of virtualisation technologies and emulators, such as VMware and QEMU, which enabled the commoditisation of building such analysis environments. The ease of use of such systems paved the way for the next era of dynamic analysis.

These advancements resulted in the establishment of various types of sandboxes, varying from extensively instrumented Windows systems used by the security industry to homebrew VMs where amateurs could study the behaviour of malicious code. These developments determine the direction of research focus in the coming decade [51], and, as we will see, the inception of sandbox evasion.

12.1.2 Analysis at Scale

With the rise of the Internet, IRC chat rooms, and email clients, the public's exposure to malware increased in the 2000s. Using exploit kits, automated SQL injections, and network worms led to increased interest in malware analysis beyond the gates of antivirus companies. Collecting malicious samples was no longer an

[1] https://userpages.umbc.edu/~dgorin1/432/example_decryptor.htm

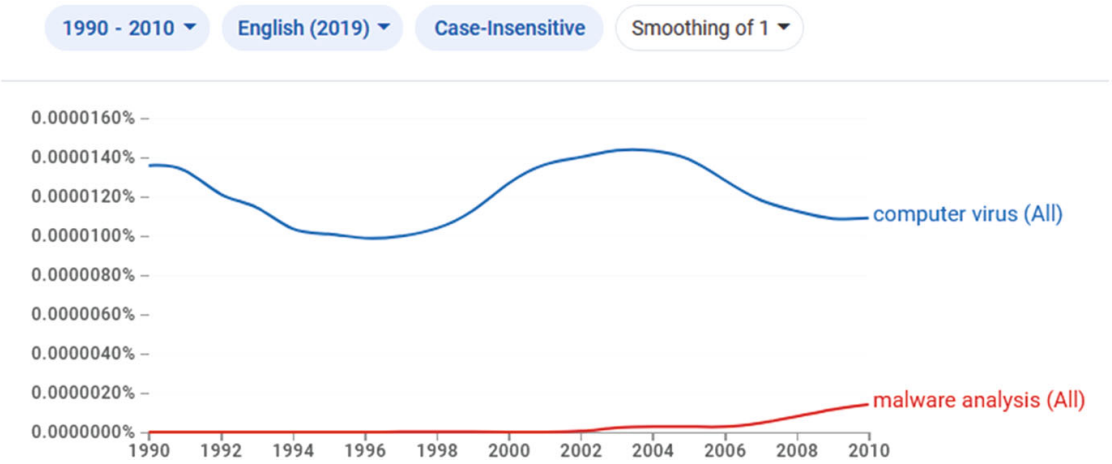


Fig. 12.1 Google n-gram for computer virus, malware analysis, 1990–2010

area of interest exclusive to specialised companies. As infection rates spiked, the malware was easy to find and collect [3]. Therefore, anyone could get their hands on a copy of ILOVEYOU, CodeRed, Nimda, or Conficker. During that decade, researchers started offering dynamic analysis services to the broader public but still assumed that human expertise was needed when assessing the outcome of the analysis report [37].

As the Google n-gram Fig. 12.1 shows, mentions of *malware analysis* began picking up traction around the same time when the term *computer virus* was at its peak. Therefore, the academic literature started looking into various practices and efforts towards automation and increased sophistication of malware analysis practices. CWSandbox [16] and Anubis [9] are the results of this quest. Nevertheless, scalability remained a big issue. Bayer et al. [10] demonstrated the value of dynamic malware analysis at scale. Moreover, due to the high number of analysis reports, there is a need to prioritise the results, and to do so, they automatically cluster based on similar behaviour. Still, the authors recognised the potential risks posed by the use of evasion.

Anubis, one of the first malware analysis platforms to be offered publicly as a service, became widely popular around the same time. Bayer et al. [11] estimated that between 2007 and 2009, the system analysed almost one million unique binaries, a very significant number for that era. Interestingly, the authors devoted a part to sandbox detection, identifying a number of real-world samples that contained explicit checks for Anubis. Nevertheless, it was not only sandboxes that malware authors targeted but also emulators and virtualised environments [45]. Indeed, notorious malware samples from the period, such as Agobot and Mydoom, included specific features to hinder analysis, such as the detection of virtual machines. Specialised packers and cryptors, such as RDG Tejon, offered options for evasion and anti-debugging, including virtualisation detection and evasion of CWSandox, Anubis, and Norman Sandbox [37].

12.1.3 Machine Learning Optimism

In the early 2010s, more research was devoted to the design, use, and limitations of behavioural analysis platforms [36] while many analysis and methodological shortcomings were identified [46]. Evasive methods pose many challenges as there is a need to tweak the analysis environments to thwart the evasion techniques [34]. Therefore, Lindorfer et al. proposed a tool that could help detect evasive malware and facilitate the discovery of previously unknown evasion techniques.

As limitations of signature-based detection became more recognised in the late 2010s and access to new machine learning tools and techniques became widely available, a new generation of solutions emerged under the generic term of "next-generation antivirus" (NGAV). In the horizontal study published of Giber et al. [27], the researchers noted an increase of 476% between 2010 and 2018 in the number of research papers on malware detection techniques that leverage machine learning. On the other side, in their longitudinal study of evasive behaviours in Windows malware, Galloro et al. [25] noted that during the decade, evasion techniques increased by 12% among malware, while Aboaoja et al. [1] report that the evasion techniques have been used by 88% of malware at the end of 2019.

These trends paved the way for new hybrid approaches, where traditional dynamic analysis methods were mixed with static ones. For instance, Raff et al. [44] proposed an ambitious architecture that moves away from analysing the behaviour of malicious samples and opts for the use of neural networks for processing the raw bytes of binaries. These advancements made the line between using sandboxes for behaviour tracking and feature extraction more blurry. The authors justify this choice due to the growing limitations of dynamic analysis and the reliance on sophisticated non-public emulation environments.

As classifiers became more prevalent, researchers became interested in adversarial examples and model poisoning [2]. For example, Anderson et al. [6] use reinforcement machine learning to evade static PE machine learning malware models through an automated framework that captures weaknesses of an anti-malware engine and produces evasive malware samples that can be used to augment a machine learning model in adversarial training.

12.1.4 The AI Revolution

As we approach the present, the commonplace cyberattacks and increased ease of access to a sophisticated malicious arsenal have made malware detection more important than ever. Dynamic analysis approaches need to consider a wide variety of methods, which must include both collection of behavioural information and feature extraction and model training [52]. From the malware authors' side, anti-analysis and evasion techniques have become a norm and are widely used by commodity malware creators, such as those behind various ransomware families [40]. We will explore some of the emerging techniques in the section dedicated to case studies.

