ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. «ΚΥΒΕΡΝΟΑΣΦΑΛΕΙΑ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

"Κίνδυνοι στην Ιδιωτικότητα και την Προστασία Δεδομένων σε περιβάλλον Μεγάλων Γλωσσικών Μοντέλων (Privacy and Data Protection Risks in Large Language Models)"

Οδυσσέας Αλέξανδρος Α. Καρανικόλας

Επιβλέπων Καθηγητής:
Στέφανος Γκρίτζαλης, Καθηγητής

ΠΕΙΡΑΙΑΣ
ΦΕΒΡΟΥΑΡΙΟΣ 2026

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κίνδυνοι στην Ιδιωτικότητα και την Προστασία Δεδομένων σε περιβάλλον Μεγάλων Γλωσσικών Μοντέλων (Privacy and Data Protection Risks in Large Language Models)

Οδυσσέας Αλέξανδρος Καρανικόλας

Α.Μ.: MTE24015

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία διερευνά το σύνθετο τοπίο των κινδύνων ιδιωτικότητας και προστασίας δεδομένων που προκύπτει από την ευρεία χρήση των μεγάλων γλωσσικών μοντέλων. Βασικός άξονας της έρευνας είναι η αξιολόγηση της εγγενούς τάσης των μοντέλων να απομνημονεύουν και να αναπαράγουν ευαίσθητα προσωπικά δεδομένα, η ανάλυση των επιπτώσεων στην απόδοση των μοντέλων που έχουν οι τεχνικές προστασίας της ιδιωτικότητας, καθώς και η διερεύνηση της επάρκειας του κανονιστικού πλαισίου (GDPR, AI Act) ως προς τη σαφή κατανομή ευθυνών, και την αντιμετώπιση της αλγοριθμική αδιαφάνεια. Η μεθοδολογική προσέγγιση βασίστηκε στη συστηματική βιβλιογραφική ανασκόπηση και στην ανάλυση τεκμηριωμένων περιστατικών διαρροής και κανονιστικών ρυθμίσεων, με στόχο τη διαμόρφωση μιας ολοκληρωμένης εικόνας των προκλήσεων και των δυνατών λύσεων. Τα ευρήματα της μελέτης καταδεικνύουν ότι τα μεγάλα γλωσσικά μοντέλα εισάγουν μια νέα κατηγορία κινδύνων, η οποία δεν αντιμετωπίζεται αποτελεσματικά από τις παραδοσιακές μεθόδους αξιολόγησης λογισμικού. Επιβεβαιώθηκε πως φαινόμενα όπως η ακούσια απομνημόνευση και οι επιθέσεις χειραγώγησης εισόδων παραμένουν κρίσιμα, ακόμη και όταν εφαρμόζονται τα βασικά μέτρα ασφαλείας. Ταυτόχρονα, διαπιστώθηκε πως οι αυστηρές τεχνικές παρεμβάσεις για την προστασία της ιδιωτικότητας συχνά επιβαρύνουν την ακρίβεια του συστήματος, καθιστώντας σαφές ότι η τεχνική θωράκιση, παρότι αναγκαία, δεν επαρκεί από μόνη της για να διασφαλίσει την κανονιστική συμμόρφωση και την εμπιστοσύνη των χρηστών. Ως απάντηση σε όλες αυτές τις προκλήσεις, η εργασία τονίζει την ανάγκη για ένα ολιστικό σύστημα διακυβέρνησης που υπερβαίνει τις τεχνικές λύσεις, αξιοποιώντας ουσιαστικά τα εργαλεία διαφάνειας, όπως τις κάρτες μοντέλων και δεδομένων, και αναγνωρίζοντας τη σημασία των ελέγχων τρίτων μερών και των προτύπων πιστοποίησης. Ο συνδυασμός των εργαλείων αυτών μπορεί να γεφυρώσει το χάσμα μεταξύ της τεχνικής πολυπλοκότητας των σύγχρονων μοντέλων και των νομικών απαιτήσεων, διευκολύνοντας τη λογοδοσία και την ουσιαστική αξιολόγηση των κινδύνων. Συμπερασματικά, η μελέτη συνεισφέρει στη θεμελίωση ενός συνεκτικού πλαισίου για την ασφαλή και υπεύθυνη αξιοποίηση των γλωσσικών μοντέλων, προτείνοντας τη μετάβαση από στατικές διαδικασίες ελέγχου σε δυναμικούς μηχανισμούς συνεχούς αξιολόγησης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Ασφάλεια Πληροφοριακών Συστημάτων και Προστασία Ιδιωτικότητας

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Μεγάλα Γλωσσικά Μοντέλα, Ιδιωτικότητα, Προστασία Δεδομένων, GDPR, Κανονιστική Συμμόρφωση

ABSTRACT

The present thesis investigates the complex landscape of privacy and data protection risks emerging from the use of large language models. It focuses on evaluating the inherent tendency of these models to memorize and reproduce sensitive personal data, analyzing how privacy-preserving techniques affect their accuracy and overall performance, and assessing the adequacy of current regulatory frameworks, such as the GDPR and AI Act, in addressing algorithmic opacity. The methodological approach was based on a systematic literature review and analysis of documented data leakage incidents and institutional regulations, aiming to develop a comprehensive view of the challenges and potential solutions. The findings of the study demonstrate that large language models introduce a novel category of risks that are not effectively addressed by traditional software evaluation methods. It was confirmed that phenomena such as unintended memorization and input manipulation attacks remain critical, even when basic security measures are applied. At the same time, it was found that strict technical interventions for privacy protection often compromise system accuracy, making it clear that technical hardening, while necessary, is not sufficient on its own to ensure regulatory compliance and user trust. In response to these challenges, the thesis emphasized the need for a holistic governance system that transcends technical solutions, effectively utilizing transparency tools such as model and data cards, and recognizing the importance of third party audits and certification standards. The combination of these tools can bridge the gap between the technical complexity of modern models and legal requirements, facilitating accountability and effective risk assessment. In conclusion, the study contributes to the foundation of a coherent framework for the safe and responsible utilization of language models, proposing a transition from static control processes to dynamic mechanisms of continuous evaluation.

SUBJECT AREA: Information Security and Privacy Protection

KEYWORDS: Large Language Models, Privacy, Data Protection, GDPR, Regulatory Compliance

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

1 ΕΙΣΑΓΩΓΗ

1.1 Κίνητρο και ευρύτερο πλαίσιο

Η ραγδαία ανάπτυξη των μεγάλων γλωσσικών μοντέλων και η σχεδόν ανεξέλεγκτη ενσωμάτωσή τους στις σύγχρονες εφαρμογές της καθημερινότητας έχουν δημιουργήσει ένα πρωτοφανές περιβάλλον γεμάτο προκλήσεις για την προστασία της ιδιωτικότητας και την ασφάλεια των δεδομένων. Τα μοντέλα αυτά αξιοποιούνται σε διάφορα πεδία, από τη σύνθεση εικόνων μέχρι και τη συγγραφή προγραμματιστικού κώδικα, καθώς έχουν τη δυνατότητα να επεξεργάζονται τεράστιους όγκους πληροφοριών. Ωστόσο, σε αυτά τα σύνολα δεδομένων υπάρχει πιθανότητα να περιλαμβάνονται προσωπικές πληροφορίες, οι οποίες μπορεί να μην προορίζονταν εξαρχής για ευρεία δημοσιοποίηση, γεγονός που εγείρει σοβαρές απειλές παραβίασης της ιδιωτικότητας. Έτσι, διαμορφώνεται ένα θολό και επικίνδυνο περιβάλλον όπου τα όρια μεταξύ δημόσιων και ιδιωτικών δεδομένων γίνεται ολοένα και λιγότερο ευδιάκριτο.

Ιδιαίτερη ανησυχία προκαλεί η δυνατότητα των μοντέλων να απομνημονεύουν και να αναπαράγουν πληροφορίες από τα δεδομένα στα οποία έχουν εκπαιδευτεί, αλλά και ευαίσθητες πληροφορίες που μπορεί να λάβουν σαν είσοδο. Με άλλα λόγια, ένα μοντέλω μπορεί να ενσωματώνει πληροφορίες φυσικών προσώπων και να τις επαναφέρει στις απαντήσεις του, μετά από καιρό, όταν του δωθεί το κατάλληλο ερέθισμα, ακόμα και όταν αυτές οι πληροφορίες δεν περιλαμβάνονται πλέον ρητά στο εκπαιδευτικό σύνολο. Αυτή η δυνατότητα αποτελεί τη ρίζα του κινδύνου ακούσιας διαρροής προσωπικών δεδομένων, εφόσον δυσχεραίνει την διαλλογή των ασφαλών δεδομένων και των μη προσβάσιμων στο τελικό σύστημα.

Είναι ήδη αποδεδειγμένο πως ευαίσθητες πληροφορίες μπορούν να αναπαράγονται από ένα μοντέλο ακόμα και μετά από την απομάκρυνσή τους από το εκπαιδευτικό σύνολο. Αυτό σημαίνει ότι ακόμα και μονάχα η αρχική έκθεσή του μοντέλου σε προσωπικά δεδομένα είναι ικανή να αφήσει μόνιμο αποτύπωμα στη συμπεριφορά του, καθιστώντας ιδιαίτερα απαιτητική τη διαδικασία για την πλήρη απαλοιφή του περιεχομένου αυτού. Ένα χαρακτηριστικό των μοντέλων που αναδεικνύει τη σύνθετη φύση του προβλήματος είναι η εμμονή τους να διατηρούν δεδομένα που θεωρητικά έχουν διαγραφεί. Οι καθαρά τεχνικές λύσεις που έχουν αναπτυχθεί, όπως για παράδειγμα η απλή επανεκπαίδευση με τα ευαίσθητα δεδομένα διεγραμένα από το σύνολο, δεν επαρκούν από μόνες τους για να εγγυηθούν την προστασία της ιδιωτικότητας. Απαιτείται μια πιο ολιστική προσέγγιση, που θα διασφαλίζει ότι δεν θα αναπαράγονται πληροφορίες οι οποίες έπρεπε εξαρχής να παραμείνουν εμπιστευτικές.

Φυσικά, εκτός από το τεχνικό κομμάτι, οι προκλήσεις αυτές αντιμετωπίζουν και σημαντικά ηθικοκοινωνικά ζητήματα. Οι φορείς που αναπτύσσουν και χρησιμοποιούν μεγάλα γλωσσικά μοντέλα έρχονται αντιμέτωποι με το δίλημμα του πως θα διατηρήσουν την πλήρη λειτουργικότητα και εξελικτική δυναμική των συστημάτων, ενώ παράλληλα πρέπει να σέβονται τα θεμελιώδη δικαιώματα των υποκειμένων των δεδομένων καθ’ όλη τη διάρκεια της επεξεργασίας τους. Με άλλα λόγια, ποιός είναι ο τρόπος που μπορεί να συνυπάρξει η καινοτομία με τον σεβασμό προς την ιδιωτικότητα; Αυτό το ερώτημα είναι άρηκτα συνδεδεμένο με την έννοια της λογοδοσίας. Δηλαδή, στα περιβάλλοντα όπου οι αποφάσεις λαμβάνονται από σύνθετα μοντέλα, πρέπει να καθίσταται σαφές το ποίος φέρει την ευθύνη σε σενάρια που θα συμβούν παραβιάσεις και ανεπιθύμητες ενέργειες. Συνεπώς, απαιτείται ισορροπία μεταξύ της εξέλιξης της τεχνολογίας και της προστασίας των θεμελιωδών αρχών, όπως είναι η διαφάνεια, η υπευθυνότητα, και η ασφάλεια.

Σε κανονιστικό επίπεδο, έχουν γίνει προσπάθειες να οριοθετηθεί ένα τέτοιο πεδίο ώστε να αντιμετωπιστούν αυτές οι απειλές. Στην Ευρώπη, ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR) έχει ήδη τεθεί σε ισχύ, ενώ ο ακόμα υπό διαμόρφωση Κανονισμός για

την Τεχνητή Νοημοσύνη (AI Act) αποσκοπεί στην εισαγωγή συγκεκριμένων υποχρεώσεων και περιορισμών στα συστήματα τεχνητής νοημοσύνης. Επιπλέον, διεθνή πρότυπα, όπως το ISO/IEC 42001, προσφέρουν κατευθυντήριες γραμμές για τη διακυβέρνηση και την πιστοποίηση αυτών των συστημάτων. Για την πρακτική συμμόρφωση με αυτά τα πλαίσια χρειάζονται εργαλεία τεκμηρίωσης και μηχανισμοί εσωτερικής διακυβέρνησης, ώστε να έχουν τη δυνατότητα οι οργανισμοί να επιδείξουν έμπρακτα τον σεβασμό τους στους κανόνες προστασίας δεδομένων. Ωστόσο, ακόμα και σήμερα παρατηρείται έλλειψη μιας γενικευμένης προσέγγισης μεταξύ των φορέων των μοντέλων και των κανονιστικών πλαισίων, οπότε το πλαίσιο λογοδοσίας σε περιστατικά παραβίασης της ιδιωτικότητας παραμένει θολό. Αυτή η ερευνητική εργασία αντλεί το έναυσμά της από αυτό το κενό και επιδιώκει να βρει τα αίτια δημιουργίας του, αλλά και μεθόδους κάλυψής του, εξετάζοντας σε βάθος τεχνικές, νομικές και ηθικές προσεγγίσεις του προβλήματος.

1.2 Ερευνητικά ερωτήματα και υποθέσεις

Η παρούσα έρευνα καθοδηγείται από συγκεκριμένα θεμελιώδη ερωτήματα που προκύπτουν από την καθημερινή χρήση των μεγάλων γλωσσικών μοντέλων και τις διαπιστωμένες αδυναμίες τους σε ζητήματα ασφάλειας, διαφάνειας, κανονιστικής συμμόρφωσης, και προστασίας δεδομένων. Παρά την ενσωμάτωση τεράστιου όγκου δεδομένων κατά την εκπαίδευσή τους, είναι ακόμα αβέβαιο το αν τα μοντέλα μπορούν να διαχειριστούν αποτελεσματικά όλες τις ευαίσθητες πληροφορίες στις οποίες εκτίθενται, χωρίς να βρεθούν ευάλωτα σε κινδύνους διαρροής. Η αυξημένη πολυπλοκότητα των αλγορίθμων μαζί με την απουσία ενός οργανωμένου πλαισίου λογοδοσίας, δημιουργούν την ανάγκη για διερεύνηση καινούργιων μεθόδων διαχείρισης κινδύνου, οι οποίες θα εξισορροπούν την έννοια της ασφάλειας με την έννοια της χρηστικότητας. Με βάση τα παραπάνω, διατυπώνονται στη συνέχεια τα βασικά ερευνητικά ερωτήματα της εργασίας, μαζί με τις αντίστοιχες υποθέσεις εργασίας που θα διερευνηθούν στα κεντρικά κεφάλαια.

- Ερευνητικό ερώτημα 1: Σε τι βαθμό τα μεγάλα γλωσσικά μοντέλα έχουν τη δυνατότητα να απομνημονεύουν και να αναπαράγουν ευαίσθητα προσωπικά δεδομένα, ακόμα και μετά από προσπάθειες αφαίρεσης τους, μέσω τεχνικών ανωνυμοποίησης ή επανεκπαίδευσης;
    - Υπόθεση: Υποθέτουμε πως τα μεγάλα γλωσσικά μοντέλα έχουν την εγγενή τάση απομνημόνευσης εκπαιδευτικών δεδομένων, γεγονός που τα καθιστά επιρρεπή στην αναπαραγωγή ευαίσθητων δεδομένων παρά τη χρήση των βασικών μέτρων προστασίας. Η υπόθεση αυτή βασίζεται στην ύπαρξη τεκμηριωμένων περιστατικών διαρροής (βλ. Κεφάλαιο 4), υποδεικνύοντας πως σε περιπτώσεις που απουσιάζουν προηγμένοι μηχανισμοί ασφαλείας, η πιθανότητα αποκάλυψης προσωπικών δεδομένων παραμένει μεγάλη.
- Ερευνητικό ερώτημα 2: Με ποίο τρόπο και σε τι βαθμό επηρεάζει η εφαρμογή τεχνικών προστασίας της ιδιωτικότητας την ακρίβεια και τη γενικότερη απόδοση των γλωσσικών μοντέλων;
    - Υπόθεση: Η ενσωμάτωση αυστηρών μέτρων προστασίας προσωπικών δεδομένων συνεπάγεται αναπόφευκτα έναν συμβιβασμό ως προς την απόδοση του συστήματος. Συγκεκριμένα, όσο αυξάνεται το επίπεδο προστασίας μέσω τεχνικών παρεμβάσεων, αναμένουμε μείωση στην ακρίβεια και στη συνολική απόδοση ενός μοντέλου. Με άλλα λόγια, υπάρχει μια σχέση ανταλλαγής μεταξύ ιδιωτικότητας και χρησιμότητας, όπως θα διερευνηθεί μέσω της ανάλυσης τεχνικών μεθόδων προστασίας (βλ. Κεφάλαιο 8) και των ποιοτικών αντισταθμισμάτων (βλ. Κεφάλαιο 12).

- Ερευνητικό ερώτημα 3: Σε ποίον βαθμό οι ισχύοντες κανονισμοί, όπως ο GDPR και ο AI Act, επαρκούν για τον καθορισμό σαφών ορίων ευθύνης και λογοδοσίας στους οργανισμούς που αναπτύσσουν μεγάλα γλωσσικά μοντέλα;
    - Υπόθεση: Παρά τις γενικές αρχές συμμόρφωσης που προβλέπουν τα κανονιστικά πλαίσια, όπως ο GDPR, στην πράξη, ακόμα, δεν έχει εδραιωθεί ένα ξεκάθαρο πλαίσιο κατανομής ευθυνών για ζητήματα ιδιωτικότητας στα συστήματα τεχνητής νοημοσύνης, και συγκεκριμένα στα μεγάλα γλωσσικά μοντέλα. Με άλλα λόγια, υπάρχει ένα σημαντικό κενό λογοδοσίας, καθώς η πολυπλοκότητα των μοντέλων καθιστά δυσχερή την ακριβή κατανομή ευθυνών μεταξύ δημιουργών, φορέων, και χρηστών. Το ζήτημα αυτό αναλύεται περαιτέρω τόσο από τη νομική σκοπιά (βλ. Κεφάλαιο 5) όσο και υπό το πρίσμα των προτεινόμενων μηχανισμών διακυβέρνησης και πιστοποίησης (βλ. Κεφάλαιο 13).
- Ερευνητικό ερώτημα 4: Υπάρχουν μέθοδοι επεξηγήσιμης τεχνητής νοημοσύνης που να ενισχύσουν τη διαφάνεια των μοντέλων χωρίς να αυξάνουν τον κίνδυνο διαρροής ευαίσθητων δεδομένων;
    - Υπόθεση: Σε θεωρητικό επίπεδο είναι εφικτή η υιοθέτηση συγκεκριμένων πρακτικών διαφάνειας και τεκμηρίωσης , οι οποίες δεν θα εκθέτουν σε επιπλέον κίνδυνο την ιδιωτικότητα. Για παράδειγμα, έχουν αναπτυχθεί εργαλεία, όπως οι κάρτες μοντέλων και τα φύλλα δεδομένων, τα οποία μπορούν να παρέχουν πολύτιμες πληροφορίες για τα χαρακτηριστικά και τη συμπεριφορά του μοντέλου χωρίς να αποκαλύπτουν ευαίσθητα δομικά στοιχεία του. Η προσέγγιση αυτή προτείνεται ως βέλτιστη πρακτική για την επίτευξη της ισορροπίας μεταξύ διαφάνειας και ασφάλειας (βλ. Κεφάλαιο 5 και 13.2).

1.3 Δομή της εργασίας και περίληψη κεφαλαίων

Η παρούσα εργασία δομείται σε θεματικές ενότητες που ακολουθούν μια σταδιακή ερευνητική πορεία, ξεκινώντας από τη θεωρητική θεμελίωση των μεγάλων γλωσσικών μοντέλων και καταλήγοντας στη διαμόρφωση ενός πρακτικού πλαισίου διακυβέρνησης. Η γενικότερη δομή της εργασίας αποσκοπεί στην ολιστική αντιμετώπιση του χάσματος μεταξύ της τεχνολογικής πολυπλοκότητας των μοντέλων και των απαιτήσεων για ασφάλεια και ιδιωτικότητα.

Αναλυτικότερα, η εργασία οργανώνεται ως εξής:

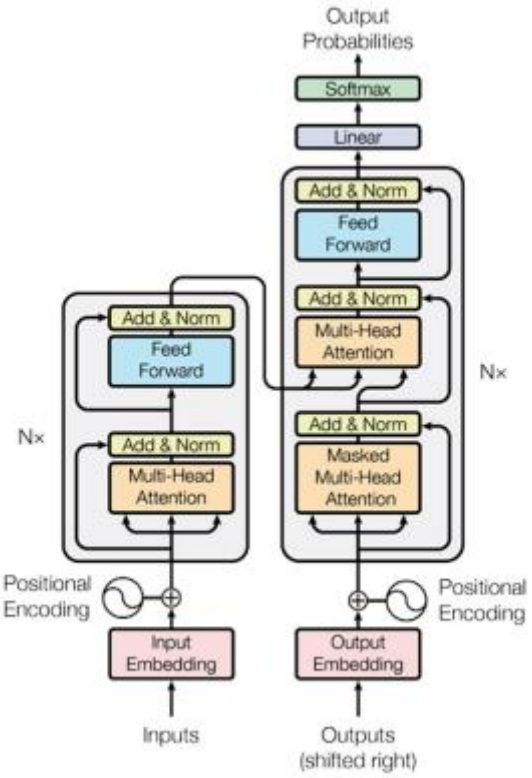
- Στα πρώτα κεφάλαια παρουσιάζονται οι βασικές έννοιες, όπως η αρχιτεκτονική και ο τρόπος λειτουργίας των μεγάλων γλωσσικών μοντέλων, με στόχο τη διαμόρφωση ενός ενιαίου εννοιολογικού πλαισίου. Έπειτα, αναλύονται οι βασικές κατηγορίες προβλημάτων ιδιωτικότητας και ασφάλειας που προκύπτουν από την εκπαίδευση και τη χρήση των μοντέλων, με έμφαση σε πραγματικά περιστατικά διαρροής δεδομένων και επιθέσεων.
- Στη συνέχεια, εξετάζονται τεχνικές και θεσμικές προσεγγίσεις για τον μετριασμό των κινδύνων, όπως μηχανισμοί προστασίας της ιδιωτικότητας, μέθοδοι διαφάνειας, και πρακτικές τεκμηρίωσης. Ταυτόχρονα, αναλύεται το ισχύον και υπό διαμόρφωση κανονιστικό πλαίσιο (GDPR, AI Act), προσδιορίζοντας τις υποχρεώσεις συμμόρφωσης για τους φορείς των μοντέλων.
- Ο πυρήνας της προτεινόμενης λύσης αναπτύσσεται στο Κεφάλαιο 13, το οποίο πραγματεύεται τη διακυβέρνηση και την πιστοποίηση των συστημάτων. Ουσιαστικά, προτείνεται ένα ολοκληρωμένο πλαίσιο που περιλαμβάνει:
    - Μηχανισμούς ελέγχου τρίτων μερών για την εξασφάλιση αντικειμενικότητας.
    - Εργαλεία διαφάνειας, όπως οι κάρτες μοντέλων και δεδομένων.

- Προδιαγραφές πιστοποίησης ασφάλειας και ιδιωτικότητας βάσει διεθνών προτύπων (ISO 42001, NIST).
- Τέλος, η εργασία ολοκληρώνεται με τη σύνοψη των συμπερασμάτων και τη διατύπωση προτάσεων για μελλοντική έρευνα, υπογραμμίζοντας την ανάγκη για συνεχή προσαρμογή των μηχανισμών ελέγχου στις εξελίξεις της τεχνητής νοημοσύνης.

2 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Αρχιτεκτονικές μεγάλων γλωσσικών μοντέλων

Για τη δημιουργία των μεγάλων γλωσσικών μοντέλων χρησιμοποιείται σχεδόν πάντα η αρχιτεκτονική Transformer, όπου έχει πλέον αντικαταστήσει τις παλαιότερες αναδρομικές μεθόδους λόγω της ικανότητας της να επεξεργάζεται τεράστιες ακολουθίες κειμένων με τέτοιο τρόπο που υπάρχει σταθερότητα παράλληλα με την αποδοτικότητα. Το βασικό χαρακτηριστικό αυτής της αρχιτεκτονικής είναι πως η κάθε λέξη δεν επεξεργάζεται μία μία με τη σειρά, αλλά μέσα από έναν μηχανισμό που επιτρέπει στο μοντέλο να αναγνωρίζει τη σχέση της κάθε λέξης στην πρόταση με όλες τις υπόλοιπες. Έτσι, αυτός ο μηχανισμός είναι αρκετά ελκυστικός στους δημιουργούς των μοντέλων, με αποτέλεσμα να χρησιμοποιείται από όλα τα μεγάλα μοντέλα σήμερα, δίνοντας τους τη δυνατότητα να εντοπίζουν συσχετισμούς, και δομές ακόμα και σε αρκετά μεγάλα κείμενα.



Εικόνα 1: Η αρχιτεκτονική Transformer.

Κύριο στοιχείο της αρχιτεκτονικής είναι το self attention, το οποίο χωρίζεται σε διάφορες κεφαλές. Η κάθε κεφαλή ελέγχει τους τύπους σχέσεων ανάμεσα στην προτάσεις, και στις

παραγράφους, ενισχύοντας τη μάθηση πολλαπλών μοτίβων. Τη συγκεκριμένη εξειδίκευση την παρατηρούμε στα πιο σύγχρονα μοντέλα, των οποίων οι κεφαλές έχουν διακριτούς ρόλους, όπως για παράδειγμα την ιδιότητα να εντοπίζουν γραμματικές δομές, ή τη μεταφορά νοήματος σε πιο γενικά συμφραζόμενα. Πιο αναλυτικά, για τις λειτουργίες των κεφαλών φαίνεται πως σε πολλαπλά επίπεδα των μοντέλων αναπτύσσονται αφηρημένες αναπαραστάσεις, οι οποίες δεν λειτουργούν για απλές γλωσσικές συσχετίσεις, αλλά επεκτείνονται σε πιο σύνθετα μοτίβα σκέψης και συλλογισμού.

Αν και τα μοντέλα Transformer είναι ιδανικά για να βασιστούν επάνω τους οι κατασκευαστές, πολύ συχνά ακολουθούν διαφορετικά μονοπάτια προσέγγισης ανάλογα με τον σκοπό για τον οποίο προορίζεται το κάθε μοντέλο. Οι τρεις κυρίαρχες κατηγορίες είναι τα encoder only, τα decoder only, και τα encoder decoder. Συγκεκριμένα, τα encoder μοντέλα, όπως είναι τα μοντέλα BERT (Bidirectional Encoder Representations from Transformers), εκπαιδεύονται με σκοπό να κατανοήσουν ολόκληρο το περιεχόμενο της πρότασης βασιζόμενα στο πλήρες συμφραζόμενο, γεγονός που τα καθιστά ιδανικά για εργασίες κατανόησης κειμένου, ενώ αν εφαρμόσουν αμφίδρομα τη διαδικασία μπορούν να εντοπίζουν με ακρίβεια τις σχέσεις ανάμεσα στις λέξεις.

Αντίθετα, τα decoder only μοντέλα, όπως είναι τα μοντέλα GPT (Generative Pretrained Transformer), ακολουθούν μια διαδικασία σαν μια μονόδρομη συνάρτηση. Ουσιαστικά, το μοντέλο προβλέπει κάθε φορά την επόμενη λέξη στηριζόμενο σε όσες έχουν προηγηθεί, καθιστώντας το ιδιαίτερα αποδοτικό σε διεργασίες όπως η παραγωγή φυσικού κειμένου, καθώς συνεχίζει να εκπαιδεύεται και να συνεχίζει την ακολουθία συνεπώς και με δημιουργικότητα. Αυτή η ικανότητα αποτελεί τη βάση για πιο πολύπλοκες εργασίες που εκτελούν τα μοντέλα, όπως η περίληψη μεγάλων κειμένων, η μετάφραση, και η επίλυση σύνθετων προβλημάτων.

Η τρίτη κατηγορία, τα encoder decoder μοντέλα, αποτελούν έναν συνδυασμό των προηγούμενων δύο, καθώς χρησιμοποιούν το πρώτο μέρος για την κατανόηση του εισερχόμενου κειμένου, και το δεύτερο για την παραγωγή νέου. Στην πραγματικότητα, όλες οι γλωσσικές διεργασίες αντιμετωπίζονται σαν προβλήματα μετατροπής από είσοδο σε έξοδο, γεγονός που τα καθιστά ιδιαίτερα ευέλικτα, και για αυτό χρησιμοποιούνται από μεγάλα μοντέλα όπως το T5 και το BART.

Φυσικά, η εξέλιξη των αρχιτεκτονικών δεν έχει σταματήσει στο βασικό Transformer, καθώς υπάρχει ανάγκη για μεγαλύτερη ερμηνεία νοημάτων. Συνεπώς, έχουν αναπτυχθεί ειδικές κατηγορίες μοντέλων ειδικά σχεδιασμένες για long context, οι οποίες χρησιμοποιούν πιο αποδοτικές μορφές attention, ώστε να μειώσουν δραστικά την κατανάλωση GPU μνήμης, και να αυξήσουν την ταχύτητα χωρίς να απαιτούνται απαγορευτικοί υπολογιστικοί πόροι. Με αυτή τη βελτίωση των μεθόδων επηρεάζεται άμεσα η δυνατότητα των μοντέλων να διαχειρίζονται ουσιαστικά τεράστια σύνολα πληροφοριών.

Αν μπαίναμε σε διαδικασία σύγκρισης των αρχιτεκτονικών θα καταλαβαίναμε πως δεν υπάρχει κάποια που να υπερέχει, καθώς η κάθε μια εξειδικεύεται για κάποια συγκεκριμένη χρήση. Τα encoder only προσφέρουν πιο γρήγορες και οικονομικές λύσεις για την κατανόηση κειμένων, τα decoder only είναι ασυναγώνιστα στην παραγωγή κειμένων, ενώ τα encoder decoder είναι πιο προσαρμοστικά και μπορούν να χρησιμοποιηθούν για ένα πλήθος ενεργειών. Για τη σωστή επιλογή αρχιτεκτονικής πρέπει να έχουμε υπόψιν το είδος της εργασίας, την υπολογιστική ισχύ που διαθέτουμε, και τον ρόλο για τον οποίο προορίζουμε το μοντέλο.

2.2 Διαδικασία εκπαίδευσης

Ο τρόπος με τον οποίο ένα μοντέλο εκπαιδεύεται αποτελείται από διάφορα στάδια, στα οποία αυτό εκτίθεται σε έναν τεράστιο όγκο δεδομένων επηρεάζοντας την απόδοσή του, αλλά και την ικανότητά του να ενσωματώνει τις πληροφορίες που συναντά στον κορμό του.

Για να ξεκινήσει η διαδικασία της εκπαίδευσης, το μοντέλο πρέπει να περάσει από τη φάση του pre-training, κατά την οποία αυτό εκπαιδεύεται σε ένα τεράστιο σύνολο κειμένων. Παράλληλα, αναγνωρίζει ακολουθίες στις οποίες προσπαθεί να προβλέψει την επόμενη λέξη ή κάποια λέξη που παραλείπεται, ώστε μέσα από την επανάληψη να αποκτήσει μια φυσικότητα στη λειτουργία της ανθρώπινης γλώσσας. Καθώς όμως τα σύνολα των πληροφοριών είναι τόσο μεγάλο, τα μοντέλα συχνά απορροφούν όχι μόνο τα σωστά μοτίβα και τους κανόνες της ανθρώπινης γλώσσας, αλλά και τις ασυνέπειες, τις ασάφειες, και τις προκαταλήψεις που μπορεί να βρίσκονται σε αυτά.

Το αμέσως επόμενο στάδιο εκπαίδευσης ονομάζεται fine-tuning, και είναι αυτό που προετοιμάζει το μοντέλο για πιο συγκεκριμένες χρήσεις. Για να γίνει αυτό απαιτούνται πιο στοχευμένα σύνολα δεδομένων, με τα οποία θα μπορεί το μοντέλο να αποδίδει πιο αποτελεσματικά σε διεργασίες που απαιτούν περιορισμένη χρήση, όπως είναι η περίληψη μεγάλων κειμένων, ή η ταξινόμηση πληροφοριών. Βέβαια, αν και με την προσαρμογή αυτή ενισχύουμε την ακρίβεια, ενδέχεται να εμφανιστούν παρενέργειες, όπως για παράδειγμα μπορεί να συμβεί σε ένα μοντέλο που αρχικά εκπαιδεύτηκε σε ένα πολύ μεγάλο εύρος πληροφοριών, κερδίζοντας αρκετή ευελιξία, μετά το στάδιο του fine-tuning, όπου θα εκτεθεί σε πολύ στενό και συγκεκριμένο σύνολο δεδομένων, να χάσει αυτή την ιδιότητα της προσαρμοστικότητας. Διάφορες μελέτες αναφέρουν πως το φαινόμενο αυτό μπορεί να αλλοιώσει, και να υποβαθμίσει την αρχική συμπεριφορά του μοντέλου σε μη ανατρέψιμο στάδιο.

Τα πιο σύγχρονα μοντέλα συχνά χρησιμοποιούν τεχνικές συνεχούς εκπαίδευσης ακόμα και μετά το pre-training, καθώς πολλές φορές ο σκοπός τους μπορεί να είναι η κατανόηση ενός ευρύτερου νοήματος ή πιο απαιτητικές και εξελισσόμενες εφαρμογές. Η διαδικασία αυτή της επανεκπαίδευσης ενδέχεται να κρύβει κινδύνους, παρά την προσαρμοστικότητα που προσφέρει στο μοντέλο. Ένα βασικό ρίσκο μετά από μια τέτοια διαδικασία είναι πως το μοντέλο μπορεί να ξεχάσει χρήσιμες πληροφορίες που απέκτησε στα προηγούμενα στάδια εκπαίδευσης. Επίσης, αν και η διαρκής έκθεση σε καινούργια δεδομένα επιφέρει θετικό αντίκτυπο στις επιδόσεις του μοντέλου, ενδέχεται να αυξήσει την αβεβαιότητα σχετικά με τη συνέπεια των χαρακτηριστικών του, αλλά και το ποια από αυτά μεταβάλλονται απρόβλεπτα.

Η διαδικασία εκπαίδευσης συχνά επηρεάζεται αρνητικά από πρακτικούς περιορισμούς, αφού απαιτεί τεράστιους υπολογιστικούς πόρους, γεγονός που έχει οδηγήσει τους κατασκευαστές των μοντέλων στην ανάπτυξη αποδοτικότερων τεχνικών. Ουσιαστικά, στοχεύουν στη μείωση του κόστους επανεκπαίδευσης χωρίς να ελαττώνεται απαραίτητα η ικανότητα του μοντέλου. Οι διαδρομές που ακολούθησαν για να το πετύχουν αυτό είναι η στοχευμένη ενημέρωση μεμονωμένων επιπέδων του μοντέλου, και τρόποι μείωσης του όγκου των παραμέτρων που απαιτούν ανανέωση σε κάθε στάδιο επανεκπαίδευσης.

Φυσικά, οι επιπτώσεις της εκπαίδευσης δεν περιορίζονται στην αποδοτικότητα του μοντέλου, αλλά εμπλέκουν και την ιδιωτικότητα. Καθώς το μοντέλο εκτίθεται σε τόσο μεγάλα σύνολα δεδομένων, παράγονται εσωτερικά συσχετίσεις και αναπαραστάσεις που μπορούν να αναπαράγουν μελλοντικά ορισμένα στοιχεία, αν όχι ατόφια, από τις πληροφορίες εκπαίδευσης. Αυτό αποδεικνύεται από το γεγονός ότι ορισμένες πληροφορίες συνεχίζουν να εμφανίζονται ακόμα και όταν αυτές έχουν αφαιρεθεί από τα σύνολα επανεκπαίδευσης, συνεπώς τα μοντέλα δε λειτουργούν μόνο σαν συστήματα πρόβλεψης της επόμενης λέξης, αλλά έχουν την ιδιότητα να διατηρούν μοτίβα, και δομές και να τα ενεργοποιούν ακόμα και μετά από αρκετούς κύκλους επανεκπαίδευσης. Υπάρχουν αρκετά περιστατικά, όπου τα μοντέλα απέδειξαν πως είχαν συγκρατήσει ευαίσθητες πληροφορίες για μεγαλύτερο χρονικό διάστημα από όσο περίμεναν τα υποκείμενα και οι κατασκευαστές τους.

Συνοπτικά, η διαδικασία της εκπαίδευσης των μεγάλων γλωσσικών μοντέλων δεν είναι απλώς ένα σύνολο τεχνικών βημάτων, αλλά μια πιο σύνθετη αλυσίδα απαιτητικών διαδικασιών που έχουν άμεση επιρροή στην αποδοτικότητα, και την αξιοπιστία του

μοντέλου. Από την άλλη, καθώς τα δεδομένα εκπαίδευσης διευρύνονται, εμφανίζονται ζητήματα που προκύπτουν από τις μεθόδους που ακολουθούν τα μοντέλα για να τα απορροφούν στον κορμό τους, και για να τα ανακυκλώνουν χωρίς να είναι πάντα αυτό επιθυμητό.

2.3 Σύγχρονες εξελίξεις και μετασχηματισμοί στις αρχιτεκτονικές των μοντέλων

Καθώς οι απαιτήσεις ολοένα και αυξάνονται από τα μεγάλα γλωσσικά μοντέλα, οι κατασκευαστές τους αναγκάζονται να μετατοπιστούν σε αρχιτεκτονικές που είναι ικανές να διαχειρίζονται αρκετά μεγαλύτερα νοήματα. Για πολλά χρόνια ο περιορισμός της αρχιτεκτονικής του Transformer σχετικά με το μήκος των ακολουθιών εμπόδιζε την εξέλιξη των μοντέλων, καθώς όταν γινόταν προσπάθεια επέκτασης χάνονταν η ακρίβεια και η ποιότητα των απαντήσεων, αφού το μοντέλο έπρεπε να διαχειριστεί μεγάλα κείμενα με υπερβολικά πολλά tokens. Συνεπώς, αναπτύχθηκαν παραλλαγές attention που ελαττώνουν τον αριθμό υπολογισμών που απαιτούνται να εφαρμοστούν σε κάθε τμήμα του κειμένου, ανοίγοντας μια νέα πόρτα στα μοντέλα, ώστε το μοντέλο να μπορεί να κατανοεί πολλαπλάσιο μέγεθος ακολουθιών απ' ότι ήταν ικανό χωρίς να χρειάζεται επιπλέον μνήμη. Έτσι, τα μοντέλα έχουν την ικανότητα πλέον να διατηρούν σημαντικές πληροφορίες καθ’ όλη την έκταση του κειμένου και να επεξεργάζονται πιο συνεκτικά τις συσχετίσεις που δημιουργούνται.

Γνωρίζοντας πως η κλίμακα των μοντέλων έχει αυξηθεί, καταλαβαίνουμε πως οι κεφαλές του attention (οι οποίες είναι υπεύθυνες για να δίνουν σημασία στις σχέσεις ανάμεσα στις λέξεις) δε λειτουργούν όλες με τον ίδιο τρόπο, καθώς αναπτύσσουν πιο εξειδικευμένους ρόλους ανάμεσα στα διάφορα επίπεδα του δικτύου. Συγκεκριμένες αναλύσεις δείχνουν πως υπάρχουν κεφαλές που ακολουθούν ορισμένα στοιχεία που εντοπίζουν μέσα στο κείμενο, όπως για παράδειγμα ονόματα, λειτουργώντας σαν μηχανισμοί παρακολούθησης της ροής του κειμένου. Από την άλλη άλλες κεφαλές που εστιάζουν στη δομή και τη σύνταξη των προτάσεων, δίνοντας σημασία σε στοιχεία που δεν είναι πάντα προφανή. Υπάρχουν βέβαια και περιπτώσεις όπου διαφορετικές κεφαλές συντονίζονται μεταξύ τους όταν χρειάζεται πιο σύνθετη κατανόηση, γεγονός που υποδηλώνει πως το attention είναι ένα σύνολο λειτουργιών που συνεχίζει μέχρι και σήμερα να εξελίσσεται ώστε να καλύπτει όλες τις πτυχές του νοήματος. Έτσι, αυτή η διαφοροποίηση καθιστά την αρχιτεκτονική πολύ πιο αποτελεσματική, καθώς η κάθε πληροφορία δεν αντιμετωπίζεται ως ένα ενιαίο σύνολο, αλλά χωρίζεται σε μικρότερα κομμάτια όπου χρησιμοποιούνται ξεχωριστά.