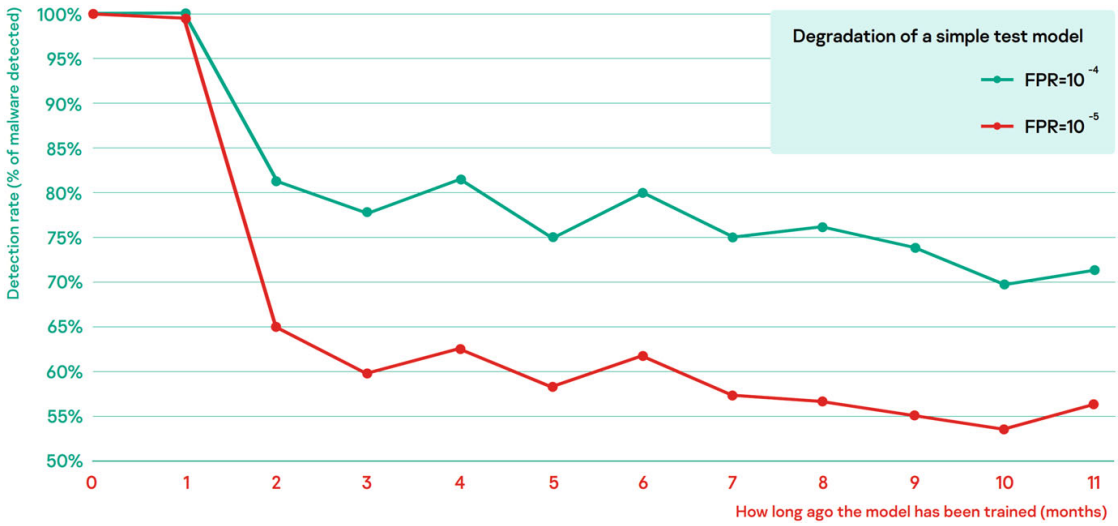


Fig. 12.2 Machine learning: test model detection rate degradation over time (Source: [17])

The early optimism around the use of machine learning for detection seems to have been toned down [33] as the detection degrades significantly over time; see Fig. 12.2. In their white paper on the use of machine learning for malware detection, Kaspersky Labs underline the challenges of using ML for detection, as active adversaries work on new ways of avoiding detection, while "*thousands of software companies produce new types of benign executables that are significantly different from previously known types*" [17]. Also, developments in Generative AI have put forward risks and opportunities for defenders.

12.2 Anatomy of Dynamic Analysis

In what follows, we focus on evading analysis and ignore more sophisticated approaches, such as those focusing on side-channel analysis (e.g. using hardware performance counters) and network simulation. The interested reader may refer to [7, 21, 26, 32, 36, 40, 41]. Note that, while nuances exist, in this section we do not perform a distinction between dynamic analysis and sandboxing.

12.2.1 Building a Sandbox

At the highest level, dynamic malware analysis focuses on analysing malicious software's behaviour in a controlled environment to assess its impact on the host system. This implies that the malicious file needs to have some impact on the underlying system. Therefore, when building a sandbox, one should consider the following aspects:

- Execution in Controlled Environments (Sandbox): dynamic analysis involves running malware in controlled, isolated, or tightly controlled systems known as sandboxes, which prevents the artefact analysed from inflicting actual harm to information systems while allowing analysts to monitor its actions. These environments may come in different forms, from physical appliances to virtualised environments or public cloud services.
- Use of Virtualisation or Emulation: virtualised environments enable the easy setup of controlled environments and faster recovery between malware execution sessions, facilitating the analysis of more samples in less time. More importantly, virtualisation enables the use of environmental variety, as different aspects of the sandbox can be easily reconfigured via orchestration tools. Such changes can include various system configurations, keyboard layouts, software packages installed, etc. VMWare is widely mentioned in multiple research papers, but VirtualBox is also popular among security researchers, malware analysts, and enthusiasts. The main point is that the analysis component needs to run at a higher privilege level than the artefact under analysis. Nevertheless, there is also substantial work on the use of bare-metal analysis environments [5, 28, 30, 31, 38] to defeat, e.g. anti-instrumentation-equipped malware.
- Monitoring System Changes: traditional dynamic analysis techniques included monitoring of various system changes, such as changes to the filesystem (creation, deletion, modification), registry modifications (adding, modifying, or deleting keys/values), network traffic (outgoing and incoming connections), process behaviour (process creation, termination), or system API calls. This category includes function call monitoring, function parameter analysis, and information flow tracking. Performing such monitoring requires system-level hooks or virtualised environments to intercept and log these activities. Nevertheless, various stand-alone tools can be used for performing these actions, such as Procmon, Regshot, and APIMonitor, and their functionality can be integrated into sandbox agents.
- Automation and Orchestration: automation and orchestration enabled more sophisticated use cases, such as the detailed tracking of system calls, logging of network traffic, screen capture, and programmable interaction with the artefact. These developments paved the way towards the commoditisation of dynamic analysis, in the form of public online malware sandboxes, such as the defunct Anubis, VirusTotal, Triage, and any.run. Open-source projects, such as Cuckoo Sandbox, can be used to automate the process of executing malware in an isolated environment and analysing its behaviour, while Dynamic Binary Instrumentation frameworks, such as DynamoRIO, allow detailed execution tracing.
- Extraction of Behavioural Signatures: the emergence of behavioural signatures came as a response to the increased use of polymorphism and obfuscation by malware authors. While many of these techniques share similarities to traditional approaches, the signatures are generally derived from tracking the sequence and context of the action performed by the artefact in the content of the analysis. This enables the detection of intent, regardless of various evasion tactics used by malware authors. For example, such signatures may include unusual patterns