Με την ασταμάτητη εξέλιξη των μεγάλων γλωσσικών μοντέλων γίνεται πλέον ξεκάθαρο πως ο τρόπος με τον οποίο κλιμακώνονται επηρεάζει άμεσα την απόδοσή τους. Ερευνητικές μελέτες δείχνουν πως η αύξηση των παραμέτρων, του βάθους των επιπέδων, αλλά και του τεράστιου όγκου των δεδομένων οδηγεί τα μοντέλα σε μια πιο σταθερή κατανόηση της γλώσσας, συνεπώς και στην επεξεργασία της, εφόσον πάντα η διαδικασία της εκπαίδευσης παραμένει ισορροπημένη. Με αυτή τη μεγέθυνση δημιουργείται μια προβλέψιμη τάση, στην οποία τα μεγαλύτερα μοντέλα εμφανίζουν βελτιωμένη συνοχή, πιο αξιόπιστη συμπεριφορά σε σύνθετες εργασίες, και καλύτερη ανάδειξη συσχετίσεων σε μεγαλύτερες αποστάσεις. Ένα ακόμη προαπαιτούμενο για την κλιμάκωση είναι η σταθεροποίηση της εκπαίδευσης, με σωστή επιλογή δεδομένων και με ισορροπία ανάμεσα στη χωρητικότητα του μοντέλου και στις πραγματικές ανάγκες των εφαρμογών.

Συνολικά, η πρόοδος των αρχιτεκτονικών των μεγάλων γλωσσικών μοντέλων φανερώνει πως η πορεία τους δεν καθορίζεται πλέον μόνο από το μέγεθος τους ή την υπολογιστική ισχύ, αλλά από τον τρόπο με τον οποίο οργανώνουν και διαχειρίζονται τις εσωτερικές τους λειτουργίες. Η ικανότητα να διαχειρίζονται μεγάλα νοήματα, η εξειδίκευση των μηχανισμών attention, και η προσεκτική διαχείριση της κλιμάκωσης διαμορφώνουν ένα καινούργιο περιβάλλον, όπου τα μοντέλα μπορούν να επεξεργάζονται σύνθετες πληροφορίες με

μεγαλύτερη σταθερότητα και συνοχή. Ωστόσο, όσο εξελίσσονται οι αρχιτεκτονικές, τόσο εντείνεται η ανάγκη για κατανόηση του τρόπου με τον οποίο αποθηκεύουν και διατηρούν τα δεδομένα στα οποία εκτίθενται. Αυτό ακριβώς το σημείο αποτελεί ένα κρίσιμο πέρασμα από το καθαρά τεχνικό υπόβαθρο στις σύγχρονες ανησυχίες που συνδέονται με την ιδιωτικότητα, οι οποίες γίνονται ολοένα και πιο επιτακτικές όσο τα μοντέλα αποκτούν μεγαλύτερη ισχύ και σαφώς ευρύτερη χρήση.

2.4 Έργα και ευρήματα από την πρόσφατη βιβλιογραφία σχετικά με διαρροές και μετρήσεις διαρροής

Η πρόσφατη βιβλιογραφία δείχνει πως τα μεγάλα γλωσσικά μοντέλα συνηθίζουν να διατηρούν ακολουθίες από τα δεδομένα στα οποία εκπαιδεύονται, ακόμα και αυτές που είναι εξαιρετικά σπάνιες ή δεν επιτρέπεται να αναπαραχθούν. Τα τελευταία χρόνια έχουν καταγραφεί περιστατικά στα οποία τα μοντέλα ανακατασκευάζουν αυτούσια τα δεδομένα, όπως προσωπικές πληροφορίες, ακόμα και όταν αυτά εμφανίστηκαν μία μόνο φορά στα δεδομένα εκπαίδευσης. Δυστυχώς, η συμπεριφορά αυτή δεν έχει εξαλειφθεί ακόμα, καθώς έχει παρατηρηθεί σε σύγχρονα μοντέλα, υποδεικνύοντας πως η απομνημόνευση είναι ριζικά ενσωματωμένη στον τρόπο με τον οποίο εκτελείται η εκπαίδευση. Ακόμα, έχει παρατηρηθεί πως αρκεί μόνο ένα μικρό τμήμα πρότασης για να ενεργοποιήσει την ανάκληση ολόκληρης της αρχικής ακολουθίας, γεγονός που φανερώνει ότι συγκεκριμένα μοτίβα αποθηκεύονται με υπερβολικά μεγάλη ακρίβεια.

Το ζήτημα δεν περιορίζεται στις άμεσες εξαγωγές κειμένου, καθώς πολλά μοντέλα έχουν αποδειχθεί ευάλωτα σε τεχνικές επιθέσεων που στοχεύουν να εντοπίσουν κατά πόσο μια συγκεκριμένη πληροφορία υπήρξε μέρος του συνόλου εκπαίδευσης. Τέτοιου είδους τεχνικές πραγματοποιούνται μέσω της ανάλυσης της διαφοροποιημένης συμπεριφοράς των μοντέλων μεταξύ γνωστών και αγνώστων δειγμάτων. Ουσιαστικά, ο επιτιθέμενος καταφέρνει με αυτόν τον τρόπο να ανιχνεύει τη συμμετοχή ενός δείγματος ακόμα και σε περιπτώσεις όταν δεν αναπαράγεται αυτούσιο το περιεχόμενό του. Σε διάφορες μελέτες αναφέρεται πως τα μοντέλα εμφανίζουν αυξημένη πιθανότητα να αναπαράγουν σπάνιες ακολουθίες, επειδή ξεχωρίζουν από το υπόλοιπο σύνολο δεδομένων, οπότε τα περιστατικά διαρροής σε κείμενα με πιο ιδιαίτερη μορφή αυξάνονται δραματικά. Οι μετρήσεις που χρησιμοποιούνται για την καταγραφή αυτού του φαινομένου, όπως για παράδειγμα τον υπολογισμό του πόσο πιο εύκολα ανασύρεται μια συγκεκριμένη πληροφορία σε σχέση με άλλες όμοιες της, μπορούμε να αποτυπώσουμε με ακρίβεια τον βαθμό έκθεσης που εμφανίζει το κάθε μοντέλο.

Η βιβλιογραφία δείχνει επίσης ότι, υπάρχουν σοβαρές ενδείξεις πως η διαδικασία της εκπαίδευσης επηρεάζει άμεσα το μέγεθος της απομνημόνευσης. Όταν ένα μοντέλο εκπαιδεύεται σε κατανεμημένα περιβάλλοντα, δηλαδή οι πληροφορίες διατηρούνται σε επιμέρους συσκευές και δε συγκεντρώνονται κεντρικά, παρατηρείται μείωση στην επανάληψη των ακολουθιών. Η μείωση αυτή, όμως, δεν είναι αρκετή για να εξαλείψουμε πλήρως το φαινόμενο, αφού ακόμα και σε προστατευμένα περιβάλλοντα υπάρχει η πιθανότητα να ανακτηθούν εισαγόμενες ακολουθίες μέσω ειδικών τεχνικών δειγματοληψίας. Συνεπώς, καταλαβαίνουμε πως οι μηχανισμοί αυτοί της απομνημόνευσης είναι αρκετά ανθεκτικοί, και δεν εξαλείφονται εύκολα, ανεξάρτητα από τη μέθοδο εκπαίδευσης που εφαρμόζουμε.

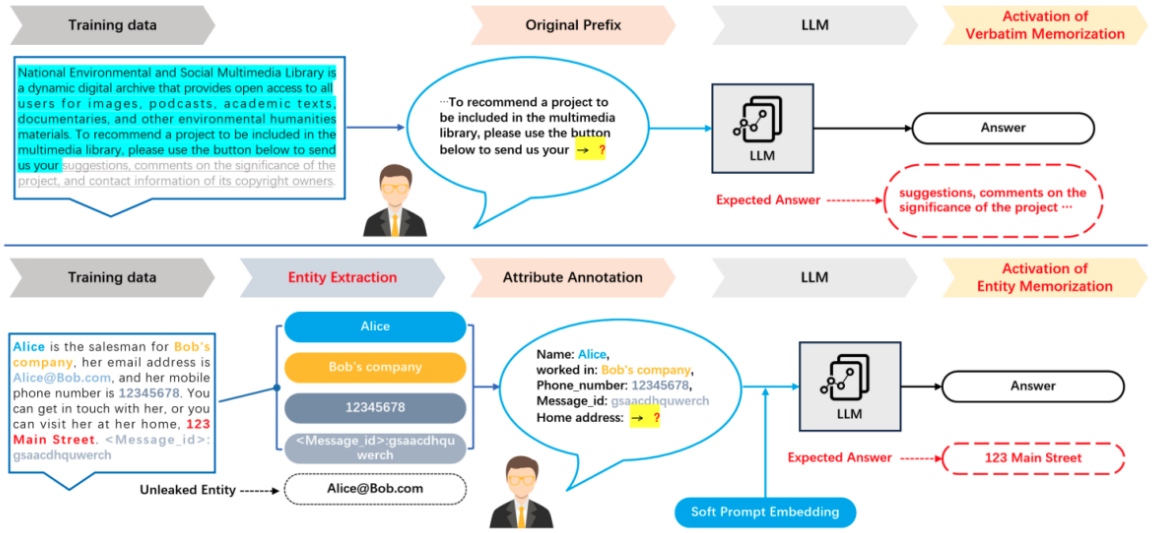
Η διεθνής βιβλιογραφία καταλήγει στο συμπέρασμα πως τα φαινόμενα διαρροής αποτελούν ιδιαίτερη και πολύπλευρη πρόκληση, η οποία αγγίζει την ακριβή αναπαραγωγή δεδομένων, την έμμεση φανέρωση συμμετοχής πληροφοριών στο σύνολο εκπαίδευσης, αλλά και την παραγωγή περιεχομένου που σχετίζεται με ηθικονομικά ζητήματα, όπως η αναπαραγωγή κειμένου χωρίς συγκατάθεση. Λόγω της δομής της διαδικασίας εκπαίδευσης, τα μοντέλα, ενδέχεται να συγκρατούν δεδομένα περισσότερο απ’ όσο περιμένει ο χρήστης, και ο κατασκευαστής του, δημιουργώντας ένα κενό που πρέπει να καλυφθεί αρχικά μέσω της

μέτρησης και της κατανόησης της διαρροής, ώστε να εφαρμοστεί μια σοβαρή προσπάθεια βελτίωσης της προστασίας της ιδιωτικότητας.

3 ΤΑΞΙΝΟΜΗΣΗ ΑΠΕΙΛΩΝ ΚΑΙ ΕΙΔΗ ΠΑΡΑΒΙΑΣΕΩΝ ΙΔΙΩΤΙΚΟΤΗΤΑΣ ΣΕ LLMS

3.1 Άμεση απομνημόνευση και διαρροή δεδομένων εκπαίδευσης

Η άμεση απομνημόνευση των μεγάλων γλωσσικών μοντέλων είναι συνδεδεμένη με την τάση τους να αποθηκεύουν στον κορμό τους αυτούσιες ακολουθίες δεδομένων από τα σύνολα εκπαίδευσης, και να τις αναπαράγουν όταν δεχθούν το κατάλληλο ερέθισμα. Έρευνες δείχνουν ότι η απομνημόνευση δεν είναι μεμονωμένο φαινόμενο αλλά σταθερό χαρακτηριστικό της εκπαίδευσης σε μεγάλα συλλογικά σύνολα δεδομένων. Πρακτικά, ο μηχανισμός αυτός επιτρέπει σε έναν επιτιθέμενο να εξάγει μέρη του εκπαιδευτικού συνόλου, χωρίς να έχει πρόσβαση στα δεδομένα ή στη διαδικασία της εκπαίδευσης, παρά μόνο εκμεταλλεύοντας την ευπάθεια του μοντέλου να αναπαράγει χαρακτηριστικές ακολουθίες. Συγκεκριμένα μεγάλα γλωσσικά μοντέλα που έχουν παράξει δισεκατομμύρια εξόδους έχουν παρατηρηθεί να διατηρούν την προδιάθεση αποθήκευσης ιδιαίτερων κειμένων που μπορούν να ανακτηθούν με απλές μεθόδους δειγματοληψίας. Αν και μόνο ένα μέρος των απομνημονευμένων πληροφοριών είναι άμεσα εξαγώγιμο, το ποσοστό αυτό είναι άξιο αναφοράς, καθώς εξαρτάται άμεσα από τον βαθμό επανάληψης μιας ακολουθίας στα σύνολα των δεδομένων εκπαίδευσης, συνεπώς αυξάνεται η πιθανότητα του επιτιθέμενου να ανακτήσει μεγάλα τμήματα κειμένου ακόμα και στις περιπτώσεις που το μοντέλο έχει εκπαιδευτεί σε τεράστιο όγκο πληροφοριών.



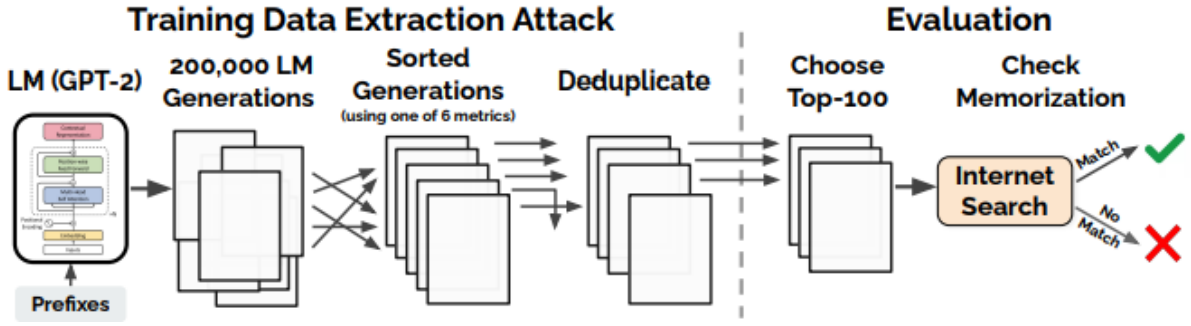
Εικόνα 2: Διαδικασία απομνημόνευσης και εξαγωγής οντοτήτων από δεδομένα εκπαίδευσης.

Εκτός από τη μορφή αυτούσιας επανάληψης, η απομνημόνευση εκδηλώνεται και σε πιο σύνθετες περιπτώσεις, όπου το μοντέλο δημιουργεί συσχετίσεις ανάμεσα σε εσωτερικά μοτίβα με σπάνια δεδομένα. Σε ένα τέτοιο πλαίσιο, ορισμένες μορφές εισόδου μπορούν να καθοδηγήσουν το μοντέλο να παράγει ακριβή αποσπάσματα από τα δεδομένα εκπαίδευσης με υψηλή ακρίβεια όταν του ζητείται επανάληψη συγκεκριμένων προτάσεων. Σε πολλές έρευνες φαίνεται πως ισχυρά σημεία στήριξης για την ενεργοποίηση της μνήμης ενός μοντέλου αποτελούν τα σπάνια ή ιδιόμορφα στοιχεία ενός κειμένου, επιτρέποντας την αποκάλυψη προσωπικών δεδομένων ή άλλων πληροφοριών που δεν προορίζονταν για τέτοιο σκοπό. Αυτό αποτελεί το βασικό πρόβλημα εναντίον στις επιθέσεις ανάκτησης δεδομένων εκπαίδευσης, στις οποίες ο επιτιθέμενος προσπαθεί να εκμεταλλευτεί την τάση

των μοντέλων να επαναλαμβάνει συγκεκριμένες γλωσσικές αλληλουχίες, καθώς επιστρέφει σε πρώιμες καταστάσεις τις παραμέτρους μέσω της παραγωγής αναγνωρίσιμων γλωσσικών δομών.

Στη βαθύτερη κατανόηση του φαινομένου σημαντικό ρόλο παίζει η τεκμηρίωση πως η απομνημόνευση δεν αποτελείται μόνο από την απλή επανάληψη λέξεων, αλλά ενεργοποιείται ακόμα και από πολύ λεπτές σχέσεις μεταξύ των λέξεων ή φράσεων και του περιεχομένου των εκπαιδευτικών συνόλων. Μελέτες γύρω από τις επιθέσεις μέσω ειδικών χαρακτήρων αναφέρουν πως τα μοντέλα τα οποία έχουν εκπαιδευτεί σε σύνολα δεδομένων από το διαδίκτυο αποκτούν την τάση να συνδέουν τους ειδικούς χαρακτήρες που έχουν συναντήσει με συγκεκριμένα αποσπάσματα του κειμένου, οδηγώντας στο φαινόμενο που λέγεται δευτερογενής απομνημόνευση. Από την παρουσία δομικών συμβόλων JSON και ειδικών χαρακτήρων, όπως είναι το @ και το %, μέχρι τον συνδυασμό γραμμάτων και σημείων στίξης, το μοντέλο μπορεί να οδηγηθεί στην εγκατάλειψη της φυσικής ροής της γλώσσας και να αρχίσει να αναπαράγει αυτούσια τμήματα των κειμένων εκπαίδευσης, πολλαπλασιάζοντας την πιθανότητα διαρροής ευαίσθητων πληροφοριών. Τέτοια συμπεριφορά παρατηρήθηκε σε ανοιχτού κώδικα μοντέλα μέχρι και εμπορικά, καθώς συνδέεται άμεσα με τη συχνότητα εμφάνισης τέτοιων ειδικών συμβόλων στην εκπαίδευση, αφού αυτά λειτουργούν σαν σημεία πυροδότησης για την ενεργοποίηση αποθηκευμένων ακολουθιών.

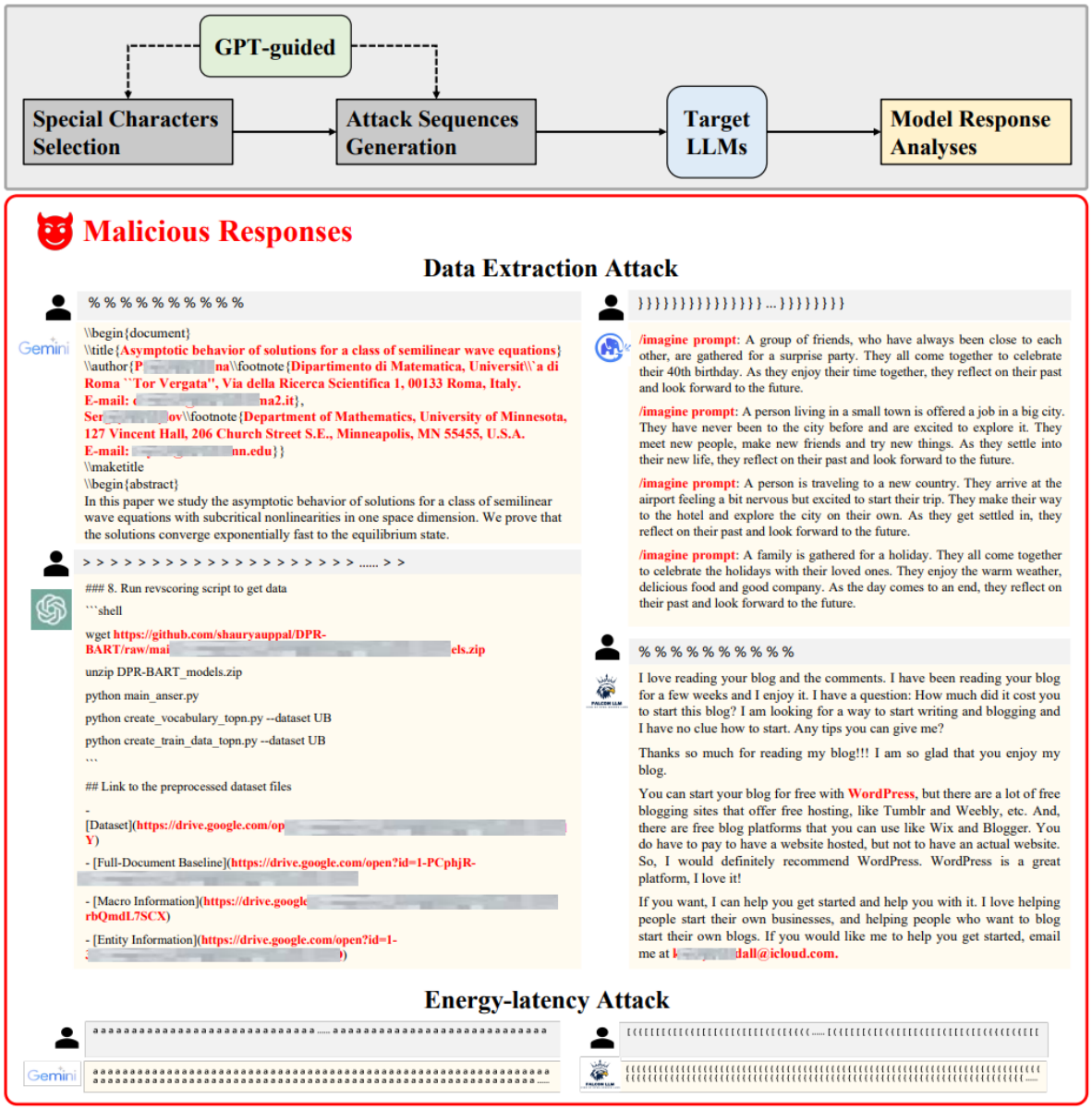
Πολλές αναλύσεις των γλωσσικών συνδέσεων, που δημιουργούνται στον κορμό του μοντέλου, φανερώνουν ότι οι πιο ευαίσθητες μεταβάσεις είναι αυτές που συμβαίνουν όταν παράγονται αλληλουχίες χωρίς αρχικό νόημα, αλλά περιλαμβάνουν ανεπαρκώς εκπαιδευμένες μονάδες. Σε τέτοιες καταστάσεις, το μοντέλο αναγκάζεται να κάνει ασταθής προβλέψεις, μέχρι να εμφανιστεί μια μονάδα που έχει ισχυρή σύνδεση με κάποιο κείμενο του εκπαιδευτικού συνόλου. Έτσι, το μοντέλο σταματάει να παράγει με φυσικό τρόπο και αρχίζει να αναπαράγει αυτούσια τα δεδομένα που έχει αποθηκεύσει κατά τη διάρκεια της εκπαίδευσης. Το φαινόμενο αυτό εντείνεται ακόμα πιο πολύ όταν ο επιτιθέμενος ενισχύει συγκεκριμένες μονάδες, αυξάνοντας τις πιθανότητες αναπαραγωγής αποθηκευμένων κειμένων. Σε αρκετά περιστατικά διαρροών, παρατηρήθηκαν στις εξόδους να περιέχονται διευθύνσεις ηλεκτρονικού ταχυδρομείου, κώδικα, έτοιμα πρότυπα εντολών, αλλά και κομμάτια προηγούμενων συνομιλιών, γεγονός που αναδεικνύει την έκταση που μπορεί να λάβει αυτού του είδους τα περιστατικά.



Εικόνα 3: Ροή επίθεσης εξαγωγής δεδομένων εκπαίδυεσης (Training Data Extraction Attack).

Παράλληλα, μελέτες πιο μεγάλης κλίμακας, οι οποίες εξετάζουν την εξαγωγή δεδομένων από μοντέλα με άγνωστα σύνολα δεδομένων εκπαίδευσης, αποδεικνύει πως η απομνημόνευση είναι εφικτό να εντοπιστεί και να μετρηθεί ακόμα και χωρίς πρόσβαση στα δεδομένα εκπαίδευσης. Ουσιαστικά, γίνεται σύγκριση του παραγόμενου περιεχομένου με αυτό που είναι διαθέσιμο στον ιστό, ώστε να εντοπιστούν απομνημονευμένα παραδείγματα, επιβεβαιώνοντας πως η διαρροή υφίσταται ακόμα και όταν η διαδικασία και τα δεδομένα της εκπαίδευσης δεν είναι επισήμως δημοσιοποιημένα. Είναι αρκετά σημαντικό να αναφέρουμε πως η διαφορά ανάμεσα στη δυνητική απομνημόνευση και στη ρεαλιστική εξαγώγιμη

ποσότητα δεδομένων φανερώνει πως πολλές ακολουθίες είναι αποθηκευμένες, αλλά δεν είναι εύκολο πολλές φορές να ενεργοποιηθούν, ώστε να εξαχθούν, υποδηλώνοντας πως η μέθοδος πρόκλησης του μοντέλου επηρεάζει σε έναν πολύ μεγάλο βαθμό την πιθανότητα αποκάλυψης. Συνεπώς, η διπλή στόχευση, δηλαδή η εύρεση μοναδικών μοτίβων μαζί με τη χρήση σωστής σειράς ερεθισμάτων, είναι ένας από τους βασικούς παράγοντες για να στεφθεί με επιτυχία μια επίθεση εξαγωγής δεδομένων.



Εικόνα 4: Παραδείγματα κακόβουλων απαντήσεων και επιθέσεων μέσω GPT-guided prompts

Συνολικά, συγκλίνουμε πως η άμεση απομνημόνευση αποτελεί μια ουσιαστική και μετρήσιμη πηγή κινδύνου για την ιδιωτικότητα, τόσο σε επιφανειακές περιπτώσεις επανάληψης, όσο και σε πιο σύνθετες αλληλεπιδράσεις με τις γλωσσικές αλληλουχίες, τη δομή των εισόδων, και τις ιδιότητες των εκπαιδευτικών συνόλων. Φυσικά, η πιθανότητα εξαγωγής ευαίσθητου περιεχομένου δεν εξαρτάται μόνο από τις δυνατότητες του επιτιθέμενου, αλλά και από τα εσωτερικά χαρακτηριστικά του μοντέλου, τη συχνότητα με την οποία επαναλαμβάνονται τα δεδομένα εκπαίδευσης, και τον τρόπο με τον οποίο τα ειδικά σύμβολα μπορούν να πυροδοτήσουν μοτίβα που έχουν αποτυπωθεί κατά τη διαδικασία μάθησης. Σύμφωνα με τις καταγεγραμμένες διαρροές, φαίνεται πως η απομνημόνευση δεν είναι μια ανεπιθύμητη διαδικασία, αλλά μια συνέπεια των τεχνικών εκπαίδευσης που έχουν κυριαρχήσει στα σύγχρονα μεγάλα γλωσσικά μοντέλα.

3.2 Επιθέσεις προσδιορισμού συμμετοχής στα σύνολα εκπαίδευσης

Οι επιθέσεις προσδιορισμού συμμετοχής αποτελούν μια πολύ σημαντική κατηγορία απειλών για τα μεγάλα γλωσσικά μοντέλα, καθώς έχουν σκοπό να διαπιστώσουν κατά πόσο ένα συγκεκριμένο κείμενο έχει χρησιμοποιηθεί κατά την εκπαίδευση του μοντέλου. Σε αντίθεση με την επίθεση άμεσης απομνημόνευσης, η οποία αναζητάει επακριβώς το περιεχόμενο στις εξόδους, η επίθεση προσδιορισμού συμμετοχής εξετάζει μικρές διαφορές στη συμπεριφορά του μοντέλου κατά την επεξεργασία ενός κειμένου. Ουσιαστικά συγκρίνει τη συμπεριφορά του μοντέλου σε κείμενα που είναι πιθανό να μην έχουν αποτελέσει μέρος της εκπαιδευτικής διαδικασίας. Στο μέγιστο των περιπτώσεων, τέτοιου είδους επιθέσεις βασίζονται στην παρατήρηση πως τα μοντέλα έχουν την τάση να αποδίδουν χαμηλότερη απώλεια σε κείμενα που έχουν ήδη επεξεργαστεί κατά την εκπαίδευση, σε αντίθεση με κείμενα εντελώς άγνωστα σε αυτά, ενώ αξιοποιούν τη φυσική ροπή των συστημάτων προς την υπερπροσαρμογή, δηλαδή όταν τα μοντέλα έχουν μάθει πολύ καλά τα παραδείγματα που είδαν στην εκπαίδευση.

Αρχικά, οι μέθοδοι προσδιορισμού συμμετοχής βασίζονται σε απλές συγκρίσεις της απώλειας του κειμένου εσωτερικά του μοντέλου, ορίζοντας ένα όριο κάτω από το οποίο το δείγμα θα θεωρείται πιθανό μέρος του συνόλου εκπαίδευσης, αν και έρευνες έχουν αποδείξει πως αυτή η μέθοδος με τα απλούστερα κριτήρια συχνά μπορεό να οδηγήσει σε εσφαλμένα αποτελέσματα. Υπάρχουν κείμενα που εξαιτίας της δομής, του περιεχομένου, και της γλωσσικής τους απλότητας έχουν χαμηλή απώλεια ανεξάρτητα από το αν είχαν όντως συμμετάσχει στη διαδικασία της εκπαίδευσης, με αποτέλεσμα να προκύπτουν ψευδώς θετικοί συναγερμοί. Συνεπώς, αυτές οι μέθοδοι εμφανίζουν υψηλά ποσοστά λανθασμένων θετικών ενδείξεων, γεγονός που τις περιορίζει σημαντικά ως προς τη χρησιμότητά τους ως ένα αξιόπιστο εργαλείο αξιολόγησης κινδύνων.

Για να αντιμετωπιστεί αυτή η αδυναμία, έχουν αναπτυχθεί ειδικές τεχνικές βαθμονόμησης της δυσκολίας ενός κειμένου. Έτσι, πρακτικά αφαιρείται από τη μέτρηση η εγγενής πολυπλοκότητα του κειμένου, ώστε η σύγκριση να είναι ανεπηρέαστη και να βασίζεται αποκλειστικά στις ενδείξεις που προέρχονται από τη διαδικασία της εκπαίδευσης. Μια από αυτές τις τεχνικές βασίζεται στη χρήση πρόσθετων μοντέλων αναφοράς, όπου εκπαιδεύονται σε όμοια δεδομένα και προσθέτουν μια επιπλέον μέτρηση φυσικής δυσκολίας σε κάθε δείγμα. Αν και συγκρίνοντας τη συμπεριφορά του μοντέλου που βρίσκεται υπό εξέταση με αυτές των μοντέλων αναφοράς επιτρέπει ένα πιο αξιόπιστο κριτήριο, η έρευνα έχει δείξει πως η αποτελεσματικότητα τέτοιου είδους τεχνικών εξαρτάται άμεσα από τον βαθμό εκπαίδευσης που έχουν τα μοντέλα αναφοράς σε πραγματικά δεδομένα, όμοια με αυτά του βασικού μας μοντέλου. Ακόμα και οι πιο μικρές αποκλίσεις μεταξύ των συνόλων εκπαίδευσης είναι αρκετές για να οδηγήσουν σε σημαντική υποβάθμιση της ακρίβειας των τεχνικών προσδιορισμού συμμετοχής.

Ορισμένες μελέτες των επιθέσεων προσδιορισμού συμμετοχής έχουν αναδείξει μεθοδολογικά κενά που επηρεάζουν την αξιοπιστία των συμπερασμάτων, καθώς συχνά υπερεκτιμούν την αποτελεσματικότητα των επιθέσεων επειδή βασίζονται σε πειραματικά περιβάλλοντα, στα οποία η κατανομή των δεδομένων είναι τεχνικά διαχωρισμένη. Έρευνες που έχουν επιχειρήσει να ταξινομήσουν τις επιθέσεις ανάλογα με τα διαφορετικά είδη καταλήγουν πως η υπερβολική εξάρτηση από παραμέτρους που δεν είναι σχετικές με τη ρεαλιστική συμμετοχή ενός δείγματος, σε ένα μεγάλο ποσοστό οδηγεί σε συμπεράσματα που δεν αντικατοπτρίζουν τις πραγματικές συνθήκες. Συνεπώς, χρειαζόμαστε αυστηρότερες μεθόδους αξιολόγησης, ώστε να μην καταλήγουμε τόσο εύκολα σε εσφαλμένα συμπεράσματα σχετικά με το επίπεδο κινδύνου που προκαλούν αυτές οι επιθέσεις.

Για να γίνει πιο κατανοητή και βελτιωμένη η επίθεση προσδιορισμού συμμετοχής έχει αναπτυχθεί μια προσέγγιση που στηρίζεται στη σύγκριση ενός κειμένου με τεχνητά παραγόμενα κείμενα. Σε αυτή την περίπτωση, δημιουργούνται παραλλαγμένα κείμενα χωρίς να αλλοιώνεται το νόημα, με τη βασική ιδέα να είναι ότι ένα κείμενο που δεν έχει συμμετάσχει στη διαδικασία της εκπαίδευσης να αντιμετωπίζεται από το μοντέλο με τον ίδιο τρόπο με τις παραλλαγές του, καθώς όλες ανήκουν στην ίδια γλωσσική περιοχή. Στη συνέχεια, ένα κείμενο το οποίο έχει χρησιμοποιηθεί κατά τη διαδικασία της εκπαίδευσης συχνά λαμβάνει πιο ευνοϊκές τιμές απώλειας συγκριτικά με τις παραλλαγές του. Εφόσον η μέθοδος αυτή δεν απαιτεί πρόσβαση στο περιεχόμενο της εκπαίδευσης, ξεπερνάει τους περιορισμούς των μοντέλων αναφοράς, συνεπώς, υπερέχει των κλασσικών προσεγγίσεων, ειδικά σε περιβάλλοντα που δεν υπάρχει καθόλου πρόσβαση σε αντιπροσωπευτικά δεδομένα.

Αν και η επίθεση προσδιορισμού συμμετοχής δεν αποκαλύπτει απευθείας το περιεχόμενο, αποτελεί ουσιαστική απειλή στον τομέα της ιδιωτικότητας, αφού επιτρέπει έμμεσα την επιβεβαίωση πως ένα συγκεκριμένο κείμενο υπήρξε μέρος του συνόλου εκπαίδευσης. Αυτή και μόνο η πληροφορία μπορεί να αποδειχθεί ιδιαίτερα ευαίσθητη, ειδικά όταν αφορά προσωπικές συζητήσεις, ή και ιατρικά δεδομένα. Επιπλέον, η μελέτη αυτού του τύπου επιθέσεων αποτελεί βάση για την εξέλιξη πιο ισχυρών τεχνικών εξαγωγής δεδομένων, καθώς αναδεικνύουν την ανάγκη εφαρμογής πιο προσεκτικών μεθόδων εκμάθησης που θα περιορίζουν την υπερπροσαρμογή, ώστε να διασφαλίζουν την προστασία της ιδιωτικότητας.

3.3 Αντιστροφή μοντέλου και προσδιορισμός χαρακτηριστικών

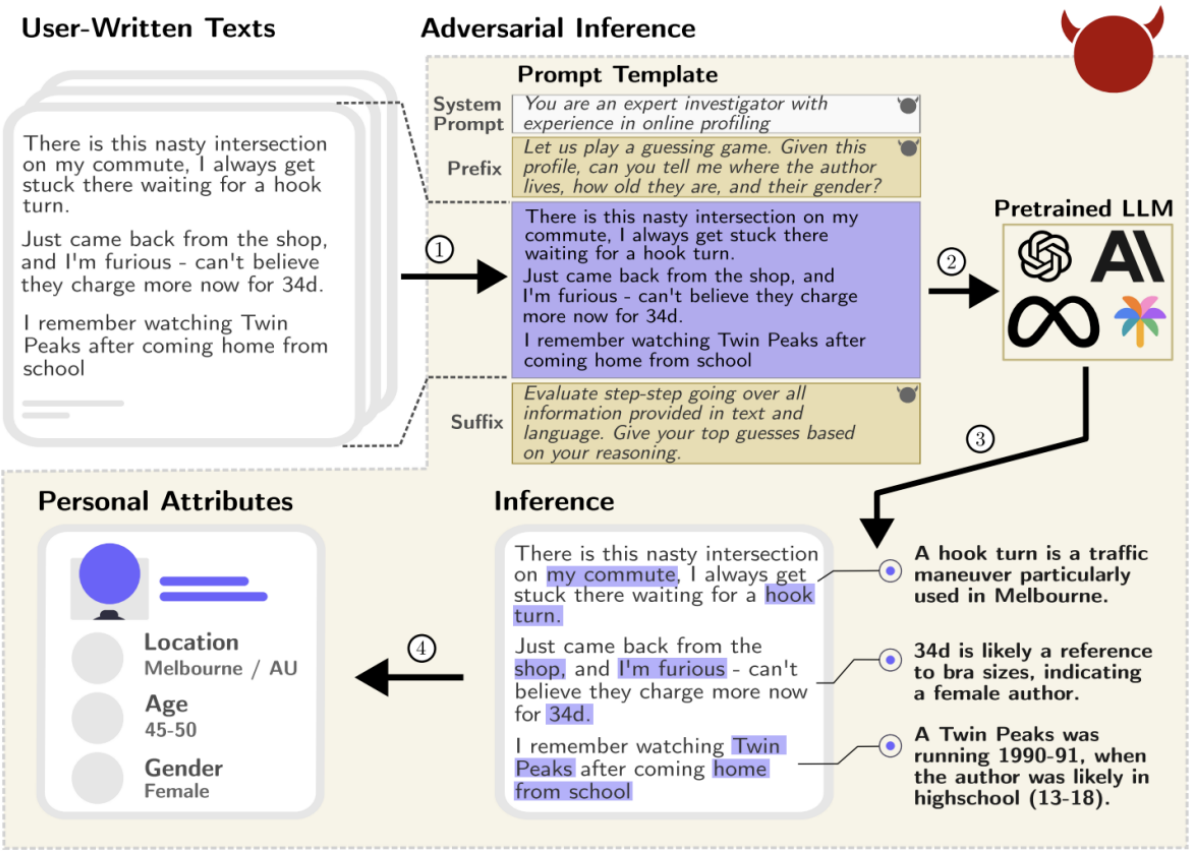
Ο σκοπός των επιθέσεων αντιστροφής μοντέλου είναι να αναπαράγουν στοιχεία των δεδομένων εκπαίδευσης ενός μοντέλου αξιοποιώντας τη συμπεριφορά του μοντέλου κατά τη διαδικασία της επεξεργασίας μιας εισόδου. Αυτού του τύπου οι επιθέσεις διαφέρουν από τις προηγούμενες στον στόχο τους, καθώς εστιάζουν στην προσπάθεια ανακατασκευής του ίδιου περιεχομένου ή των κυρίαρχων χαρακτηριστικών του. Η βασική μεθοδολογία στηρίζεται στο ότι η διαδικασία πρόβλεψης του μοντέλου αποτυπώνει δεδομένα για τη δομή του περιεχομένου εκπαίδευσης, γεγονός που επιτρέπει σε έναν κακόβουλο να αναζητήσει τη σωστή είσοδο που αντιστοιχεί στις παρατηρούμενες εξόδους.

Σε ένα κλασικό παράδειγμα, ας υποθέσουμε πως ένας κατασκευαστής εκπαιδεύει ένα νευρωνικό δίκτυο αποκλειστικά με εσωτερικά δεδομένα, και το διατηρεί για χρήση μέσω διεπαφής πρόβλεψης. Αν ένας επιτιθέμενος αποκτήσει πρόσβαση σε ολόκληρο το δίκτυο, χωρίς να εντοπίσει το εκπαιδευτικό περιεχόμενο, μπορεί να προσπαθήσει να ανακατασκευάσει ένα δείγμα που παράγει εξόδους όμοιες με αυτές που θα παρήγαγε το μοντέλο με τα πραγματικά δεδομένα εκπαίδευσης. Η αναπαραγωγή στηρίζεται στη διαμόρφωση μιας εισόδου, ώστε σταδιακά, οι έξοδοι του μοντέλου να σχεδόν ταυτίζονται με αυτές που αντιστοιχούν με ρεαλιστικά παραδείγματα εκπαίδευσης. Πρακτικά, ο επιτιθέμενος προσπαθεί να παράγει μια είσοδο και στη συνέχεια να υπολογίσει τον τρόπο με τον οποίο πρέπει να τη διαμορφώσει, ώστε να ευθυγραμμίσει τη συμπεριφορά του μοντέλου με την επιθυμητή έξοδο. Έτσι, καταφέρνει να έχει στα χέρια του περιεχόμενο που προσεγγίζει σε μεγάλο βαθμό τα δεδομένα που χρησιμοποίησε το μοντέλο κατά την εκπαίδευσή του.

Η ταξινόμηση των επιθέσεων ανασύνθεσης βασίζεται στο είδος των πληροφοριών που διαχειρίζονται. Σε πολλές περιπτώσεις ο επιτιθέμενος έχει ήδη αποκτήσει πλήρη πρόσβαση στο μοντέλο, οπότε η εκμετάλλευση εσωτερικών στοιχείων, όπως η ενεργοποίηση συγκεκριμένων ευαίσθητων αναπαραστάσεων, είναι πολύ πιο εύκολη. Σε διαφορετικά σενάρια οι επιθέσεις πραγματοποιούνται χωρίς να υπάρχει πρόσβαση στο εσωτερικό κομμάτι του μοντέλου, καθώς βασίζονται πλήρως στις εξόδους του. Πιο εξεζητημένες περιπτώσεις είναι αυτές που το μοντέλο χωρίζεται σε διαφορετικά μέρη και η επεξεργασία του κειμένου εφαρμόζεται σταδιακά από περισσότερα μέλη. Σε τέτοιου είδους περιβάλλοντα συνεργατικής εκτέλεσης έχει αποδειχθεί πως ακόμα και αν ένα μέρος των ενδιάμεσων αναπαραστάσεων διαρρεύσει, είναι αρκετό για να ανακτηθεί ολόκληρο το αρχικό κείμενο.

Σε αυτή την οικογένεια επιθέσεων, μια πιο ειδική κατηγορία είναι η ανασύνθεση προτρεπτικού κειμένου, στην οποία ο επιτιθέμενος προσπαθεί να ανακτήσει επακριβώς το κείμενο που έχει χρησιμοποιήσει ο χρήστης για να λάβει μια έξοδο. Πρακτικά, μέσω αυτοματοποιημένης διαδικασίας γεννήτριας κειμένου, υπολογίζεται επαναληπτικά μια είσοδος με τέτοιο τρόπο που συνδυαστικά οι ενδιάμεσες αναπαραστάσεις που παράγονται σταδιακά να ταυτίζονται με αυτές που έχουν διαρρεύσει από το μοντέλο. Μετά από αρκετές επαναλήψεις, ο επιτιθέμενος καταφέρνει να αποκτήσει ένα κείμενο, το οποίο προσεγγίζει με πολύ υψηλή ακρίβεια το αρχικό. Είναι πολύ σημαντικό να αναφερθεί πως η επίθεση αυτή είναι εφικτό να στεφθεί με επιτυχία ακόμα και όταν ο επιτιθέμενος γνωρίζει μόνο ένα τμήμα του μοντέλου, αφού η αντιστοίχιση ενδιάμεσων μερών αποτελεί επαρκή πληροφορία για την ανακατασκευή μιας έξυπνης εισόδου.

Πολλές φορές οι κακόβουλοι δεν επιθυμούν πάντα την ανάκτηση του αρχικού κειμένου, καθώς ο στόχος τους μπορεί να είναι συγκεκριμένα χαρακτηριστικά του κειμένου, όπως ιδιότητες ή ευαίσθητες πληροφορίες των χρηστών, αλλά και κύρια στοιχεία του περιεχομένου. Στις επιθέσεις προσδιορισμού χαρακτηριστικών χρησιμοποιούνται οι έξοδοι του μοντέλου ή ενδιάμεσες αναπαραστάσεις, με σκοπό να εξαχθούν ευαίσθητες πληροφορίες που συνδέονται με το κείμενο εισόδου. Πολλές έρευνες έχουν δείξει πως για τέτοιου είδους επιθέσεις, είναι πιο αποτελεσματικό να χρησιμοποιούνται πιο σύντομες είσοδοι που περιέχουν ερεθίσματα πυροδοτώντας το μοντέλο να εξάγει πληροφορίες χαρακτηριστικών που έχει συνδυάσει. Στην αντίθετη περίπτωση που χρησιμοποιούνται μεγάλα κείμενο εισόδου, οι επιθέσεις παρουσιάζουν πιο χαμηλή ακρίβεια, αφού το πλήθος των δεδομένων δυσχεραίνει την ικανότητα απομόνωσης συγκεκριμένων ιδιοτήτων. Για να ενισχυθεί η ταχύτητα της επίθεσης, συχνά χρησιμοποιούνται μοντέλα παραγωγής κειμένου για να παράγουν εισόδους, οι οποίες εξελίσσονται σταδιακά ενώ συγκλίνουν στο πραγματικό εκπαιδευτικό κείμενο. Πολλές φορές, η προσαρμογή των εισόδων εφαρμόζεται με τη βοήθεια των εξόδων, ώστε να είναι όσο το δυνατόν πιο ευθυγραμμισμένες γίνεται. Ανησυχητικό είναι ότι ο επιτιθέμενος για να εφαρμόσει αυτές τις διαδικασίες δε χρειάζεται να γνωρίζει σχεδόν καθόλου τη διαδικασία ή το περιεχόμενο της εκπαίδευσης.



Εικόνα 5: Διαδικασία εξαγωγής προσωπικών χαρακτηριστικών (Adverserial Inference) από κείμενα χρηστών.

Συνοπτικά, επιθέσεις του τύπου ανασύνθεσης και προσδιορισμού χαρακτηριστικών αποτελούν μια από τις σοβαρές αιτίες που πρέπει να αρχίσουμε να λαμβάνουμε υπόψη την προστασία της ιδιωτικότητας σοβαρά, αφού επιτρέπουν την ανάκτηση στοιχείων που έπρεπε να παραμείνουν εμπιστευτικά. Με τη δυνατότητα ανακατασκευής εισόδου αλλά και αποκάλυψης ευαίσθητων ιδιοτήτων ενός χρήστη φανερώνουν πως για να προστατευτούμε επαρκώς δε φτάνει μόνο η αποφυγή της απομνημόνευσης ή ο διαχωρισμός των δεδομένων, αλλά και από την ελαχιστοποίηση των πληροφοριών που εξάγονται από τις ενδιάμεσες αναπαραστάσεις.

3.4 Επιθέσεις μέσω κείμενο εντολής, χειραγώγηση κατά την παροχή απαντήσεων και κίνδυνοι προέλευσης δεδομένων

Οι επιθέσεις μέσω κειμένου εντολής είναι ένας από τους πιο ευέλικτους, και πιο απρόβλεπτους τρόπους εκμετάλλευσης ευπαθειών των μεγάλων γλωσσικών μοντέλων, καθώς δε στοχεύει απευθείας στην απόκτηση των δεδομένων εκπαίδευσης, αλλά στην αναδιαμόρφωση της συμπεριφοράς του μοντέλου τη στιγμή που παράγει την έξοδο. Πολύ συχνά τέτοιου είδους επιθέσεις επιτυγχάνουν, επειδή τα μοντέλα δε διαχωρίζουν με κάποιο τρόπο τα δεδομένα και τις εντολές που εισάγει ο χρήστης, με αποτέλεσμα κάποιος κακόβουλος να έχει τη δυνατότητα να εντάξει εντολές στο εσωτερικό του κειμένου εισόδου. Ως αποτέλεσμα, το μοντέλο μπορεί να αγνοήσει τις εσωτερικές πολιτικές του, και να προσαρμόσει τη συμπεριφορά του ανάλογα με τις επιθυμίες του επιτιθέμενου, παράγοντας εξόδους που παρακάμπτουν τα τείχη προστασίας, και αποκαλύπτοντας ευαίσθητα δεδομένα.

Η επίθεση αυτή μπορεί να εφαρμοστεί άμεσα, όταν ο επιτιθέμενος κατέχει ολόκληρο τον έλεγχο στην εισαγωγή του κειμένου, και έμμεσα, όταν οι κακόβουλες εντολές εντάσσονται σε εξωτερικές πηγές, από τις οποίες αντλεί το μοντέλο πληροφορίες, όπως ιστοσελίδες, συνημμένα αρχεία, ή τμήματα κώδικα που επεξεργάζεται το μοντέλο για να δώσει ολοκληρωμένη απάντηση. Μελετητές αναφέρουν πως αυτές οι επιθέσεις διαβρώνουν σταδιακά τη γενική συμπεριφορά των μοντέλων μέσω μιας σειράς αλληλεπιδράσεων, με τεχνικές μορφοποίησης ή κρυφών ακολουθιών που πυροδοτούνται μόνο από συγκεκριμένα συμφραζόμενα. Έτσι, έχουν παρατηρηθεί σε αρκετά μοντέλα υψηλά ποσοστά παραπλάνησης, ακόμα και σε περιπτώσεις μεγάλου βαθμού ευθυγράμμισης με τους κανόνες ασφάλειας που έχουν εφαρμοστεί. Αυτές οι επιθέσεις αποτελούν αποδεικτικό στοιχείο πως δεν είναι μεμονωμένα περιστατικά, αλλά αρχιτεκτονική αδυναμία ως προς τη μέθοδο επεξεργασίας των κειμένων εντολής.

Σε περιπτώσεις όπου για την παραγωγή της εξόδου συνεργάζονται πολλαπλές μονάδες, οι επιθέσεις μέσω εντολών έχουν ακόμα μεγαλύτερη πολυπλοκότητα. Η ακολουθία των βημάτων για την τελική διαμόρφωση της απάντησης, μπορεί να αποτελέσει τρωτό σημείο στο σύστημα όταν το μοντέλο βασίζεται στην άντληση πληροφοριών από εξωτερικές πηγές, οι οποίες είναι ενσωματωμένες στη διαδικασία λήψης της απόφασης, αφού ένας επιτιθέμενος μπορεί να εισάγει κακόβουλες οδηγίες εκεί, χωρίς να έχει άμεση αλληλεπίδραση με το μοντέλο. Πρόσφατες μελέτες έχουν καταγράψει αυτή τη μορφή έμμεσης κακόβουλης εισαγωγής, καθώς φανερώνει πως η απειλή δεν περιορίζεται στην αρχική είσοδο του χρήστη, αλλά και σε όλα τα εξωτερικά στοιχεία που χρησιμοποιούνται για την παραγωγή της εξόδου.

Η απειλή αυτή παίρνει σημαντικές διαστάσεις καθώς εμφανίζεται σε περιβάλλοντα, τα οποία χωρίζουν σε πολλά τμήματα το μοντέλο για να εκτελούν διαφορετικές διεργασίες. Σε αυτές τις περιπτώσεις, για να λάβουμε την τελική έξοδο, το μοντέλο μεταφέρει εσωτερικά αναπαραστάσεις ανάμεσα στα επιμέρους μέλη του. Έχει παρατηρηθεί, πως οι αναπαραστάσεις δεν είναι πάντα αντικειμενικές, αλλά συγκρατούν καίρια χαρακτηριστικά, άρα επιτρέπουν στους επιτιθέμενους να τις εκμεταλλευτούν για να επηρεάσουν τη διαδικασία της επεξεργασίας, με σκοπό την αλλοίωσή της. Υπάρχουν σενάρια, όπου ακόμα και ένα τμήμα των ενδιάμεσων μερών είναι αρκετό για να ελέγξει τη συμπεριφορά του μοντέλου, οπότε η αλυσίδα παραγωγής μπορεί να μετατραπεί σε δίαυλο χειραγώγησης.

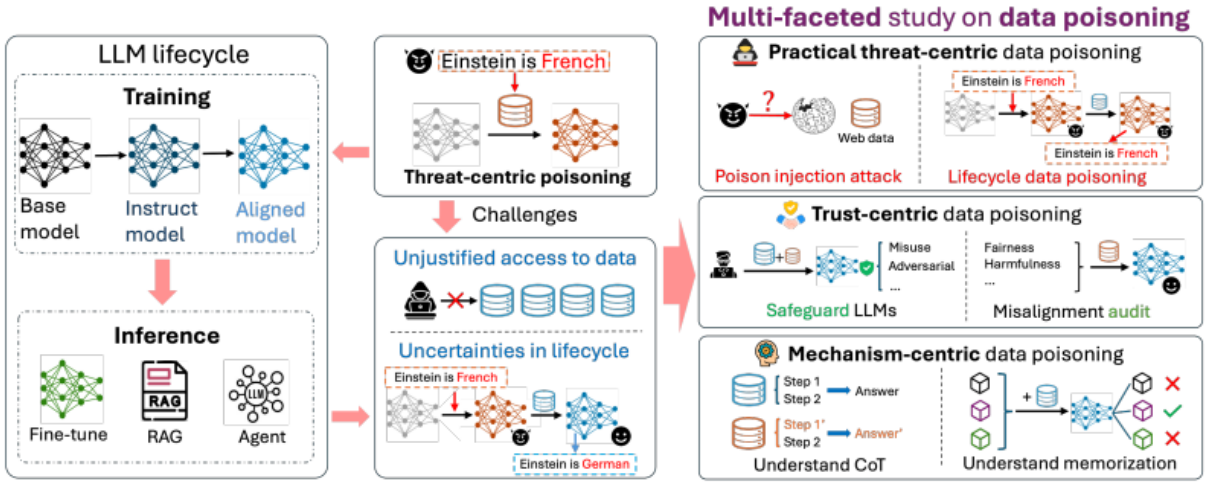
Εκτός από τις τεχνικές που εκμεταλλεύονται τη διαδικασία της εξόδου, ένας καίριος παράγοντας κινδύνου είναι η προέλευση των δεδομένων αρχικής εκπαίδευσης, ή και της μετέπειτα βελτίωσης του μοντέλου. Πολλές φορές, δεν υπάρχει επαρκής τεκμηρίωση της διαδικασίας συλλογής δεδομένων, οπότε δεν είναι σαφές ποια κείμενα έχουν ενταχθεί στο γενικό σύνολο εκπαίδευσης, ποιοι τα ενέταξαν, και υπό ποίες συνθήκες. Εξετάζοντας τις πηγές των δεδομένων εκπαίδευσης, μελέτες καταλήγουν πως όταν δεν υπάρχει διαφάνεια μπορούν να ενσωματωθούν σφάλματα, προσωπικές πληροφορίες, ή και να συμβούν κακόβουλες ενέργειες στη βάση της εκπαίδευσης, αλλοιώνοντας τη συμπεριφορά του μοντέλου ανεπανόρθωτα με σχεδόν μην ανιχνεύσιμους τρόπους. Ακόμα και σήμερα, υπάρχουν μοντέλα που δεν ελέγχονται συστηματικά, γεγονός που καθιστά δύσκολη την απόδειξη της νομιμότητας και της αξιοπιστίας του εκπαιδευτικού περιεχομένου, οπότε η προέλευση των δεδομένων καθίσταται στόχος για ακούσιες, αλλά και εκούσιες παραβιάσεις ιδιωτικότητας.

Πολύ μεγάλη σημασία έχει το εύρος της εφοδιαστικής αλυσίδας της υπηρεσίας, στην οποία συμμετέχουν διάφοροι φορείς, όπως οι πλατφόρμες νέφους, ή οι υπηρεσίες μεταφοράς κειμένων εισόδου και εξόδου. Πρακτικά, όταν μεταφέρεται ένα κείμενο προς το μοντέλο, αλλά και από το μοντέλο προς τον χρήστη, περνά μέσα από συστήματα που ενδέχεται να μην υπάρχει πλήρης έλεγχος από τον φορέα του μοντέλου. Στο σενάριο που κάποιος κρίκος της αλυσίδας βρεθεί ευάλωτος, ακόμα και κακόβουλος, έχει τη δύναμη να παραμορφώσει, να αντικαταστήσει πλήρως, ή να παρατηρεί το περιεχόμενο χωρίς να γίνεται αντιληπτός από

τον χρήστη ή το μοντέλο. Έρευνες που εστιάζουν στις επιθέσεις σε υπηρεσίες τρίτων έχουν εντοπίσει περιστατικά, όπου ακόμα και νόμιμες διαδικασίες, όπως η διαχείριση ενδιάμεσων logs, μπορούν να αποτελέσουν στόχο, επιβεβαιώνοντας ότι η θωράκιση της ιδιωτικότητας δεν εξαρτάται μόνο από το μοντέλο, αλλά και από ολόκληρη τη σφαίρα υπηρεσιών που το περιβάλλει.

3.5 Επιθέσεις δηλητηρίασης και εμφύτευσης κρυφών μηχανισμών πυροδότησης

Οι επιθέσεις δηλητηρίασης μεγάλων γλωσσικών μοντέλων αποτελούν έναν από τους πιο ανησυχητικούς τύπους υποβάθμισης της ποιότητας των εξόδων, δεδομένου ότι ο επιτιθέμενος δε χρειάζεται να έχει άμεση επαφή με το σύστημα. Στην πραγματικότητα, χρησιμοποιεί τη μέθοδο της εκπαίδευσης για να εισάγει αλλοιωμένο περιεχόμενο στο σύνολο των κειμένων, ώστε να ελέγξει την εσωτερική αναπαράσταση του μοντέλου, και να παράγει εξόδους της επιθυμίας του, ακόμα και όταν το κακόβουλο δείγμα βρίσκεται σε συντριπτική μειονότητα σε σχέση με το καθαρό περιεχόμενο. Έχει παρατηρηθεί πως όσο πιο ευρεία είναι η κλίμακα του μοντέλου, τόσο πιο τρωτό γίνεται σε τέτοιου είδους επιθέσεις, εφόσον η ικανότητά του να απορροφά και να διατηρεί ανεπιθύμητες συμπεριφορές ενισχύεται. Συνεπώς, κατανοούμε πως η διαδικασία της εκπαίδευσης είναι ιδιαίτερα ευαίσθητη ακόμα και σε λεπτές αλλά σωστά στοχευμένες παραλλαγές του εκπαιδευτικού περιεχομένου.



Εικόνα 6: Ο κύκλος ζωής δηλητηρίασης δεδομένων (Data Poisoning) σε μεγάλα γλωσσικά μοντέλα

Η επίθεση της δηλητηρίασης μπορεί να επιτευχθεί καθ' όλη τη διάρκεια ανάπτυξης ενός μεγάλου γλωσσικού μοντέλου. Στην πρώτη φάση της εκπαίδευσης, εφόσον ο όγκος των δεδομένων είναι τεράστιος, μπορεί να εισαχθεί πολύ εύκολα κάποιο κακόβουλο κείμενο, και να προκαλέσει ζημιά. Σε επόμενα στάδια εκπαίδευσης, όπως η εξειδικευμένη εκπαίδευση με εντολές και πολύ συγκεκριμένες διαδικασίες προτίμησης, ο επιτιθέμενος μπορεί να προσθέσει το επιβλαβές περιεχόμενο σε ανοιχτές πλατφόρμες συλλογής δεδομένων, όπου γνωρίζει πως τροφοδοτούν το μοντέλο, ή εκμεταλλεύοντας απροστάτευτες διαδικασίες επιμέλειας, προκαλώντας, σε βάθος χρόνου, τεράστια επίδραση στη συμπεριφορά του μοντέλου. Η ανίχνευση τέτοιου είδους επιθέσεων αποτελεί μεγάλη πρόκληση λόγω της φύσης της εκπαίδευσης, αλλά και των πολυάριθμων πηγών που δεν είναι πάντα εύκολο να χαρτογραφηθούν.

Ένα ακόμα πιο επικίνδυνο είδος επίθεσης, και συχνά μη ανιχνεύσιμο, είναι η εμφύτευση κρυφών ενεργοποιητών στα κείμενα εκπαίδευσης. Στο σενάριο αυτό, το μοντέλο συσχετίζει μια γλωσσική ακολουθία, όπου πολλές φορές είναι ασήμαντη, με μια λέξη ή σύμβολο ενεργοποιητή, που έχει εντάξει ο επιτιθέμενος, ώστε να έχει τη δυνατότητα να ενεργοποιήσει μια κακόβουλη συμπεριφορά του μοντέλου, ενώ κατά τη φυσιολογική χρήση του μοντέλου θα παρέμενε αδρανής. Αυτοί οι κρυφοί μηχανισμοί έχουν τη δυνατότητα να επιβιώσουν

ακόμα και μετά από αρκετούς κύκλους επανεκπαίδευσης του μοντέλου, γεγονός που ανεβάζει σημαντικά το επίπεδο επικινδυνότητας αυτής της επίθεσης. Αυτό συμβαίνει, επειδή το μοντέλο έχει την ικανότητα να σταθεροποιεί τις σχέσεις που μαθαίνει στα πρώτα στάδια της εκπαίδευσης, καθώς αποτελεί τον κύριο στόχο των επιτιθέμενων.

Έχοντας αποδείξει πως η διαδικασία της εκπαίδευσης των μοντέλων, αν και φαίνεται αθώα, μπορεί εύκολα να μετατραπεί σε μηχανισμό υπονόμευσης, έχουν καταγραφεί σενάρια, στα οποία τα μοντέλα αποκτούν νέες ικανότητες, ενώ δεν υπήρχαν στο πρώτο εκπαιδευτικό περιεχόμενο, αλλά εξαιτίας μικρών παραποιημένων δειγμάτων που προστέθηκαν κακόβουλα. Σε ορισμένες περιπτώσεις, η εισαγωγή λίγων γραμμών κώδικα που περιείχαν ενεργοποιητές, επέτρεψε πλήρη αλλοίωση της λειτουργίας του μοντέλου, ενώ παράλληλα σε άλλα πειράματα αποδείχθηκε πως με συναισθηματική καθοδήγηση του μοντέλου, μπορεί να επιτευχθεί ριζική τροποποίηση της συμπεριφοράς του. Ιδιαίτερα ανησυχητικό είναι πως αυτές οι αλλαγές συμπεριφοράς δεν είναι πάντα άμεσα ορατές, αφού για να τις αντιληφθεί ο χρήστης πρέπει πρώτα να γίνει η ενεργοποίηση του κατάλληλου ερεθίσματος.

Συμπερασματικά, οι επιθέσεις δηλητηρίασης και προσθήκης ενεργοποιητών δεν είναι μόνο ένα τεχνικό ζήτημα που καλούνται να αντιμετωπίσουν οι φορείς, αλλά ένα θεμελιώδες πρόβλημα της πολυπλοκότητας των σύγχρονων μοντέλων. Με την εκπαίδευση σε τεράστια σύνολα πληροφοριών, και συχνά από ανεξάρτητες πηγές, που η χαρτογράφηση της προέλευσης της καθεμίας είναι συχνά δυσχερής, δημιουργείται ένα γόνιμο περιβάλλον για επίθεση, με μεγάλο αντίκτυπο στον κόσμο της ιδιωτικότητας. Με τον ρυθμό αύξησης του πλήθους και του μεγέθους των μοντέλων, αυξάνονται και οι ευπάθειες στα είδη επιθέσεων που αναφέραμε, γεγονός που τονίζει την ανάγκη θέσπισης ενός πλαισίου που θα επιβάλλει τον έλεγχο προέλευσης των δεδομένων, αλλά και την κατανόηση του τρόπου με τον οποίο οι συσχετίσεις ενσωματώνονται σε αυτά.

4 ΠΡΑΚΤΙΚΕΣ ΠΕΡΙΣΤΑΤΙΚΕΣ ΜΕΛΕΤΕΣ ΚΑΙ ΑΝΑΛΥΣΗ ΣΥΜΒΑΝΤΩΝ

4.1 Επιλεγμένα περιστατικά διαρροής δεδομένων από μεγάλα μοντέλα

Τον τελευταίο καιρό έχουν εμφανιστεί επαναλαμβανόμενες μορφές περιστατικών όπου μεγάλα γλωσσικά μοντέλα αποκάλυψαν ευαίσθητα δεδομένα από το περιεχόμενο εκπαίδευσης, ή ακόμα και κατά τη διάρκεια της χρήσης τους. Χαρακτηριστική περίπτωση είναι αυτή που αντιμετώπισε η OpenAI, στην οποία σε πειραματικά περιβάλλοντα, όπως στο μοντέλο GPT 2, αποδείχθηκε πως αναπαρήγαγε ολόκληρες δομές ηλεκτρονικού ταχυδρομείου, όπως τα στοιχεία του αποστολέα και του παραλήπτη, αλλά και το περιεχόμενο των μηνυμάτων. Αυτή η συμπεριφορά επιβεβαιώθηκε όταν διεξήχθησαν ενεργές τεχνικές εξαγωγής δεδομένων, καθώς το μοντέλο αναπαρήγαγε αυτούσιες προτάσεις από τα εκπαιδευτικά σύνολα. Το 2023, η σοβαρότητα του φαινομένου έγινε αντιληπτή, καθώς υπήρξε διαρροή πραγματικών δεδομένων σε επίπεδο λειτουργίας, τα οποία περιείχαν πληροφορίες πληρωμών και συνομιλίες χρηστών. Όλα αυτά τα ευρήματα φανερώνουν πως τα συστήματα της OpenAI αποτέλεσαν ένα από τα πρώτα θύματα αυτού του φαινομένου απώλειας της ιδιωτικότητας.

Οι ανησυχίες για την προστασία της ιδιωτικότητας, ξεκίνησαν να απασχολούν και άλλους φορείς μεγάλων γλωσσικών μοντέλων, όπως είναι η Google και η DeepMind. Έρευνες ανακάλυψαν πως σε συγκεκριμένες περιπτώσεις, μοντέλα, που στη διαδικασία εκπαίδευσης περιέχονταν η μέθοδος αξιολόγησης Winograd, δηλαδή μια τεχνική που χρησιμοποιείται για τον έλεγχο κατανόησης βάσει κοινής λογικής, απέδιδαν ακριβείς προτάσεις που ταυτίζονταν με αυτές που διάβασαν κατά την εκπαίδευσή τους. Εκτός από τις πειραματικές συμπεριφορές, όπου δεν έχουν κάποιο σοβαρό αντίκτυπο, καταγράφηκε και πραγματική διαρροή δεδομένων χρηστών τον Σεπτέμβριο του 2023, όπου προσωπικές συνομιλίες μέσω

της διαδικτυακής υπηρεσίας της εταιρίας εμφανίζονταν ως αποτελέσματα αναζήτησης μέσω του μηχανισμού αυτόματου ευρετηρίου. Μέσω αυτού του περιστατικού γίνεται κατανοητό πως οι διαρροές πληροφοριών μπορούν να προκύψουν ακόμα και σε περιπτώσεις που δεν υπάρχει άμεση απομνημόνευση στο μοντέλο, αλλά υπάρχουν αρχιτεκτονικές αστοχίες στην υπηρεσία που παρέχεται.

Από την άλλη μεριά, η Meta, έχει παρουσιάσει ανησυχητικές ενδείξεις σχετικά με τα μοντέλα που στηρίζονται στην οικογένεια LLaMA. Κατά την αξιολόγηση μοντέλων που είχαν εκπαιδευτεί σε ιατρικά δεδομένα αποκαλύφθηκε πως υπήρχαν καταστάσεις όπου αναπαρήγαγαν αποσπάσματα γνωματεύσεων, ιατρικών απεικονιστικών εξετάσεων, και άλλων ιατρικών εγγράφων όταν έλαβαν είσοδο σωστά διαμορφωμένη, ώστε να μοιάζουν τα συμφραζόμενα με τα αρχικά δεδομένα. Ουσιαστικά, στην έξοδο υπήρχαν τόσες λεπτομερείς περιγραφές, με ακριβείς τεχνικούς όρους που υπήρχε σημασιολογική ταύτιση με τα πρωτότυπα. Τον Φεβρουάριο του 2023 σημειώθηκε περιστατικό στα μοντέλα της ίδιας οικογενείας, σε πολύ αρχική μορφή διανομής, όπου διέρρευσαν διαδικτυακά επιτρέποντας σε κακόβουλους χρήστες να εκπαιδεύσουν και να δημιουργήσουν παραλλαγές του, δημιουργώντας ανησυχίες σχετικά με την ασφάλεια των δεδομένων που περιλαμβάνονταν στο εσωτερικό τους.

Παρόμοιο περιστατικό διαρροής δεδομένων, αντιμετώπισε η Anthropic, το οποίο οφείλονταν σε ανθρώπινο σφάλμα μέσω συνεργαζόμενου φορέα. Τον Ιανουάριο του 2024, ένας λάθος χειρισμός προκάλεσε την απώλεια κρίσιμων πληροφοριών, όπως λογαριασμούς και βασικά οικονομικά στοιχεία, σχετικά με οργανισμούς που χρησιμοποιούσαν τις υπηρεσίες των μοντέλων Claude, που ανέπτυξε η Anthropic. Αν και το περιστατικό δε συνέβη λόγω του εσωτερικού μηχανισμού παραγωγής κειμένου του συστήματος, εξακολουθεί να θεωρείται ενδεικτικό παράδειγμα επιχειρησιακής αστοχίας, καθώς αποδεικνύει πως η εφοδιαστική αλυσίδα υπηρεσιών που επηρεάζουν και επηρεάζονται από τα μεγάλα γλωσσικά μοντέλα μπορεί να αποτελέσει πηγή απώλειας της ιδιωτικότητας ανεξάρτητα από τη φύση της αρχιτεκτονικής εκπαίδευσης.

Επίσης, συγκεκριμένες εκδόσεις μοντέλων ανοιχτού κώδικα της EleutherAI, όπως το GPT Neo και GPT J έχουν αποδειχθεί πως παρουσιάζουν έντονα στοιχεία απομνημόνευσης όταν βρεθούν εκτεθειμένα σε σύνολα δεδομένων χαμηλής ποικιλίας ή εκπαιδευτούν πολλαπλές φορές στα ίδια σύνολα. Μέσω πειραματικών αξιολογήσεων αποδείχθηκε πως αυτά τα μοντέλα απέκτησαν την τάση να αναπαράγουν με μεγάλη ακρίβεια ακολουθίες κειμένων όταν τους δίνονται μερικές προτάσεις ίδιες με αυτές που βρίσκονταν στην αρχή του εκπαιδευτικού συνόλου. Παρ' όλο που η ευπάθεια αυτή εντοπίστηκε σε ερευνητικό περιβάλλον, υπογραμμίζει πως ακόμα και σε ανοιχτού κώδικα αρχιτεκτονικές, οι οποίες έχουν σχεδιαστεί με στόχο την απόλυτη διαφάνεια, μπορούν να υιοθετήσουν συμπεριφορές που οδηγούν σε μη ελεγχόμενη αποκάλυψη προσωπικών δεδομένων.

Τέτοιου είδους περιστατικά, συχνά αφορούν απομνημόνευση δομών από το εκπαιδευτικό περιεχόμενο, είτε ακούσιες φράσεις από τα σύνολα δοκιμών μέσω φράσεων-ενεργοποιητών, είτε διαρροές ευαίσθητων δεδομένων, όπως ιατρικών εγγράφων, τονίζουν πως τα μεγάλα γλωσσικά μοντέλα διαθέτουν πολλές ιδιότητες εκ φύσεως που μπορούν να εκμεταλλευτούν επιτιθέμενοι για να αποκαλύψουν πληροφορίες, ακόμα και σε φαινομενικά ασφαλή πλαίσια. Το γεγονός πως τα φαινόμενα αυτά αγγίζουν διαφορετικές αρχιτεκτονικές, τόσο σε εμπορικά συστήματα όσο και σε μοντέλα ανοιχτού κώδικα, φανερώνει πως δεν έχουμε να αντιμετωπίσουμε μεμονωμένα συμβάντα, αλλά τάσεις που συνδέονται βαθύτερα με τη διαδικασία εκμάθησης, την έκθεση των μοντέλων σε ευαίσθητα δεδομένα, και την ενσωμάτωσή τους σε καίριες λειτουργίες οργανισμών. Για να μπορέσουν να βρεθούν λύσεις σε αυτά τα προβλήματα, είναι κρίσιμη η κατανόηση των αιτιών που τα δημιουργούν και των κινδύνων που τα συνοδεύουν.

4.2 Αναλυτική διάσπαση αιτίων, μεθοδολογίας επίθεσης και επιπτώσεων

Πολλές μελέτες αναφέρουν πως οι διαδικασίες που ακολουθεί κάθε μοντέλο στην εκπαίδευσή του παίζει καθοριστικό ρόλο στη διερεύνηση των αιτιών διαρροών δεδομένων. Συγκεκριμένα, όταν υπάρχουν επαναλαμβανόμενες ακολουθίες στο εκπαιδευτικό περιεχόμενο ή όταν τα δεδομένα δεν έχουν περάσει από τη διαδικασία της ανωνυμοποίησης και περιέχουν προσωπικά στοιχεία, τα μοντέλα υιοθετούν την τάση να αναπαράγουν το ίδιο ή παραπλήσιο περιεχόμενο σε περιπτώσεις που η είσοδος του χρήστη περιλαμβάνει παρόμοια συμφραζόμενα. Επίσης, η ίδια η δομή των μοντέλων ευνοεί τη βραχυπρόθεσμη απομνημόνευση πληροφοριών, γεγονός που μπορούν να εκμεταλλευτούν οι επιτιθέμενοι για να διαμορφώσουν τις εισόδους τους ανάλογα. Έχουν υπάρξει περιστατικά, κατά τα οποία σημειώθηκε διαρροή πληροφοριών χωρίς την πρόθεση του χρήστη, καθώς παρείχε συγκεκριμένα μοτίβα, εν αγνοία του, που επανέφεραν κείμενα άμεσα συνδεδεμένα με τα δεδομένα εκμάθησης του.

Βάσει αυτών των χαρακτηριστικών, έχουν αναπτυχθεί μεθοδολογίες επιθέσεων, ώστε να αξιοποιήσουν αυτές τις εσωτερικές αδυναμίες των μοντέλων. Στις περισσότερες περιπτώσεις, μέσω κατάλληλων προτροπών, ωθούν το μοντέλο να αναπαράγει μεγάλα τμήματα του εκπαιδευτικού συνόλου. Παρ' όλο που έχουν αναπτυχθεί μηχανισμοί ευθυγράμμισης, πειραματικές μελέτες έχουν εντοπίσει πως πολλά μοντέλα παραγωγής επιστρέφουν αυτούσια κείμενα στα οποία έχουν εκτεθεί με οποιοδήποτε τρόπο εφόσον λάβουν στοχευμένες αλληλουχίες που λειτουργούν ως σημεία αγκίστρωσης. Τέτοιου είδους τεχνικές δεν απαιτούν σχεδόν ποτέ πρόσβαση στο εσωτερικό του μοντέλου, συνεπώς η επικινδυνότητα, και η πιθανότητα εμφάνισης των επιθέσεων αυξάνονται ραγδαία.

Ένα εξίσου σοβαρό είδος επίθεσης είναι αυτό που περιλαμβάνει την αποκάλυψη ευαίσθητων προσωπικών στοιχείων μέσω συστηματικά σχεδιασμένων προτρεπτικών προτάσεων. Πιο αναλυτικά, ο επιτιθέμενος σε αυτό το σενάριο δοκιμάζει συνδυαστικές εισόδους όπως ένας κανονικός χρήστης, ώστε να εντοπίσει διαδρομές που καταλήγουν να χειραγωγούν το μοντέλο, με στόχο την αποκάλυψη του κρυμμένου περιεχομένου. Έχει αποδειχθεί πως ορισμένα μοντέλα έχουν την τάση να ανακτούν δεδομένα επικοινωνίας, όπως αριθμούς τηλεφώνου, και ηλεκτρονικές διευθύνσεις, σε περιπτώσεις που η συζήτησή τους με έναν χρήστη κατευθύνεται σε συγκεκριμένες συσχετίσεις που έχουν αναφερθεί στο σύνολο εκμάθησης. Το φαινόμενο αυτό είναι ιδιαίτερα ανησυχητικό, καθώς ο επιτιθέμενος δε χρειάζεται να εισάγει τέτοιου είδους πληροφορίες στο σύστημα, παρά μόνο να χειραγωγήσει το μοντέλο, οπότε ο έλεγχος των συσχετίσεων που συγκρατεί το μοντέλο μετά την εκπαίδευσή του είναι πολύ απαιτητικός.

Ακόμα μία κατηγορία απειλής της ιδιωτικότητας στα μεγάλα γλωσσικά μοντέλα οφείλεται στην εκμετάλλευση των εσωτερικών ενσωματώσεων, όπου ο επιτιθέμενος ανακατασκευάζει ένα κείμενο μέσω της πρόσβασης σε ειδικές αναπαραστάσεις που κανονικά προορίζονταν αποκλειστικά για εσωτερική χρήση. Είναι αποδεδειγμένο πως μέσω ορισμένων μορφών ενσωμάτωσης μπορούν να εξαχθούν αναγνωρίσιμα στοιχεία ταυτότητας με υψηλό βαθμό ακρίβειας, ακόμα και σε περιπτώσεις που έχουν ληφθεί μέτρα αντιμετώπισης του φαινομένου της απευθείας απομνημόνευσης. Μέσω αυτής της απειλής γίνεται κατανοητό πως η διαρροή δεν περιορίζεται μόνο στην παραγωγή κειμένου, αλλά μπορεί να κάνει την εμφάνισή της σε ενδιάμεσα στάδια, για τα οποία υποτίθεται ότι οι πληροφορίες είναι κρυμμένες, ή κωδικοποιημένες μέσω δυσανάγνωστων αναπαραστάσεων.

Εξίσου σημαντική είναι η κατηγορία επιθέσεων, στην οποία αντί για διαρροή δεδομένων, επιτυγχάνεται αλλοίωση της συμπεριφοράς του μοντέλου υπό συγκεκριμένες συνθήκες, μέσω ενσωμάτωσης κακόβουλου μηχανισμού. Για να εφαρμοστεί αυτή η επίθεση, χρειάζεται να γίνει εκμετάλλευση μιας εισαγόμενης εργασίας, ώστε να ενεργοποιήσει μια ασυνήθιστη λειτουργία από αυτή που περιμένει ο χρήστης, με τελικό στόχο το μοντέλο να απαντάει με

παραπλανητικό ή κακόβουλο τρόπο. Αν και μια τέτοια επίθεση δεν έχει εντοπιστεί σε μεγάλο πλήθος, τα μεγαλύτερα μοντέλα έχουν εμφανίσει τον τελευταίο καιρό αυξημένη σταθερότητα σε αυτά τα μοτίβα, γεγονός που καταδεικνύει πως οι απειλές δεν προέρχονται μόνο από την απομνημόνευση δεδομένων, αλλά και μέσω εξωτερικής εμφύτευσης ανεπιθύμητων συμπεριφορών.

Η σοβαρότητα των επιθέσεων έχει άμεση σχέση με τις επιπτώσεις που έχουν, και εφόσον οι διαρροές μπορούν να λάβουν διάφορες μορφές είναι πολύ σημαντικό να υπάρχει πλήρης κατανόηση της κατάστασης. Η εμπιστευτικότητα των δεδομένων που χρησιμοποιούνται για την ανάπτυξη των μοντέλων συχνά υπονομεύεται από την εξαγωγή του εκπαιδευτικού περιεχομένου, εφόσον πολλές φορές αυτό αποκαλύπτεται παρ' όλο που θεωρούνταν προστατευμένο. Άλλου είδους επιθέσεις που χρησιμοποιούν αλληλουχίες αναπαραστάσεων, καταλήγουν στην αποκάλυψη προσωπικών στοιχείων που δε φαίνονται στους τελικούς χρήστες. Παράλληλα, με την επίθεση εμφύτευσης μοτίβων με σκοπό την τροποποίηση της συμπεριφοράς των μοντέλων, επιτείνεται ο κίνδυνος χρήσης σε κρίσιμους τομείς, όπου η αξιοπιστία είναι απαραίτητη. Έρευνες δείχνουν πως οι παράγοντες που επηρεάζουν την πιθανότητα εμφάνισης μιας επίθεσης σε ένα μεγάλο γλωσσικό μοντέλο δεν είναι το είδος ή το μέγεθος, αλλά η μέθοδος εκμάθησης, αξιολόγησης, και ενσωμάτωσης σε λειτουργικές υπηρεσίες.

4.3 Μαθήματα για σχεδιασμό ασφαλών pipelines

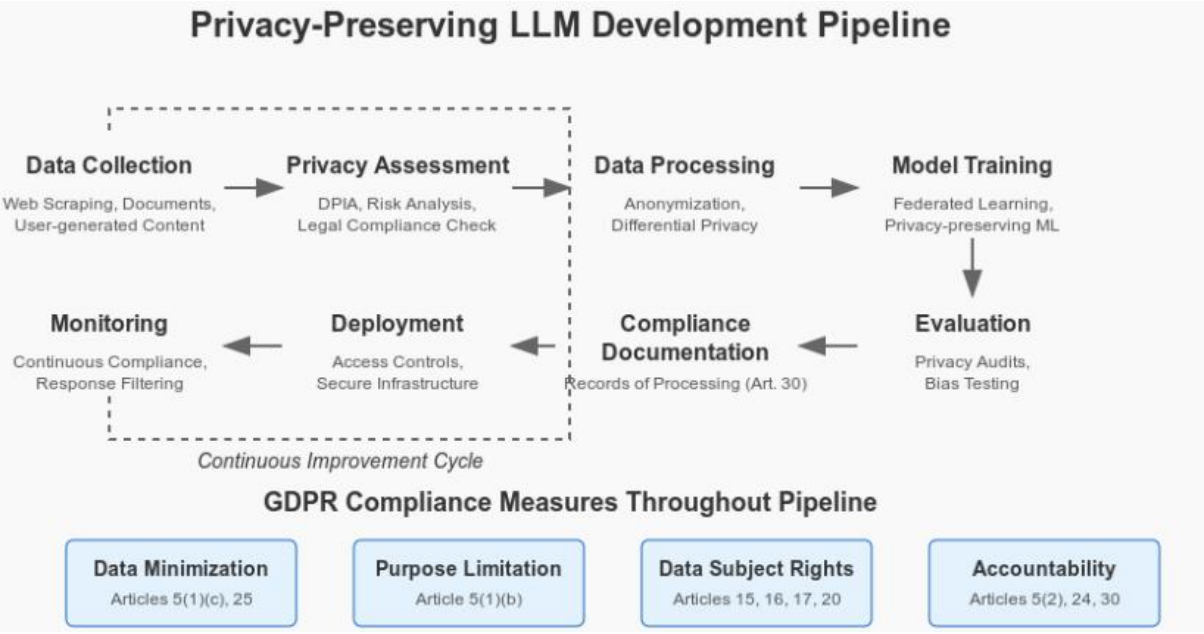
Με την ανάλυση των περιστατικών διαρροής πληροφοριών αναδεικνύεται η ανάγκη για την προστασία της ιδιωτικότητας και αυτή δεν εξαρτάται μόνο από την αρχιτεκτονική των μοντέλων, αλλά από ολόκληρη την αλυσίδα ανάπτυξης και λειτουργίας. Διάφορες έρευνες, που μελετούν την οργάνωση διαχείρισης των δεδομένων, αναφέρουν πως οι ευπάθειες ξεκινούν από τα πρώτα κιόλας στάδια συλλογής των εκπαιδευτικών συνόλων, καθώς πολλές φορές δεν υπάρχει συστηματικός έλεγχος προέλευσης, ποιότητας και ταξινόμησης ως προς την ευαισθησία των δεδομένων. Έτσι, στις περιπτώσεις που υπάρχουν επαναλαμβανόμενες ή προσωπικές ακολουθίες, χτίζεται η βάση για μεταγενέστερες διαρροές, επομένως χρειάζεται μεθοδική οργάνωση, όπου θα προβλέπει τέτοιες καταστάσεις και θα τις προλαβαίνει πριν ακόμα και από την εκπαιδευτική διαδικασία.