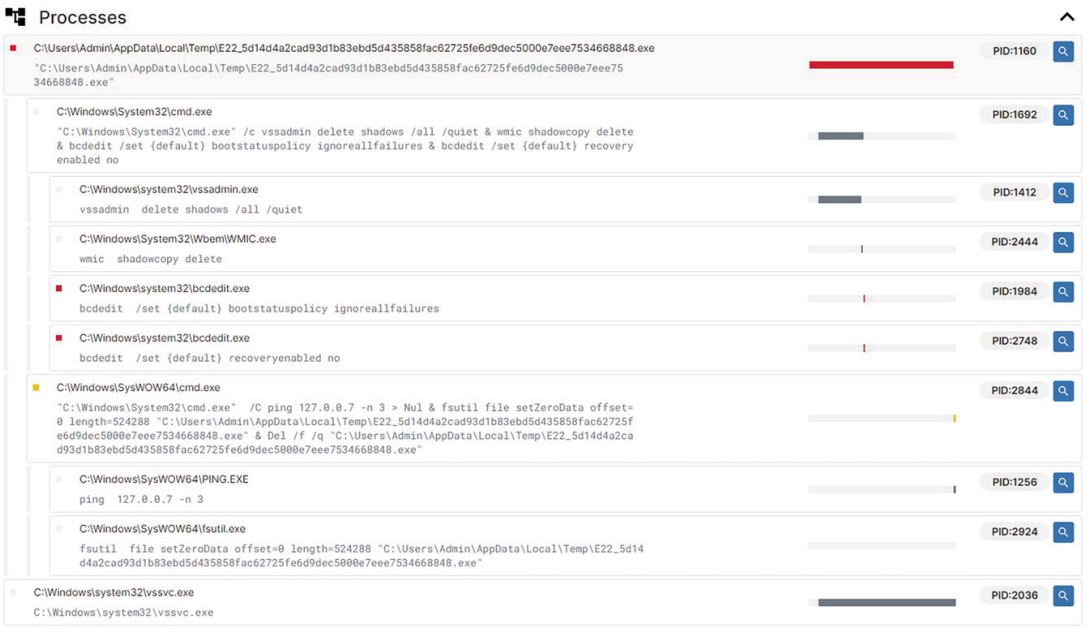


Fig. 12.3 A LockBit sample trying deleting shadow copies using `vssadmin` (Source: https://tria.ge/220202-vc31labdgm/behavioral1)

of writing or reading files, changes to the registry that facilitate persistence, accessing cryptographic libraries while listing directory contents, attempts to execute processes using high privileges from unusual locations, such as the TEMP directory, etc. These behavioural signatures can be used to detect malware families as a whole. For instance, many ransomware use `vssadmin` to delete the shadow copies and prevent recovery. Therefore, the attempt to use it can be a very accurate detection method to distinguish ransomware from other malware; see Fig. 12.3.

While integrating the aforementioned features into a custom solution may seem relatively straightforward, the sophistication of evasion techniques makes such an endeavour non-trivial. Thus, the use of behavioural analysis should be done in conjunction with static analysis techniques. This enables a more complete view of an artefact's behaviour, including its raw structure, binary features, and other characteristics, to identify potential hidden functionalities that might not be triggered during a single execution. Such hybrid approaches, along with the use of machine learning, enable the creation of more advanced signatures and the detection of new malicious patterns. Thus, an effective sandbox may require substantial investment and access to industry-specific resources and know-how that may no longer be available to security researcher enthusiasts.

Table 12.1 The benefits and drawbacks of closed sandboxes and public malware-analysis platforms

Aspect	Closed sandbox	Public platform
Privacy	High, as data remains internal	Lower, due to potential data exposure
Resource control	Complete control over resources	Dependent on external resources
Customisation	High, can tailor environment	Limited, must use as provided
Cost	Higher due to infrastructure	Lower, as infrastructure is shared
Scalability	Limited by own resources	High, leverages cloud resources
Collaboration	Limited to internal team	Easier, through shared findings
Evasion detection	Can customise to detect evasion	Malware may recognise and evade

12.2.2 On-prem vs the Cloud

Due to the increased cost of entry, there are some emerging trends in sandbox dynamic analysis beyond the integration of artificial intelligence and machine learning. The importance of scale, including the availability of ample computing resources and quick access to large malware sample sets, transformed the dynamic analysis landscape. As a result, there is a growing emphasis on cloud-based analysis platforms for scalability and accessibility. As previously underlined, maintaining an in-house sandbox comes with many challenges and may not be suitable for most organisations. Vendors are also migrating to Cloud solutions, abandoning the on-prem appliance product lines. Still, there might be cases when having such a setup may be needed. A telling example is a recent VirusTotal data leak that impacted 5,600 users, including the employees of the US Department of Justice, the US NSA, and the US FBI, alongside German intelligence agencies [19]. In Table 12.1, we provide an overview of the pros and cons of using an in-house sandbox vs a public cloud-hosted one.

12.2.3 Agent-Based vs Agentless Sandboxes

A significant development was the growing shift from agent-based dynamic analysis to agentless systems, such as those used by VMRay.[2] Ali et al. [4] performed an empirical comparative analysis between agent-based and agentless analysis platforms. The aim of the study was to measure the effectiveness of these two approaches when it came to detecting sophisticated malicious software; see Table 12.2. The main drawback of agent-based sandboxes is that they lack evasion resistance due to the deployment of artefacts in the same environment where the agent resides. Observing the behaviour at lower levels (out-of-guest) enables substantial advancements, such as using memory introspection techniques [15].

[2] https://www.vmray.com/vmray-analyzer/

Table 12.2 Capacities of agent-based vs agentless sandboxes

Features	Cuckoo sandbox (agent-based)	VMRay (agentless)
Zero data detection		✓
Visualisation of the parent-child relationship		✓
Evasion of anti-analysis technique possible or not?		✓
Files-related activities	✓	✓
Registry-related activities	✓	✓
API calls	✓	✓
HTTP requests	✓	✓
DNS requests	✓	✓
Process tree	✓	✓
IRC commands	✓	✓
Service activities	✓	✓
IDS signatures		✓

12.3 Anatomy of Anti-analysis

As automation in malware analysis becomes more sophisticated to handle the increasing volume of malware samples, the need for evasion tactics by malware authors becomes more pronounced. This is a constant dynamic that we observe, where advancements in detection prompt new strategies for evasion. Evasion or anti-analysis techniques became commonly found in more prolific types of malware, such as spyware and bots, in the 2010s. Still, recently they are embedded in all variants of malicious software [40], including banking trojans and ransomware spanning across various operating systems [18]. By using anti-analysis techniques, malware can evade analysis environments and alter their behaviour to escape or hinder detection. Not surprisingly, this capability can lead analysts to false conclusions and disorient the full incident-handling process.

Nevertheless, we need to understand what an evasion is. Evasion or anti-analysis is the use of any technique that employs the use of techniques that will evade detection and/or identification. Defence evasion is a broader category for which the MITRE ATT&CK has a dedicated tactic.[3] MITRE distinguishes between Masquerading (T1036) and Virtualization/Sandbox Evasion (T1497), the first being employed by attackers to "*manipulate features of their artefacts to make them appear legitimate or benign to users and/or security tools*". At the same time, the latter is explicitly linked to malware's author intention to avoid detection and analysis environments.

In the literature, there is a clear distinction between *evading analysis* and *avoiding detection*. The latter considers aspects such as polymorphism, network

[3] https://attack.mitre.org/tactics/TA0005/

evasion, and fileless malware, which we deem out of scope for this work. Also, while privilege escalation is a possible threat (the move from an inferior trust ring to a superior one, such as from the guest VM to the hypervisor), we will focus on the detection of the analysis framework, as it is "*the most effective ways for malware to evade analysis is by hiding when it detects that it is under analysis*" [41].