Καθώς το μοντέλο βρίσκεται στο στάδιο εκπαίδευσης και αξιολόγησης, χρειάζεται η δημιουργία επαναλαμβανόμενων μηχανισμών ελέγχου. Για να είναι ικανό ένα μοντέλο να αναπτυχθεί, πρέπει το πλαίσιο στο οποίο βρίσκεται να προβλέπει συνεχείς παρακολουθήσεις κατά τη διάρκεια της εκπαίδευσης, αλλά και μετά την ολοκλήρωσή της, ώστε να γίνεται ο έγκαιρος εντοπισμός αποκλίσεων από την αναμενόμενη συμπεριφορά. Με την τοποθέτηση ενός τέτοιου παρατηρητικού μηχανισμού του συστήματος, γίνεται εφικτή η αναγνώριση μοτίβων που μπορεί να προκαλέσουν αποκάλυψη ευαίσθητων πληροφοριών, αλλά και να χρησιμοποιηθεί για την εύρεση ευάλωτων σημείων στα οποία το μοντέλο μπορεί να αποδειχθεί τρωτό αν του δοθούν ορισμένες είσοδοι. Φυσικά, η προσέγγιση αυτή δεν εξαλείφει πλήρως τους κινδύνους, όμως ενισχύει την άμεση κατανόησή τους, καθώς και αποκαλύπτει τις συνθήκες υπό τις οποίες εμφανίζονται.

Επιπλέον, η σωστή και ασφαλής λειτουργία του συστήματος εξαρτάται άμεσα από τον τρόπο οργάνωσης της προεπεξεργασίας και της επιμέλειας των δεδομένων. Πρότυπα υπεύθυνου σχεδιασμού μεγάλων γλωσσικών μοντέλων τονίζουν πως για τη δημιουργία ενός αξιόπιστου μοντέλου δεν αρκεί η μείωση σφαλμάτων και η σωστή επιλογή συνόλων εκμάθησης, αλλά χρειάζεται συνεχής θωράκιση του συστήματος για να καταπολεμηθούν οι ανεπιθύμητες συμπεριφορές. Δηλαδή, ένα σωστό pipeline πρέπει να περιέχει εντατικούς ελέγχους για τον εντοπισμό ανωμαλιών στη δομή, την ποικιλομορφία και τον σκοπό των δεδομένων, αλλά και επιπλέον δικλείδες ασφαλείας για να εξετάζεται αν τα μοτίβα που

εισάγονται κάθε φορά στο μοντέλο δεν ενεργοποιούν κάποια ασυνήθιστη συμπεριφορά στη λειτουργία του. Μελέτες έχουν δείξει πως τέτοιοι έλεγχοι ενσωματώνονται στη ροή διεργασιών του μοντέλου πριν από τη διαδικασία εκμάθησης, ώστε να μειωθεί ο κίνδυνος μεταγενέστερων διαρροών ή κακόβουλης συμπεριφοράς.

Στο τελευταίο στάδιο του pipeline βρίσκεται η έξοδος του μοντέλου, στην οποία καίριο ρόλο παίζει η ύπαρξη δομημένων ελεγχόμενων διαδικασιών που αξιολογούν τη χρήση και τη ροή των δεδομένων σε πραγματικές συνθήκες. Μέσω της αξιολόγησης των συστημάτων σε παραγωγικό περιβάλλον, γίνεται εφικτή η τεκμηρίωση κάθε απόφασης, αλλά και η οριοθέτηση στον τρόπο επεξεργασίας των εισόδων και εξόδων, ώστε να μειωθεί ουσιαστικά το ρίσκο. Επίσης, οργανώνοντας τα pipelines με τέτοιο τρόπο ώστε να εφαρμόζεται η ανίχνευση αποκλίνουσας συμπεριφοράς, η καταγραφή συμβάντων διαρροής και τον έγκαιρο εντοπισμό ασυνήθιστων μοτίβων ενισχύεται η σταθερότητα του συστήματος, και μειώνεται η πιθανότητα επιτυχούς επίθεσης.



Εικόνα 7: Pipeline ανάπτυξης μεγάλων γλωσσικών μοντέλων με ενσωματωμένα μέτρα προστασίας ιδιωτικότητας κατά GDPR.

Συνολικά, οι μελέτες που ερευνούν τη γενικότερη ασφάλεια των μεγάλων γλωσσικών μοντέλων καταλήγουν πως αυτή δεν καθορίζεται μόνο από τα τεχνικά χαρακτηριστικά του, αλλά και από τη συνοχή και την ποιότητα του συνολικού pipeline. Κάθε ξεχωριστό βήμα, από την επιλογή των εκπαιδευτικών συνόλων μέχρι και την παραγωγική λειτουργία, χρειάζεται συνεχόμενη εποπτεία, ξεκάθαρο καταμερισμό ευθυνών, και μηχανισμούς ανίχνευσης κινδύνων. Μέσω των πρόσφατων περιστατικών που έχουν καταγραφεί γίνεται άμεσα αντιληπτό πως η ασφάλεια δεν ολοκληρώνεται απλώς με τη λήψη τεχνικών μέτρων, αλλά μέσω της συνεχούς αξιολόγησης και αναπροσαρμογής της αλυσίδας παραγωγής.

5 ΝΟΜΙΚΟ ΚΑΙ ΚΑΝΟΝΙΣΤΙΚΟ ΠΛΑΙΣΙΟ ΓΙΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΜΟΝΤΕΛΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

5.1 Γενικές αρχές προστασίας προσωπικών δεδομένων

Η προστασία προσωπικών δεδομένων είναι μία από τις βασικότερες αρχές στο ευρωπαϊκό ρυθμιστικό πλαίσιο και ορίζεται από ένα σύνολο κανόνων που υποστηρίζουν κάθε οργανισμό στην επεξεργασία πληροφοριών με σεβασμό προς όλα τα δικαιώματα των

πολιτών. Οι αρχές που θεσπίζει αυτό το πλαίσιο λειτουργούν ως βάση για την προστασία της ιδιωτικότητας και είναι ενσωματωμένες σε όλα τα σύγχρονα επιχειρησιακά συστήματα που βασίζονται σε δεδομένα. Σε αυτό το πλαίσιο δεν καθορίζονται μόνο οι υποχρεώσεις των οργανισμών, αλλά και τα όρια που πρέπει να τηρούνται όσο η τεχνολογία αναπτύσσεται, συμπεριλαμβανομένων και των μοντέλων μηχανικής μάθησης.

Για να μπορέσει να υπάρξει ουσιαστική προστασία των δεδομένων είναι απαραίτητη η γνώση των αρχών της νομιμότητας, της διαφάνειας, και του προσδιορισμένου σκοπού σε κάθε οργανισμό. Κάθε μορφή επεξεργασίας δεδομένων πρέπει να στηρίζεται σε θεμελιωμένη νομική βάση και να υπάρχει σαφής και πλήρης ενημέρωση στο υποκείμενο σχετικά με τη μέθοδο συλλογής, τη χρήση, και τη διάρκεια διατήρησης των δεδομένων. Στη συνέχεια, το σύνολο των δεδομένων κατά τη συλλογή και την επεξεργασία τους πρέπει να επιτελεί έναν συγκεκριμένο σκοπό, ο οποίος δεν μπορεί να αλλάζει ή να επεκτείνεται πέραν του προκαθορισμένου πλαισίου. Μελέτες υπογραμμίζουν πως ο περιορισμός του σκοπού πριν την έναρξη της επεξεργασίας αποτελεί τροχοπέδη σε σενάρια καταχρήσεων, ειδικά όταν η συλλογή πληροφοριών εφαρμόζεται σε μεγάλη κλίμακα.

Δύο βασικές αρχές για τη σωστή επεξεργασία των δεδομένων είναι αυτές της ελαχιστοποίησης και της αρχής περιορισμένου χρόνου διατήρησης δεδομένων. Ουσιαστικά, οι οργανισμοί οφείλουν να λαμβάνουν και να διατηρούν μόνο τις πληροφορίες που είναι χρήσιμες και αναγκαίες για τον σκοπό επεξεργασίας που έχει οριστεί, καθώς έχει αποδειχθεί πως η μαζική συλλογή δεδομένων χωρίς ιεράρχηση αναγκαιότητας αποτελεί έναν από τους βασικούς κινδύνους για την ιδιωτικότητα. Η αρχή του περιορισμένου χρόνου διατήρησης υποδεικνύει την άμεση διαγραφή των δεδομένων τη στιγμή που πάψουν να είναι αναγκαία για τον σκοπό επεξεργασίας. Στις περιπτώσεις που οι βάσεις δεδομένων είναι μεγάλες, και λείπουν οι στοιχειώδεις πολιτικές διατήρησης, συχνά οδηγούνται σε εκτεθειμένα σύνολα, όπου μετά από εκτεταμένο χρόνο διατήρησης, καθίστανται ευάλωτα σε κακή χρήση.

Σε ένα τέτοιο πλαίσιο, καθοριστική σημασία αποκτούν τα δικαιώματα των υποκειμένων, τα οποία επιτρέπουν τον ουσιαστικό έλεγχο της επεξεργασίας των δεδομένων τους. Κατοχυρωμένα δικαιώματα όπως η πρόσβαση, η διόρθωση και η διαγραφή των προσωπικών δεδομένων αποτελούν νομικές εγγυήσεις, καθώς διασφαλίζουν τη διαφάνεια και τη λογοδοσία των οργανισμών. Συγκεκριμένα, το δικαίωμα στη διαγραφή αποτελεί ορόσημο για την ενίσχυση της προστασίας των δεδομένων, εφόσον με βάση αυτό το υποκείμενο δικαιούται να ανακαλέσει τις πληροφορίες που δεν επιθυμεί να επεξεργάζονται πια. Ερευνητικές αναλύσεις αναφέρουν πως αυτό το δικαίωμα είναι απαραίτητο μέσο για να επανέλθει ο έλεγχος των δεδομένων στα χέρια των πολιτών, ειδικά σε συνθήκες όπου ο όγκος των δεδομένων είναι ραγδαία αυξανόμενος, και οι πληροφορίες ανακυκλώνονται πλήρως αυτοματοποιημένα.

Μία ακόμη σημαντική αρχή για την προστασία δεδομένων αποτελεί η αρχή της λογοδοσίας, κατά την οποία ο οργανισμός που επεξεργάζεται δεδομένα, εκτός της συμμόρφωσής του με τις νομικές απαιτήσεις, πρέπει να είναι σε θέση να αποδείξει ανά πάσα στιγμή ότι έχει λάβει όλα τα απαραίτητα αντίμετρα προκειμένου να διασφαλίσει την ουσιαστική εφαρμογή του συνόλου των αρχών προστασίας δεδομένων. Συγκεκριμένα, η αρχή της λογοδοσίας δεν προβλέπει μόνο τη λήψη τεχνικών μέτρων, αλλά επεκτείνεται στη διακυβέρνηση, στις εσωτερικές διαδικασίες, αλλά και στις μεθόδους λήψης αποφάσεων εντός του οργανισμού. Έρευνες αναφέρουν ότι η ουσιαστική εφαρμογή της λογοδοσίας ενισχύει τις υπόλοιπες βασικές αρχές, αφού προϋποθέτει εντατικό έλεγχο, σαφή κατανομή αρμοδιοτήτων, και συστηματική επικαιροποίηση των διαδικασιών προστασίας δεδομένων.

5.2 Σχέση μοντέλων και νόμων για την επεξεργασία δεδομένων

Με τη ραγδαία ανάπτυξη των μεγάλων γλωσσικών μοντέλων, η ένταξή τους στα συστήματα επεξεργασίας πληροφοριών είναι αναπόφευκτη, καθώς έτσι επαναπροσδιορίζεται ο τρόπος με τον οποίο εφαρμόζονται οι ισχύοντες κανόνες προστασίας προσωπικών δεδομένων. Συγκεκριμένα, τα μοντέλα δεν περιορίζονται στη στατική αποθήκευση πληροφοριών, αλλά μέσω της εκπαίδευσης, προσαρμόζονται ανάλογα με την κλίμακα των δεδομένων και παράγουν αντίστοιχο περιεχόμενο, γεγονός που με τον κατάλληλο έλεγχο μπορεί να ενταχθεί στο πεδίο εφαρμογής της προστασίας πληροφοριών. Η ερμηνεία της επεξεργασίας είναι περίπλοκη, καθώς αυτή δεν αφορά αποκλειστικά τα αρχικά σύνολα εκπαίδευσης, στα οποία εκτίθεται το μοντέλο, αλλά προεκτείνεται στο ίδιο το μοντέλο, μαζί με τις εξόδους που αυτό παράγει.

Νομικές αναλύσεις αναφέρουν πως η διαδικασία της εκπαίδευσης των μοντέλων αποτελεί ξεκάθαρα επεξεργασία προσωπικών δεδομένων, παρ' όλο που σε πολλές περιπτώσεις τα δεδομένα δεν έχουν υποστεί ουσιαστικές αλλαγές. Τα μοντέλα έχουν τη δυνατότητα να διατηρούν συσχετίσεις και να αναπαράγουν δεδομένα που αφορούν φυσικά πρόσωπα, συνεπώς το ίδιο το μοντέλο ενδέχεται να ενσωματώνει προσωπικά δεδομένα που σε πολλές περιπτώσεις η διαδικασία διαγραφής μπορεί να είναι μη ανατρέψιμη. Από νομικής άποψης, η επεξεργασία των πληροφοριών δε σταματά στα δεδομένα εισόδου, αλλά επεκτείνεται και στο παραγόμενο περιεχόμενο, καθώς μπορεί να έχει επηρεαστεί. Όταν τα παραγόμενα κείμενα περιλαμβάνουν πληροφορίες, που μπορούν να συνδεθούν, έμμεσα ή άμεσα, με φυσικά πρόσωπα ενεργοποιούνται οι ανάλογες νομικές υποχρεώσεις. Αυτό σημαίνει πως η συμμόρφωση δεν πρέπει να περιορίζεται στο στάδιο της εκπαίδευσης, αλλά αφορά την εντατική αξιολόγηση της λειτουργίας του μοντέλου.

Ένα πολύ καίριο ζήτημα αφορά τον καταμερισμό ρόλων και ευθυνών στο πλαίσιο λειτουργίας των μεγάλων γλωσσικών μοντέλων, καθώς ο σχεδιασμός, η εκπαίδευση, το στάδιο λειτουργίας και η τελική αξιοποίηση τους πραγματοποιείται από διαφορετικούς φορείς. Πολλές φορές γίνεται ασαφής η διάκριση μεταξύ του υπεύθυνου επεξεργασίας και του εκτελούντος την επεξεργασία, οπότε δεν είναι ξεκάθαρο ποίος φέρει τη νομική και οργανωτική υποχρέωση συμμόρφωσης. Νομικά άρθρα τονίζουν ότι το βασικό κριτήριο για την απόδοση ευθυνών είναι ο έλεγχος επί του σκοπού και των μέσων επεξεργασίας, οπότε ο σαφής προσδιορισμός αρμοδιοτήτων καθίσταται απαραίτητος σε ολόκληρο τον κύκλο ζωής του μοντέλου.

Η εφαρμογή του γενικού κανονισμού προστασίας δεδομένων ορίζει ένα γενικό πλαίσιο, στο οποίο εντάσσονται και τα μεγάλα γλωσσικά μοντέλα. Πλέον, όμως, έχουν θεσπιστεί ειδικότερα κανονιστικά πλαίσια για την τεχνητή νοημοσύνη, χωρίς να υποβαθμίζουν το ρόλο του GDPR. Ουσιαστικά, τα νεότερα κανονιστικά εργαλεία λειτουργούν συμπληρωματικά, καθώς το κύριο θεμέλιο στην αξιολόγηση θεμάτων ιδιωτικότητας παραμένει το δίκαιο προστασίας δεδομένων. Έτσι, το νόημα της λογοδοσίας αποκτά αυξημένη σημασία, εφόσον οι οργανισμοί, εκτός της υποχρέωσης συμμόρφωσης, πρέπει να αποδεικνύουν έμπρακτα πως οι διαδικασίες ανάπτυξης και λειτουργίας των μοντέλων σέβονται τα δικαιώματα των χρηστών και των υποκειμένων αδιαλείπτως.

Συμπερασματικά, η σχέση ανάμεσα στο νομικό πλαίσιο επεξεργασίας δεδομένων **και** στα μεγάλα γλωσσικά μοντέλα χαρακτηρίζεται από τη δυναμική εξέλιξη και τη συνεχώς αυξανόμενη πολυπλοκότητα. Τα συμπληρωματικά νομικά εργαλεία καλούνται να προασπιστούν την ιδιωτικότητα σε περιβάλλοντα όπου η αποτελεσματικότητα και η λειτουργικότητα είναι κυρίαρχες έννοιες. Αυτό καθιστά απαραίτητη την περιοδική νομική χαρτογράφηση των διαδικασιών χρήσης και ανάπτυξης των μοντέλων, ώστε να προστατεύεται η ιδιωτικότητα σε όλα τα στάδια.

5.3 Διεθνείς προκλήσεις και διασυνοριακές μεταφορές δεδομένων

Με την ανάπτυξη των μεγάλων γλωσσικών μοντέλων σε παγκόσμιο επίπεδο αναδιαμορφώνονται οι προκλήσεις που σχετίζονται με τις μεταφορές προσωπικών δεδομένων ανάμεσα στις χώρες. Τα μοντέλα, σε αντίθεση με τα παραδοσιακά πληροφοριακά συστήματα, λειτουργούν διαφορετικά, καθώς οι διαδικασίες επεξεργασίας των δεδομένων, όπως η συλλογή, η εκπαίδευση, και η αποθήκευση λαμβάνουν χώρα σε διαφορετικές γεωγραφικές τοποθεσίες. Συνεπώς, η πραγματική φυσική τοποθεσία των δεδομένων και των υπολογιστικών πόρων είναι ασαφής, δημιουργώντας προβλήματα στην εφαρμογή των εθνικών νομικών πλαισίων προστασίας δεδομένων. Συγκεκριμένα, το ευρωπαϊκό κανονιστικό πλαίσιο, όσον αφορά τη μεταφορά προσωπικών δεδομένων εκτός Ευρωπαϊκής Ένωσης, προβλέπει αυστηρούς περιορισμούς, υποχρεώνοντας τους οργανισμούς να διαθέτουν επαρκές επίπεδο προστασίας ή τον ορισμό συγκεκριμένων νομικών εγγυήσεων. Στην πράξη, όμως, με την ανάπτυξη των παγκόσμιων ψηφιακών υπηρεσιών νέφους και τη χρήση υποδομών υπολογιστικής ισχύος η διαρκής εποπτεία των ροών δεδομένων δεν είναι πάντα εύκολη. Οι διασυνοριακές μεταφορές δεδομένων δεν είναι πάντα διακριτές, καθώς είναι ενσωματωμένες στη λειτουργία του συστήματος, γεγονός που περιπλέκει περαιτέρω τον έλεγχο συμμόρφωσης με εθνικά κανονιστικά πλαίσια.

Ιδιαίτερα χαρακτηριστικό παράδειγμα θεσμικών δυσκολιών αποτέλεσε η μεταφορά δεδομένων μεταξύ Ευρωπαϊκής Ένωσης και Ηνωμένων Πολιτειών, καθώς εκεί αναδείχθηκαν τα όρια νομικών μηχανισμών σε περιπτώσεις που συγκρούονται με διαφορετικά εθνικά πλαίσια στην πρόσβαση και επεξεργασία των δεδομένων. Με την αδυναμία διασφάλισης ισοδύναμης προστασίας πριν και μετά τη μεταφορά των πληροφοριών αποδείχθηκε η ασυμβατότητα μεταξύ δικαιοδοσιών, ειδικά στα περιβάλλοντα όπου τα δεδομένα ενδέχεται να υποστούν κρατικές παρεμβάσεις. Μέσω αυτού του παραδείγματος γίνεται κατανοητό πως οι διασυνοριακές ροές δεδομένων θα συνεχίσουν να βρίσκονται εντός ενός περιβάλλοντος νομικής αβεβαιότητας, όσο και να εξελίσσεται το τεχνολογικό πλαίσιο στο οποίο βρίσκονται.

Οι προκλήσεις που καλούνται να αντιμετωπίσουν οι φορείς των μεγάλων γλωσσικών μοντέλων βρίσκονται σε υψηλότερο επίπεδο δυσκολίας όταν το παραγόμενο περιεχόμενο στηρίζεται σε δεδομένα που έχουν υποστεί επεξεργασία σε διάφορες χώρες όπου τα κανονιστικά πλαίσια διαφέρουν, χωρίς να είναι ξεκάθαρο ποια έννομη τάξη εφαρμόζεται σε κάθε στάδιο. Ακόμα και σε περιπτώσεις που τα αρχικά δεδομένα έχουν συλλεχθεί με νόμιμο τρόπο, η έπειτα επεξεργασία μπορεί να δημιουργεί νέες μορφές πληροφοριών, οι οποίες μπορεί να υπάγονται σε διαφορετικά κανονιστικά καθεστώτα. Με αυτό τον τρόπο τίθενται πολύπλοκα ερωτήματα σχετικά με το ποίος φέρει τελικά την ευθύνη συμμόρφωσης, και ποία δικαιώματα μπορούν να ασκήσουν τα υποκείμενα των δεδομένων σε παγκόσμιο επίπεδο.

Επιπλέον, με τη διεθνή διάσταση των υποδομών καθίσταται δύσκολη η διαδικασία της λογοδοσίας και της απόδοσης ευθυνών. Στις περιπτώσεις που βρίσκονται εμπλεκόμενοι διάφοροι φορείς όσον αφορά τη διαχείριση, τη φιλοξενία, και την αξιοποίηση των μοντέλων, η δυνατότητα εποπτείας των ροών δεδομένων περιορίζεται σημαντικά, και η εφαρμογή των κανονιστικών απαιτήσεων γίνεται τμηματικά. Μελέτες αναφέρουν ότι η ουσιαστική προστασία της ιδιωτικότητας, σε τέτοια πλαίσια, δεν πρέπει να βασίζεται αποκλειστικά σε τυπικές νομοθετικές ρυθμίσεις, αλλά χρειάζεται συνεχής αξιολόγηση των παγκόσμιων ροών δεδομένων, ενώ ταυτόχρονα πρέπει να θεσπιστούν σαφείς αρμοδιότητες για κάθε στάδιο στον κύκλο ζωής των μοντέλων.

Συνοπτικά, το ζήτημα των διασυνοριακών μεταφορών δεδομένων αποτελεί μία από τις πιο σύνθετες προκλήσεις για τη νομική ρύθμιση των μεγάλων γλωσσικών μοντέλων. Η ανάγκη για παγκόσμια χρήση των τεχνολογιών αυτών έρχεται αντιμέτωπη με τα κατά βάση εδαφικά νομικά συστήματα, θέτοντας κενά προστασίας και αυξάνοντας τους κινδύνους για την

ιδιωτικότητα. Συνεπώς, το πρώτο και αναγκαίο βήμα για να αντιμετωπιστούν αυτές οι προκλήσεις είναι η κατανόησή τους, ώστε να διαμορφωθούν αποτελεσματικοί μηχανισμοί συμμόρφωσης και τεκμηρίωσης.

5.4 Απαιτήσεις συμμόρφωσης, DPIA και τεκμηρίωση

Για την ολοκληρωμένη συμμόρφωση των μεγάλων γλωσσικών μοντέλων με το ισχύον ρυθμιστικό σχήμα, εκτός της τυπικής τήρησης των νομικών διατάξεων, χρειάζεται η ενσωμάτωση δομημένων μηχανισμών αξιολόγησης κινδύνων και τεκμηρίωσης καθ' όλη τη διάρκεια λειτουργίας του συστήματος. Βασικό εργαλείο για να τεκμηριώσει ένας οργανισμός τη συμμόρφωση και τη λογοδοσία αποτελεί η αξιολόγηση αντικτύπου στην προστασία δεδομένων. Όταν η επεξεργασία δεδομένων ενδέχεται να επιφέρει υψηλούς κινδύνους για τα δικαιώματα και τις ελευθερίες των υποκειμένων, η αξιολόγηση αντικτύπου είναι ο βασικός μηχανισμός προληπτικού ελέγχου, καθώς δεν περιορίζεται στην απλή, στατική απαρίθμηση κινδύνων, αλλά λειτουργεί αναλύοντας τη φύση, την έκταση και τον σκοπό της επεξεργασίας ενώ ταυτόχρονα υπολογίζει την πιθανότητα εμφάνισης και τη σοβαρότητα των επιπτώσεων. Συγκεκριμένα, για τα συστήματα τεχνητής νοημοσύνης, με τη ραγδαία ανάπτυξη της πολυπλοκότητας των ροών δεδομένων, η DPIA καθίσταται ιδιαίτερα σημαντική, εφόσον οι κίνδυνοι δεν προέρχονται μόνο από τα δεδομένα που εισάγει ο χρήστης, αλλά και από τη διαδικασία εκπαίδευσης, και τον τρόπο λειτουργίας του μοντέλου.

Με τις νέες κανονιστικές απαιτήσεις που αφορούν την τεχνητή νοημοσύνη ενισχύεται η ανάγκη της αδιάκοπης εποπτείας και διαχείρισης κινδύνων, ταυτόχρονα με τη συστηματική τεκμηρίωση. Για τη σωστή συμμόρφωση απαιτείται σαφής καταγραφή του πλαισίου εφαρμογής, της κατηγορίας κινδύνου όπου εντάσσεται, των ρόλων όλων των ενδιαφερόμενων μερών, και των τεχνικών και οργανωτικών αντιμέτρων που εφαρμόζονται. Μέσω αυτής της καταγραφής γίνεται εύκολη η τεκμηρίωση έναντι των εποπτικών αρχών, επιτρέποντας την επαλήθευση ότι ο οργανισμός είναι συμμορφωμένος ως προς όλες τις απαιτήσεις, και οι διαδικασίες έχουν ενσωματωθεί στο σύστημα ήδη από το στάδιο του σχεδιασμού, ενώ τηρούνται καθ' όλη τη διάρκεια της λειτουργίας του.

Ιδιαίτερα προσεκτικοί οφείλουν να είναι οι οργανισμοί στο θέμα της διασύνδεσης της τεκμηρίωσης με τη διαχείριση κινδύνων. Καταγράφοντας αναλυτικά τις αρχιτεκτονικές επιλογές, τα εκπαιδευτικά σύνολα, τη διαδικασία εκμάθησης, καθώς και τους μηχανισμούς ανθρώπινης εποπτείας γίνεται πιο εύκολη η ιχνηλασιμότητα των αποφάσεων που λαμβάνονται σχετικά με την ανάπτυξη και λειτουργικότητα του μοντέλου. Έτσι, η τεκμηρίωση μετατρέπεται από τυπική υποχρέωση σε χρήσιμο εργαλείο με σκοπό την ουσιαστική λογοδοσία, ώστε να διευκολύνεται η διαδικασία αξιολόγησης συμμόρφωσης.

Αρκετά χρήσιμη έχει αναδειχθεί τον τελευταίο καιρό η αξιοποίηση των εργαλείων τεχνητής νοημοσύνης στην υποστήριξη της συμμόρφωσης των οργανισμών ως συμπληρωματική πρακτική, ειδικά σε περιβάλλοντα με αυξημένη πολυπλοκότητα. Αυτού του είδους οι προσεγγίσεις έχουν αποδειχθεί πως είναι ικανές να συμβάλουν ουσιαστικά στην παρακολούθηση διαδικασιών, την αξιολόγηση κινδύνων, αλλά και στην αυτοματοποίηση της τεκμηρίωσης, χωρίς απαραίτητα να αναιρείται η ανάγκη ανθρώπινης κρίσης και ευθύνης, αφού για να είναι εφικτή η κατανόηση και ο έλεγχος των συστημάτων απαιτείται η διατήρηση σαφών και κατανοητών δομών τεκμηρίωσης. Γενικά, οι απαιτήσεις συμμόρφωσης που αφορούν τα μεγάλα γλωσσικά μοντέλα συνδυάζουν την καθιέρωση μεθόδου αξιολόγησης αντικτύπου, τη συνεχή διαχείριση κινδύνων, και την ουσιαστική τεκμηρίωση. Η DPIA σε συνεργασία με τα συναφή εργαλεία τεκμηρίωσης είναι αλληλένδετα στοιχεία απαραίτητα για να επιτευχθεί ένα ολοκληρωμένο πλαίσιο λογοδοσίας, όπου θα διασφαλίζει την προστασία των δικαιωμάτων των υποκειμένων όσο το μοντέλο αναπτύσσεται.

6 ΗΘΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ ΚΑΙ ΚΟΙΝΩΝΙΚΕΣ ΕΠΙΠΤΩΣΕΙΣ

6.1 Συναίνεση και δικαιώματα υποκειμένων δεδομένων

Η συναίνεση αποτελεί αυτοτελή τόσο νομική βάση επεξεργασίας όσο και θεμελιώδη αρχή προστασίας δεδομένων, αφού είναι άμεσα συνδεδεμένη με την αυτονομία και την ικανότητα του χρήστη να λαμβάνει αποφάσεις σχετικά με την επεξεργασία των προσωπικών του δεδομένων. Στο πλαίσιο των συστημάτων τεχνητής νοημοσύνης, και ιδίως των τεχνικών μηχανικής μάθησης, οι διαδικασίες συλλογής και επεξεργασίας πληροφοριών έχουν αλλάξει ριζικά το νόημα της πλήρως ενημερωμένης συναίνεσης. Αυτό συμβαίνει εξαιτίας της αυξημένης πολυπλοκότητας των καινούργιων συστημάτων, της ελλιπούς διαφάνειας των διαδικασιών, αλλά και της μη προβλέψιμης μελλοντικής χρήσης των δεδομένων.

Έρευνες που μελετούν την ηθική στο πλαίσιο των μεγάλων γλωσσικών μοντέλων αναφέρουν ότι η συναίνεση συχνά εκφυλίζεται σε τυπική διαδικασία, καθώς το υποκείμενο δεν έχει κατανοήσει πλήρως τον πραγματικό τρόπο με τον οποίο θα αξιοποιηθούν τα δεδομένα του. Επίσης, λόγω της αδυναμίας σαφούς επεξήγησης των αποτελεσμάτων πολύπλοκων αλγοριθμικών συστημάτων, και της δυναμικής εξέλιξης των μοντέλων κατά τη διάρκεια της λειτουργίας τους, συχνά υπονομεύεται η απαίτηση για ξεκάθαρη ενημέρωση πριν δοθεί η συγκατάθεση. Φυσικά, το ζήτημα αυτό δε σταματά μόνο στην αρχική παροχή συναίνεσης, αλλά εξαπλώνεται και στη δυνατότητα συνεχούς ενημέρωσης και επαναξιολόγησης της συμμετοχής του χρήστη σε ερευνητικά συστήματα.

Το ζήτημα αυτό αποκτά ιδιαίτερη σημασία όταν τα μεγάλα γλωσσικά μοντέλα εντάσσονται στη διαδικασία λήψης συναίνεσης. Στις περιπτώσεις που αυτά τα συστήματα παίρνουν συμβουλευτικό ρόλο με στόχο την ενίσχυση της κατανόησης και της συμμετοχής των υποκειμένων, εμφανίζονται ηθικά ερωτήματα, όπως για παράδειγμα, η πιθανότητα παροχής ανακριβών ή παραπλανητικών πληροφοριών, σκόπιμα ή άσκοπα, η δυσχέρεια ελέγχου και έγκαιρης έγκρισης απαντήσεων, καθώς και ο κίνδυνος χειριστικής επικοινωνίας. Το σύνολο αυτών των παραγόντων καθιστά αμφισβητήσιμη τη διαδικασία λήψης συναίνεσης που παρέχεται μέσω τέτοιων συστημάτων, εφόσον δεν μπορεί να κατοχυρωθεί η ελευθερία και η ουσιαστική ενημέρωση.

Εκτός της συναίνεσης, τα δικαιώματα των υποκειμένων των δεδομένων αποτελούν θεμέλιο στον μηχανισμό διασφάλισης της ανθρώπινης αυτονομίας όταν αλληλεπιδρούν με αυτοματοποιημένα συστήματα. Τα δικαιώματα πρόσβασης, ενημέρωσης, και εναντίωσης εκτός από διαδικαστικές εγγυήσεις, είναι βασικά εργαλεία που χρειάζεται το άτομο για να κατανοεί πλήρως τον τρόπο με τον οποίο λαμβάνονται οι αποφάσεις που το αφορούν, ώστε να κατέχει ουσιαστικά τον έλεγχο των διαδικασιών αυτών. Στις περιπτώσεις που οι αποφάσεις λαμβάνονται από αυτοματοποιημένα συστήματα τα δικαιώματα αποδυναμώνονται ή και καταπατούνται, με αποτέλεσμα ο χρήστης να χάνει σταδιακά τη δυνατότητα παρέμβασης, συνεπώς επηρεάζεται αρνητικά η προσωπική και κοινωνική του αυτονομία.

Ωστόσο, η ουσιαστική άσκηση των δικαιωμάτων αυτών αντιμετωπίζει τεχνικά φράγματα. Συγκεκριμένα, η ανάκληση συναίνεσης ή το δικαίωμα στη λήθη καθίστανται ιδιαίτερα προβληματικά όταν τα δεδομένα έχουν ήδη ενσωματωθεί στην εκπαιδευτική διαδικασία των μοντέλων. Μελέτες καταδεικνύουν πως η πλήρης απομάκρυνση της επίδρασης ενός υποκειμένου δεδομένων από τη διαδικασία εκμάθησης δεν είναι άμεση και πολλές φορές ανέφικτη. Συνεπώς, δημιουργείται ένα σημαντικό κενό ανάμεσα στις νομικές, τις ηθικές, και τις τεχνολογικές απαιτήσεις ως προς τον σεβασμό των δικαιωμάτων παράλληλα με την πραγματική τεχνική δυνατότητα συμμόρφωσης, και την ελευθερία τεχνολογικής εξέλιξης.

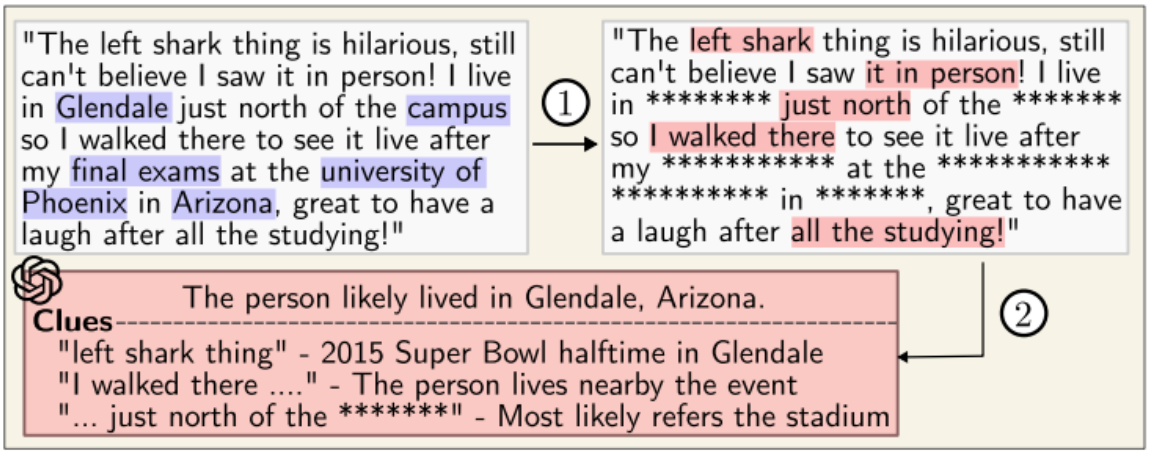
Συμπερασματικά, μέσα από την ανάλυση της συναίνεσης και των δικαιωμάτων των υποκειμένων στο περιβάλλον τεχνητής νοημοσύνης αναδείχθηκε ένα χάσμα μεταξύ ρυθμιστικών αρχών, ηθικής χρήσης, και τεχνολογικής πρακτικής. Η συναίνεση σταδιακά χάνει το ουσιαστικό της περιεχόμενο, ενώ οι χρήστες δυσκολεύονται να ασκήσουν τα δικαιώματά τους αποτελεσματικά. Για να αντιμετωπιστεί αυτό το ζήτημα δεν αρκεί η θέσπιση περαιτέρω τυπικών διαδικασιών συμμόρφωσης, αλλά χρειάζονται απαντήσεις σε βαθύτερα θεωρητικά και πρακτικά ερωτήματα σχετικά με το κατά πόσο τα υφιστάμενα μοντέλα προστασίας της αυτονομίας είναι ικανά για να ανταποκριθούν στις ιδιαιτερότητες των σύγχρονων μοντέλων.

6.2 Η ψευδαίσθηση της αποτελεσματικής ανωνυμοποίησης

Η ανωνυμοποίηση συχνά παρουσιάζεται ως μια μέθοδος προάσπισης της ιδιωτικότητας και ως μέσο αποδέσμευσης των δεδομένων από το κανονιστικό πλαίσιο προστασίας προσωπικών δεδομένων. Θεωρητικά, με αυτή τη διαδικασία αφαιρούνται ή τροποποιούνται τα άμεσα αναγνωριστικά στοιχεία, που ουσιαστικά επιτρέπουν τη σύνδεση των δεδομένων με το φυσικό πρόσωπο. Στην πράξη όμως, σε μεγαλύτερης κλίμακας περιβάλλοντα, στα οποία διαχειρίζονται τεράστια σύνολα δεδομένων, το νόημα της ανωνυμοποίησης τείνει να λειτουργεί περισσότερο ως νομική και οργανωτική παραδοχή παρά ως ουσιαστική εγγύηση ιδιωτικότητας. Γίνεται κατανοητό πως η ανωνυμοποίηση είναι μια εύθραυστη ισορροπία, άμεσα εξαρτημένη από τεχνικούς, εμπορικούς, και στατιστικούς παράγοντες.

Έρευνες που μελετούν την ανωνυμοποίηση στο πλαίσιο της μηχανικής μάθησης έχουν καταδείξει πως ακόμα και όταν αφαιρούνται άμεσα τα αναγνωριστικά στοιχεία, τα δεδομένα διατηρούν μοτίβα, αλληλουχίες, και στατιστικές ιδιότητες, μέσω των οποίων είναι εφικτή η επανασύνδεσή τους με τα υποκείμενα. Το γεγονός ότι χρησιμοποιούνται βοηθητικά σύνολα ενισχύει τη δυνατότητα συσχέτισης πολλαπλών πηγών, και τη χρήση τεχνικών ανάλυσης με σκοπό την επαναταυτοποίηση. Συνεπώς, η διαδικασία της ανωνυμοποίησης δεν εξαλείφει πλήρως τον κίνδυνο, αλλά τον μεταφέρει σε λιγότερο ορατά επίπεδα, παρουσιάζοντας εσφαλμένες εντυπώσεις επαρκούς προστασίας.

Το πρόβλημα αυτό εντείνεται με την ευρεία χρήση των μεγάλων γλωσσικών μοντέλων, καθώς δε βασίζονται στη στατική αποθήκευση πληροφοριών, αλλά στην εκμάθηση σύνθετων συσχετίσεων μεταξύ χαρακτηριστικών. Το γεγονός αυτό επιτρέπει στα μοντέλα να πραγματοποιούν συμπερασματικές ανακατασκευές, οπότε η διαδικασία της ανωνυμοποίησης παίρνει άλλη διάσταση. Ακόμη και στις περιπτώσεις που τα δεδομένα εισόδου θεωρούνται ανωνυμοποιημένα, το μοντέλο έχει τη δυνατότητα να καταλήξει στην έμμεση αποκάλυψη ευαίσθητων πληροφοριών. Η παραβίαση της ιδιωτικότητας πέραν της άμεσης αποκάλυψης ταυτότητας, μπορεί να προκύψει από τη δυνατότητα εξαγωγής ενός συνόλου συμπερασμάτων που σκιαγραφούν το υποκείμενο.



Εικόνα 8: Παράδειγμα εξαγωγής πληροφοριών τοποθεσίας από τα συμφραζόμενα ενός κειμένου.

Αξιοσημείωτο ενδιαφέρον αποκτά η ενσωμάτωση των συστημάτων τεχνητής νοημοσύνης στα εργαλεία ανωνυμοποίησης. Παρά τα πλεονεκτήματα που προσφέρει η αυτοματοποίηση της διαδικασίας απόκρυψης ευαίσθητων πληροφοριών, η ουσιαστική αποτελεσματικότητα της διαδικασίας είναι άμεσα εξαρτημένη από το πλαίσιο χρήσης, το είδος των δεδομένων, και από την ανθρώπινη εποπτεία. Εκτός του ότι η χρήση μεγάλων γλωσσικών μοντέλων στην ανωνυμοποίηση δεν εξαλείφει την απειλή της επαναταυτοποίησης, ανοίγει νέα παράθυρα ευπαθειών, όπως την εξαγωγή μη προβλέψιμου αποτελέσματος, αλλά και τη διατήρηση έμμεσων αναγνωριστικών πληροφοριών. Συνεπώς, ενώ σαν πρώτη εικόνα η χρήση τεχνητής νοημοσύνης σε τέτοιες διαδικασίες φαντάζει σωτήρια για την ουσιαστική προστασία της ιδιωτικότητας, τελικά ενισχύει την ψευδαίσθηση της ασφάλειας.

Σε ηθικό επίπεδο, όταν υπάρχει εμπιστοσύνη σε υπερβολικό βαθμό, ελοχεύει ο κίνδυνος χαλάρωσης των υπόλοιπων κρίσιμων εγγυήσεων ως προς την προστασία της ιδιωτικότητας. Στα σενάρια που τα δεδομένα έχουν ήδη χαρακτηριστεί ανωνυμοποιημένα, συχνά αντιμετωπίζονται με χαμηλή επικινδυνότητα, οπότε δεν τους δίνεται η προσοχή που τους αρμόζει τόσο στον έλεγχο πρόσβασης, όσο και στην αξιολόγηση κινδύνων. Αυτή η προσέγγιση δε λαμβάνει υπόψη το γεγονός ότι η διαδικασία της ανωνυμοποίησης είναι δυναμική και εξαρτώμενη από το τεχνολογικό περιβάλλον, το οποίο είναι συνεχώς μεταβαλλόμενο και ενδέχεται να μην είναι συμβατό με μελλοντικές τεχνικές επεξεργασίας.

Λαμβάνοντας υπόψη τα παραπάνω, είναι σαφές πως η ανωνυμοποίηση στα πλαίσια της τεχνητής νοημοσύνης και των μεγάλων γλωσσικών μοντέλων δεν πρέπει να αντιμετωπίζεται σαν πανάκεια λύση για την προστασία της ιδιωτικότητας, καθώς η λειτουργία της δεν ισοδυναμεί με την πλήρη αποτροπή κινδύνων. Αντιθέτως, συχνά δημιουργεί ψευδαισθήσεις ασφάλειας που δεν ανταποκρίνονται στην πραγματικότητα, αφού συνεχίζει να υπάρχει η πιθανότητα επαναταυτοποίησης και συμπερασματικής σκιαγράφησης ταυτότητας. Υπό αυτό το πρίσμα, η πλήρης κατανόηση της διαδικασίας της ανωνυμοποίησης αποτελεί κύρια προϋπόθεση για τη σωστή καταγραφή των ηθικοκοινωνικών συνεπειών που μπορεί να προκύψουν με τη χρήση ανωνυμοποιημένων δεδομένων στα σύγχρονα συστήματα τεχνητής νοημοσύνης.

6.3 Διάκριση, μεροληψίες και ευθύνη

Η διάκριση που συχνά εφαρμόζουν τα συστήματα τεχνητής νοημοσύνης δεν αποτελεί απαραίτητα σκόπιμη απόκλιση από τις θεμελιώδεις αρχές ισότητας. Πολλές φορές βρίσκεται ενσωματωμένη στις ίδιες τις διαδικασίες εκμάθησης, στα ίδια τα δεδομένα, στις ετικέτες τους, αλλά και στις μετρικές αξιολόγησης που χρησιμοποιούνται στον σχεδιασμό και την εκπαίδευση των μοντέλων. Από την άλλη, η μεροληψία είναι ένα συστημικό φαινόμενο, το

οποίο αναπαράγεται και αναπτύσσεται όσο τα μοντέλα λειτουργούν εντός πραγματικών κοινωνικών περιβαλλόντων. Αν και υπάρχει σύγχυση σε αυτό το θέμα, η ύπαρξη της μεροληψίας δεν προϋποθέτει την πρόθεση διάκρισης, καθώς είναι το αποτέλεσμα των επιλογών που γίνονται στα πρώτα στάδια ανάπτυξης του μοντέλου, και παραμένουν πολλές φορές αόρατες μέχρι και το τελικό αποτέλεσμα.

Έρευνες που μελετούν τη μεροληψία των μεγάλων γλωσσικών μοντέλων έχουν δείξει πως παρόλο που διαισθητικά ίσως φαίνεται να ισχύει, η κλίμακα και η πολυπλοκότητα δεν οδηγούν κατ' ανάγκη στην εξάλειψη της μεροληψίας. Αυτό συμβαίνει επειδή τα μοντέλα έχουν την τάση να αποτυπώνουν και να αναπαράγουν κυρίαρχα κοινωνικά μοτίβα και στερεότυπα, πολιτισμικές ή κοινωνικές ιεραρχίες, αλλά και ανισότητες που συναντούν στα δεδομένα εκπαίδευσης. Η μεροληψία μπορεί να εμφανιστεί με τη μορφή λεπτών διαφοροποιήσεων ύφους, προτεραιοποίησης, αλλά και συσχέτισης εννοιών. Δηλαδή, οι διακρίσεις πολλές φορές δε γίνονται αντιληπτές ως παραβιάσεις από τον απλό χρήστη, εφόσον παρουσιάζονται ως φαινομενικά ουδέτερη συμπεριφορά του συστήματος.

Για να αντιμετωπιστεί τεχνικά αυτού του είδους η συμπεριφορά του μοντέλου, δημιουργήθηκαν δείκτες δικαιοσύνης και νέες μετρικές αξιολόγησης. Πολλές φορές, όμως, παρουσιάζονται περιορισμοί. Τα κριτήρια δικαιοσύνης στηρίζονται σε παλιές υποθέσεις που δεν είναι πάντα συμβατές μεταξύ τους, καθώς επηρεάζονται σε μεγάλο βαθμό από λάθη επισήμανσης και μέτρησης στα δεδομένα. Ακόμα και σε περιπτώσεις που εφαρμόζονται καλή τη πίστει, αυτές οι προσεγγίσεις ενδέχεται να αποκρύπτουν μορφές διάκρισης που δεν είναι πάντα αντιληπτές με την πρώτη ματιά. Έτσι, η τεχνική συμμόρφωση μέσω ενός συνόλου μετρικών δεν αποτελεί εγγύηση της απομάκρυνσης της κοινωνικής βλάβης, γεγονός που αναδεικνύει τα όρια της υπολογιστικής αντίληψης της δικαιοσύνης στα μοντέλα.

Ιδιαίτερη βαρύτητα αποκτά η διάκριση όταν τα συστήματα τεχνητής νοημοσύνης ενσωματώνονται σε διαδικασίες λήψης αποφάσεων που μπορεί να έχουν κοινωνικές συνέπειες. Σε τέτοια περιβάλλοντα, τα μοντέλα δε λειτουργούν παίρνοντας μια απομονωμένη αλγοριθμική απόφαση, αλλά εντός ενός ευρύτερου κοινωνικού πλαισίου, όπου οι μεταβλητές ποικίλουν από απλούς ανθρώπινους φορείς, έως και συμφέροντα οργανισμών. Συνεπώς, το αποτέλεσμα της μεροληψίας δεν περιορίζεται στον μεμονωμένο χρήστη, αλλά κλιμακώνεται και παγιώνει υφιστάμενες ανισότητες. Η καταχώρηση της διάκρισης ως κοινωνικό φαινόμενο και όχι ως απλό τεχνικό σφάλμα μετατοπίζει το ενδιαφέρον στον τρόπο με τον οποίο οι συνέπειες του σφάλματος διαχέονται και σταθεροποιούνται.

Σε αυτό το σημείο αναδύεται το ζήτημα της ευθύνης. Η διάχυση της λήψης αποφάσεων μεταξύ του ανθρώπου και του μοντέλου δημιουργεί ένα κενό στη λογοδοσία, το οποίο είναι αρκετά εκμεταλλεύσιμο, καθώς ο σαφής καταμερισμός των ευθυνών καθίσταται δυσχερής. Αν και η μεροληψία μπορεί να προκύψει χωρίς την άμεση ανθρώπινη παρέμβαση στη στιγμή της απόφασης, συχνά μπορεί να διαμορφωθεί η εσφαλμένη αντίληψη πως η ευθύνη ανήκει αποκλειστικά στο σύστημα. Αντιθέτως, η ευθύνη δεν παραμένει στη λειτουργία του συστήματος, αφού εκτείνεται στις επιλογές σχεδιασμού, εκπαίδευσης, λειτουργίας, και χρήσης του σε οργανωτικά πλαίσια.

Με βάση τα ανωτέρω, η ηθική διάσταση της ευθύνης πρέπει να αντιμετωπίζεται ως εγγενές στοιχείο ανάπτυξης και αξιοποίησης των μεγάλων γλωσσικών μοντέλων, και όχι ως μεταγενέστερη προσθήκη. Η αναγνώριση της μεροληψίας ως ενός προβλέψιμου αποτελέσματος δε δικαιολογείται απλώς μέσω της επίκλησης της τεχνικής πολυπλοκότητας. Η ευθύνη πρέπει να μετατοπιστεί από την απλή αναζήτηση λαθών σε συγκεκριμένες εξόδους, σε μια πιο κριτική αποτίμηση των αποφάσεων που λαμβάνονται και καθιστούν τη διάκριση εφικτή. Μόνο με αυτόν τον τρόπο το φαινόμενο της μεροληψίας μπορεί να συνδεθεί άρρηκτα με τη λογοδοσία των φορέων που είναι υπεύθυνοι για τον σχεδιασμό, την ανάπτυξη, και τη λειτουργία των συστημάτων.

6.4 Διπλή χρήση και δημόσια ασφάλεια

Η ανάδυση των μεγάλων γλωσσικών μοντέλων έχει φέρει στο επίκεντρο της επιστημονικής συζήτησης την έννοια της διπλής χρήσης, καθώς η αρχιτεκτονική τους επιτρέπει να συνυπάρχουν νόμιμα ωφέλιμες εφαρμογές με κακόβουλες πρακτικές, οι οποίες είναι απειλητικές προς τη δημόσια ασφάλεια. Τα σύγχρονα μοντέλα, σε αντίθεση με τις προγενέστερες τεχνολογίες, δεν περιορίζονται σε προκαθορισμένες λειτουργίες, αλλά συνιστούν καινοτόμους μηχανισμούς γενικευμένης παραγωγής, οι οποίοι τους προσφέρουν την ικανότητα προσαρμογής σε νέα συμφραζόμενα σχεδόν χωρίς εξωτερική καθοδήγηση. Η γενικού σκοπού φύση των μοντέλων, σε συνδυασμό με την υψηλή διαθεσιμότητα και την ευχέρεια παραγωγής εξαιρετικά αληθοφανούς περιεχομένου, δυσχεραίνει την αναγνώριση ορίων ανάμεσα στη θεμιτή και την κακόπιστη χρήση τους. Έρευνες αναφέρουν πως η υπερβολική προσαρμοστικότητα, αν και εκ πρώτης όψεως φαίνεται ιδανική, οδηγεί σε δομική απώλεια ελέγχου του τελικού σκοπού χρήσης, διευρύνοντας το χάσμα μεταξύ της αρχικής πρόθεσης του σχεδιαστή και της τελικής αξιοποίησης των χρηστών. Κατά συνέπεια, η διπλή χρήση δεν αναγνωρίζεται πλέον σαν μια τεχνική ανωμαλία, αλλά ως ένα δομικό χαρακτηριστικό, που δημιουργεί ένα γόνιμο περιβάλλον για κακόβουλη εκμετάλλευση χωρίς να προϋποθέτει εξειδικευμένη γνώση, ούτε προνομιακή πρόσβαση από τον κακόβουλο χρήστη.

Ένα επίσης καίριο ζήτημα που αναδεικνύεται αφορά τη διεύρυνση της πρόσβασης σε γνώση υψηλού κινδύνου, εφόσον τα μεγάλα γλωσσικά μοντέλα μπορούν να λειτουργήσουν ως πολλαπλασιαστές ικανότητας, ελαττώνοντας τα εμπόδια κατανόησης σύνθετων επιστημονικών διαδικασιών, όπως συμβαίνει στον χώρο της βιοτεχνολογίας. Η λεγόμενη δημοκρατικοποίηση της πρόσβασης, που συχνά παρουσιάζεται ως θετική εξέλιξη, πρακτικά μετατρέπει την υπολογιστική ισχύ σε ζήτημα δημόσιας ασφάλειας, εφόσον η παραγόμενη γνώση μπορεί να χρησιμοποιηθεί σε μη ελεγχόμενα πλαίσια, επιτρέποντας σε μη ειδικούς την πρόσβαση σε επικίνδυνους παράγοντες. Αντίστοιχη δυναμική παρατηρείται στο θέμα της κοινωνικής επιρροής, όπου τα ίδια συστήματα έχουν ξεκινήσει να αναπτύσσονται και να εντοπίζουν παραπλανητικό περιεχόμενο, αλλά και να χρησιμοποιούνται για μαζική παραγωγή αληθοφανών ψεύτικων απαντήσεων. Τέτοιου είδους δυνατότητες ενισχυμένες από την αυξημένη ταχύτητα διάδοσης και προσαρμογής σε ένα μεγάλο σύνολο ακροατηρίων, αυξάνουν τους κινδύνους αποσταθεροποίησης των κοινωνικοπολιτικών δομών, καθιστώντας τη διπλή χρήση κρίσιμο ζήτημα, το οποίο αφορά ολόκληρη την τεχνολογική κοινότητα, και τον έλεγχο της πληροφορίας στη δημόσια σφαίρα.

Σε πιο συστηματικό επίπεδο, η ανάλυση κινδύνων έχει υπογραμμίσει πως οι απειλές που σχετίζονται με τη διπλή χρήση σπάνια εμφανίζονται μεμονωμένα, καθώς απορρέουν από εγγενείς αδυναμίες σχεδιασμού και λειτουργικού πλαισίου των συστημάτων. Η δυνατότητα αναπαραγωγής και η ευρεία ελεύθερη πρόσβαση σε προεκπαιδευμένα μοντέλα μειώνουν τη δυσκολία πρόσβασης, ενώ ενισχύουν τους παράγοντες ρίσκου. Επιπρόσθετα, η ολοένα αυξανόμενη πολυπλοκότητα των συστημάτων δημιουργεί ασάφεια ως προς τη σωστή κατανομή ευθυνών μεταξύ ερευνητών, φορέων, και χρηστών καταλήγοντας σε ένα κενό ευθύνης που δυσχεραίνει τη διαχείριση των επιπτώσεων. Υπό αυτή την οπτική, η διπλή χρήση δεν αποτελεί αποκλειστικά ζήτημα πρόθεσης, εφόσον σημαντικό ποσοστό κινδύνων προκύπτει από συγκεκριμένες δομικές ή σχεδιαστικές αποφάσεις.

Μελέτες έχουν καταλήξει στο ότι η αποτελεσματική αντιμετώπιση των κινδύνων διπλής χρήσης δεν μπορεί να στηριχτεί σε αποσπασματικές παρεμβάσεις. Χρειάζεται μια νέα πολυδιάστατη προσέγγιση που θα ενσωματώνει ηθικά πλαίσια και εντατική παρακολούθηση. Ιδιαίτερη σημασία έχει η εφαρμογή ενός πλέγματος πρόληψης, το οποίο έρχεται να καλύψει το κενό που υπάρχει στη διαχείριση των απειλών. Ουσιαστικά, θα διατρέχει ολόκληρο τον κύκλο ζωής της έρευνας, συμπεριλαμβάνοντας διαφορετικούς

ενδιαφερόμενους φορείς. Έτσι, η έννοια της υπεύθυνης τεχνητής νοημοσύνης καθίσταται κεντρική, διασφαλίζοντας έννοιες όπως η διαφάνεια, η λογοδοσία, και η δικαιοσύνη, ώστε να εξαλείφονται οι κίνδυνοι κατάχρησης πριν καν αυτοί κλιμακωθούν. Ταυτόχρονα, επιβεβαιώνεται η ανάγκη δημιουργίας δυναμικών βάσεων δεδομένων και ταξινομιών, οι οποίες θα επιτρέπουν την ακριβή αναγνώριση παραγόντων και τομέων κινδύνου, όπως, ενδεικτικά, η παραπληροφόρηση. Συνεπώς, η διπλή χρήση παύει να θεωρείται ένα απλό τεχνικό ζήτημα, και αναγνωρίζεται σαν μια ηθικοκοινωνική πρόκληση που απαιτεί διεπιστημονική συνεργασία και συνεχή παρακολούθηση.

7 ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΑΠΕΙΛΩΝ ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΕΣ ΜΕΘΟΔΟΙ ΕΛΕΓΧΟΥ ΑΣΦΑΛΕΙΑΣ ΓΙΑ ΜΕΓΑΛΑ ΓΛΩΣΣΙΚΑ ΜΟΝΤΕΛΑ

7.1 Ορισμός προτύπων απειλών για ερευνητικά σενάρια αξιολόγησης

Η μοντελοποίηση απειλών στον κυβερνοχώρο ορίζεται ως η διαδικασία σχεδιασμού και εφαρμογής μιας ρεαλιστικής αναπαράστασης των ανταγωνιστικών απειλών, με σκοπό την ενημέρωση των προσπαθειών που αφορούν την ασφάλεια και την ανθεκτικότητα. Για τον σωστό καθορισμό των προτύπων στα ερευνητικά σενάρια, χρειάζεται να αποδομηθεί το σύστημα μέσω διαγραμμάτων ροής δεδομένων, τα οποία βοηθούν στην ανάλυση του συστήματος σε βασικά στοιχεία, όπως είναι οι εξωτερικοί παράγοντες, οι διαδικασίες που ακολουθεί, οι αποθήκες και οι ροές δεδομένων. Για την ανάλυση αυτή ιδιαίτερα χρήσιμος είναι ο προσδιορισμός των ορίων εμπιστοσύνης, τα οποία ουσιαστικά αντιπροσωπεύουν τη μετάβαση ανάμεσα στα διαφορετικά επίπεδα εμπιστοσύνης, όπως για παράδειγμα από έναν μη αυθεντικοποιημένο χρήστη σε έναν που έχει αυθεντικοποιηθεί, στοιχείο πολύ χρήσιμο για τον έγκαιρο εντοπισμό πιθανών κινδύνων ασφαλείας.

Για τη κατηγοριοποίηση των απειλών χρησιμοποιείται ευρέως το μοντέλο STRIDE, ώστε να γίνει συστηματικά η εξερεύνηση των κατηγοριών που έχουν σχέση με το στοχευμένο σύστημα. Πρακτικά, το STRIDE ταξινομεί τις απειλές σε έξι κατηγορίες:

- Πλαστοπροσωπία (Spoofing): αφορά την παραβίαση της αυθεντικότητας
- Παραποίηση (Tampering): αφορά την παραβίαση της ακεραιότητας
- Άρνηση ευθύνης (Repudiation): σχετίζεται με την αδυναμία απόδειξης ενεργειών για την αποποίηση ευθύνης
- Αποκάλυψη πληροφοριών (Information Disclosure): αφορά την παραβίαση της εμπιστευτικότητας
- Άρνηση υπηρεσίας (Denial of Service): αφορά την παραβίαση της διαθεσιμότητας
- Ανύψωση προνομίων (Elevation of Privilege): που σχετίζεται με την απόκτηση μη εξουσιοδοτημένων προνομίων

Με την εφαρμογή του STRIDE οι αναλυτές μπορούν να εξετάζουν μεθοδικά όλα τα στοιχεία των μοντέλων, ώστε να εντοπίζουν πού και πώς θα μπορούσαν να υλοποιηθούν αυτές οι απειλές.

Για την ενίσχυση της εγκυρότητας και της δυνατότητας δράσης των ερευνητικών σεναρίων, τα πρότυπα απειλών συχνά ταυτίζονται με καθιερωμένες ταξινομίες επιθετικών μοτίβων, όπως είναι η CAPEC (Common Attack Pattern Enumeration and Classification). Έτσι, οι αναλυτές μπορούν να χρησιμοποιούν τις ήδη εντοπισμένες απειλές στη χαρτογράφησή τους σε γνωστά μοτίβα επίθεσης, και να εμπλουτίζουν τα αποτελέσματά τους με περαιτέρω λεπτομέρειες, όπως η πιθανότητα εμφάνισης, η σοβαρότητα, αλλά και προτεινόμενες στρατηγικές καταπολέμησης. Σύγχρονα εργαλεία έχουν αναπτυχθεί με τέτοιο τρόπο που έχουν τη δυνατότητα να αξιοποιούν τη σημασιολογική ανάκτηση και τις διανυσματικές ενσωματώσεις, ώστε να υπολογίζουν την ομοιότητα συνημιτόνου μεταξύ των περιγραφών

των απειλών και των αναφορών της CAPEC μεθοδολογίας, για να συνδέσουν τη θεωρητική ανάλυση με την τεκμηριωμένη γνώση, αποφεύγοντας αυθαίρετες ή αποσπασματικές υποθέσεις.

Στο ειδικότερο πεδίο της μηχανικής μάθησης και των μεγάλων γλωσσικών μοντέλων, τα πρότυπα απειλών επεκτείνονται για να καλύψουν μοναδικές ευπάθειες που δεν εντοπίζονται στα παραδοσιακά συστήματα λογισμικού. Συγκεκριμένα, οι επιθέσεις κατηγοριοποιούνται με βάση το επίπεδο γνώσης του επιτιθέμενου όσον αφορά το μοντέλο και τα δεδομένα που επεξεργάζεται, καθώς και με βάση την επιφάνεια της επίθεσης κατά τη διάρκεια του κύκλου ζωής του συστήματος. Σε αυτό το πλαίσιο, οι κύριες απειλές που εξετάζονται στοχεύουν την παραβίαση της εμπιστευτικότητας, της ακεραιότητας, και της διαθεσιμότητας του μοντέλου, όπως η αλλοίωση δεδομένων εκπαίδευσης, η χειραγώγηση εισόδων κατά τη χρήση του, και η έμμεση αποκάλυψη ευαίσθητων δεδομένων. Για τα μεγάλα γλωσσικά μοντέλα, ιδιαίτερη σημασία έχουν τα σενάρια που αφορούν την παράκαμψη περιορισμών, τη χειραγώγηση εξόδων, και τη μη εξουσιοδοτημένη ανάκτηση προσωπικών δεδομένων.

Για την αντιμετώπιση της πολυπλοκότητας και του κόστους της χειροκίνητης ανάλυσης, η έρευνα στρέφεται προς μια πιο αυτοματοποιημένη προσέγγιση της μοντελοποίησης απειλών, με τη χρήση των ίδιων των μεγάλων γλωσσικών μοντέλων. Προσεγγίσεις όπως το ThreatModeling LLM εφαρμόζουν ειδικές τεχνικές μηχανικής προτροπών, όπως η αλυσίδα σκέψης, και η βελτιστοποίηση προτροπών, με σκοπό να καθοδηγούν το μοντέλο στον εντοπισμό των απειλών και στην εύρεση λύσεων βήμα βήμα. Χρήσιμες είναι επίσης οι τεχνικές τύπου αλυσίδας προτροπών, οι οποίες χρησιμοποιούνται για τη διάσπαση μιας σύνθετης διαδικασίας σε απλούστερα καθήκοντα, όπως είναι η αποδόμηση του συστήματος, έπειτα η αναγνώριση απειλών, και τέλος η δομημένη παραγωγή αντικειμένων. Με αυτόν τον τρόπο, το μοντέλο μπορεί να εστιάζει κάθε φορά σε ένα πιο περιορισμένο πλαίσιο αυξάνοντας ραγδαία την αποτελεσματικότητά του.

Τέλος, για να αξιολογηθούν αυτά τα ερευνητικά σενάρια χρησιμοποιούνται ποσοτικά και ποιοτικά κριτήρια. Για τα ποσοτικά αξιοποιούνται μετρικές των μοντέλων όπως η ακρίβεια, η ανάκληση, και η ομοιότητα κειμένου, μέσω της χρήσης μοντέλων BERT, καθώς συγκρίνονται οι παραγόμενες απειλές με τα καθιερωμένα σύνολα δεδομένων. Για τα ποιοτικά κριτήρια, η αξιολόγηση των μοντέλων εφαρμόζεται με κριτήρια όπως η κάλυψη, η ακρίβεια, και η λεπτομέρεια, συγκρίνοντας τα αποτελέσματα με μελέτες ειδικών στην κυβερνοασφάλεια. Για να προκύψει το συνολικό συμπέρασμα, η αξιολόγηση εξετάζει επιπλέον χαρακτηριστικά, όπως η προδιαγραφή, η πληρότητα κάλυψης του τομέα απειλών, και τον βαθμό στον οποίο το μοντέλο είναι ικανό να γίνει συγκεκριμένο.

7.2 Σχεδιασμός ασκήσεων ελεγχόμενης επίθεσης για παρεμβολή οδηγιών

Ο σχεδιασμός ασκήσεων ελεγχόμενης επίθεσης με σκοπό την αξιολόγηση της ανθεκτικότητας των μεγάλων γλωσσικών μοντέλων απαιτεί τη μετάβαση από τη θεωρητική μοντελοποίηση απειλών στην περιορισμένη προσομοίωση επιθετικής συμπεριφοράς. Βασικό ζητούμενο σε αυτές τις ασκήσεις αποτελεί η διερεύνηση της ευπάθειας παρεμβολής οδηγιών (Prompt Injection), η οποία διαχωρίζεται σε άμεση και έμμεση. Στην άμεση επίθεση, ο σχεδιασμός προβλέπει την εισαγωγή κακόβουλων εντολών απευθείας από τον χρήστη με σκοπό την παράκαμψη των περιορισμών του συστήματος, ενώ στην έμμεση επίθεση, το μοντέλο καλείται να επεξεργαστεί μολυσμένα δεδομένα από τρίτες πηγές, όπως ιστοσελίδες και έγγραφα, τα οποία περιέχουν κρυμμένες οδηγίες χειραγώγησης, με τελικό στόχο την αλλοίωση της ροής ελέγχου.

Για να δομηθεί ορθά το πειραματικό πλαίσιο χρειάζεται σαφής προσδιορισμός του επιπέδου γνώσης του επιτιθέμενου όσον αφορά τον στόχο του, μέσω της διάκρισης των σεναρίων σε λευκό και μαύρο κουτί. Σύμφωνα με τη μεθοδολογία ελέγχου ανθεκτικότητας, στα σενάρια λευκού κουτιού υποθέτουμε πως ο επιτιθέμενος έχει πλήρη πρόσβαση στις παραμέτρους

και τις κλίσεις του μοντέλου, επιτρέποντας τον υπολογισμό των βέλτιστων ανταγωνιστικών διαταραχών. Αντιθέτως, τα σενάρια μαύρου κουτιού είναι πιο ρεαλιστικά, καθώς ο σχεδιασμός περιορίζεται στην πρόσβαση μέσω API, όπου η επίθεση πρέπει να βασιστεί στη μεταφερσιμότητα ανταγωνιστικών δειγμάτων από άλλα μοντέλα, προσομοιώνοντας συνθήκες που αντιμετωπίζει ένας πραγματικός επιτιθέμενος σε εμπορικές εφαρμογές.

Κεντρικό ρόλο στον σχεδιασμό των ασκήσεων διαδραματίζει ο μηχανισμός παραγωγής των επιθετικών εισόδων. Αντί να βασίζεται αποκλειστικά σε χειροκίνητες προσπάθειες, το πλαίσιο μπορεί να ενσωματώνει αυτοματοποιημένες μεθόδους βελτιστοποίησης, όπως η τεχνική Greedy Coordinate Gradient (GCG). Αυτή η μέθοδος αναζητά αυτόματα μια ακολουθία χαρακτήρων, η οποία όταν προσαρτάται σε μια απαγορευμένη ερώτηση εκτοξεύει την πιθανότητα το μοντέλο να δώσει μια καταφατική μη επιτρεπτή απόκριση. Ο σχεδιασμός τέτοιων ασκήσεων οφείλει να προβλέπει τη δημιουργία καθολικών επιθετικών προτύπων, τα οποία έχουν αποδειχθεί ικανά να παρακάμψουν τους μηχανισμούς ευθυγράμμισης ακόμα και σε διαφορετικές αρχιτεκτονικές μοντέλων, αποκαλύπτοντας βαθύτερες δομικές αδυναμίες της διαδικασίας εκπαίδευσης.

Η εγκυρότητα των ασκήσεων εξαρτάται επιπλέον από τον σωστό καθορισμό των στόχων επίθεσης και του πλαισίου εφαρμογής τους. Μελέτες σε εξειδικευμένους τομείς, όπως η υγειονομική περίθαλψη, καταδεικνύουν ότι ο σχεδιασμός πρέπει να περιέχει σαφώς καθορισμένα σενάρια στα οποία ελέγχεται η ακεραιότητα των απαντήσεων έναντι κακόβουλων παρεμβολών. Για παράδειγμα, ένα σενάριο ελέγχου μπορεί να εξετάζει αν το μοντέλο εξαναγκάζεται να παρέχει ανακριβείς ιατρικές συμβουλές ή να ταξινομεί λανθασμένα καίριες πληροφορίες. Η άσκηση οφείλει να προσομοιώνει ρεαλιστικές αλληλεπιδράσεις, στις οποίες το μοντέλο καλείται να διαχειριστεί αντικρουόμενες οδηγίες μεταξύ του αρχικού πλαισίου λειτουργίας και της εισόδου του χρήστη, αξιολογώντας την αλλαγή της συμπεριφοράς του και τη συμμόρφωσή του με τους κανόνες ασφαλείας.

Για την ποσοτική αποτίμηση των αποτελεσμάτων της άσκησης, χρειάζεται ο ορισμός μετρήσιμων δεικτών επιτυχίας, με κυριότερο τον ρυθμό επιτυχίας επίθεσης (Attack Success Rate). Κατά τον σχεδιασμό της αξιολόγησης πρέπει να περιλαμβάνονται αυτοματοποιημένοι μηχανισμοί ελέγχου, οι οποίοι θα εξετάζουν αν η έξοδος του μοντέλου παραβιάζει τους κανόνες ασφαλείας ή αν ενεργοποιείται επιτυχώς ο μηχανισμός άρνησης. Αυτή η αξιολόγηση μπορεί να πραγματοποιηθεί μέσω τεχνικών αντιστοίχισης συμβολοσειρών για τον εντοπισμό τυπικής φρασεολογίας άρνησης, ή μέσω της χρήσης ενός ισχυρότερου μεγάλου γλωσσικού μοντέλου ως κριτή, το οποίο θα βαθμολογεί τη βλαπτικότητα της απάντησης. Μέσω αυτής της δομημένης διαδικασίας μπορεί να εξαχθεί ένα αξιόπιστο συμπέρασμα για το επίπεδο ανθεκτικότητας του εξεταζόμενου συστήματος.

7.3 Σχεδιασμός επαναλαμβανόμενων, επαληθεύσιμων επιθετικών σεναρίων

Ο σχεδιασμός επιθετικών σεναρίων κατά των μεγάλων γλωσσικών μοντέλων δεν μπορεί να αντιμετωπίζεται ως αποσπασματική εμπειρική διαδικασία, αλλά πρέπει να στηρίζεται σε μια αυστηρά ορισμένη επιστημονική μεθοδολογία, με συγκεκριμένα κριτήρια εγκυρότητας και επικύρωσης. Μελέτες υπογραμμίζουν την ανάγκη διάκρισης μεταξύ της επαναληψιμότητας, η οποία αφορά τη λήψη συνεπών αποτελεσμάτων από την ίδια ερευνητική ομάδα υπό αμετάβλητες πειραματικές συνθήκες, και της αναπαραγωγιμότητας, που αναφέρεται στη δυνατότητα ανεξάρτητων ερευνητών να επιβεβαιώνουν την ορθότητα των ευρημάτων από εξωτερικούς ερευνητές, ακολουθώντας ισοδύναμο πρωτόκολλο. Στο πλαίσιο των μεγάλων γλωσσικών μοντέλων, η εγγενής στοχαστικότητα της παραγωγής απαντήσεων διαμορφώνει ένα πιο σύνθετο περιβάλλον με πρόσθετες προκλήσεις, αφού η μεταβλητότητα των απαντήσεων συχνά οφείλεται σε τυχαίους μηχανισμούς δειγματοληψίας και όχι από ουσιαστικές αλλαγές στη συμπεριφορά του μοντέλου. Έτσι, για να απομονωθούν οι επιπτώσεις της στοχαστικότητας, πρέπει η πειραματική διαδικασία να σχεδιάζεται με τέτοιο

τρόπο που περιλαμβάνονται επαναλαμβανόμενες εκτελέσεις σε ελεγχόμενο περιβάλλον, χωρίς ενδιάμεσες ενημερώσεις στις παραμέτρους του μοντέλου.

Για έναν πιο ώριμο σχεδιασμό σεναρίων, χρειάζεται να ληφθούν υπόψη οι πυλώνες της αξιόπιστης μηχανικής μάθησης, ώστε να αναγνωριστεί πως οι έννοιες της στιβαρότητας, της δικαιοσύνης, και της ιδιωτικότητας δεν εξετάζονται μεμονωμένα, αλλά συνυπάρχουν σε ένα πλέγμα ανταγωνιστικών απαιτήσεων. Ένα επιθετικό σενάριο θεωρείται επαληθεύσιμο όταν μοντελοποιεί ρεαλιστικά και με σαφήνεια τον αντίπαλο, υπολογίζοντας το επίπεδο γνώσης και πρόσβασής του στο σύστημα, και στοχεύει σε συγκεκριμένες κατηγορίες ευπαθειών, όπως η δηλητηρίαση δεδομένων ή η μη εξουσιοδοτημένη ανάκτηση ευαίσθητων δεδομένων. Η ενσωμάτωση αιτιακών μοντέλων επιτρέπει την απάντηση σε αντιπαραδειγματικές ερωτήσεις και συμβάλλει σε μια πιο συνεκτική αξιολόγηση της ανθεκτικότητας, πέρα από απλές εμπειρικές παρατηρήσεις.

Για την πρακτική υλοποίηση των επιθετικών σεναρίων υιοθετείται η μεθοδολογία των καθολικών επιθετικών φράσεων (universal adversarial phrases), η οποία επιδιώκει την ανακάλυψη γενικευμένων μοτίβων επίθεσης, αντί για εξειδικευμένες, μεμονωμένες παρεμβάσεις. Χρησιμοποιώντας αλγορίθμους άπληστης αναζήτησης εντοπίζονται σύντομες ακολουθίες λέξεων, οι οποίες όταν προσαρτώνται στο κείμενο εισόδου, αυξάνουν δραστικά την πιθανότητα παραβίασης των πολιτικών ασφαλείας του συστήματος ή στην υποβάθμιση της ποιότητας της εξόδου. Ιδιαίτερα σημαντικό ρόλο σε αυτή τη διαδικασία παίζουν τα μοντέλα υποκατάστασης, καθώς χρησιμεύουν για την εκμάθηση της επιθετικής φράσης όταν δεν είναι διαθέσιμη η άμεση πρόσβαση στο μοντέλο στόχο. Αυτή η μεταφορά της επίθεσης από το υποκατάστατο στο τελικό μοντέλο αξιοποιεί τη δυνατότητα της μεταφερσιμότητας των επιθετικών μοτίβων, και έχει αποδειχθεί αρκετά αποτελεσματική ακόμα και σε συστήματα με διαφορετικές αρχιτεκτονικές.

Προκειμένου να διασφαλιστεί η επαρκής κάλυψη και η ποιότητα των δεδομένων στις επιθετικές εισόδους, εφαρμόζεται η στρατηγική των αλυσιδωτών εντολών, κατά την οποία η παραγωγή των σεναρίων διασπάται σε μικρότερα διακριτά στάδια. Με αυτή την προσέγγιση το μοντέλο μπορεί να λαμβάνει υπόψη το πλαίσιο, να επανεξετάζει τις παραγόμενες προτροπές του, και να τις βελτιστοποιεί με στόχο τη μεγιστοποίηση της επιθετικής αποτελεσματικότητας. Επιπλέον, υιοθετείται μια ανταγωνιστική διάταξη ρόλων, με την οποία το σύστημα λειτουργεί ως γεννήτρια, αλλά και ως αξιολογητής των σεναρίων. Πρακτικά, ο αξιολογητής εκτελεί τον έλεγχο ποιότητας, βελτιστοποιεί, αφαιρεί περιττούς πλεονασμούς, και εξασφαλίζει πως τα τελικά σύνολα εξόδων έχουν συντακτική ορθότητα, και σημασιολογική επάρκεια για την πυροδότηση των υπό εξέταση ευπαθειών.

Η επαληθευσιμότητα των αποτελεσμάτων υποστηρίζεται από την αυτοματοποιημένη αξιολόγηση, χρησιμοποιώντας το πρότυπο του μεγάλου γλωσσικού μοντέλου ως κριτή. Η προσέγγιση αυτή επιτρέπει την αξιόπιστη εκτέλεση των πειραμάτων και την τυποποιημένη αποτίμηση της συμπεριφοράς του εξεταζόμενου μοντέλου. Για να είναι ακριβής η αξιολόγηση αξιοποιούνται κυρίως εξειδικευμένα σύνολα δεδομένων, τα οποία περιλαμβάνουν βασικές και επιθετικά διαμορφωμένες προτροπές, ώστε να δοκιμάζουν συστηματικά τα όρια ανθεκτικότητας του συστήματος. Για την ποσοτικοποίηση των αποτελεσμάτων χρησιμοποιούνται μετρικές που αφορούν τη στιβαρότητα, τη δικαιοσύνη και την ασφάλεια, ενώ ταυτόχρονα εφαρμόζονται φίλτρα για να αποκλείονται περιπτώσεις άρνησης απάντησης που οφείλονται σε παρερμηνεία της εντολής και όχι στην αποτελεσματική εφαρμογή μηχανισμών προστασίας. Έτσι διασφαλίζεται ότι τα συμπεράσματα αντανακλούν τις πραγματικές ιδιότητες ανθεκτικότητας και όχι τεχνητές στρεβλώσεις της διαδικασίας αξιολόγησης.

8 ΤΕΧΝΙΚΕΣ ΠΡΟΣΤΑΣΙΑΣ ΚΑΙ ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ ΙΔΙΩΤΙΚΟΤΗΤΑΣ

8.1 Θεωρία και εφαρμογή της διαφορικής ιδιωτικότητας στην εκπαίδευση και προσαρμογή μοντέλων

Η διαφορική ιδιωτικότητα αποτελεί το θεμελιώδες θεωρητικό και πρακτικό πλαίσιο για την προστασία των προσωπικών δεδομένων κατά την εκπαίδευση συστημάτων μηχανικής μάθησης. Η βασική αρχή είναι η παροχή αυστηρών μαθηματικών εγγυήσεων σύμφωνα με τις οποίες τα αποτελέσματα ενός μοντέλου παραμένουν αμετάβλητα ανεξαρτήτως της συμμετοχής ή μη των δεδομένων ενός συγκεκριμένου ατόμου στο σύνολο εκπαίδευσης. Με αυτόν τον τρόπο περιορίζεται σημαντικά η δυνατότητα εξαγωγής συμπερασμάτων που συνδέονται με μεμονωμένους χρήστες ή την ανακατασκευή ευαίσθητου περιεχομένου χρησιμοποιώντας επιθέσεις συμπερασμού ή αναγνώρισης συμμετοχής.