12.3.1 Evasion Classes

Malware authors know that security professionals often rely on identifiable signatures and behavioural patterns to analyse their code. Naturally, the first step in the anti-analysis arsenal is carefully identifying and understanding these telltale signs. Once these analysis fingerprints are recognised, the next phase involves implementing detection mechanisms to identify these fingerprints. Malware authors' aim is not only to evade the immediate threat of detection but also to abstract the detection mechanism itself, which generally comes in the form of detection signatures. The goal is to camouflage the malware's malicious intent while preserving its ability to execute undetected in victim environments.

For this section, we focus on some of the main evasion techniques used by malware authors [14].

12.3.1.1 Environment Sensing

One of the most common techniques used to detect an analysis framework is through fingerprinting of one or more key components of the analysis infrastructure itself. The use of virtualisation detection is one such example of fingerprinting. As previously discussed, platforms like VMware, VirtualBox, and Hyper-V are used in both open-source and paid malware analysis solutions. This common knowledge has led malware authors to attempt fingerprinting of these platforms to hinder malware analysis.

Various open-source tools can be used to stress test analysis platforms. For example, pafish[4] can be used "*to detect virtual machines and malware analysis environments in the same way that malware families do*", al-khaser,[5] cover a wide range of tricks, such as timing attacks, anti-VM and human interaction. In contrast, DebugOff[6] is a Rust library to make both static and dynamic analyses more difficult in Linux environments.

[4] https://github.com/a0rtega/pafish

[5] https://github.com/LordNoteworthy/al-khaser

[6] https://github.com/0xor0ne/debugoff

Due to broad research interest, many of these techniques are very well documented. In what follows, we explore some examples; however, a more comprehensive list of techniques is maintained by CheckPoint's Research.[7]

CPUID Fingerprint

CPUID is an instruction present in almost all modern CPU architectures. It is intended to allow the software to discover various details about the processor of the machine it is being executed on. Using the code in Listing 12.1, one can check the vendor ID string, e.g. `Microsoft Hv` for Hyper-V, `KVMKVMKVM` for KVM, and `VBoxVBoxVBox` for VirtualBox.

```
__declspec(naked) void get_cpuid_vendor(char *vendor_id) {
  __asm {
    ; save non-volatile register
    push ebx

    ; nullify output registers
    xor ebx, ebx
    xor ecx, ecx
    xor edx, edx

    ; call cpuid with argument in EAX
    mov eax, 0x40000000
    cpuid

    ; store vendor_id ptr to destination
    mov edi, vendor_id

    ; move string parts to destination
    mov eax, ebx  ; part 1 of 3 from EBX
    stosd
    mov eax, ecx  ; part 2 of 3 from ECX
    stosd
    mov eax, edx  ; part 3 of 3 from EDX
    stosd

    ; restore saved non-volatile register
    pop ebx

    ; return from function
    retn
  }
}
```

Listing 12.1 Check vendor ID string via CPUID instruction

[7] https://evasions.checkpoint.com/

MAC Address Fingerprint

A MAC address is a unique identifier assigned to Network Interface Controllers on a machine. By checking MAC address prefixes, the first 24 bits, it is possible to identify the manufacturer of the network adapter. This information can be easily extracted with the use of `ipconfig /all` or `wmic nic list` command. For instance, network adapters from Broadcom, a common supplier of networking hardware to many PC manufacturers, may start with a prefix such as `E0:3E:44`, `00:10:18`, `D4:01:29`, and `00:0D:E6`. Similarly, VMware network adapters will start with one of the following prefixes: `00:1C:14`, `00:50:56`, `00:05:69`, and `00:0C:29`, VirtualBox network adapters will start with `08:00:27`, and Hyper-V with starts with`00:03:FF`. Therefore, using public information,[8] one can easily distinguish MAC addresses of real network appliances from virtualised.

Registry Checks

Virtualised Windows environments will often contain various registry entries not commonly found on physical machines, the presence of which can be a tip-off to the fact the malware is running on a VM. For instance, VMware uses the following:

- `\HKEY_CURRENT_USER\Software\VMware, Inc.\*`
- `\HKEY_LOCAL_MACHINE\SOFTWARE\VMware, Inc.\*`
- `\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Enum\SCSI \*VMware*\*`
- `\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services \*VMware*\*`

Similarly, for VirtualBox, we have the following entries:

- `\HKEY_LOCAL_MACHINE\HARDWARE\ACPI\DSDT\VBOX__`
- `\HKEY_LOCAL_MACHINE\HARDWARE\ACPI\FADT\VBOX__`
- `\HKEY_LOCAL_MACHINE\HARDWARE\ACPI\RSDT\VBOX__`
- `\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\VBox*`
- `\HKEY_LOCAL_MACHINE\SOFTWARE\Oracle\VirtualBox Guest Additions\*`

Process Checks

To facilitate their operation, many virtualised Windows environments will run specific processes not usually found on physical machines. This information can be easily extracted using `wmic process` or `tasklist` commands. For instance, VMware

[8] https://gist.github.com/aallan/b4bb86db86079509e6159810ae9bd3e4

would use the following processes `VGAuthService.exe`, `vmacthlp.exe`, and `vmtoolsd.exe`, while VirtualBox may use `VBoxService.exe` and `VBoxTray.exe`.

Service Checks

Similarly to the above, virtualised Windows environments will often have certain services that are not present on physical machines. This information can be easily extracted using `sc query` or `wmic service list` commands. For instance, VMware may have the services `VMTools`, `vmvss`, `VGAuthService`, and `VMware Physical Disk Helper Service` running, while VirtualBox may have the `VBoxService` service running.

File Checks

In addition to the above, virtualised Windows environments will also often contain certain files (usually drivers), which can indicate a VM. For instance, for VMware one may find the following files:

- `C:\Windows\System32\drivers\vmhgfs.sys`
- `C:\Windows\System32\drivers\vmmemctl.sys`
- `C:\Windows\System32\drivers\vmmouse.sys`
- `C:\Windows\System32\drivers\vmrawdsk.sys`

Similarly, for VirtualBox, one can find:

-
- `C:\Windows\System32\drivers\VBoxMouse.sys`
- `C:\Windows\System32\drivers\VBoxGuest.sys`
- `C:\Windows\System32\drivers\VBoxSF.sys`
- `C:\Windows\System32\drivers\VBoxVideo.sys`

Timing and Event-Based Evasion

One of the simplest and longest-lasting methods of analysis and sandbox avoidance is based on the simple premise that any sandbox will only allocate a finite amount of time (usually only several minutes) for each given analysis before moving to the next sample in its queue. By delaying its execution beyond this timeframe, malware can hide its malicious actions and activities from the sandbox. This is commonly accomplished by using either the `Sleep` or `NtDelayExecution` Windows APIs, which some sandboxes patch to circumvent this behaviour.

As a result, some malware has evolved to detect these patches as an additional indicator of an analysis/sandbox environment. This can be done by taking a timestamp, going to sleep, and checking the timestamp upon waking up. If the time

difference from the previously taken timestamp substantially differs from when the malware was programmed to sleep, the malware will avoid or adjust its execution.

Anti-debugging and Anti-emulation

Over time, malware developers have gained insight into the tools commonly used by malware researchers and analysts and have modified their malware accordingly to avoid or adjust its execution if such tools are detected as present or active on the executing machine. These tools include debuggers, disassemblers, and networking monitoring utilities/sniffers. Their existence in an environment signals a red flag for executing a malicious sample. Malware has been spotted actively looking for running processes linked to the malware analyst's arsenal. Typical such processes are the following:

- Wireshark.exe
- Fiddler.exe
- Procmon.exe / Procmon64.exe
- Procexp.exe / Procexp64.exe
- Sysmon.exe / Sysmon64.exe
- ProcessHacker.exe
- OllyDbg.exe
- ImmunityDebugger.exe
- Windbgx86.exe / Windbgx64.exe
- x32dbg.exe / x64dbg.exe
- Python.exe

12.3.2 Reverse Turing Tests and Cyber Deception

A particular category of evasion techniques includes reverse Turing tests (RTTs). An RTT is an inversion of the traditional Turing test and implies that the computer attempts to determine whether it is interacting with a human. As many malware authors aim to infect end-user devices, they can check for traces of human activity, which can also be employed as a tactic against sandboxes.

Bulazel et al. [14] distinguish between three categories of RTTs:

- passive observation of live human computer use (e.g. mouse movement, typing cadence, process creation),
- active interaction with a human user (e.g. making the user click a dialogue box created by the malware before it runs),
- passive wear-and-tear artefacts, such as limited changes to default OS file systems.

Such checks can include checking for mouse cursor position changes over time. For instance, if the mouse cursor position is found to be permanent, the machine is much more likely to be a sandbox. Similarly, the recent file count/and Desktop file count can be used to determine whether a machine is a sandbox. For instance, if the executing machine is found to have a low count of recently opened files or files on the desktop, it is much more likely to be a sandbox. Similarly, suppose the count of currently running applications or currently active windows or currently running processes is found to be unusually low. In that case, the machine is much more likely to be a sandbox, as a regular user would usually have at least one or two applications running at any given time.

Other approaches may involve probing the hardware, e.g. screen resolution. In this case, if the executing machine's display resolution is unusually low (e.g. 800 × 600), something a contemporary user would not most likely use, the machine is more likely to be a sandbox. Likewise, suppose the CPU core count and/or amount of available system RAM is found to be low (e.g. a single CPU core and 1GB of RAM). In that case, the machine is much more likely to be a sandbox, as a contemporary user is unlikely to tolerate such a machine [32].

Interestingly, these checks also offer opportunities for defenders, in particular in the context of cyber deception strategies. While traditional efforts are to evade the evasion, the use of fingerprinting for cyber deception is to enable the evasion. More precisely, one can use the fact that many malware would seize if it noticed that it is running in a virtualised environment. Note that in a 2019 report, Symantec[9] estimated that up to 20% of malicious code attempts to detect virtual environments, which leads to a change of their behaviour. Therefore, leaving specific artefacts that would trigger sandbox fingerprinting can be used to avoid infection.

There are such projects, e.g. Fake Sandbox Processes[10] and its derivative, Fake Sandbox Artifacts.[11] The latter is feature-rich, covering the creation of dummy files, folders and registries, spawning dummy processes with names related to analysis tools, and many more. The authors showcase the performance of the script by comparing it against sandbox stress testers, such as Pafish and al-khaser. When tested against three real-world samples, the tactic prevented the deployment of AgentTesla and ZLoader, while it had no impact on the execution of TrickBot.

Although the concept is straightforward and elegant, it is not evident how impactful it is in practice, as studies on the effectiveness of cyber deception are still in their infancy [24]. Moreover, such a tactic should be seen as a last resort since it implies that the malicious code has reached the point where it could be successfully executed on the victim's machine.

[9] https://docs.broadcom.com/doc/istr-24-2019-en

[10] https://github.com/NuclearPhoenixx/fake-sandbox

[11] https://github.com/NavyTitanium/Fake-Sandbox-Artifacts

12.4 Evasion in the Mobile World

The study of dynamic analysis tends to focus on conventional computing systems, particularly those that use the x86 architecture and Microsoft Windows operating system. While this particular operating system may still hold an important market share, the ubiquity of mobile devices has led to significant shifts in usage over the last decade. Indeed, currently, there are more than three billion Android devices,[12] and as a result, this has attracted a lot of interest from cybercriminals. Moreover, Android's relatively open design and the diversity of third-party app stores make it a perfect target for malicious app developers. In what follows, we will solely focus on the dynamic analysis and sandbox evasion for Android apps.

Similar to the early developments in the personal computer space, the early initiatives to establish dynamic analysis environments came from the industry,[13] but there were also some early academic research initiatives [12].

In 2012, Google introduced Bouncer,[14] a service that enabled automated scanning of the Android Market for potential malicious software. The announcement claimed that since its introduction, Bouncer led to "*a 40% decrease in the number of potentially-malicious downloads from Android Market*". However, in the same year, Oberheide and Miller [39] presented various techniques for fingerprinting Bouncer and its environment, such as its underlying system, the Android Framework used and specific behaviours, such as timing attacks and input automation. Some of these points showcase that the previous experience in detecting malware evasion in the PC world was not translated to the mobile space.

Then, Android malware analysis started to emerge with DroidBox,[15] DordiScope [50], and TaintDroid [22], and a multitude of online platforms, such as Andrubis [35]. Petsas et al. [43] underline the challenges of dynamic analysis for Android, as mobile devices are equipped with a variety of sensors, including an accelerometer, gyroscope, GPS, and gravity sensors, which are hard to simulate [13]. Such challenges set the world of mobile dynamic analysis apart from those of classical computing.

Also, due to its architecture, Android allocates to each application its own disk space, which enables these applications to download updates and run new code in what is called update attacks [20]. This implies that malware authors may distribute malicious behaviour in future updates of what may seem benign apps. This technique renders dynamic analysis even more challenging as the app at the time of analysis does not actually perform any malicious action.

Moreover, unlike traditional sandboxes, relying on emulators offers substantial benefits to attackers, as they can investigate their payload's interaction with default

[12] https://blog.google/products/android/io22-multideviceworld/

[13] https://www.trendmicro.com/en_us/research/12/g/a-look-at-google-bouncer.html

[14] https://googlemobile.blogspot.com/2012/02/android-and-security.html

[15] https://github.com/pjlantz/droidbox

Android emulators, such as QEMU and Genymotion [23]. Some of the strategies employed by the attackers, in addition to the traditional fingerprinting techniques, include programme interaction detection and detection based on context switching.