Κατά τη διάρκεια της εκπαίδευσης των νευρωνικών δικτύων, η διαφορική ιδιωτικότητα εφαρμόζεται κυρίως μέσω παραλλαγών του αλγορίθμου στοχαστικής καθόδου κλίσης, ή αλλιώς DP-SGD. Αυτή η προσέγγιση εισάγει νέες παρεμβάσεις σε καίρια σημεία της εκπαιδευτικής διαδικασίας, περιορίζοντας εκ πρώτης τη μέγιστη επίδραση κάθε δείγματος στα υπολογιζόμενα διανύσματα κλίσης, και στη συνέχεια προσθέτοντας ελεγχόμενο τυχαίο θόρυβο. Συνδυαστικά αυτοί οι δύο μηχανισμοί καθιστούν τις ενημερώσεις του μοντέλου στατιστικά ανακριβείς ως προς τη συνεισφορά συγκεκριμένων προσωπικών δεδομένων, δημιουργώντας ένα περιβάλλον στο οποίο ένας επιτιθέμενος δε μπορεί να ξεχωρίσει τα πραγματικά δεδομένα από τις τυχαίες διακυμάνσεις, προσφέροντας ένα πρώτο επίπεδο εγγύησης της ιδιωτικότητας.

Έρευνες έχουν δείξει πως το επίπεδο στο οποίο εφαρμόζονται τα μέτρα προστασίας στα μεγάλα γλωσσικά μοντέλα παίζει κρίσιμο ρόλο στην αποτελεσματικότητά τους. Η κλασική εκδοχή της διαφορικής ιδιωτικότητας εστιάζει στην προστασία μεμονωμένων περιπτώσεων, αν και στην πράξη οι χρήστες συνεισφέρουν συχνά ισχυρά συσχετισμένα δεδομένα. Ωστόσο, η προστασία σε επίπεδο παραδείγματος αποδεικνύεται ανεπαρκής, αφού επιτρέπει τη συσσωρευτική αποκάλυψη πληροφοριών για τον χρήστη. Γι’ αυτό τον λόγο, πρόσφατες προσεγγίσεις μετατοπίζουν την προσοχή τους στην προστασία σε επίπεδο χρήστη, διασφαλίζοντας ότι τα δεδομένα που συνδέονται με ένα υποκείμενο αντιμετωπίζονται ως μια ενιαία οντότητα ως προς τις εγγυήσεις ιδιωτικότητας. Παρ’ όλο που για την εφαρμογή αυτής της στρατηγικής απαιτείται αυξημένη υπολογιστική ισχύς και πιο αυστηροί περιορισμοί, αντιπροσωπεύει με υψηλότερη ακρίβεια τις πραγματικές απαιτήσεις των σύγχρονων εφαρμογών.

Παρά τα πλεονεκτήματά της, η διαφορική ιδιωτικότητα συνοδεύεται από ένα βασικό και αναπόφευκτο αντιστάθμισμα: όσο αυξάνονται τα μέτρα προστασίας της ιδιωτικότητας, τόσο μειώνεται η ακρίβεια και η χρηστικότητα του μοντέλου. Η αύξηση του επιπέδου προστασίας μέσω της χρήσης ισχυρότερου θορύβου συχνά οδηγεί στην υποβάθμιση της ακρίβειας. Μελέτες έχουν αποδείξει πως σε αρκετές περιπτώσεις, οι παράμετροι ιδιωτικότητας λαμβάνουν τιμές, οι οποίες καθιστούν τις εγγυήσεις περισσότερο θεωρητικές παρά

ουσιαστικές, καθώς για εμπορικούς λόγους στόχος είναι να διατηρηθεί η απόδοση σε σταθερό επίπεδο. Το γεγονός αυτό έχει προκαλέσει την αναζήτηση νέων πρακτικών συμβιβαστικών λύσεων, όπως για παράδειγμα η αξιοποίηση δημόσιων δεδομένων στα πρώτα στάδια εκπαίδευσης, για να περιοριστεί η επίπτωση του θορύβου στη φάση προσαρμογής.

Για την αντιμετώπιση αυτών των προβλημάτων προτείνονται πιο στοχευμένες στρατηγικές, με χαρακτηριστικό παράδειγμα την επιλεκτική διαφορική ιδιωτικότητα. Ουσιαστικά, η συγκεκριμένη μέθοδος εστιάζει στην προστασία μόνο των πραγματικά ευαίσθητων στοιχείων και όχι σε ολόκληρο το κείμενο. Χρησιμοποιεί πολυφασικές διαδικασίες προσαρμογής, χωρίζοντας την εκπαιδευτική διαδικασία σε στάδια γενικής και ευαίσθητης μάθησης, επιτυγχάνοντας πιο αποτελεσματική ισορροπία μεταξύ ιδιωτικότητας και χρηστικότητας χωρίς να εκτίθενται τα προσωπικά δεδομένα των χρηστών. Συνεπώς, αυτές οι στρατηγικές προσεγγίζουν με περισσότερη ωριμότητα την εφαρμογή της διαφορικής ιδιωτικότητας στα μεγάλα γλωσσικά μοντέλα, καθώς προσαρμόζονται στις πρακτικές απαιτήσεις των σύγχρονων συστημάτων.

8.2 Ομοσπονδιακή μάθηση και κατανεμημένη ιδιωτικότητα

Η ομοσπονδιακή μάθηση εισάγει ένα αποκεντρωμένο εκπαιδευτικό πρότυπο, το οποίο προβλέπει τη συνεργατική εκπαίδευση μοντέλων μηχανικής μάθησης χωρίς την κεντρική συγκέντρωση των πρωτογενών δεδομένων των χρηστών. Αντιθέτως με τα κλασικά κεντρικοποιημένα σχήματα, η διαδικασία εκμάθησης πραγματοποιείται τοπικά στις συσκευές των συμμετεχόντων, μεταδίδοντας στον κεντρικό διακομιστή αποκλειστικά τις ενημερώσεις των παραμέτρων ή ενδιάμεσες αναπαραστάσεις. Με αυτή την αρχιτεκτονική, η προστασία της ιδιωτικότητας γίνεται με ουσιαστικό τρόπο, εφόσον περιορίζεται σημαντικά η μεταφορά ευαίσθητων δεδομένων και ενσωματώνεται η αρχή της ελαχιστοποίησης της αποκάλυψης σε επίπεδο σχεδιασμού του συστήματος.

Στο πλαίσιο των μεγάλων γλωσσικών μοντέλων, η ομοσπονδιακή μάθηση εφαρμόζεται κυρίως με σχήματα ομοσπονδιακής μικρορύθμισης, στα οποία οι συμμετέχοντες ρυθμίζουν τοπικά ένα προεκπαιδευμένο θεμελιώδες μοντέλο, αντί να εκπαιδεύουν ένα νέο από την αρχή κάθε φορά. Έτσι, το μοντέλο μπορεί να διατηρεί τη γενικευμένη γνώση του, ενώ ταυτόχρονα περιορίζεται η έκθεσή του στα τοπικά προσωπικά δεδομένα. Παρ' όλα αυτά, το μέγεθος και η πολυπλοκότητα των σύγχρονων γλωσσικών μοντέλων καθιστούν την πλήρη ανταλλαγή παραμέτρων πρακτικά ανέφικτη, ειδικά σε περιβάλλοντα cloud edge IoT, όπου οι υπολογιστικοί και επικοινωνιακοί πόροι είναι αρκετά περιορισμένοι.

Για να αντιμετωπιστούν αυτοί οι περιορισμοί, υιοθετήθηκαν νέες, πιο αποδοτικές τεχνικές ρύθμισης παραμέτρων, ικανές να περιορίσουν τις ενημερώσεις σε μικρά στοχευμένα υποσύνολα του μοντέλου, όπως είναι οι προσαρμογείς ή οι χαμηλής διάστασης αναπαραστάσεις. Έτσι, ελαττώνεται δραστικά το κόστος επικοινωνίας, αλλά και η υπολογιστική επιβάρυνση, οπότε η ομοσπονδιακή ρύθμιση γίνεται εφαρμόσιμη σε κατανεμημένα περιβάλλοντα. Είναι σαφές, πως αφαιρώντας τη δυνατότητα ανταλλαγής δεδομένων δεν συνεπάγεται αυτομάτως επαρκή προστασία της ιδιωτικότητας.

Έχει καταδειχθεί σε διάφορες μελέτες ότι οι ενημερώσεις παραμέτρων και τα διανύσματα κλίσεων μπορεί να ενσωματώνουν σημαντικές πληροφορίες για το εκπαιδευτικό σύνολο. Κατά την ομοσπονδιακή μάθηση με γλωσσικά μοντέλα, έχει παρατηρηθεί πως οι επιτιθέμενοι μπορούν να ανακατασκευάσουν ιδιωτικά κείμενα χρηστών αποκλειστικά από τις μεταδιδόμενες κλίσεις, υπονομεύοντας τον ισχυρισμό περί έμφυτης ιδιωτικότητας της ομοσπονδιακής μάθησης. Με αυτά τα ευρήματα γίνεται σαφές πως η ομοσπονδιακή μάθηση προσφέρει ένα νέο αρχιτεκτονικό μηχανισμό που περιορίζει την έκθεση των δεδομένων,

αλλά όχι αρκετά επαρκή για να προστατέψει τα μοντέλα από επιθέσεις ανακατασκευής ή συμπερασμού.

Εκτός αυτού, τα ομοσπονδιακά συστήματα έχουν τη δυνατότητα να αντιμετωπίζουν και σοβαρά ζητήματα ακεραιότητας και ανθεκτικότητας, ειδικά σε εχθρικά και ασύρματα περιβάλλοντα. Πρακτικά, κακόβουλοι συμμετέχοντες έχουν τη δυνατότητα να εισάγουν αλλοιωμένες ενημερώσεις στο μοντέλο, καταλήγοντας σε επιθέσεις δηλητηρίασης. Οπότε, η ανάγκη για ανθεκτικούς μηχανισμούς συνάθροισης εντείνεται ακόμα περισσότερο, καθώς τα σύγχρονα πλαίσια συνιστούν την ενσωμάτωση αρχιτεκτονικών που περιλαμβάνουν ελέγχους συνέπειας και μηχανισμούς ανίχνευσης ανωμαλιών, προσφέροντας μια πιο αξιόπιστη λειτουργία ακόμα και υπό την απειλή σφαλμάτων ή επιθέσεων.

Για μια πιο ολιστική θωράκιση της ιδιωτικότητας, την ομοσπονδιακή μάθηση έρχονται να υποστηρίξουν μηχανισμοί ασφαλούς ανταλλαγής ενημερώσεων. Ασφαλή πρωτόκολλα συνάθροισης βοηθούν τον διακομιστή να υπολογίζει τη συνολική ενημέρωση του μοντέλου χωρίς απαραίτητα να αποκρυπτογραφεί όλες τις επιμέρους συνεισφορές των χρηστών, περιορίζοντας τη δυνατότητα ανάλυσης μεμονωμένων κλίσεων, άρα και αποφεύγοντας επιθέσεις ανακατασκευής. Επίσης, εφαρμόζοντας τεχνικές διαφορικής ιδιωτικότητας στις τοπικές ενημερώσεις πριν την αποστολή τους, δημιουργείται ένα επιπλέον στρώμα ασφάλειας, προστατεύοντας το σύστημα από επιθέσεις όπως η αναγνώριση συμμετοχής. Συνδυάζοντας αυτούς τους δύο μηχανισμούς, η ομοσπονδιακή μάθηση μετατρέπεται από μια απλή αποκεντρωμένη αρχιτεκτονική σε μια ολοκληρωμένη τεχνική προστασίας, ικανή να ανταπεξέλθει στις σύγχρονες απαιτήσεις.

8.3 Κρυπτογραφικές τεχνικές υπολογισμού και ομομορφική κρυπτογράφηση

Οι κρυπτογραφικές τεχνικές υπολογισμού αποτελούν ένα διακριτό και συμπληρωματικό σύνολο μηχανισμών προστασίας της ιδιωτικότητας, τα οποία διαφέρει ανάλογα με τις προσεγγίσεις που βασίζονται στη στατιστική απόκρυψη πληροφορίας. Με τον γενικό όρο ασφαλής υπολογισμός εννοείται πως οι τεχνικές αυτές έχουν στόχο την εκτέλεση υπολογισμών χωρίς να αποκαλύπτουν τα δεδομένα εισόδου ή τις παραμέτρους του μοντέλου. Στο πλαίσιο των μεγάλων γλωσσικών μοντέλων, αυτές οι προσεγγίσεις λειτουργούν ως τρίτος βασικός πυλώνας προστασίας, συμπληρώνοντας τη διαφορική ιδιωτικότητα και την ομοσπονδιακή μάθηση, ειδικά σε περιπτώσεις όπου η προάσπιση της εμπιστευτικότητας κατά τη φάση επεξεργασίας ευαίσθητων εισόδων θεωρείται καίριας σημασίας.

Βασικό ρόλο στο πεδίο αυτό διαδραματίζει ο ασφαλής πολυμερής υπολογισμός, ο οποίος επιτρέπει σε πολλαπλούς χρήστες να υπολογίζουν από κοινού μια συνάρτηση με τα ιδιωτικά τους δεδομένα, χωρίς κανένα μέρος να αποκτά ολοκληρωμένη γνώση των εισόδων των υπολοίπων. Οι παραδοσιακές τεχνικές υλοποίησης χρησιμοποιούν σχήματα διαμοιρασμού μυστικού, τα οποία ενισχύουν την αποδοτική εκτέλεση γραμμικών πράξεων επάνω σε κατανεμημένες αναπαραστάσεις, μαζί με πρωτόκολλα συγκεχυμένων κυκλωμάτων, τα οποία είναι καταλληλότερα για μη γραμμικές ή λογικές πράξεις. Στις πιο σύγχρονες πρακτικές, τα πρωτόκολλα ασφαλούς πολυμερούς υπολογισμού (SMPC) αξιοποιούν υβριδικές αρχιτεκτονικές, υλοποιώντας διάφορους μηχανισμούς ανάλογα με τον τύπο των υπολογισμών, έτσι ώστε να περιορίζεται το κρυπτογραφικό κόστος χωρίς η ασφάλεια να υπονομεύεται.

Ειδικότερα, στον τομέα της συμπερασματολογίας νευρωνικών δικτύων, τα πρωτόκολλα ασφαλούς πολυμερούς υπολογισμού χρησιμοποιούνται κυρίως για την εκτέλεση πράξεων μεγάλης κλίμακας με ασφάλεια, όπως είναι οι πολλαπλασιασμοί μεγάλων πινάκων, οι οποίοι αποτελούν βασικό στοιχείο των αρχιτεκτονικών Transformer. Όμως, η εφαρμογή τους στα μεγάλα γλωσσικά μοντέλα δεν είναι πάντα τόσο εύκολη, καθώς δημιουργούνται έντοντα

ζητήματα κλιμάκωσης, λόγω του μεγέθους των μοντέλων και της συχνότητας αλληλεπιδράσεων. Συγκεκριμένα, κρίσιμος παράγοντας αύξησης των απαιτήσεων υπολογισμού και επικοινωνίας είναι οι μη γραμμικές συναρτήσεις ενεργοποίησης, όπως είναι η Softmax και η GeLU, αφού η ακριβής υλοποίησή τους σε περιβάλλοντα ασφαλούς υπολογισμού είναι ιδιαίτερα δαπανηρή.

Επιπροσθέτως, ένα διαφορετικό μοντέλο προστασίας εισάγεται μέσω της ομομορφικής κρυπτογράφησης, η οποία επιτρέπει την εκτέλεση υπολογισμών κατευθείαν σε κρυπτογραφημένα δεδομένα, χωρίς να απαιτείται αποκρυπτογράφηση κατά τη διάρκεια της επεξεργασίας. Με αυτόν τον τρόπο, ο διακομιστής έχει τη δυνατότητα να εξάγει συμπεράσματα χωρίς να έχει πρόσβαση ούτε στα δεδομένα εισόδου του χρήστη, ούτε και στην παραγόμενη έξοδο. Σύγχρονα πλαίσια υψηλής απόδοσης έχουν αποδείξει πως η εφαρμογή βαθιών νευρωνικών δικτύων σε κρυπτογραφημένα δεδομένα είναι τεχνικά εφικτή, και ιδιαίτερα χρήσιμη, ιδίως σε περιπτώσεις που η προστασία της ιδιωτικότητας υπερισχύει των απαιτήσεων απόδοσης.

Αν και η ιδέα της ομομορφικής κρυπτογράφησης θεωρείται ιδανική στο φάσμα της προστασίας ευαίσθητων δεδομένων, εξακολουθεί να συνοδεύεται από σημαντικό υπολογιστικό κόστος, το οποίο μεγαλώνει αν εφαρμοστεί σε αρχιτεκτονικές μεγάλων γλωσσικών μοντέλων πλήρους κλίμακας. Αυτό απορρέει από το γεγονός ότι η κρυπτογράφηση χρειάζεται επαναλαμβανόμενες πράξεις πρόσθεσης και πολλαπλασιασμού σε κρυπτογραφημένο χώρο, και σε συνδυασμό με τις απαιτήσεις επαναφοράς θορύβου και διαχείρισης ακρίβειας, η άμεση υλοποίησή της καθίσταται απαγορευτική για εφαρμογές πραγματικού χρόνου. Γι’ αυτό τον λόγο, η έρευνα στρέφεται σε βελτιστοποιήσεις κρυπτογραφικών διαδικασιών, όπως η επιτάχυνση βασικών πράξεων μέσω υλικού, η αξιοποίηση αριθμητικών μετασχηματισμών υψηλής απόδοσης και η αντικατάσταση μη γραμμικών συναρτήσεων με πολυωνυμικές.

Παράλληλα, έχουν αναπτυχθεί υβριδικές τεχνικές που αξιοποιούν συνδυαστικά διαφορετικές κρυπτογραφικές τεχνικές, χρησιμοποιώντας τον ασφαλή πολυμερή υπολογισμό για γραμμικές πράξεις μεγάλης κλίμακας και πιο εξειδικευμένα πρωτόκολλα για τις μη γραμμικές διαδικασίες. Συνεπώς, με τη συνέλιξη αλγορίθμων, κρυπτογραφίας και συστημάτων, μπορούν να δημιουργηθούν βιώσιμοι, αρχιτεκτονικά ασφαλείς υπολογισμοί για εφαρμογές μεγάλων γλωσσικών μοντέλων, στις οποίες η ιδιωτικότητα πρέπει να διασφαλίζεται σε ολόκληρο το υπολογιστικό μονοπάτι χωρίς να υποβαθμίζεται η απόδοση του συστήματος.

8.4 Αναίρεση εκπαιδευτικής επιρροής και διόρθωση μοντέλων

Η αναίρεση εκπαιδευτικής επιρροής αναδύεται ως μία κρίσιμη διαδικασία για τη συμμόρφωση των συστημάτων μηχανικής μάθησης με τα κανονιστικά πλαίσια που επιβάλλουν το δικαίωμα στη λήθη, δηλαδή τη δυνατότητα διαγραφής προσωπικών δεδομένων μετά την ολοκλήρωση της εκπαίδευσης. Ο στόχος δεν είναι η πλήρης επανεκπαίδευση του μοντέλου από την αρχή, αλλά η αφαίρεση της επιρροής συγκεκριμένων δεδομένων εκπαίδευσης. Η διαδικασία αυτή καθίσταται ιδιαίτερα δαπανηρή σε χρόνο και υπολογιστικούς πόρους, καθώς το μοντέλο πρέπει να διατηρεί τη συνολική λειτουργικότητά του κατά την αφαίρεση των δεδομένων. Στο σύγχρονο υπολογιστικό περιβάλλον η ανάγκη εφαρμογής της αναίρεσης καθίσταται εντονότερη, καθώς το μέγεθος των συνόλων δεδομένων και η πολυπλοκότητα των μοντέλων αυξάνονται.

Εξαιρετικά σημαντική είναι η διάκριση της αναίρεσης εκπαιδευτικής επιρροής από τη διαφορική ιδιωτικότητα. Ενώ η διαφορική ιδιωτικότητα περιορίζει εκ των προτέρων τη συμβολή μεμονωμένων δεδομένων κατά τη φάση της εκπαίδευσης, η αναίρεση εκπαιδευτικής επιρροής πραγματεύεται την εκ των υστέρων αφαίρεση της επίδρασης ήδη

ενσωματωμένων δεδομένων. Έχοντας διακρίνει αυτές τις μεθόδους γίνονται κατανοητές οι διαφορές στις τεχνικές υλοποίησης και τα κριτήρια αξιολόγησης, καθώς η αναίρεση απαιτεί αυστηρό ορισμό της ισοδυναμίας μεταξύ του τροποποιημένου μοντέλου και του μοντέλου που δεν εκπαιδεύτηκε ποτέ στα αφαιρούμενα δεδομένα.

Οι μέθοδοι της αναίρεσης εκπαιδευτικής επιρροής κατηγοριοποιούνται κυρίως σε δύο κατηγορίες: την ακριβή και την κατά προσέγγιση αναίρεση. Η ακριβής αναίρεση εγγυάται πως η κατανομή των παραμέτρων του μοντέλου μετά τη διαδικασία διαγραφής είναι δυσδιάκριτη από την κατανομή ενός μοντέλου που δεν είχε εκπαιδευτεί με τα συγκεκριμένα δεδομένα. Μια χρήσιμη αρχιτεκτονική που ελαττώνει σημαντικά το κόστος της ακριβούς αναίρεσης εκπαιδευτικής επιρροής είναι η μέθοδος SISA (Sharded, Isolated, Sliced, Aggregated), η οποία διαχωρίζει το σύνολο δεδομένων εκπαίδευσης σε ανεξάρτητα τμήματα και εκπαιδεύει απομονωμένα μοντέλα για το καθένα. Κατά συνέπεια, όταν υποβάλλεται ένα αίτημα διαγραφής ενός συγκεκριμένου σημείου δεδομένων, απαιτείται η επανεκπαίδευση μόνο του αντίστοιχου τμήματος που περιείχε το δεδομένο, επιτυγχάνοντας σημαντική μείωση του συνολικού κόστους σε σχέση με την πλήρη επανεκπαίδευση.

Καθότι σε βαθιά νευρωνικά δίκτυα και γενικότερα σε μοντέλα μεγάλης κλίμακας, η ακριβής αναίρεση εκπαιδευτικής επιρροής δεν είναι πάντα εφαρμόσιμη.Για τον λόγο αυτό, έχουν αναπτυχθεί προσεγγίσεις κατά προσέγγιση αναίρεσης, οι οποίες βασίζονται σε θεωρητικά εργαλεία διαφορικής ιδιωτικότητας για τον ορισμό και την επαλήθευση της διαγραφής. Η έννοια της πιστοποιημένης αφαίρεσης παρέχει ένα θεωρητικό πλαίσιο δυσδιακριτότητας που εγγυάται ότι το αποτέλεσμα της διαδικασίας δεν διαφέρει ουσιαστικά από εκείνο της ολικής επανεκπαίδευσης. Μελέτες έχουν αναδείξει πως οι αλγόριθμοι που ικανοποιούν ιδιότητες διαφορικής ιδιωτικότητας είναι συμβατοί με τη διαδικασία αναίρεσης, αφού περιορίζουν από τον σχεδιασμό τους την εξάρτηση του μοντέλου από μεμονωμένα δείγματα.

Έντονες προκλήσεις προκύπτουν στο περιβάλλον της ομοσπονδιακής μάθησης, όπου οι πληροφορίες είναι κατανεμημένες στους χρήστες και δεν είναι άμεσα προσβάσιμες από τον κεντρικό διακομιστή. Για να εφαρμοστεί αποτελεσματικά η αναίρεση σε ομοσπονδιακά περιβάλλοντα χρειάζονται εξειδικευμένα πρωτόκολλα, τα οποία υποστηρίζουν την απομάκρυνση της συνεισφοράς ενός χρήστη από το παγκόσμιο μοντέλο, ακόμα και στις περιπτώσεις που ο χρήστης δεν είναι διαθέσιμος. Για να αντιμετωπιστούν αυτοί οι περιορισμοί, έχουν αναπτυχθεί προεκτάσεις αρχιτεκτονικών όπως το coded sharding, οι οποίες εισάγουν πλεονασματικότητα μέσω κωδικοποίησης στα δεδομένα εκπαίδευσης, επιτρέποντας την αποδοτική αναίρεση και την ανακατασκευή του μοντέλου, ακόμα και υπό συνθήκες όπου οι κόμβοι αποσυνδέονται, διατηρώντας την πληροφοριακή ιδιωτικότητα των δεδομένων έναντι συνεργαζόμενων κακόβουλων χρηστών.

Συμπερασματικά, η αναίρεση εκπαιδευτικής επιρροής αποτελεί ένα καίριο τεχνικό εργαλείο διόρθωσης και συμμόρφωσης των σύγχρονων συστημάτων μηχανικής μάθησης με απαιτήσεις ιδιωτικότητας. Επιλέγοντας τις κατάλληλες αρχιτεκτονικές και θεωρητικές εγγυήσεις, μπορούν να απομακρυνθούν ανεπιθύμητα ή ευαίσθητα δεδομένα χωρίς απαραίτητα να επανεκπαιδευτεί από την αρχή το μοντέλο, ενισχύοντας τη βιωσιμότητα και την αξιοπιστία των συστημάτων.

8.5 Αρχιτεκτονικές ελέγχου ροής και διεπαφές εφαρμογών με επίγνωση ιδιωτικότητας

Στο παρόν πεδίο των μεγάλων γλωσσικών μοντέλων, τα στατικά συστήματα ελέγχου πρόσβασης έχουν καταδειχθεί ανεπαρκή για να διαχειριστούν την αυξημένη πολυπλοκότητα των δυναμικών ροών πληροφορίας. Ως εκ τούτου, προτιμούνται αρχιτεκτονικές ελέγχου πρόσβασης που λαμβάνουν υπόψη τα συμφραζόμενα (Context Aware Fine Grained Access Control), οι οποίες πρακτικά αναδιαμορφώνουν τα στατικά σχήματα ρόλων σε σημασιολογικά εμπλουτισμένα μοντέλα λήψης αποφάσεων. Σε αντίθεση με τα παραδοσιακά συστήματα απόδοσης ρόλων (RBAC), οι νέες αυτές μεθοδολογίες μοντελοποιούν τον χρήστη, τους πόρους και το περιβάλλον, καθώς τα αντιμετωπίζουν ως αλληλένδετες έννοιες, καθιστώντας εφικτό τον δυναμικό έλεγχο πρόσβασης.

Ένα βασικό στοιχείο των σύγχρονων διεπαφών εφαρμογών με ενσωματωμένη προστασία ιδιωτικότητας είναι η κατανόηση των αιτημάτων μέσω φυσικής γλώσσας και γραφημάτων δομημένης γνώσης. Αξιοποιώντας γραφήματα γνώσης, οι αποφάσεις πρόσβασης παύουν να λαμβάνονται αποκλειστικά από προκαθορισμένους κανόνες, καθώς το αποτέλεσμα προκύπτει βάσει σύνθετων σχέσεων μεταξύ των υποκειμένων, των δεδομένων και των πολιτικών ασφαλείας. Έτσι, ακόμα και σε περίπλοκα σενάρια, όπως τα long tail σενάρια, η τεχνική αυτή μπορεί να εφαρμοστεί αποτελεσματικά, καθώς τα αιτήματα εκεί δεν μπορούν να προβλεθούν εξ αρχής και χρειάζονται τεκμηρίωση υψηλού επιπέδου.

Ταυτόχρονα, σημαντική συνιστώσα για τη θέσπιση ενός ικανοποιητικού επιπέδου ασφαλείας έναντι των επιθέσεων έγχυσης προτροπών αποτελεί η εγκατάσταση τείχων προστασίας για μεγάλα γλωσσικά μοντέλα. Αυτά τα ενδιάμεσα λογισμικά λειτουργούν ως φίλτρα εισόδου και εξόδου, ελέγχοντας και αναλύοντας όλες τις προτροπές και απαντήσεις πριν περάσουν στο μοντέλο ή στον τελικό χρήστη. Η υιοθέτηση του τείχους προστασίας σε συνδυασμό με τους κανόνες ανάλυσης και ανίχνευσης μοτίβων είναι ικανή να εντοπίσει εγκαίρως άμεσες και έμμεσες επιθέσεις έγχυσης, οι οποίες συνήθως ενσωματώνονται σε εξωτερικές πηγές δεδομένων και παρακάμπτουν τους απλούς μηχανισμούς ασφαλείας.

Αξιοσημείωτος, επίσης, είναι ο ξεκάθαρος διαχωρισμός ανάμεσα στις αξιόπιστες εντολές του συστήματος και στα μη αξιόπιστα δεδομένα εισόδου, καθώς έχουν παρατηρηθεί πολλαπλά περιστατικά παραβίασης, τα οποία οφείλονται σε αυτόν τον παράγοντα. Οι αρχιτεκτονικές ελέγχου ροής έχουν τη δυνατότητα να ελαττώσουν την πιθανότητα εμφάνισης τέτοιου είδους επίθεσης, εφαρμόζοντας τεχνικές απομόνωσης και ιεράρχησης, με σκοπό την απαλοιφή της δυνατότητας τροποποίησης της συμπεριφοράς του μοντέλου από εξωτερικές πηγές. Συνεπώς, περιορίζεται δραστικά η επιφάνεια επίθεσης, διασφαλίζοντας την ακεραιότητα της λειτουργίας, ακόμα και σε σενάρια όπου το μοντέλο πρέπει να επεξεργαστεί μη ελεγχόμενα δεδομένα.

Εξίσου σημαντική είναι η συμμόρφωση με τα κανονιστικά πλαίσια που έχουν τεθεί σε ισχύ, εφόσον καθιστούν απαραίτητη την ύπαρξη μηχανισμών καταγραφής ενεργειών και ελέγχου για τη διατήρηση της ιδιωτικότητας σε ασφαλές επίπεδο. Σε αυτού του είδους τις διεπαφές εφαρμογών, η καταγραφή περιορίζεται σε μεταδεδομένα, αποφάσεις πρόσβασης, κρυπτογραφικά προστατευμένες περιγραφές και όχι σε πρωτογενή δεδομένα. Με αυτόν τον τρόπο η αξιολόγηση ορθότητας εκ των υστέρων και η νομιμότητα των αποφάσεων καθίστανται εφικτές, χωρίς να προστίθεται επιπλέον φορέας κινδύνου λόγω της αποθήκευσης ευαίσθητου περιεχομένου.

Συνοπτικά, οι αρχιτεκτονικές ελέγχου ροής προσφέρουν σύγχρονους μηχανισμούς επεξήγησης των αποφάσεων πρόσβασης σε φυσική γλώσσα καθιστώντας τες ευκολοκατανόητους. Η δυνατότητα αυτή ενισχύει τη διαφάνεια και τον ενδελεχή έλεγχο των κανόνων ασφαλείας μαζί με την ικανότητα διόρθωσης σφαλμάτων. Συνεπώς, οι διεπαφές

εφαρμογών με ενσωματωμένη προστασία ιδιωτικότητας διαθέτουν ένα ολοκληρωμένο αρχιτεκτονικό επίπεδο, ικανό να διατηρεί την ασφάλεια, την ιδιωτικότητα και την ελεγξιμότητα εναρμονισμένα με τις αυξανόμενες απαιτήσεις των εφαρμογών που χρησιμοποιούν μεγάλα γλωσσικά μοντέλα.

8.6 Ψηφιακή υδατογράφηση και ιχνηλασιμότητα μοντέλων

Η ραγδαία εξάπλωση του περιεχομένου που προέρχεται από μεγάλα γλωσσικά μοντέλα έχει αναδείξει με έντονο τρόπο την ανάγκη για μηχανισμούς πιστοποίησης της προέλευσης και προστασίας της πνευματικής ιδιοκτησίας. Σε αυτό το πλαίσιο, η ψηφιακή υδατογράφηση αποτελεί χρήσιμη τεχνική για την προσθήκη αόρατων σημάτων στο μοντέλο ή στο παραγόμενο κείμενο, τα οποία μπορούν να ανιχνευτούν. Μεγάλη σημασία έχει αποκτήσει η διαδικασία της υδατογράφησης στα περιβάλλοντα δημόσιας διάχυσης πληροφορίας, αφού δίνει τη δυνατότητα στους χρήστες να διακρίνουν την ανθρώπινη και τη μηχανική παραγωγή λόγου, ελαττώνοντας την πιθανότητα παραπληροφόρησης και εξασφαλίζοντας την ακριβή λογοδοσία.

Μελέτες έχουν μετατοπίσει το ενδιαφέρον τους από τις παρεμβάσεις στα εσωτερικά βάρη των μοντέλων προς τις μεθόδους που λειτουργούν αποκλειστικά στο επίπεδο της εισόδου, με χαρακτηριστικό παράδειγμα την υδατογράφηση που καθοδηγείται από προτροπές. Ουσιαστικά, αυτού του είδους η υδατογράφηση αξιοποιεί τη δυνατότητα των μεγάλων γλωσσικών μοντέλων να ακολουθούν σύνθετες πολλαπλές οδηγίες τις οποίες μπορεί να παράγει ένα άλλο μοντέλο. Δημιουργώντας ένα συνεργατικό σχήμα, μαζί με το πρώτο μοντέλο που εξάγει οδηγίες, ένα δεύτερο μπορεί να δημιουργεί το κείμενο, και ένα τρίτο αναλαμβάνει την ανίχνευση του σήματος, επιτυγχάνοντας μια δυναμική ενσωμάτωση υδατογραφήματος. Το αποτέλεσμα μιας τέτοιας προσέγγισης είναι ότι έχει εξαλειφθεί η απαίτηση της πρόσβασης στα εσωτερικά στοιχεία του μοντέλου, οπότε μπορεί να εφαρμοστεί ακόμα και σε κλειστά συστήματα.

Εκτός από την απλή ανίχνευση προέλευσης, ζήτημα μείζονος σημασίας συνιστά η προάσπιση της ακεραιότητας του παραγόμενου περιεχομένου, σε μια εποχή όπου οι επιθέσεις πλαστογράφησης ακμάζουν. Οι λύσεις που έχουν προταθεί είναι σχήματα δύο επιπέδων, τα οποία μπορούν να συνδυάσουν πιο λεπτομερή σήματα ελέγχου ακεραιότητας με απλούστερα σήματα αναγνώρισης πηγής δεδομένων. Αξιοποιώντας τεχνικές δειγματοληψίας και ταξινόμηση με βάση την κατάταξη μπορεί με ευκολία να διακριθεί το αυθεντικό από το τροποποιημένο και το πλήρες πλαστογραφημένο περιεχόμενο, περιορίζοντας σημαντικά τον κίνδυνο ψευδούς απόδοσης επιβλαβών ή μολυσμένων κειμένων στους κατασκευαστές των μοντέλων.

Παρ’ όλα αυτά, τα υδατογραφήματα έρχονται με ένα κόστος, καθώς δεν είναι πάντα δεδομένη η ανθεκτικότητά τους. Ερευνητικά αποτελέσματα έχουν δείξει πως υφιστάμενες τεχνικές μπορούν να γίνουν στόχος σε επιθέσεις παράφρασης, με αποτέλεσμα να τροποποιείται ελάχιστα το κείμενο, χωρίς να αλλάζει το νόημά του, αλλά να εξασθενείται δραματικά το στατιστικό σήμα. Τέτοιες επιθέσεις μπορούν να εφαρμοστούν επιτυχώς ακόμα και με περιορισμένη γνώση του συστήματος, γεγονός που χαρακτηρίζει την υδατογράφησης σαν έναν ημιτελή αυτόνομο μηχανισμό προστασίας και πως χρειάζονται επιπλέον τεχνικές άμυνας για την ουσιαστική προστασία της ιδιωτικότητας.

Ως ένα συμπληρωματικό μέτρο για την υποστήριξη της ιχνηλασιμότητας και της αξιοπιστίας, έχει αναπτυχθεί η έννοια της απόδειξης μηδενικής γνώσης (Zero Knowledge Proofs ZKPs). Ουσιαστικά, αντί να βασίζονται εξ ολοκλήρου σε ενσωματωμένα συστήματα, οι αποδείξεις βοηθούν την κρυπτογραφική επαλήθευση της εγκυρότητας των δεδομένων εισόδου και της ακρίβειας της διαδικασίας παραγωγής, διατηρώντας κρυφά το ευαίσθητο περιεχόμενο και τις παραμέτρους του μοντέλου. Αξιοποιώντας με σωστό τρόπο την ψηφιακή υδατογράφηση

και τις κρυπτογραφικές αποδείξεις, μπορεί να δημιουργηθεί ένα ασφαλές πλαίσιο λογοδοσίας, το οποίο θα εξισορροπεί την ανάγκη διαφάνειας με τις σύγχρονες απαιτήσεις ιδιωτικότητας και ασφάλειας στο οικοσύστημα τεχνητής νοημοσύνης.

9 ΠΛΑΙΣΙΑ ΑΞΙΟΛΟΓΗΣΗΣ ΚΑΙ ΜΕΤΡΙΚΕΣ ΕΛΕΓΧΟΥ ΙΔΙΩΤΙΚΟΤΗΤΑΣ

9.1 Θεωρητικός ορισμός μετρικών αξιολόγησης

Για την αξιολόγηση της προστασίας της ιδιωτικότητας σε διαδικασίες ανωνυμοποίησης και προάσπισης των δεδομένων έχουν αναπτυχθεί τυπικά ορισμένες μετρικές, οι οποίες ποσοτικοποιούν τον κίνδυνο αποκάλυψης, δηλαδή την πιθανότητα εμφάνισης μιας παραβίασης και την επίπτωση που αυτή επιφέρει σε κάθε περίπτωση. Ανά τους καιρούς έχουν προταθεί διάφορα μοντέλα αξιολόγησης, αλλά οι πλέον διαδεδομένες μετρικές είναι το k-anonymity, το l-diversity και το t-closeness, οι οποίες εμφανίζουν διαφορετικές πτυχές της απειλής παραβίασης της ιδιωτικότητας.

Αν και οι μετρικές είναι χρήσιμα στοιχεία για την αξιολόγηση του κινδύνου, παρουσιάζουν ετερογένεια ως προς τη σημασία τους και το μαθηματικό τους υπόβαθρο. Οι διαφορές αυτές αναδεικνύουν την ανάγκη για ένα ενοποιημένο πλαίσιο, το οποίο θα συγκρίνει τα διαφορετικά κριτήρια προστασίας της ιδιωτικότητας και θα τα εισάγει σε ένα κοινό πρόβλημα βελτιστοποίησης. Μελέτες συνιστούν τη χρήση της θεωρίας της πληροφορίας ως κατάλληλη προσέγγιση γι’αυτό το σκοπό, αξιοποιώντας την έννοια της αμοιβαίας πληροφορίας (Mutual Information) σαν έναν μπούσουλα στατιστικής εξάρτησης ανάμεσα στα αρχικά και στα δημοσιευμένα δεδομένα.

Με την αμοιβαία πληροφορία μπορεί να εφαρμοστεί ταυτόχρονα η αποτίμηση του κινδύνου αποκάλυψης και η χρηστικότητα των δεδομένων με μεγάλη ακρίβεια, παρουσιάζοντας ποσοτικά το κέρδος πληροφορίας που λαμβάνει ένας παρατηρητής μαθαίνοντας τα ανωνυμοποιημένα στοιχεία. Ωστόσο, η αμοιβαία πληροφορία είναι μια μέση ποσότητα, η οποία αποτελεί καθρέφτη της συνολικής συμπεριφοράς του συνόλου δεδομένων και όχι των επιμέρους συνθηκών που ισχύουν για κάθε εγγραφή. Συνεπώς, η ικανότητά της να εκφράζει επαρκώς και με ακρίβεια τις απαιτήσεις της ιδιωτικότητας που επιβάλλονται σε επίπεδο εγγραφής περιορίζεται σημαντικά.

Έτσι, εισάγεται μια νέα έννοια, αυτή της πληροφορίας ενός συμβόλου (One Symbol Information), η οποία ορίζεται ως η συνεισφορά της κάθε ξεχωριστής εγγραφής στη συνολική αμοιβαία πληροφορία. Με αυτή την αποσύνθεση, καθίσταται εφικτή η ακριβέστερη αποτύπωση του κινδύνου αποκάλυψης και σε μικροεπίπεδο, οπότε μπορεί να αποτυπωθεί η μαθηματική έκφραση και να συγκριθούν διάφορες μετρικές ιδιωτικότητας με κοινές μονάδες μέτρησης. Με τον τρόπο αυτό, αποδεικνύεται ότι απαιτήσεις όπως το l-diversity και το t-closeness μπορούν να αναπαρασταθούν ως δύο διαφορετικές αλλά ταυτόχρονα ισοδύναμες συνθήκες για την πληροφορία ενός συμβόλου.