12.5 Case Studies of Cunning Code

The following section presents a series of recent cases in which malware authors have used old and new techniques to evade analysis. The three examples we chose aim to highlight and put into perspective some of the concepts presented in the previous sections. For each case study, we provide a short overview of the technique used and what we think sets it apart.

12.5.1 Vectored Exception Handling Enabled Runtime Padding in GULoader

GuLoader, also known as CloudEyE, is a Visual Basic Script (VBS) downloader used to distribute remote access trojans on infected machines. It was first detected in the wild in 2019 but evolved over time to utilise novel evasive tricks. The shellcode employed several anti-analysis tricks at every step of execution, including anti-debugging and anti-disassembling checks to detect the presence of remote debuggers and breakpoints, along with scans for virtualisation software.

The novelty of the malware is that its latest version deployed a less-known anti-analysis technique dubbed Vector Exception Handling (VEH). Implementing the exception handling by the operating system provides an opportunity for the adversary to take over the execution flow. The Vectored Exception Handling on Windows provides the user with the ability to register a custom exception handler, effectively allowing for the execution of additional code in the event of an exception. The interesting thing about handling exceptions is the way in which the system resumes its normal execution flow of the programme after the event of exception.

Adversaries exploit this mechanism and take ownership of the execution flow. When an exception occurs, malware can divert the flow to the code under its control. Normally, malware employs this mechanism for hooking, covert code execution, and anti-analysis.

The GuLoader employs the VEH mainly to obfuscate the execution flow and slow down the analysis. In-depth analysis of recent samples[16] exposed an interesting feature of this technique seen in GULoader shellcode, runtime padding. Runtime padding is an evasive behaviour that beats automated scanners and other security

[16] https://www.mcafee.com/blogs/other-blogs/mcafee-labs/guloader-campaigns-a-deep-dive-analysis-of-a-highly-evasive-shellcode-based-loader/

checks employed at runtime. It delays the malicious activities performed by the malware on the target system. To do so, it follows this flow:

- An egg value is set.
- A function initiates a search for this value in its own data segment of the shellcode.
- This is implemented via the VEH handler, so the search itself adds 0.3 million VEH iterations on top of regular VEH control manipulation employed in the code.
- The loader ends this search when it retrieves the address location of the egg value. To ensure the value is not being manipulated by the researcher, it performs two additional checks to validate the egg location.
- If the check fails, the search continues.

12.5.2 LummaC2 Reverse Turing Test

LummaC2 is an information stealer written in C language and sold in underground forums since December 2022.[17] The sample uses different obfuscation techniques (such as Windows API hashing and encoded strings) but does not include any anti-sandbox techniques. However, in a newer report, it was identified that it uses a novel Anti-Sandbox technique.[18]

More precisely, the new sample performs an interesting reverse Turing test that takes into account different positions of the mouse cursor in short intervals of time. The researchers point out that malware code uses trigonometry to detect human-like behaviour. It starts by collecting mouse coordinates at short intervals, which will be used as Euclidean vector points. Then, it will calculate the angles formed between consecutive vectors and check them against a hardcode value (45°). The malware authors seem to consider that any value below this threshold does not constitute human-like behaviour.

While interesting, this technique appears rather over-engineered. While a naive sandbox may not support mouse movements or use simplistic, linear paths, industry-grade sandboxes would not be tricked by such evasion techniques. Nevertheless, it shows the level of sophistication of evasion techniques, as a specific model using vector geometry was developed.

12.5.3 Unsupported Decompression Breaks Sandboxes

Attackers will not only employ well-known techniques to disable the analysis arsenal, but they will also study subtle changes in the behaviour of the target

[17] https://outpost24.com/blog/everything-you-need-to-know-lummac2-stealer/

[18] https://outpost24.com/blog/lummac2-anti-sandbox-technique-trigonometry-human-detection/

operating system and identify the correct technique that can be exploited in the context of this change.

For instance, recently, Zimperium revealed 3,000 Android malware samples using multiple techniques to bypass detection.[19] The original discovery was made by Joe Sandbox, who detected an Android APK file that could be installed on mobile devices but could not be analysed by most anti-decompilation tools. They discovered that the attackers were abusing an old method regarding the handling of ZIP files [42]. By tampering with the compression algorithm used for packaging APKs, they could block automatic script analysis. The attackers abuse Android's default behaviour for unknown or unsupported compression methods. More precisely, as of Android version 9, applications using unsupported or unknown compression methods work properly, a behaviour which was not present in previous versions of the operating system.

The researchers identified additional tampering of the APK files, such as filenames longer than 256 bytes, malformed `AndroidManifest.xml` files, and String Pool. While many of these impact the static analysis efforts, attacking the unpacking sandbox becomes more important, as hybrid approaches are now the norm.

12.6 Conclusions and Future Directions

As already discussed, the fight between malware authors and defenders continuously evolves, with both sides using ingenious methods to bypass detections or identify malicious behaviour. In dynamic analysis, this translates to a battle where malware authors strive to find artefacts that distinguish analysis environments while the other side tries to hide them. To this end, both sides resort to using low-level processor and operating system features, as well as the identification of human interaction or other hardware features and execution traces.

The above leads us to speculate on several possible future directions related to dynamic analysis from both the attackers' and defenders' perspectives.

12.6.1 Emergence of GPT Malware

The emergence of artificial intelligence, in particular Generative AI and large language models, introduces a new dimension of the race for superiority between malware creators and cybersecurity professionals. In 2023, we already witnessed the emergence of proof of concepts, such as BlackMamba [48] and LLMorpher,[20]

[19] https://www.zimperium.com/blog/over-3000-android-malware-samples-using-multiple-techniques-to-bypass-detection/

[20] https://github.com/SPTHvx/SPTH/blob/master/articles/files/LLMorpher.txt

the first malware observed which uses GPT to rewrite its code. As a result, malware writers will leverage AI to enhance their ability to abstract analysis fingerprints more efficiently. Through machine learning algorithms, they can systematically identify and analyse patterns within the detection mechanisms, allowing for the rapid extraction of key features that signify scrutiny.

Moreover, AI empowers malware writers to generate a myriad of samples by systematically combining various evasion techniques. Through generative models, they can create diverse iterations of malicious payloads, each possessing a unique blend of evasion tactics. This dynamic approach significantly amplifies the adaptability and resilience of malware, making it increasingly challenging for traditional analysis platforms to keep pace with the evolving threat landscape.

12.6.2 AI-Enabled Sandboxes

Conversely, cybersecurity analysts leverage AI to fortify their defence mechanisms against these sophisticated anti-analysis techniques. By employing machine learning algorithms, analysts can create less fingerprintable systems, thereby mitigating the effectiveness of the abstraction phase in the malware development cycle. We predict that AI-driven dynamic analysis platforms will deploy agents that emulate natural human interaction, effectively defeating human detection fingerprints employed by malware creators. These AI-driven agents will simulate user behaviour, making it harder for malicious entities to distinguish between genuine user interactions and automated analysis processes.