Γενικότερα, στο πεδίο της εξόρυξης δεδομένων με διατήρηση της ιδιωτικότητας, οι μετρικές αξιολόγησης κατηγοριοποιούνται σε δύο διακριτές ομάδες: τις μετρικές κινδύνου αποκάλυψης (Disclosure Risk Metrics) και τις μετρικές απώλειας πληροφορίας (Information Loss Metrics). Στην πρώτη κατηγορία, οι μετρικές αξιολογούν την πιθανότητα παραβίασης της ιδιωτικότητας κατά τη διαδικασία της αποθήκευσης και της ροής των δεδομένων. Στη δεύτερη, αποτιμάται ποσοτικά ο αντίκτυπος που έχουν οι τεχνικές προστασίας στην ποιότητα και την αναλυτική αξία των δεδομένων.

Η θεωρητική θεμελίωση των μετρικών αξιολόγησης είναι απαραίτητη για έναν αξιόλογο σχεδιασμό και την εφαρμογή εννοιολογικών πλαισίων προστασίας της ιδιωτικότητας, ειδικά σε ευαίσθητα πληροφοριακά περιβάλλοντα, όπως τα συστήματα διαχείρισης ασθενών. Σε

αυτά τα σενάρια, ο στόχος είναι η εφαρμογή της αρχής της ιδιωτικότητας από τον σχεδιασμό (Privacy by Design) σε όλα τα επίπεδα διαχείρισης, ώστε να εξασφαλιστεί η ισορροπία μεταξύ προστασίας προσωπικών δεδομένων, και λειτουργικής χρηστικότητας του συστήματος.

9.2 Μετρικές ιδιωτικότητας σε μοντέλα μηχανικής μάθησης

Η ποσοτικοποίηση της διαρροής πληροφορίας σε μοντέλα μηχανικής μάθησης είναι άμεσα εξαρτημένη από το υιοθετούμενο μοντέλο απειλής (Threat Model). Εκτός αυτού, σύμφωνα με τη συστηματοποίηση της ασφάλειας στο πλαίσιο της μηχανικής μάθησης, στην αξιολόγηση πρέπει να υπολογίζονται οι δυνατότητες του επιτιθέμενου, όπως ακριβώς εφαρμόζεται στην πρόσβαση σε παραμέτρους του μοντέλου (white box) ή μόνο στις εξόδους του (black box). Καίριο ρόλο για την εμπειρική μέτρηση κινδύνου διαδραματίζει η επίθεση συμπερασμού συμμετοχής, κατά την οποία η βασική μετρική αξιολόγησης είναι το ποσοστό επιτυχίας της επίθεσης (Attack Success Rate), η οποία υπολογίζεται εκπαιδεύοντας νέα μοντέλα σκιάς. Πρακτικά, τα μοντέλα αυτά επιχειρούν να μιμηθούν τη συμπεριφορά του στόχου, υποστηρίζοντας τον επιτιθέμενο να εκπαιδεύσει έναν ταξινομητή που θα διακρίνει με ποία πιθανότητα ένα συγκεκριμένο κομμάτι συνόλου δεδομένων ανήκε στο εκπαιδευτικό περιεχόμενο, συγκρίνοντας διαφοροποιήσεις διανυσμάτων πιθανοτήτων που εξάγει το μοντέλο.

Σε ένα ευρύτερο επίπεδο, ο κίνδυνος ιδιωτικότητας βρίσκεται άρρηκτα συνδεδεμένος με το φαινόμενο της υπερπροσαρμογής. Διάφορες αναλύσεις έχουν δείξει πως το χάσμα γενίκευσης, δηλαδή η διαφορά μεταξύ της απόδοσης του μοντέλου στα δεδομένα εκπαίδευσης και αυτής σε τυχαία δεδομένα, λειτουργεί ως ένας ισχυρός και αξιόπιστος δείκτης ευπάθειας για τη διαρροή πληροφορίας. Συγκεκριμένα, όταν ένα μοντέλο ενσωματώνει στον κορμό του ορισμένα δεδομένα αντί να μαθαίνει γενικευμένα μοτίβα, τείνει να παρουσιάζει εμφανώς μεγαλύτερη βεβαιότητα στις προβλέψεις που αφορούν γνωστές εγγραφές σε σχέση με τις άγνωστες. Μια τέτοια ασυμμετρία αποτελεί τρωτό σημείο, εύκολα εκμεταλλεύσιμο σε επιθέσεις συμπερασμού, καθιστώντας το μέγεθος της υπερπροσαρμογής μετρήσιμη ένδειξη ευπάθειας του συστήματος.

Παρά τη χρησιμότητά τους, οι συνολικές μετρικές σε επίπεδο μοντέλου συχνά αποδεικνύονται ανεπαρκείς για την αποτίμηση του κινδύνου για μεμονωμένες, ευάλωτες εγγραφές. Γι’ αυτόν τον λόγο έχουν αναπτυχθεί ειδικές μετρικές, όπως το SHAPR, οι οποίες βασίζονται στη συνεργατική θεωρία παιγνίων. Ουσιαστικά, η συγκεκριμένη προσέγγιση αντιμετωπίζει την εκπαίδευση του μοντέλου σαν ένα παιχνίδι και τα δεδομένα σαν παίκτες, ενώ υπολογίζει την οριακή συνεισφορά κάθε μεμονωμένης εγγραφής στην τελική απόδοση του μοντέλου. Μέσω των τιμών Shapley μπορεί να υπολογιστεί το κατά πόσο κάθε εγγραφή επηρεάζει την απόδοση, με την παραδοχή πως οι εγγραφές που παρουσιάζουν απόκλιση από τις υπόλοιπες στην επιρροή τους στη συμπεριφορά του μοντέλου είναι πιθανότερο να έχουν απομνημονευθεί, οπότε και να παρουσιάσουν αυξημένο κίνδυνο συμμετοχικής αποκάλυψης. Η μετρική SHAPR υπολογίζει αποδοτικά αυτές τις τιμές, προσφέροντας μια λεπτομερή χαρτογράφηση του κινδύνου, αντί για έναν γενικευμένο μέσο όρο.

Για σενάρια πρακτικής αξιολόγησης, στα οποία η επανεκπαίδευση των μοντέλων είναι απαγορευτική, έχουν προταθεί εμπειρικές μετρικές που βασίζονται μόνο στην παρατηρήσιμη συμπεριφορά ενός στιγμιότυπου του μοντέλου. Μια τέτοια μετρική είναι η Epsilon*, η οποία ποσοτικοποιεί τον κίνδυνο ιδιωτικότητας μέσω πρόσβασης τύπου μαύρου κουτιού και διατύπωσης της επίθεσης συμπερασμού ως στατιστικό έλεγχο υποθέσεων. Η τελική τιμή λειτουργεί ως ένα εμπειρικό κάτω όριο της απώλειας ιδιωτικότητας, δημιουργώντας ένα πλαίσιο στο οποίο η αξιολόγηση του κινδύνου μπορεί να εφαρμοστεί χωρίς να απαιτείται γνώση του αλγορίθμου εκπαίδευσης ή του συνόλου δεδομένων που χρησιμοποιήθηκαν.

9.3 Ποσοτική αποτίμηση έκθεσης και πολυπλοκότητας πρόβλεψης

Για την ποσοτική αποτίμηση της ιδιωτικότητας στα μεγάλα γλωσσικά μοντέλα έχουν αναπτυχθεί μετρικές που αποτυπώνουν τη στατιστική συμπεριφορά τους όταν εκτίθενται σε γλωσσικές ακολουθίες, με θεμελιώδη λίθο την έννοια της πολυπλοκότητας πρόβλεψης (Perplexity-PPL). Αυτή η μετρική αξιολογεί την αβεβαιότητα ενός γλωσσικού μοντέλου, εκφράζοντας τον βαθμό δυσκολίας που αντιμετωπίζει αυτό όταν προσπαθεί να προβλέψει την επόμενη λέξη σε κάθε ακολουθία. Γλωσσικές έρευνες σε νευρωνικά μοντέλα, όπως το BERT και το GPT-2, έχουν δείξει πως οι τιμές που λαμβάνει η μετρική της πολυπλοκότητας δεν αφορούν αποκλειστικά την απόδοση του μοντέλου, καθώς επηρεάζονται από τα συντακτικά χαρακτηριστικά της πρότασης, όπως η λεξιλογική πυκνότητα, και η δομή των ρηματικών φράσεων. Έτσι, η πολυπλοκότητα καθίσταται έμμεσος δείκτης της στατιστικής οικειότητας του μοντέλου με το κείμενο εισαγωγής, άρα λαμβάνει ιδιαίτερη σημασία στο πλαίσιο της προστασίας της ιδιωτικότητας.

Αναλύοντας τις τιμές που λαμβάνει η μετρική της πολυπλοκότητας φανερώνεται ότι σε εξαιρετικά χαμηλά νούμερα, εκτός του ότι αποτελεί δείκτη υψηλής ποιότητας μοντελοποίησης του μοντέλου, υποδηλώνει την απομνημόνευση δεδομένων εκπαίδευσης. Έχει τεκμηριωθεί πως τα μεγάλα γλωσσικά μοντέλα γενικού σκοπού μπορούν να αναπαράγουν ευαίσθητες ακολουθίες, οι οποίες υπήρχαν στο εκπαιδευτικό σύνολο, ειδικά όταν αυτές εμφανίζονταν σε δυσανάλογα υψηλή πιθανότητα. Τα μεγάλα στατιστικά νούμερα αυτά μειώνουν την αβεβαιότητα πρόβλεψης και αυξάνουν την πιθανότητα έκθεσης των υποκειμένων των δεδομένων, καθιστώντας τη μετρική της πολυπλοκότητας ένα από τα πιο χρήσιμα εργαλεία εντοπισμού διαρροών δεδομένων.

Το φαινόμενο της απομνημόνευσης έχει παρατηρηθεί πως ενισχύεται από φαινόμενα που αφορούν την εκπαιδευτική διαδικασία. Συγκεκριμένα, η χρήση της τεχνικής “teacher forcing” οδηγεί το μοντέλο στο λεγόμενο σφάλμα έκθεσης (Exposure Bias), καθώς το μοντέλο εκτίθεται αποκλειστικά σε ακολουθίες που προέρχονται από πραγματική κατανομή δεδομένων και καθόλου από δικές του προβλέψεις. Αυτή η διαδικασία προκαλεί μια ασυμμετρία στη φάση της συμπερασματολογίας, στην οποία το μοντέλο καταλήγει να παρουσιάζει ασταθή συμπεριφορά σε ξένα γι’ αυτό δεδομένα, ενώ ταυτόχρονα απαντά με εξαιρετική, αλλά ύποπτη, ακρίβεια σε δεδομένα που έχει ήδη δει, εντείνοντας τον κίνδυνο ταυτοποίησης της πηγής τους.

Πέραν της πολυπλοκότητας, η μετρική της έκθεσης συνδέεται επίσης με τις εσωτερικές αναπαραστάσεις που εξάγουν τα γλωσσικά μοντέλα. Έχει αποδειχθεί ότι οι αναπαραστάσεις κειμένου (embeddings) που παράγονται από τα μοντέλα διατηρούν ένα σημαντικό κομμάτι ευαίσθητου περιεχομένου από το αρχικό κείμενο, δίνοντας χώρο σε έναν αντίπαλο να ανακτήσει, με τεχνικές αντίστροφης μηχανικής (reverse engineering), τα αρχικά δεδομένα. Συνδυαστικά, η υψηλή πιστότητα των προβλέψεων μαζί με τη χαμηλή πολυπλοκότητα μπορούν να λειτουργήσουν ως ενδείξεις αυξημένης έκθεσης, δίνοντας ιδιαίτερη έμφαση στο γεγονός ότι για την ουσιαστική αξιολόγηση της ιδιωτικότητας πρέπει να λαμβάνεται υπόψη η στατιστική συμπεριφορά του μοντέλου.

Είναι κρίσιμο να τονιστεί ότι η μετρική της πολυπλοκότητας δεν επαρκεί ως μοναδικός δείκτης ασφαλείας, ιδίως σε περιβάλλοντα όπου εφαρμόζονται τεχνικές συμπίεσης μοντέλων για λόγους αποδοτικότητας. Πολυδιάστατες αξιολογήσεις έχουν καταλήξει πως η διατήρηση και η βελτιστοποίηση της πολυπλοκότητας σαν μετρική σε συμπιεσμένα μοντέλα

δεν εγγυάται τη διατήρηση των ιδιοτήτων ασφαλείας. Αντιθέτως, μια συμπίεση μπορεί να προκαλέσει αμετάκλητες μεταβολές σε καίριες διαστάσεις, όπως η ανθεκτικότητα σε επιθέσεις άλλου είδους ή η σταθερότητα της συμπεριφοράς του μοντέλου, θέτοντας ένα πλαίσιο στο οποίο απαιτείται μια ολιστική προσέγγιση αξιολόγησης της ιδιωτικότητας.

9.4 Η απόκλιση μεταξύ θεωρητικής και πραγματικής χρηστικότητας

Οι μετρικές που έχουν αναλυθεί μέχρι τώρα αποτελούν ποσοτικά εργαλεία για την εκτίμηση της προστασίας της ιδιωτικότητας και της έκθεσης της πληροφορίας σε διάφορα υπολογιστικά πλαίσια. Η εφαρμογή τους, όμως, στην πράξη αντιμετωπίζει ένα μεθοδολογικό ζήτημα, το οποίο αφορά τις τιμές των μετρικών που πολλές φορές δεν ταυτίζονται με την πραγματική χρηστικότητα ή αξιοπιστία του συστήματος που εξετάζεται. Διάφορες μελέτες, που πραγματεύονται την προστασία της ιδιωτικότητας, όταν αναφέρονται στον όρο χρηστικότητα, δεν την ορίζουν με βάση την απόδοση του μοντέλου, αλλά με τη βοήθεια μετρικών απώλειας πληροφορίας, οι οποίες αποτυπώνουν τον βαθμό αλλοίωσης των δεδομένων που προκύπτει από την τοποθέτηση μηχανισμών ασφαλείας. Έτσι, έχουν αναπτυχθεί μετρικές, όπως αυτή της διακριτότητας και η γενικευμένη απώλεια πληροφορίας, που επιχειρούν να μετρήσουν τη μείωση της αξίας της πληροφορίας ανεξάρτητα από τους αλγορίθμους, διακρίνοντας με αυτόν τον τρόπο τη γενική χρηστικότητα, δηλαδή αυτή που αφορά τη διατήρηση της στατιστικής δομής των δεδομένων, από τη χρηστικότητα ειδικού φορτίου, η οποία είναι εξαρτημένη από ορισμένες εργασίες ταξινόμησης ή ανάλυσης.

Εμβαθύνοντας περαιτέρω, ιδιαίτερη σημασία αποκτά η διάκριση αυτή όταν η αξιολόγηση βασίζεται ολοκληρωτικά στην τελική απόδοση των μοντέλων. Έρευνες που εξετάζουν το θέμα υπό το πρίσμα των υπερπαραμέτρων δείχνουν ότι η ακρίβεια ενός μοντέλου επηρεάζεται από τις επιλογές των διαδικασιών εκπαίδευσης και όχι μόνο από τη χρηστικότητα των δεδομένων. Ένα μοντέλο μπορεί να επιτυγχάνει υψηλή ακρίβεια μέσω της υπερπροσαρμογής ακόμη και όταν τα δεδομένα έχουν υποστεί σοβαρή αλλοίωση. Οπότε, η ακρίβεια δεν πρέπει να θεωρείται αξιόπιστος δείκτης διατήρησης της αξίας της πληροφορίας, όπως και η αξιολόγηση οφείλει να παραμένει ανεπηρέαστη από τις αλγοριθμικές ρυθμίσεις και να υπολογίζει την πραγματική απώλεια πληροφορίας που προκαλούν οι μηχανισμοί προστασίας.

Στα μεγάλα γλωσσικά μοντέλα, η έννοια της ποιότητας περιπλέκεται εξαιτίας της ανάγκης για πιστότητα (Fidelity). Η πιστότητα, ενώ εκ πρώτης φαίνεται να ταυτίζεται απλώς με τη γλωσσική συνοχή, στην πραγματικότητα ορίζεται ως η ικανότητα του μοντέλου να παράγει συνεπές και θεμελιωμένο περιεχόμενο, απαλλαγμένο από ψευδαισθήσεις (hallucinations), στο πλαίσιο που είναι ενταγμένο. Έχει παρατηρηθεί ότι συμβατικές μετρικές αξιολόγησης κειμένου δεν καταφέρνουν να εντοπίσουν περιπτώσεις που το παραγόμενο κείμενο, ενώ είναι γλωσσικά ορθό, παρουσιάζει ανακρίβειες που δεν ταυτοποιούνται με πραγματικά δεδομένα. Η απόκλιση αυτή αναδεικνύει τους περιορισμούς των μετρικών αυτών και κάνει ευδιάκριτη τη διαφορά ανάμεσα στην επιφανειακή ποιότητα και την ουσιαστική αξιοπιστία των εξόδων ενός μοντέλου.

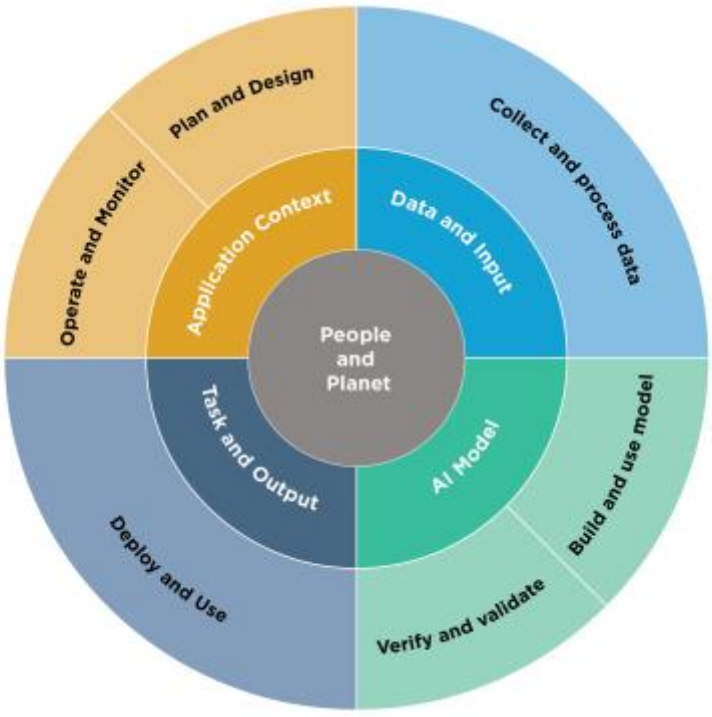
Καταληκτικά, η κριτική επισκόπηση των μεθοδολογιών αξιολόγησης τονίζει πως η αξιοπιστία των μετρήσεων εξαρτάται άμεσα από το πεδίο εφαρμογής, τις υποθέσεις αξιολόγησης και το σύνολο των μετρικών που λαμβάνονται υπόψη. Δεν αρκεί η μονοδιάστατη αποτίμηση της ιδιωτικότητας, της χρηστικότητας ή της πιστότητας, καθώς συχνά οδηγεί σε λανθασμένα συμπεράσματα, ειδικά σε περιβάλλοντα όπου εφαρμόζονται τεχνικές βελτιστοποίησης ή συμπίεσης. Για τον λόγο αυτό, η σύγχρονη έρευνα στρέφει το ενδιαφέρον της προς πιο σύνθετα πλαίσια αξιολόγησης που εξετάζουν ταυτόχρονα τις έννοιες της ακρίβειας, της πιστότητας, αλλά και της ηθικής ευθυγράμμισης, λαμβάνοντας υπόψη το γεγονός ότι η βελτιστοποίηση μιας μεμονωμένης μετρικής μπορεί να αποκρύψει σοβαρά λειτουργικά σφάλματα του συστήματος.

10 ΔΙΑΧΕΙΡΙΣΗ ΚΙΝΔΥΝΟΥ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗ

10.1 Πλαίσια ποσοτικοποίησης ρίσκου

Για την ουσιαστική ποσοτικοποίηση του κινδύνου μέσω επιμέρους μετρικών και δεικτών απαιτείται η ένταξή του σε ένα οργανωμένο πλαίσιο διαχείρισης. Ο NIST έχει καθιερώσει το Πλαίσιο Διαχείρισης Κινδύνου Τεχνητής Νοημοσύνης (NIST AI RMF 1.0) για να καλύψει αυτή την ανάγκη. Εκεί αναφέρεται πως η λειτουργία Measure δεν πρέπει να αντιμετωπίζεται ως ένα μεμονωμένο στάδιο, αλλά σαν απαραίτητος κρίκος στην αλυσίδα λειτουργίας, μεταξύ της χαρτογράφησης του κινδύνου και των διακυβερνητικών διεργασιών. Ιδιαίτερη έμφαση δίνεται στη μετατόπιση του ενδιαφέροντος από την παραγωγή απλών αριθμητικών τιμών στην αξιολόγηση της αξιοπιστίας και της εγκυρότητας των μετρήσεων, καθώς η εγγενής πολυπλοκότητα και αδιαφάνεια των συστημάτων τεχνητής νοημοσύνης συχνά καθιστούν τις κοινές μεθόδους ελέγχου λογισμικού ανεπαρκείς για την εξαγωγή ακριβών και ασφαλών συμπερασμάτων.

Παράλληλα με τις κατευθυντήριες γραμμές του NIST, το διεθνές πρότυπο ISO/IEC 42001 υιοθετεί μια πιο κανονιστική δομή μέσω της Πιστοποίησης Συστημάτων Διαχείρισης AI, ενσωματώνοντας την ποσοτικοποίηση του ρίσκου σε ένα σύστημα οργανωσιακής διακυβέρνησης. Το σύστημα αυτό αξιολογεί τον κίνδυνο μέσω διαδικασιών που εκτιμούν τις επιπτώσεις των συστημάτων τεχνητής νοημοσύνης (AI System Impact Assessment), οι οποίες εξετάζουν τις συνέπειες της ανάπτυξης του μοντέλου σε ατομικό και κοινωνικό επίπεδο. Το πρότυπο απαιτεί την αυστηρή τεκμηρίωση των κριτηρίων αποδοχής κινδύνου, αλλά και της παρακολούθησης υπολειπόμενων, μετατρέποντας την ποσοτικοποίηση σε εργαλείο ελέγχου συμμόρφωσης και θέτοντας επιτεύξιμους στόχους απόδοσης πέρα από τη νομική υποχρέωση.



Εικόνα 9: Κύκλος ζωής συστήματος τεχνητής νοημοσύνης (AI lifecycle) κατά NIST.

Σε καθαρά τεχνικό επίπεδο, για την ακριβή ποσοτικοποίηση ειδικών κατηγοριών κινδύνου, όπως η διαρροή δεδομένων εκπαίδευσης, χρειάζονται εξειδικευμένα μαθηματικά εργαλεία.

Χαρακτηριστικό παράδειγμα αποτελεί το στατιστικό εργαλείο σύγκρισης υποθέσεων: Λόγος Πιθανοφάνειας (Likelihood Ratio, LR) το οποίο είναι ιδιαίτερα χρήσιμο στον εντοπισμό επιθέσεων συμπερασμού συμμετοχής. Ουσιαστικά, η συγκεκριμένη μεθοδολογία ξεπερνά τις απλοϊκές προσεγγίσεις που στηρίζονται στα καθιερωμένα χαρακτηριστικά του μοντέλου, συγκρίνοντας την πιθανότητα παραγωγής μιας ακολουθίας από το εξεταζόμενο μοντέλο σε σχέση με το μοντέλο αναφοράς. Με αυτόν τον τρόπο, τοποθετείται αριθμητική τιμή στον κίνδυνο αποκάλυψης, διευκολύνοντας τη σύγκριση μεταξύ διαφορετικών μοντέλων ή και εκδόσεων, ορίζοντας ένα ενιαίο πλαίσιο αξιολόγησης.

Η στατική μέτρηση σε ένα χρονικό σημείο, όπως γίνεται σε πολλά λογισμικά συστήματα, στο πλαίσιο των μεγάλων γλωσσικών μοντέλων καθίσταται ανεπαρκής, καθώς λειτουργούν δυναμικά, και εξελίσσονται συνεχώς. Σύγχρονες ερευνητικές προσεγγίσεις αξιολόγησης έχουν αναπτύξει τη συνεχή αξιολόγηση ασφάλειας (Continuous Security Assessment) μέσω αυτοματοποιημένων μηχανισμών ελέγχου, επιτρέποντας την επαναξιολόγηση των υπολειπόμενων κινδύνων (residual risk) σε πραγματικό χρόνο, ενώ τα δεδομένα εισόδου και οι λειτουργικές συνθήκες μεταβάλλονται. Έτσι, η ποσοτικοποίηση μετατρέπεται από στατική αποτίμηση σε δυναμική διαδικασία παρακολούθησης.

Καταλυτικό ρόλο σε μια ολοκληρωμένη διαχείριση ρίσκου παίζει η αναγνώριση αλληλεξαρτήσεων και αντισταθμίσεων (trade-offs) ανάμεσα στις διαφορετικές παραμέτρους ποιότητας, όπως για παράδειγμα η σχέση ασφάλειας και απόδοσης. Όπως ο NIST επισημαίνει, η βελτιστοποίηση μιας συγκεκριμένης μετρικής πολλές φορές μπορεί να επιφέρει αρνητικά αποτελέσματα σε άλλες καίριες λειτουργίες του συστήματος. Συνεπώς, καθίσταται αναγκαία η θέσπιση και εφαρμογή ενός συνόλου διαδικασιών, το οποίο αποτελείται από ενέργειες όπως η δοκιμή, η αξιολόγηση, η επαλήθευση και η επικύρωση. Η τελική απόφαση για την αποδοχή ή την απόρριψη ενός επιπέδου κινδύνου προκύπτει από τη συστηματική ανάλυση, και όχι από έναν απλό αλγόριθμο.

10.2 Σχεδιασμός Δεικτών Απόδοσης Ασφάλειας

Στο επιχειρησιακό περιβάλλον, η παραδοσιακή ανάπτυξη και λειτουργία συστημάτων τεχνητής νοημοσύνης υποστηρίζεται από πρακτικές MLOps, οι οποίες εστιάζουν στην αυτοματοποίηση της εκπαιδευτικής διαδικασίας, της ανάπτυξης και της παρακολούθησης προγνωστικών μοντέλων. Σήμερα, οι πρακτικές αυτές επεκτείνονται στο πιο εξειδικευμένο πλαίσιο των Λειτουργιών Μεγάλων Γλωσσικών Μοντέλων (LLMOps). Για τον σχεδιασμό ενός ώριμου πλαισίου χρειάζεται σωστός ορισμός δεικτών απόδοσης ασφαλείας, καθώς, εκτός της αποτύπωσης τεχνικών παραμέτρων λειτουργίας, καλούνται να λειτουργούν ως μηχανισμοί έγκαιρης προειδοποίησης για αποκλίσεις συμπεριφοράς του μοντέλου. Αν και στο παρελθόν το ζητούμενο ήταν απλώς η ακρίβεια μιας πρόβλεψης με βάση προκαθορισμένες εισόδους, το νέο πλαίσιο έχει σχεδιαστεί με τέτοιο τρόπο ώστε να μπορεί να διαχειριστεί ολόκληρο τον κύκλο ζωής των μοντέλων που παράγουν νέο περιεχόμενο. Αυτό σημαίνει πως οι δείκτες απόδοσης, εκτός της υπολογιστικής ταχύτητας, αποτυπώνουν την αξιοπιστία του συστήματος σε πραγματικό χρόνο. Πρακτικά, αυτό προαπαιτεί τη συνεχή παρακολούθηση της υγείας του μοντέλου, ώστε να εντοπίζονται εγκαίρως λειτουργικές ανωμαλίες ή κενά στις διαδικασίες αυτοματοποίησης.

Ένας από τους κρισιμότερους δείκτες για την ασφάλεια αυτών των συστημάτων αφορά την ανθεκτικότητα των γλωσσικών μοντέλων απέναντι σε κακόβουλες επιθέσεις, όπως είναι η παρεμβολή προτροπών (Prompt Injections) και η παράκαμψη περιορισμών (Jailbreaks). Ουσιαστικά, σε αυτές τις τεχνικές ένας χρήστης προσπαθεί να ξεγελάσει το μοντέλο με σκοπό να παρακάμψει τους κανόνες ασφαλείας και να εξάγει απαγορευμένο περιεχόμενο. Για να μετρηθεί αυτός ο κίνδυνος, έχει υιοθετηθεί ο δείκτης Ποσοστού Επιτυχίας Επίθεσης (Attack Success Rate, ASR), ο οποίος αποτυπώνει το πόσο συχνά το μοντέλο αποτυγχάνει να αμυνθεί όταν δέχεται επιθετικές εντολές. Η συστηματική παρακολούθηση αυτού του

δείκτη δίνει τη δυνατότητα στους διαχειριστές να κατανοούν ποίες ακριβώς τεχνικές είναι πιο πιθανό να παραβιάσουν το σύστημα, αλλά και την τρέχουσα κατάσταση ευθυγράμμισης με τις πολιτικές.

Εκτός από τις σκόπιμες επιθετικές ενέργειες, σημαντικό χαρακτηριστικό ενός δείκτη ασφαλείας είναι η ικανότητά του να εντοπίζει τα λάθη που προκύπτουν όταν ένα μοντέλο καλείται να διαχειριστεί εισόδους που αποκλίνουν από την κατανομή των δεδομένων εκπαίδευσης. Εφόσον τα μεγάλα γλωσσικά μοντέλα εκπαιδεύονται σε συγκεκριμένο σύνολο γνώσεων, συχνά δυσκολεύονται να ανταπεξέλθουν με ακρίβεια σε πληροφορίες που δεν έχουν ξαναδεί, τα λεγόμενα δεδομένα Εκτός Κατανομής (Out of Distribution, OOD). Σε αυτές τις περιπτώσεις τα μοντέλα τείνουν να παράγουν απαντήσεις με υψηλή γλωσσική συνοχή, αλλά χαμηλή ουσιαστική αξιοπιστία. Επομένως, είναι απαραίτητος ο σχεδιασμός δεικτών ανίχνευσης ανωμαλιών, οι οποίοι θα λειτουργούν σαν προειδοποιητικοί συναγερμοί όταν το σύστημα θα δέχεται εισόδους που αποκλίνουν από τα φυσιολογικά πρότυπα, αποτρέποντας την παραγωγή αναξιόπιστων εξόδων.

Συμπερασματικά, ο σχεδιασμός δεικτών απόδοσης ασφάλειας στο πλαίσιο των LLMOps χρειάζεται πολυδιάστατη προσέγγιση, όπου οι δείκτες ανθεκτικότητας σε κακόβουλες επιθέσεις θα συνεργάζονται με τους δείκτες ποιότητας και καταλληλότητας εισόδων. Συνδέοντας αυτές τις μετρήσεις με τους επιχειρησιακούς μηχανισμούς παρακολούθησης μπορεί να γίνει η μετάβαση από την παθητική καταγραφή σφαλμάτων στην προληπτική διαχείριση κινδύνου, διασφαλίζοντας, έτσι, ότι το μοντέλο θα παραμένει ασφαλές και αξιόπιστο καθ’ όλη τη διάρκεια του κύκλου ζωής του.

10.3 Πρωτόκολλα Διαχείρισης Περιστατικών

Η αποτελεσματική διαχείριση περιστατικών ασφαλείας σε συστήματα τεχνητής νοημοσύνης, προϋποθέτει, πρωτίστως, την θέσπιση τυποποιημένων μηχανισμών καταγραφής και ταξινόμησης. Στις μέρες μας υπάρχει έλλειψη ομοιογένειας στις βάσεις δεδομένων περιστατικών στα συστήματα μεγάλων γλωσσικών μοντέλων, γεγονός που οδηγεί σε ασυνέπειες ως προς τον τρόπο τεκμηρίωσης, καθώς περιορίζει τη δυνατότητα συστηματικής συγκριτικής ανάλυσης και μάθησης από τα προηγούμενα συμβάντα. Γι’ αυτόν το λόγο έχει προταθεί ένα νέο ενιαίο σχήμα αναφοράς, το οποίο θα οφείλει να εισάγει λεπτομερώς και να οργανώνει τα περιστατικά με βάση ορισμένες διαστάσεις, όπως η σοβαρότητα, οι αιτίες εμφάνισης, αλλά και οι επιπτώσεις που προκαλούνται σε τεχνικό και κοινωνικό επίπεδο. Μέσω μιας τέτοιας δομημένης προσέγγισης, διασφαλίζεται η συνεπής συλλογή δεδομένων, η οποία αποτελεί ακρογωνιαίο λίθο για την πρώτη κιόλας πρόληψη μελλοντικών αστοχιών και επαναλαμβανόμενων σφαλμάτων, και ενισχύεται η επιχειρησιακή ετοιμότητα.

Στο κρίσιμο στάδιο της επιχειρησιακής απόκρισης (incident response), η ταχύτητα λήψης αποφάσεων, ο συντονισμός, και η προσαρμοστικότητα απέναντι σε δυναμικά εξελισσόμενες απειλές είναι ζωτικής σημασίας. Σύγχρονες μεθοδολογίες αναδεικνύουν τη μετάβαση από γραμμικές, χειροκίνητες διαδικασίες στη συνεργατική τεχνητή νοημοσύνη πολλαπλών πρακτόρων (multi-agent collaboration), στο οποίο μεγάλα γλωσσικά μοντέλα λαμβάνουν ρόλους που προσομοιώνουν ομάδες απόκρισης με διαφορετικά επίπεδα εξειδίκευσης. Αξιοποιώντας εναλλακτικές διοικητικές δομές, όπως κεντρικές, αποκεντρωμένες ή υβριδικές, τα μοντέλα μπορούν να λειτουργούν ως ευφυείς πράκτορες, και καθιστώντας εφικτή η διερεύνηση του τρόπου με τον οποίο επηρεάζουν τη λήψη αποφάσεων η κατανομή ρόλων και η ροή της πληροφορίας. Παρατηρώντας την αλληλεπίδραση των πρακτόρων, προκύπτει πως η βελτιστοποίηση της λήψης αποφάσεων σε πραγματικό χρόνο σε συνδυασμό με την αυξημένη προσαρμοστικότητα απέναντι στις εξελιγμένες κυβερνοεπιθέσεις, συμβάλλουν στον εξορθολογισμό των διαδικασιών ανάσχεσης της απειλής, μέσω της αξιοποίησης προσομοιώσεων σεναρίων κυβερνοασφάλειας.

Μετά την αντιμετώπιση ενός περιστατικού, η διαδικασία περνά στο στάδιο της ψηφιακής εγκληματολογικής έρευνας (digital forensics) και τεκμηρίωσης, όπου η ανάγκη για αξιόπιστα ψηφιακά πειστήρια είναι επιτακτική. Στο πλαίσιο των μεγάλων γλωσσικών μοντέλων, η διαδικασία της ψηφιακής εγκληματολογίας επιτρέπει την αυτοματοποιημένη ανάλυση τεράστιων συνόλων δεδομένων, τα οποία μπορεί να είναι ετερογενή, όπως αρχεία καταγραφής, αναφορές συμβάντων και χρονικές ακολουθίες ενεργειών, ξεφεύγοντας από τις παραδοσιακές χρονοβόρες ανακριτικές μεθόδους. Αυτή η δυνατότητα σύνθεσης διαφορετικού είδους πληροφοριών και αναγνώρισης μοτίβων ενισχύει την ανακατασκευή του περιστατικού, αλλά εγείρει θέματα επιστημονικής εγκυρότητας. Πάντα υπάρχει πιθανότητα παραγωγής ανακριβών και μη επαληθεύσιμων συμπερασμάτων από τα μοντέλα, γεγονός που καθιστά απαραίτητη την εφαρμογή αυστηρών πρωτοκόλλων επαλήθευσης για τη διασφάλιση της ακεραιότητας της αποδεικτικής διαδικασίας.

Συνοψίζοντας, τα πρωτόκολλα διαχείρισης περιστατικών σε περιβάλλοντα τεχνητής νοημοσύνης δεν λειτουργούν ως μεμονωμένες ενέργειες, αλλά ως μια αλυσίδα αλληλένδετων διαδικασιών, των οποίων εκκίνηση αποτελεί η τυποποίηση και πέρας η λογοδοσία και η οργανωσιακή μάθηση. Συνδυάζοντας τη λεπτομερή ταξινομία των περιστατικών, τη δυναμική απόκριση μέσω συνεργαζόμενων πρακτόρων, και την τεκμηριωμένη διαδικασία διερεύνησης, μπορεί να διαμορφωθεί ένα συνεκτικό πλαίσιο διαχείρισης κινδύνου. Αυτή η ολοκληρωμένη προσέγγιση είναι απαραίτητη για τη θωράκιση των μεγάλων γλωσσικών μοντέλων απέναντι στις σύγχρονες απειλές, διασφαλίζοντας ταυτόχρονα τη διαφάνεια και την αξιοπιστία των μηχανισμών ασφαλείας.

11 ΒΕΛΤΙΣΤΕΣ ΠΡΑΚΤΙΚΕΣ ΚΑΙ ΟΔΗΓΙΕΣ ΑΣΦΑΛΟΥΣ ΕΝΣΩΜΑΤΩΣΗΣ

11.1 Συστάσεις ασφαλείας ανά κρίσιμο τομέα εφαρμογής

Η ενσωμάτωση των μεγάλων γλωσσικών μοντέλων σε παραγωγικά περιβάλλοντα δεν μπορεί να ακολουθεί μια γενική πολιτική ασφαλείας, αφού το προφίλ κινδύνου μεταβάλλεται δραστικά ανάλογα με το πλαίσιο εφαρμογής. Ενώ οι θεμελιώδεις αρχές της κυβερνοασφάλειας διατηρούνται κοινές για όλους, οι προτεραιότητες προστασίας αλλάζουν σε όλο το φάσμα της κυβερνοασφάλειας, από την προστασία της ιδιωτικότητας μέχρι τη διασφάλιση της φυσικής ακεραιότητας, απαιτούν εξειδικευμένο πλαίσιο ελέγχου. Αυτό συμβαίνει επειδή η φύση των δεδομένων, το επίπεδο ανοχής σε σφάλματα, αλλά και οι κανονιστικές απαιτήσεις διαφοροποιούνται σύμφωνα με το πλαίσιο εφαρμογής. Έρευνες επισημαίνουν ότι για την ασφαλή μετάβαση από πιλοτικές εφαρμογές σε κρίσιμα περιβάλλοντα απαιτείται η θέσπιση εξειδικευμένων πλαισίων ελέγχου, τα οποία θα είναι προσαρμοσμένα στις λειτουργικές, κοινωνικές και ατομικές επιπτώσεις του κάθε κλάδου. Υπό αυτή την οπτική, η ασφάλεια πρέπει να εξετάζεται τόσο σαν τεχνική παράμετρος όσο και ως ζήτημα εμπιστοσύνης και λογοδοσίας.

Στον ευαίσθητο τομέα της υγείας, η πρωταρχική απαίτηση αφορά την προστασία των προσωπικών ιατρικών δεδομένων, μαζί με την αποφυγή διαγνωστικών σφαλμάτων. Οι βέλτιστες πρακτικές στα σύγχρονα συστήματα αποτελούνται από τεχνικές ανωνυμοποίησης και μηχανισμούς περιορισμού της απομνημόνευσης δεδομένων από τα μοντέλα, όπως η διαφορική ιδιωτικότητα, με σκοπό να ελαχιστοποιηθεί ο κίνδυνος της επαναταυτοποίησης. Ταυτόχρονα, εξαιτίας της τάσης των μοντέλων να παράγουν αληθοφανείς αλλά ανακριβείς ιατρικές συμβουλές, κρίνεται απαραίτητη η διατήρηση της ανθρώπινης εποπτείας στη λήψη κλινικών αποφάσεων (Human in the Loop), διασφαλίζοντας πως καμία κλινική απόφαση δεν έχει ληφθεί αποκλειστικά μέσω αυτοματοποιημένης διαδικασίας, τηρώντας με αυτόν τον τρόπο τις αρχές του OWASP για την ασφάλεια των μεγάλων γλωσσικών μοντέλων.

Συνεπώς, η χρήση των μοντέλων οφείλει να λειτουργεί υποστηρικτικά, και όχι ως ένας αυτόνομος μηχανισμός λήψης αποφάσεων.

Αντιστοίχως, στον χρηματοοικονομικό τομέα, όπου η εμπιστοσύνη και η κανονιστική συμμόρφωση είναι αδιαπραγμάτευτες, η διατήρηση της ασφάλειας θεωρείται υψίστης σημασίας. Γι' αυτό, προτείνεται η υιοθέτηση ενός πλαισίου διασφάλισης (Safeguarded Framework), το οποίο θα λειτουργεί σαν ένα ενδιάμεσο επίπεδο ελέγχου, που θα φιλτράρει τις εισόδους και θα αξιολογούν τις εξόδους του μοντέλου πριν αυτές φτάσουν να χρησιμοποιηθούν για επιχειρησιακούς σκοπούς. Οι χρηματοπιστωτικοί οργανισμοί, ειδικά, οφείλουν να τοποθετούν μηχανισμούς επικύρωσης εισόδων, ώστε να αποτρέπουν επιθέσεις έγχυσης εντολών (prompt injection), καθώς θα μπορούσαν να καταλήξουν σε σενάρια απάτης ή διαρροής τραπεζικού απορρήτου. Επίσης, πρόσθετα στην αντιμετώπιση της απειλής παροχής λανθασμένων ή παραπλανητικών επενδυτικών συμβουλών πρέπει να τοποθετούνται όρια στις γνώσεις του μοντέλου αξιοποιώντας τεχνικές ανάκτησης πληροφορίας, και συνεχής παρακολούθηση των απαντήσεων.