12.6.3 EDR Goes Mobile

As the industry has moved away from signature-based detection, a new class of security solutions emerged: endpoint detection and response (EDR) solutions. At their core, EDRs behave similarly to dynamic analysis sandboxes: they monitor systems and their behaviour passively, and extract features that match known or potentially dangerous behaviour. The key difference is that the live device, where the EDR agent runs, collects the needed telemetry. The device is the sandbox.

This is an essential breakaway from previous approaches, where the sandbox was deemed a tightly controlled environment. This new approach follows an assumed breach mentality, and as shown, there is still a lot of room for improvement [29]. Still, EDRs are a more robust alternative to traditional endpoint protection tools.

Considering the recent increase in the number of zero-day vulnerabilities for mobile operating systems and the growing number of mobile device exploitation cases, an EDR solution for iOS and Android seems like a reasonable value proposition. Currently, there are very few players in the mobile EDR space, so it

is likely that we will see an increase in this offering in the near future. This will likely lead to an increased interest in mobile EDR evasion.

Finally, as we have noted in the agent-based vs agentless subsection, there are some inherent challenges when running an agent next to a malicious process that may escalate their privileges to the system level. It is possible that some mobile manufacturers may choose to embed the behavioural monitoring features in the device hardware. While such an approach would likely bring more assurance to the monitoring process, it may induce substantial privacy risks as well.

References

1. Aboaoja, F.A., Zainal, A., Ali, A.M., Ghaleb, F.A., Alsolami, F.J., Rassam, M.A.: Dynamic extraction of initial behavior for evasive malware detection. Mathematics **11**(2), 416 (2023)
2. Al-Dujaili, A., Huang, A., Hemberg, E., O'Reilly, U.M.: Adversarial deep learning for robust detection of binary encoded malware. In: 2018 IEEE Security and Privacy Workshops (SPW), pp. 76–82. IEEE (2018)
3. Alenezi, M.N., Alabdulrazzaq, H., Alshaher, A.A., Alkharang, M.M.: Evolution of malware threats and techniques: A review. Int. J. Commun. Netw. Inf. Secur. **12**(3), 326–337 (2020)
4. Ali, M., Shiaeles, S., Papadaki, M., Ghita, B.V.: Agent-based vs agent-less sandbox for dynamic behavioral analysis. In: 2018 Global Information Infrastructure and Networking Symposium (GIIS), pp. 1–5. IEEE (2018)
5. Alptekin, H., Yildizli, C., Savas, E., Levi, A.: Trapdroid: Bare-metal android malware behavior analysis framework. In: 2019 21st International Conference on Advanced Communication Technology (ICACT), pp. 664–671. IEEE (2019)
6. Anderson, H.S., Kharkar, A., Filar, B., Evans, D., Roth, P.: Learning to evade static pe machine learning malware models via reinforcement learning. arXiv preprint arXiv:1801.08917 (2018)
7. Apostolopoulos, T., Katos, V., Choo, K.K.R., Patsakis, C.: Resurrecting anti-virtualization and anti-debugging: Unhooking your hooks. Future Gener. Comput. Syst. **116**, 393–405 (2021)
8. Bassham, L.E., Polk, W.T.: Threat assessment of malicious code and external attacks (1992) https://csrc.nist.gov/pubs/ir/4939/final
9. Bayer, U., Krügel, C., Kirda, E.: Ttanalyze: A tool for analyzing malware. In: Proceedings of the 15th Annual Conference of the European Institute for Computer Antivirus Research (EICAR), pp. 180–192 (2006)
10. Bayer, U., Comparetti, P.M., Hlauschek, C., Kruegel, C., Kirda, E.: Scalable, behavior-based malware clustering. In: NDSS, vol. 9, pp. 8–11 (2009)
11. Bayer, U., Habibi, I., Balzarotti, D., Kirda, E., Kruegel, C.: A view on current malware behaviors. In: Proceedings of the 2nd USENIX Conference on Large-Scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More, pp. 8–8 (2009)
12. Bläsing, T., Batyuk, L., Schmidt, A.D., Camtepe, S.A., Albayrak, S.: An android application sandbox system for suspicious software detection. In: 2010 5th International Conference on Malicious and Unwanted Software, pp. 55–62 (2010). https://doi.org/10.1109/MALWARE.2010.5665792
13. Boomgaarden, J., Corney, J., Whittaker, H., Dinolt, G., McEachen, J.: Challenges in emulating sensor and resource-based state changes for android malware detection. In: 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1–10 (2015). https://doi.org/10.1109/ICSPCS.2015.7391733
14. Bulazel, A., Yener, B.: A survey on automated dynamic malware analysis evasion and counter-evasion: Pc, mobile, and web. In: Proceedings of the 1st Reversing and Offensive-Oriented Trends Symposium, ROOTS. Association for Computing Machinery, New York, NY (2017). https://doi.org/10.1145/3150376.3150378

15. Caragea, R.: Telescope-real-time peering into the depths of tls traffic from the hypervisor. Bitdefender Labs (2016)
16. CWSandbox: CWSandbox. https://cwsandbox.org/
17. Cybersecurity, K.E.: Machine learning for malware detection. https://media.kaspersky.com/en/enterprise-security/Kaspersky-Lab-Whitepaper-Machine-Learning.pdf (2021)
18. Damri, G., Vidyarthi, D.: Automatic dynamic malware analysis techniques for linux environment. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 825–830. IEEE (2016)
19. Dark Reading: VirusTotal Data Leak Affects 5K+ Users. https://www.darkreading.com/cyberattacks-data-breaches/virustotal-data-leak-affects-5k-users (2023). Accessed 31 Mar 2024
20. De Lorenzo, A., Martinelli, F., Medvet, E., Mercaldo, F., Santone, A.: Visualizing the outcome of dynamic analysis of android malware with vizmal. J. Inf. Secur. Appl. **50**, 102423 (2020)
21. Egele, M., Scholte, T., Kirda, E., Kruegel, C.: A survey on automated dynamic malware-analysis techniques and tools. ACM Comput. Surv. **44**(2), 1–42 (2008)
22. Enck, W., Gilbert, P., Han, S., Tendulkar, V., Chun, B.G., Cox, L.P., Jung, J., McDaniel, P., Sheth, A.N.: Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. ACM Trans. Comput. Syst. **32**(2), 1–29 (2014)
23. Faruki, P., Bhan, R., Jain, V., Bhatia, S., El Madhoun, N., Pamula, R.: A survey and evaluation of android-based malware evasion techniques and detection frameworks. Information **14**(7), 374 (2023)
24. Ferguson-Walter, K.J., Major, M.M., Johnson, C.K., Johnson, C.J., Scott, D.D., Gutzwiller, R.S., Shade, T.: Cyber expert feedback: experiences, expectations, and opinions about cyber deception. Comput. Secur. **130**, 103268 (2023)
25. Galloro, N., Polino, M., Carminati, M., Continella, A., Zanero, S.: A systematical and longitudinal study of evasive behaviors in windows malware. Comput. Secur. **113**, 102550 (2022)
26. Geng, J., Wang, J., Fang, Z., Zhou, Y., Wu, D., Ge, W.: A survey of strategy-driven evasion methods for pe malware: transformation, concealment, and attack. Comput. Secur. **137**, 103595 (2024)
27. Gibert, D., Mateu, C., Planes, J.: The rise of machine learning for detection and classification of malware: research developments, trends and challenges. J. Netw. Comput. Appl. **153**, 102526 (2020)
28. Guan, L., Jia, S., Chen, B., Zhang, F., Luo, B., Lin, J., Liu, P., Xing, X., Xia, L.: Supporting transparent snapshot for bare-metal malware analysis on mobile devices. In: Proceedings of the 33rd Annual Computer Security Applications Conference, pp. 339–349 (2017)
29. Karantzas, G., Patsakis, C.: An empirical assessment of endpoint detection and response systems against advanced persistent threats attack vectors. J. Cybersecur. Privacy **1**(3), 387–421 (2021)
30. Kirat, D., Vigna, G., Kruegel, C.: Barebox: efficient malware analysis on bare-metal. In: Proceedings of the 27th Annual Computer Security Applications Conference, ACSAC '11, p. 403–412. Association for Computing Machinery, New York, NY (2011). https://doi.org/10.1145/2076732.2076790
31. Kirat, D., Vigna, G., Kruegel, C.: BareCloud: Bare-metal analysis-based evasive malware detection. In: 23rd USENIX Security Symposium (USENIX Security 14), pp. 287–301. USENIX Association, San Diego, CA (2014). https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/kirat
32. Koutsokostas, V., Patsakis, C.: Python and malware: Developing stealth and evasive malware without obfuscation. In: Proceedings of the 18th International Conference on Security and Cryptography (SECRYPT 2021) (2021)
33. Li, X., Li, Q.: An irl-based malware adversarial generation method to evade anti-malware engines. Comput. Secur. **104**, 102118 (2021)

34. Lindorfer, M., Kolbitsch, C., Milani Comparetti, P.: Detecting environment-sensitive malware. In: Recent Advances in Intrusion Detection: 14th International Symposium, RAID 2011, Menlo Park, CA, September 20–21, 2011. Proceedings 14, pp. 338–357. Springer (2011)
35. Lindorfer, M., Neugschwandtner, M., Weichselbaum, L., Fratantonio, Y., Van Der Veen, V., Platzer, C.: Andrubis–1,000,000 apps later: A view on current android malware behaviors. In: 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), pp. 3–17. IEEE (2014)
36. Maffia, L., Nisi, D., Kotzias, P., Lagorio, G., Aonzo, S., Balzarotti, D.: Longitudinal study of the prevalence of malware evasive techniques. arXiv preprint arXiv:2112.11289 (2021)
37. Mandl, T., Bayer, U., Nentwich, F.: Anubis analyzing unknown binaries the automatic way. In: Virus Bulletin Conference, vol. 1, p. 02 (2009)
38. Mutti, S., Fratantonio, Y., Bianchi, A., Invernizzi, L., Corbetta, J., Kirat, D., Kruegel, C., Vigna, G.: Baredroid: Large-scale analysis of android apps on real devices. In: Proceedings of the 31st Annual Computer Security Applications Conference, ACSAC '15, pp. 71–80. Association for Computing Machinery, New York, NY (2015). https://doi.org/10.1145/2818000.2818036
39. Oberheide, J., Miller, C.: Dissecting the android bouncer. SummerCon2012 New York **95**, 110 (2012)
40. Olaimat, M.N., Maarof, M.A., Al-rimy, B.A.S.: Ransomware anti-analysis and evasion techniques: A survey and research directions. In: 2021 3rd International Cyber Resilience Conference (CRC), pp. 1–6. IEEE (2021)
41. Or-Meir, O., Nissim, N., Elovici, Y., Rokach, L.: Dynamic malware analysis in the modern era—a state of the art survey. ACM Comput. Surv. **52**(5), 1–48 (2019). https://doi.org/10.1145/3329786
42. Panakkal, G.R.: Leaving our zip undone: How to abuse zip to deliver malware apps. In: Virus Bulletin Conference (2014)
43. Petsas, T., Voyatzis, G., Athanasopoulos, E., Polychronakis, M., Ioannidis, S.: Rage against the virtual machine: Hindering dynamic analysis of android malware. In: Proceedings of the Seventh European Workshop on System Security, EuroSec '14. Association for Computing Machinery, New York, NY (2014). https://doi.org/10.1145/2592791.2592796
44. Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., Nicholas, C.K.: Malware detection by eating a whole exe. In: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
45. Raffetseder, T., Kruegel, C., Kirda, E.: Detecting system emulators. In: Information Security: 10th International Conference, ISC 2007, Valparaíso, Chile, October 9–12, 2007. Proceedings 10, pp. 1–18. Springer (2007)
46. Rossow, C., Dietrich, C.J., Grier, C., Kreibich, C., Paxson, V., Pohlmann, N., Bos, H., Van Steen, M.: Prudent practices for designing malware experiments: Status quo and outlook. In: 2012 IEEE Symposium on Security and Privacy, pp. 65–79. IEEE (2012)
47. Sikorski, M., Honig, A.: Practical Malware Analysis: The Hands-on Guide to Dissecting Malicious Software. No Starch Press, San Francisco (2012)
48. Sims, J.: BlackMamba: Using AI to Generate Polymorphic Malware. https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware (2023). Accessed 31 Mar 2024
49. Stoll, C.: The Cuckoo's Egg: Tracking a Spy Through the Maze of Computer Espionage. Simon and Schuster, New York (2005)
50. Yan, L.K., Yin, H.: {DroidScope}: Seamlessly reconstructing the {OS} and dalvik semantic views for dynamic android malware analysis. In: 21st USENIX Security Symposium (USENIX Security 12), pp. 569–584 (2012)
51. You, I., Yim, K.: Malware obfuscation techniques: A brief survey. In: 2010 International Conference on Broadband, Wireless Computing, Communication and Applications, pp. 297–300. IEEE (2010)
52. Zhang, Z., Qi, P., Wang, W.: Dynamic malware analysis with feature engineering and feature learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 1210–1217 (2020)

“Let’s use AI to prepare for the future we hope never comes.”

— Alexandros Zacharis