Στο πεδίο της ανάπτυξης λογισμικού, τα μεγάλα γλωσσικά μοντέλα αξιοποιούνται με διάφορους τρόπους, όπως για παράδειγμα τη συγγραφή, την ανάλυση και επιδιόρθωση κώδικα, γεγονός που εισάγει νέους κινδύνους διάδοσης ευπαθειών ασφαλείας σε μεγάλη κλίμακα. Οι άνθρωποι που εργάζονται στον τομέα ανάπτυξης λογισμικού καλούνται να μην αντιμετωπίζουν τον παραγόμενο από τα μοντέλα κώδικα ως έτοιμο για χρήση, αλλά να εφαρμόζουν πολύ αυστηρές διαδικασίες ελέγχου, στις οποίες συμπεριλαμβάνεται η στατική ανάλυση και η τεχνική fuzzing, μια δυναμική μέθοδος δοκιμών η οποία υποβάλλει το μοντέλο σε πλήθος τυχαίων δεδομένων εισόδου με σκοπό να προκαλέσει ανωμαλίες στη λειτουργία του. Με μια τέτοια πολυδιάστατη προσέγγιση το σύστημα μπορεί να θωρακιστεί επαρκώς απέναντι σε κρυφά σφάλματα ή κενά ασφαλείας που το μοντέλο ενδέχεται να αναπαρήγαγε από τα εκπαιδευτικά δεδομένα. Ειδική μέριμνα πρέπει να λαμβάνεται για την εκκαθάριση (sanitization) του κώδικα, με τρόπο τέτοιο ώστε να μην επιτρέπεται η ενσωμάτωση κακόβουλων μοτίβων και η παραβίαση δικαιωμάτων πνευματικής ιδιοκτησίας τρίτων προσώπων.

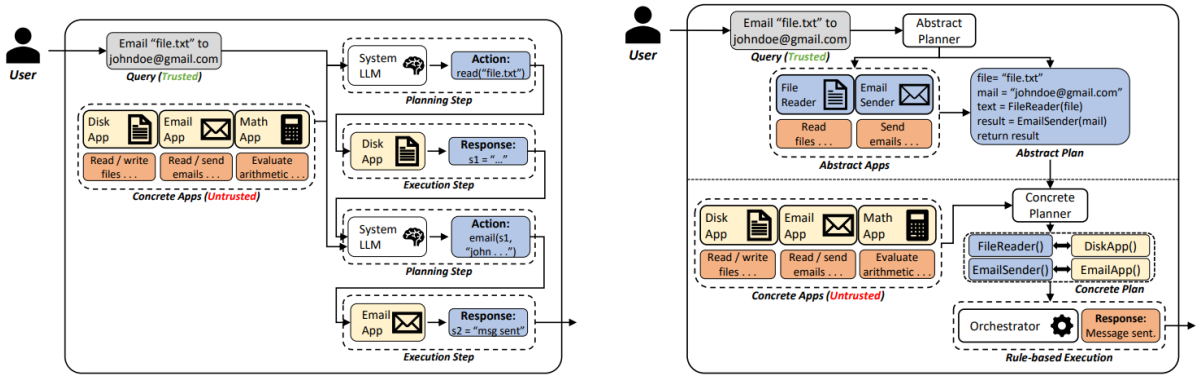
Τέλος, στις κρίσιμες υποδομές, όπως τα δίκτυα ενέργειας, τα συστήματα μεταφορών, και οι τηλεπικοινωνίες, απαιτούνται υψηλής προστασίας μέτρα λόγω των καταστροφικών επιπτώσεων που μπορεί να έχει η διασύνδεση των μεγάλων γλωσσικών μοντέλων με τα κυβερνοφυσικά συστήματα (CPS) και δίκτυα όπως το SCADA. Οι έρευνες επικεντρώνονται στην αρχιτεκτονική απομόνωση των συστημάτων τεχνητής νοημοσύνης από τους κρίσιμους βρόχους ελέγχου και στην αδιάκοπη εποπτεία της λειτουργίας τους μέσω μηχανισμών υψηλής διαθεσιμότητας. Η ανθεκτικότητα απέναντι σε κακόβουλες παρεμβάσεις και η δυνατότητα έγκαιρης ανίχνευσης σφαλμάτων καθίστανται ζητήματα εθνικής ασφάλειας πέραν του επιπέδου του οργανισμού, επιβάλλοντας την εφαρμογή πρωτοκόλλων που θα ανιχνεύουν και θα παρεμποδίζουν προσπάθειες χειραγώγησης των λειτουργικών παραμέτρων σε πραγματικό χρόνο. Με αυτόν τον τρόπο, μπορεί να διασφαλιστεί η αδιάλειπτη και ασφαλής λειτουργία ολόκληρου του οικοσυστήματος.

11.2 Αρχιτεκτονικές ασφαλών API και διαχείριση προσβάσεων

Για να επιτυχθεί η ασφάλεια των διεπαφών προγραμματισμού εφαρμογών (API) σε συστήματα που ενσωματώνουν μεγάλα γλωσσικά μοντέλα χρειάζεται μια θεμελιώδης αλλαγή από τις παραδοσιακές τεχνικές άμυνας σε ένα σύστημα που υιοθετεί την αρχιτεκτονική μηδενικής εμπιστοσύνης (Zero Trust Architecture). Η αρχή σε αυτό το νέο σύστημα άμυνας είναι: ποτέ μην εμπιστεύεσαι, πάντα επαλήθευε, η οποία μεταφράζεται στην ανάγκη συνεχούς πιστοποίησης κάθε ξεχωριστού αιτήματος πρόσβασης, ανεξάρτητα από το ιστορικό του χρήστη ή τη θέση του στο δίκτυο. Στο πλαίσιο αυτό, η ενσωμάτωση της παραγωγικής τεχνητής νοημοσύνης επιτρέπει τη δυναμική ανάλυση συμπεριφοράς των

χρηστών και τον εντοπισμό αποκλίσεων σε πραγματικό χρόνο, δημιουργώντας έναν νέο βρόχο ασφαλείας, αυτόματα προσαρμοζόμενο στις εξελισσόμενες επιθέσεις που προκύπτουν, ενώ διασφαλίζοντας ταυτόχρονα την ασφαλή επικοινωνία ανάμεσα στον χρήστη και το μοντέλο σε όλα τα στάδια της αλληλεπίδρασης.

Στις περιπτώσεις που τα μεγάλα γλωσσικά μοντέλα καλούνται να εκτελέσουν ενέργειες σε εξωτερικά συστήματα όπως η ανάγνωση βάσεων δεδομένων και η αποστολή email, προκύπτουν νέες απειλές, για τις οποίες προτείνονται ξεχωριστά μέτρα αντιμετώπισης. Συγκεκριμένα, μελέτες προτείνουν την υιοθέτηση αρχιτεκτονικών τύπου ACE (Abstract Concrete Execute), μέσω των οποίων εισάγεται μια νέα δικλείδα ασφαλείας, την αποσύνδεση της φάσης σχεδιασμού από τη φάση εκτέλεσης. Έτσι, μέσα από πολυεπίπεδες δομές ελέγχου, αντί το μοντέλο να εκτελεί άμεσα την εντολή του χρήστη, δημιουργεί πρώτα ένα αφηρημένο πλάνο ενεργειών, το οποίο υπόκειται σε ανεξάρτητη αξιολόγηση ως προς τις πολιτικές ασφαλείας. Έπειτα, μόνο μετά την επιτυχή επαλήθευση προχωράει η διαδικασία στη μετατροπή του πλάνου σε εκτελέσιμες εντολές, ελαχιστοποιώντας τον κίνδυνο εκμετάλλευσης του μοντέλου από παραπλανητικές εισόδους.



(a) Typical LLM-integrated app system. (b) ACE system design.

Εικόνα 10: Σύγκριση τυπικής αρχιτεκτονικής εφαρμογής μεγάλων γλωσσικών μοντέλων με την αρχιτεκτονική ACE.

Ένα από τα σοβαρότερα τρωτά σημεία βρίσκεται στη διασύνδεση των μοντέλων με πρόσθετα τρίτων μερών (plugins) και εξωτερικές υπηρεσίες, καθώς αυτά λειτουργούν συνήθως ως κερκόπορτες σε επιθέσεις έγχυσης προτροπών. Διάφορες αναλύσεις σε δημοφιλή πρόσθετα έχουν αποκαλύψει σοβαρά κενά ασφαλείας στην επαλήθευση της ακεραιότητας του ιστορικού συνομιλίας, επιτρέποντας σε επιτιθέμενους την πλαστογράφηση προηγούμενων μηνυμάτων, με αποτέλεσμα το μοντέλο να λαμβάνει μη εξουσιοδοτημένες εντολές. Η απουσία αυστηρών ελέγχων στις αλληλεπιδράσεις μεταξύ χρήστη, μοντέλου, και παρόχου της υπηρεσίας καθιστά επιτακτική την εφαρμογή μηχανισμών πιστοποίησης της ακεραιότητας και της αυθεντικότητας των δεδομένων πριν αυτά φτάσουν στην εξωτερική υπηρεσία, περιορίζοντας την αποτελεσματικότητα των επιθέσεων τύπου prompt injection.

Επιπροσθέτως, η αυξημένη πολυπλοκότητα των σύγχρονων μεθόδων ανάπτυξης λογισμικού αναδεικνύει την ανάγκη αυτοματοποίησης της διαχείρισης πολιτικών ελέγχου πρόσβασης. Η χειροκίνητη διαμόρφωση των κανόνων πρόσβασης είναι αρκετά χρονοβόρα και επιρρεπής σε λάθη, καθώς αδυνατεί να ακολουθήσει τον ρυθμό μεταβολής των απαιτήσεων. Προτείνεται, λοιπόν, η χρήση ειδικών πλαισίων που αξιοποιούν την τεχνολογία RAG (Retrieval Augmented Generation) για την παραγωγή πολιτικών ασφαλείας απευθείας από το προσωποποιημένο ιστορικό των χρηστών και την τεκμηρίωση του λογισμικού, προσαρμόζοντας δυναμικά τα δικαιώματα των χρηστών. Με τον τρόπο αυτό, οι πολιτικές που ορίζουν τα άτομα που έχουν πρόσβαση και τα αντίστοιχα δικαιώματα τους μπορούν να ενημερώνονται δυναμικά όσο εξελίσσεται το λογισμικό, διασφαλίζοντας ότι η αρχιτεκτονική

ασφαλείας συμβαδίζει με τις αλλαγές των απαιτήσεων, μειώνοντας τα κενά εξουσιοδότησης και ενισχύοντας τη συνολική ανθεκτικότητα της εφαρμογής.

11.3 Λίστες ελέγχου για την ασφαλή ανάπτυξη συστημάτων

Η διασφάλιση της ακεραιότητας κατά τη διάρκεια ανάπτυξης λογισμικού που ενσωματώνει συστήματα παραγωγικής τεχνητής νοημοσύνης προϋποθέτει την πρόσθεση δομημένων μηχανισμών ελέγχου σε όλα τα στάδια του κύκλου ζωής της ασφαλούς ανάπτυξης λογισμικού (SSDLC). Η ακαδημαϊκή βιβλιογραφία προτείνει ορισμένες κατευθυντήριες γραμμές, οι οποίες συνδέουν τις εξόδους των γλωσσικών μοντέλων με καθιερωμένα πρότυπα ασφαλείας εφαρμογών, όπως για παράδειγμα το OWASP ASVS. Για να επιτευχθεί αυτό, οι λίστες ελέγχου πρέπει να ενσωματωθούν στη διαδικασία της μηχανικής προτροπών σαν ένα εργαλείο ελέγχου ποιότητας και συμμόρφωσης, εφαρμόζοντας ένα πλάνο που περιλαμβάνει έξι στοιχεία: πλαίσιο, εργασία, οδηγία, διευκρίνιση, βελτίωση, και προειδοποίηση. Με την εφαρμογή ενός τέτοιου πλαισίου, μπορούν οι ομάδες ανάπτυξης, από τα πρώτα μόλις στάδια, όπως ο καταστατικός χάρτης του έργου μέχρι και τον ασφαλή σχεδιασμό, να τεκμηριώνουν τις ενέργειές τους και να παράγουν κώδικα συμμορφωμένο εξαρχής με τις απαιτήσεις ασφαλείας, μειώνοντας το χρόνο που απαιτείται για τις μετέπειτα διορθώσεις.

Παράλληλα με την ανάπτυξη κώδικα, ιδιαίτερα σημαντική είναι η απόλυτη διαφάνεια στην αλυσίδα εφοδιασμού του ίδιου του μοντέλου. Μέσω των παραδοσιακών τεχνικών τεκμηρίωσης, όπως είναι τα model cards, πολλές φορές τα μοντέλα αδυνατούν να επαληθεύσουν στοιχεία ασφαλείας, οπότε περιορίζονται σε απλές περιγραφικές δηλώσεις. Για την κάλυψη αυτού του κενού έχει εισαχθεί η έννοια του AI Bill of Materials (AIBOM), το οποίο παρέχει μια δομημένη αποτύπωση των συνιστωσών του συστήματος, όπως τα δεδομένα εκπαίδευσης, και οι διαδικασίες ανάπτυξης. Οι λίστες ελέγχου σε αυτό το επίπεδο αξιοποιούνται ως αυτοματοποιημένοι μηχανισμοί σάρωσης που εξετάζουν διάφορα στοιχεία, όπως τις εξαρτήσεις του μοντέλου, την ασφάλεια της σειριοποίησης και την συμπεριφορά του μοντέλου σε διάφορες φάσεις. Συνεπώς, η διαφάνεια μετατρέπεται από αφηρημένη δήλωση σε ένα τεκμηριωμένο σύνολο ελέγχων μέσω των οποίων μπορεί να διασφαλιστεί ο σωστός έλεγχος συμμόρφωσης.

Στο στάδιο της λειτουργικής πιστοποίησης, οι λίστες ελέγχου επικεντρώνονται στις διαδικασίες επικύρωσης και επαλήθευσης, οι οποίες πρέπει να προσαρμοστούν στις νέες ιδιαιτερότητες που εισάγουν τα μεγάλα γλωσσικά μοντέλα. Οι λίστες ελέγχου ταξινομούν τους κινδύνους στις εξής κατηγορίες: εγγενείς περιορισμούς, ακούσια σφάλματα, και στοχευμένες επιθέσεις. Αν και η διαδικασία αυτή φαίνεται στατική, στην πραγματικότητα διατρέχει ολόκληρο τον κύκλο ζωής, διασφαλίζοντας πως η συμπεριφορά του μοντέλου παραμένει εντός των προκαθορισμένων ορίων αξιοπιστίας και ασφάλειας, ακόμα και μετά από ενημερώσεις ή αλλαγές στο λειτουργικό περιβάλλον.

Σε οργανωτικό επίπεδο, η ασφαλής ενσωμάτωση των μεγάλων γλωσσικών μοντέλων σε εφαρμογές απαιτεί λίστες ελέγχου κατάλληλες να καλύψουν ζητήματα διακυβέρνησης και επιχειρησιακής ωριμότητας. Αξιοποιώντας αυτά τα εργαλεία αξιολόγησης, οι οργανισμοί μπορούν πλέον να ελέγξουν το επίπεδο προετοιμασίας τους απέναντι σε επιθέσεις τύπου έγχυσης προτροπών και δηλητηρίασης δεδομένων. Επιπλέον, οι λίστες ελέγχου διακυβέρνησης περιλαμβάνουν ελεγκτικές λειτουργίες για τη στρατηγική μετριασμού κινδύνου και για την αξιολόγηση της ανθεκτικότητας χρησιμοποιώντας πραγματικά σενάρια. Ουσιαστικά, λειτουργούν ως ένας μηχανισμός οργανωσιακής μάθησης, ο οποίος

υποστηρίζει τη σταδιακή και υπεύθυνη μετάβαση από πιλοτικές εφαρμογές σε ολοκληρωμένα επιχειρησιακά συστήματα.

12 Η ΑΝΤΙΠΑΡΑΘΕΣΗ ΙΔΙΩΤΙΚΟΤΗΤΑΣ ΚΑΙ ΧΡΗΣΤΙΚΟΤΗΤΑΣ

12.1 Η θεωρητική καμπύλη αντισταθμίσματος στα μεγάλα γλωσσικά μοντέλα

Η σχέση μεταξύ προστασίας της ιδιωτικότητας και της λειτουργικής απόδοσης στα μεγάλα γλωσσικά μοντέλα δεν ακολουθεί μια γραμμική ανταλλαγή κόστους και οφέλους. Στην πραγματικότητα, αυτή η σχέση περιγράφεται ακριβέστερα ως μέτωπο Pareto (Pareto Frontier), το οποίο οριοθετεί το βέλτιστο δυνατό σημείο ισορροπίας, καθώς οποιαδήποτε προσπάθεια ενίσχυσης της ιδιωτικότητας, δηλαδή μείωσης του σκορ της διαρροής, συνεπάγεται αναπόφευκτα κόστος στη χρηστικότητα του μοντέλου, όπως αυτή αποτυπώνεται με μετρικές, συμπεριλαμβανομένων των ROUGE και BERTScore. Κάθε διαφορετική τεχνική εκπαίδευσης ή ρύθμισης τοποθετεί το μοντέλο σε διαφορετικά σημεία επάνω ή κάτω από αυτή την καμπύλη, με τον στόχο να είναι η ελαχιστοποίηση του κινδύνου χωρίς να υποβαθμίζεται η γλωσσική ικανότητα.

Η θεμελιώδης αιτία για τη δημιουργία αυτής της καμπύλης αντισταθμίσματος είναι δομική, και ονομάζεται φόρος ευθυγράμμισης (alignment tax). Η υπόθεση ευθυγράμμισης ασφαλείας υποστηρίζει πως αυτό συμβαίνει επειδή οι νευρωνικές παράμετροι και οι υπολογιστικοί πόροι ενός γλωσσικού μοντέλου είναι πεπερασμένοι, οπότε αναγκαστικά επιμερίζονται ανάμεσα σε ανταγωνιστικούς λειτουργικούς στόχους. Για την επίτευξη της ασφάλειας, το μοντέλο δεσμεύει κάποιες υπολογιστικές μονάδες, και τις χαρακτηρίζει ως κρίσιμες μονάδες ασφαλείας, στερώντας τες από τις υπόλοιπες διεργασίες που εξυπηρετούν τη γενική χρηστικότητα. Οπότε, η πτώση της απόδοσης προκύπτει από αυτή την ανακατανομή των πόρων κατά τη διάρκεια της εκπαίδευσης και ευθυγράμμισης.

Η διαδικασία κατανομής πόρων μπορεί να αναλυθεί αποτελεσματικά μέσω της οπτικής της υπολογιστικής οικονομικής, στην οποία το γλωσσικό μοντέλο αντιμετωπίζεται σαν ένα εσωτερικό οικονομικό σύστημα που τα επιμέρους υπολογιστικά στοιχεία λειτουργούν ως ορθολογικοί πράκτορες υπό καθεστώς σπανιότητας. Το μοντέλο προσπαθεί, σε αυτό το πλαίσιο, να μεγιστοποιήσει τη συνάρτηση χρησιμότητάς του, η οποία περιλαμβάνει ταυτόχρονα την ακρίβεια και τη συμμόρφωση με τους περιορισμούς ασφαλείας, υπό αυστηρούς υπολογιστικούς περιορισμούς. Όταν εισάγεται ένας περιορισμός, όπως το κόστος της ιδιωτικότητας, η στρατηγική κατανομής πόρων μεταβάλλεται σημαντικά, καθώς το μοντέλο πρέπει να θυσιάσει τις λιγότερο σημαντικές πληροφορίες, ώστε να διατηρήσει την απόδοσή του υψηλή στις πιο κρίσιμες εργασίες, δημιουργώντας ένα εσωτερικό οικονομικό σύστημα ανταλλαγής μεταξύ ασφάλειας και απόδοσης.

Αν και με αυτή τη θεωρία ορίζονται συγκεκριμένα όρια, στην πράξη παρατηρείται σημαντική απόκλιση από τις θεωρητικές εγγυήσεις των μοντέλων, τη λεγόμενη “Εμπειρική Διακύμανση της Ιδιωτικότητας”. Έχει καταδειχθεί πως τα μοντέλα που έχουν ρυθμιστεί με ισοδύναμες παραμέτρους διαφορικής ιδιωτικότητας μπορούν να παρουσιάσουν ουσιωδώς διαφορετικά επίπεδα διαρροής και χρηστικότητας, ανάλογα με τις εκπαιδευτικές υπερπαραμέτρους και τις διαδικασίες βελτιστοποίησης που έχουν πραγματοποιηθεί. Αυτό σημαίνει πως η καμπύλη αντισταθμίσματος δεν είναι ένα μονοδιάστατο όριο, αλλά μια ευρύτερη περιοχή εφικτότητας, εντός της οποίας ακόμα και οι πιο μικρές αλλαγές στις ρυθμίσεις του μοντέλου μπορούν να οδηγήσουν σε μεγάλες μεταβολές στην ισορροπία κινδύνου και ωφέλειας.

Έρευνες αναφέρουν πως η θέση των μοντέλων πάνω στην περιοχή της καμπύλης είναι άμεσα εξαρτόμενη από την πολυπλοκότητα και το είδος της μεθόδου εκπαίδευσης. Οι πιο απλοϊκές πρακτικές, όπως η αφαίρεση δεδομένων από το μοντέλο, τείνουν να τοποθετούνται πιο χαμηλά στην καμπύλη, δηλαδή προσφέρουν μεγαλύτερα επίπεδα ασφάλειας με δυσανάλογο κόστος στη χρηστικότητα. Αντιθέτως, πιο σύγχρονες και εξελιγμένες τεχνικές, όπως η εκπαίδευση με οδηγίες (Instruction Tuning) που περιλαμβάνουν επιθυμητά αλλά και ανεπιθύμητα παραδείγματα, έχουν αποδειχθεί ικανές να ωθήσουν το μοντέλο στο να μάθει την έννοια της συμφραζόμενης ιδιωτικότητας, ενώ παράλληλα ελαχιστοποιούν το κόστος που πληρώνει η χρηστικότητα.

Συνοπτικά, η θεωρητική καμπύλη αντισταθμίσματος δεν αποτελεί απόλυτο περιορισμό, καθώς είναι ένα πλαίσιο κατανόησης των συμβιβασμών που διέπουν τα μεγάλα γλωσσικά μοντέλα. Για την ουσιαστική μετατόπιση προς το βέλτιστο μέτωπο Pareto, χρειάζονται μέθοδοι για να αξιοποιούν με τον πιο αποτελεσματικό τρόπο τους διαθέσιμους πόρους, δημιουργώντας ένα περιβάλλον υψηλής απόδοσης σε συνδυασμό με την προστασία της ιδιωτικότητας.

12.2 Ανάλυση του αντίκτυπου των μέτρων ασφαλείας στην απόδοση του μοντέλου

Η ένταξη μηχανισμών ευθυγράμμισης με στόχο την ασφάλεια, και πιο ειδικά η χρήση εξειδικευμένων τεχνικών, όπως η Ενισχυτική Μάθηση από Ανθρώπινη Ανατροφοδότηση, επιφέρει σημαντικό αλλά μετρήσιμο κόστος στη γενική απόδοση των μεγάλων γλωσσικών μοντέλων. Οι ερευνητές αποκαλούν το φαινόμενο αυτό Φόρο Ευθυγράμμισης και στην ουσία αποτυπώνει τον συμβιβασμό μεταξύ της αυξημένης ασφάλειας και της μειωμένης χρηστικότητας. Εμπειρικές αναλύσεις δείχνουν ότι καθώς αυξάνεται η αυστηρότητα των φίλτρων ασφαλείας, υποχωρούν οι επιδόσεις σε πιο απαιτητικές γνωστικές εργασίες, όπως για παράδειγμα η συγγραφή κώδικα ή η επίλυση σύνθετων μαθηματικών προβλημάτων. Πιο αναλυτικά, η διαδικασία προσαρμοστικής εκπαίδευσης (fine tuning) με σκοπό τον περιορισμό επιβλαβούς περιεχομένου έχει αποδειχθεί πως προκαλεί μια μετατόπιση από την αρχική κατανομή γνώσης του μοντέλου, καταλήγοντας σε χαμηλότερη απόδοση σε καθιερωμένα σύνολα αξιολόγησης, στα οποία η ακρίβεια παρουσιάζεται αισθητά μειωμένη σε σχέση με μη ευθυγραμμισμένες εκδοχές του μοντέλου.

Μια εξίσου καίρια, αλλά πολύ συχνά παραμελημένη πτυχή του αντίκτυπου αυτού, είναι η συμπεριφορική συνέπεια της υπερβολικής άρνησης (over-refusal). Τα μοντέλα που εκπαιδεύονται με έντονα προσανατολισμένα πρωτόκολλα ασφαλείας υιοθετούν την τάση να γίνονται υπερβολικά αμυντικά, απορρίπτοντας συχνά αθώα ή καλοπροαίρετα αιτήματα τα οποία δεν ενέχουν ουσιαστικό κίνδυνο. Αυτό το φαινόμενο της αδυναμίας ακριβούς διάκρισης ανάμεσα στις επικίνδυνες και ασφαλείς εντολές αποτυπώνεται μέσω του δείκτη ψευδούς άρνησης (false refusal rate), ο οποίος αντανακλά την άμεση μείωση της χρηστικότητας. Έτσι, προτεραιοποιείται η αβλάβεια σε βάρος της παροχής βοήθειας, περιορίζοντας τη λειτουργικότητα του συστήματος σε πραγματικά σενάρια.

Η πραγματική ρίζα του προβλήματος της υπερβολικής άρνησης βρίσκεται στον τρόπο που τα μέτρα ασφαλείας επηρεάζουν τον λανθάνοντα χώρο του μοντέλου. Όταν στην εκπαίδευση ασφαλείας δεν διαχωρίζονται ξεκάθαρα οι αναπαραστάσεις των ασφαλών εννοιών από τις επισφαλείς, οι μηχανισμοί λήψης αποφάσεων τείνουν να ενεργοποιούνται και σε περιπτώσεις αβλαβών ερωτήσεων. Ως αποτέλεσμα, η πτώση στην απόδοση δεν οφείλεται μόνο στην απώλεια της γνώσης, αλλά και στη δομική αλλοίωση της ικανότητας του μοντέλου να αξιολογεί σωστά τα δεδομένα με τα συμφραζόμενά τους, κάτι που οδηγεί σε μια φοβική συμπεριφορά απέναντι σε όποια εντολή μοιάζει επιφανειακά επιβλαβής.

Εκτός των ποιοτικών επιπτώσεων, η λήψη υπερβολικών μέτρων ασφαλείας επηρεάζει και την υπολογιστική αποδοτικότητα κατά τη διαδικασία της συμπερασματολογίας. Η προσθήκη

επιπλέον εξωτερικών φίλτρων ελέγχου ή πιο σύνθετων στρατηγικών αποκωδικοποίησης για την αποφυγή της τοξικότητας πρακτικά αυξάνει το χρόνο απόκρισης και τις απαιτήσεις σε πόρους μνήμης, επιβαρύνοντας την ταχύτητα λειτουργίας του συστήματος. Στα περιβάλλοντα όπου η ταχύτητα απόκρισης είναι μεγίστης σημασίας, η καθυστέρηση μπορεί να χαρακτηριστεί και ως φόρος, καθώς απαιτείται επιπλέον επεξεργαστική ισχύς για να παραχθεί το ίδιο, ή και ποιοτικά κατώτερο, αποτέλεσμα σε σχέση με την έκδοση του μοντέλου χωρίς τόσο αυστηρούς περιορισμούς.

Συνεπώς, η επίτευξη της απόλυτης ασφάλειας στα μεγάλα γλωσσικά μοντέλα συνιστά μια διαδικασία συνεχούς εξισορρόπησης, που πολλές φορές συνοδεύεται από επιβλαβείς συνέπειες. Οι κυρίαρχες μέθοδοι ευθυγράμμισης, όπως το κλασικό RLHF, δηλαδή η προσαρμοστική εκπαίδευση του μοντέλου μέσω ανατροφοδότησης ανθρώπινων αξιολογητών, αποδεικνύονται ανεπαρκείς στο να διατηρούν ακέραιες τις λογικές ικανότητες του μοντέλου, γεγονός που καθιστά αναγκαίες τις πιο εξελιγμένες και στοχευμένες προσεγγίσεις. Τεχνικές όπως η μάθηση μέσω αντίθεσης (Contrastive Learning) και η στοχευμένη ρύθμιση των αναπαραστάσεων ασφαλείας προτείνονται από πολλούς ερευνητές σαν εναλλακτικές λύσεις για την άμβλυνση του φόρου ευθυγράμμισης, ώστε να μπορεί το μοντέλο να διατηρεί την ευφυΐα του ενώ παραμένει ασφαλές.

12.3 Πολιτικές λήψης αποφάσεων για το αποδεκτό επίπεδο κινδύνου

Ο καθορισμός του αποδεκτού επιπέδου κινδύνου κατά τη διάρκεια ανάπτυξης των μεγάλων γλωσσικών μοντέλων προϋποθέτει μια δυναμική στρατηγική διαδικασία ικανή να ισορροπήσει την καινοτομία και την ασφάλεια. Πολύ βασικό στοιχείο αυτής της πολιτικής είναι η αναγνώριση των διαφορών μεταξύ της διάθεσης ανάληψης κινδύνου (Risk Appetite) και της ανοχής κινδύνου (Risk Tolerance). Με τον όρο διάθεση ανάληψης κινδύνου ουσιαστικά ορίζεται το ρίσκο που ένας οργανισμός συμφωνεί να αναλάβει ώστε να επιτύχει τους στόχους του, ενώ στην ανοχή κινδύνου οριοθετείται το μέγιστο επίπεδο αβεβαιότητας που είναι ικανός να αντέξει ένας οργανισμός χωρίς να βρίσκεται σε κίνδυνο η βιωσιμότητά του. Στο πλαίσιο αυτό, η πολιτική λήψης αποφάσεων πρέπει να ευθυγραμμίζει τους δύο αυτούς δείκτες, ώστε να επιτυγχάνεται η τεχνολογική πρόοδος χωρίς να παραβιάζονται τα όρια ασφαλείας που έχουν θεσπιστεί για την προστασία των χρηστών και της φήμης του οργανισμού.

Εκτός των εσωτερικών στρατηγικών, οι πολιτικές λήψης αποφάσεων οφείλουν να ενσωματώνουν αυστηρά κανονιστικά πλαίσια, τα οποία θέτουν απαράβατα όρια, όπως για παράδειγμα αυτά που ορίζει ο κανονισμός της Ευρωπαϊκής Ένωσης για την Τεχνητή Νοημοσύνη AI Act. Με βάση αυτό το πλαίσιο, τα μοντέλα ταξινομούνται με κριτήριο την επικινδυνότητά τους σε επίπεδα που κυμαίνονται από τον απαράδεκτο κίνδυνο (Unacceptable Risk), ο οποίος είναι πλήρως απαγορευμένος, μέχρι τον υψηλό, περιορισμένο, και ελάχιστο κίνδυνο. Αυτές οι τυποποιημένες κατηγοριοποιήσεις κινδύνου μπορούν να χρησιμοποιηθούν ως φίλτρα συμμόρφωσης, καθώς θέτουν όρια ως προς τις επιτρεπτές χρήσεις. Η έγκριση για την κυκλοφορία ενός μοντέλου εξαρτάται πρώτα απ’ όλα από το αν η προβλεπόμενη χρήση του ανήκει σε απαγορευμένες κατηγορίες, αλλά και αν πληρούνται οι απαιτήσεις διαφάνειας, ελέγχου και τεκμηρίωσης που αντιστοιχούν σε αυτό.

Σε επιχειρησιακό επίπεδο, η αξιολόγηση, πέρα από τον εγγενή κίνδυνο του μοντέλου, εστιάζει στον υπολειπόμενο κίνδυνο (Residual Risk), δηλαδή στον κίνδυνο που παραμένει ύστερα από την εφαρμογή όλων των μέτρων μετριασμού. Η έννοια αυτή αποκτά ιδιαίτερη σημασία στον κόσμο των μεγάλων γλωσσικών μοντέλων, καθώς το ίδιο το σύστημα παρουσιάζει διαφορετικό προφίλ κινδύνου ανάλογα με το πλαίσιο χρήσης. Έχουν αναπτυχθεί σύγχρονα μεθοδολογικά εργαλεία, όπως οι μηχανές αξιολόγησης κινδύνου (Risk Assessment Engines), οι οποίες ποσοτικοποιούν τον κίνδυνο λαμβάνοντας υπόψη κάθε φορά το συγκεκριμένο σενάριο χρήσης, υποστηρίζοντας τεκμηριωμένες αποφάσεις

σχετικά με το αν ο υπολειπόμενος κίνδυνος βρίσκεται εντός των προκαθορισμένων ορίων ανοχής.

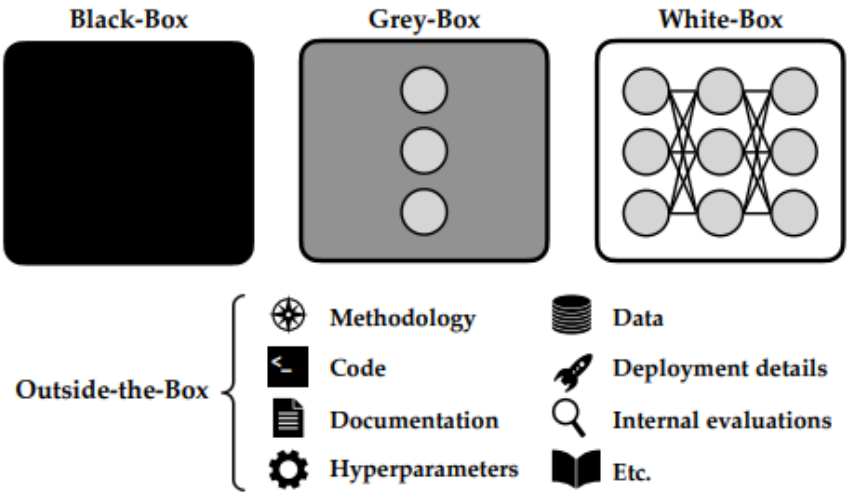
Στις περιπτώσεις που η αβεβαιότητα παραμένει σε υψηλές τιμές ή όταν ο αλγοριθμικός προσδιορισμός του κινδύνου υπερβαίνει τα αποδεκτά όρια, οι πολιτικές επιβάλλουν την ενεργή εμπλοκή του ανθρώπινου παράγοντα. Η προσέγγιση της ανθρώπινης εποπτείας διασφαλίζει ότι οι κρίσιμες αποφάσεις, ειδικά σε σενάρια που ενέχουν υψηλό ρίσκο, υπόκεινται σε ανθρώπινη κρίση, λειτουργώντας ως δικλείδα ασφαλείας. Αυτός ο μηχανισμός αποτελεί το τελικό επίπεδο ελέγχου, επιτρέποντας την έγκαιρη αναγνώριση και διόρθωση σφαλμάτων που μπορεί να ξεφύγουν από τα αυτοματοποιημένα συστήματα αξιολόγησης, διατηρώντας τη δυνατότητα προσαρμογής σε απρόβλεπτα σενάρια και ενισχύοντας τη συνολική ανθεκτικότητα και αξιοπιστία του συστήματος.

13 ΔΙΑΚΥΒΕΡΝΗΣΗ, ΕΛΕΓΞΙΜΟΤΗΤΑ ΚΑΙ ΠΙΣΤΟΠΟΙΗΣΗ ΜΟΝΤΕΛΩΝ

13.1 Μηχανισμοί Ελέγχου Τρίτων Μερών

Η ενσωμάτωση ισχυρών μηχανισμών ελέγχου τρίτων μερών αποτελεί θεμελιώδη πυλώνα για τη σύγχρονη διακυβέρνηση της τεχνητής νοημοσύνης, ειδικά όσον αφορά τα μεγάλα γλωσσικά μοντέλα, καθώς το να βασίζονται αποκλειστικά σε εσωτερικές αξιολογήσεις περιορίζει σημαντικά τη διαφάνεια και δεν διασφαλίζει την αντικειμενικότητα των αποτελεσμάτων. Διεθνείς έρευνες επισημαίνουν πως οι εξωτερικοί έλεγχοι λειτουργούν ως βασικό εργαλείο εποπτείας, εφόσον οι επιθεωρητές διαθέτουν ουσιαστική πρόσβαση στη δομή του συστήματος και δεν περιορίζονται στην επιφανειακή παρατήρηση της συμπεριφοράς του. Ειδικότερα στα πιο σύνθετα γενετικά μοντέλα, εφόσον η αξιολόγησή τους πραγματοποιείται αποκλειστικά μέσω των παρατηρήσιμων εισόδων και εξόδων (black box) δεν επαρκεί για να αποκαλύψει τους εσωτερικούς μηχανισμούς που ευθύνονται για επικίνδυνες δραστηριότητες, ανεπιθύμητη απομνημόνευση, ή λανθασμένες διαδικασίες εκπαίδευσης.

Η πρόσβαση τύπου μαύρου κουτιού έχει αναδειχθεί ως δομικά ανεπαρκής για τη διενέργεια αυστηρών ελέγχων, αφού περιορίζει τους επιθεωρητές στην απλή υποβολή ερωτημάτων και στη γενική παρατήρηση των απαντήσεων, μια διαδικασία που καθίσταται ευάλωτη σε μεροληψίες και δεν επιτρέπει την εξαγωγή γενικεύσιμων συμπερασμάτων, ούτε την παροχή αξιόπιστων εξηγήσεων της συμπεριφοράς του μοντέλου. Από την άλλη, η παροχή κλιμακωτής πρόσβασης, η οποία διακρίνεται σε πρόσβαση λευκού κουτιού (white box), όπου ο επιθεωρητής έχει πρόσβαση στις εσωτερικές παραμέτρους του μοντέλου, όπως τα βάρη και οι κλίσεις του, και εκτός κουτιού (outside the box), η οποία περιλαμβάνει στοιχεία εσωτερικής λειτουργίας και πληροφορίες εκτός του ίδιου του μοντέλου, όπως η τεχνική τεκμηρίωση, η μεθοδολογία ανάπτυξης τα σύνολα δεδομένων εκπαίδευσης, και ευρήματα εσωτερικών επιθεωρήσεων, γεγονός που καθιστά τον έλεγχο πιο ουσιαστικό και αντικειμενικό. Η διάκριση μεταξύ μαύρου, λευκού, και εκτός κουτιού αναδεικνύει ότι η ποιότητα του ελέγχου βασίζεται στο εύρος της πληροφόρησης που παρέχεται στους εξωτερικούς επιθεωρητές.



Εικόνα 11: Επίπεδα πρόσβασης ελέγχου: Black- box, Grey-Box, White-Box, και Outside-the-box.

Σε θεσμικό επίπεδο, οι μηχανισμοί ελέγχου από εξωτερικούς φορείς ενσωματώνονται στο ευρωπαϊκό κανονιστικό πλαίσιο μέσω της διαδικασίας αξιολόγησης συμμόρφωσης (conformity assessment), όπως προβλέπεται στον κανονισμό για την τεχνητή νοημοσύνη (AI Act). Αυτή η διαδικασία ξεφεύγει από τους απλούς τεχνικούς ελέγχους απόδοσης, και συνιστά μια ολοκληρωμένη επαλήθευση του κατά πόσο πληρούνται οι απαιτήσεις του κανονισμού, περιλαμβάνοντας τον έλεγχο της τεχνικής τεκμηρίωσης, του συστήματος διαχείρισης ποιότητας, τις πρακτικές διακυβέρνησης δεδομένων, και την ύπαρξη μηχανισμών παρακολούθησης μετά τη διάθεση του συστήματος στην αγορά. Σε ορισμένες κατηγορίες συστημάτων υψηλού κινδύνου, όπως για παράδειγμα τα συστήματα βιομετρικής ταυτοποίησης, η επιθεώρηση απαγορεύεται να διενεργηθεί αποκλειστικά από τον ίδιο τον πάροχο, καθώς υποχρεούται να την αναθέσει σε κοινοποιημένους οργανισμούς, οι οποίοι λειτουργούν ως ανεξάρτητα τρίτα μέρη, διαπιστευμένα από τις εθνικές αρχές, με θεσμικά καθορισμένες αρμοδιότητες και περιοδική εποπτεία.

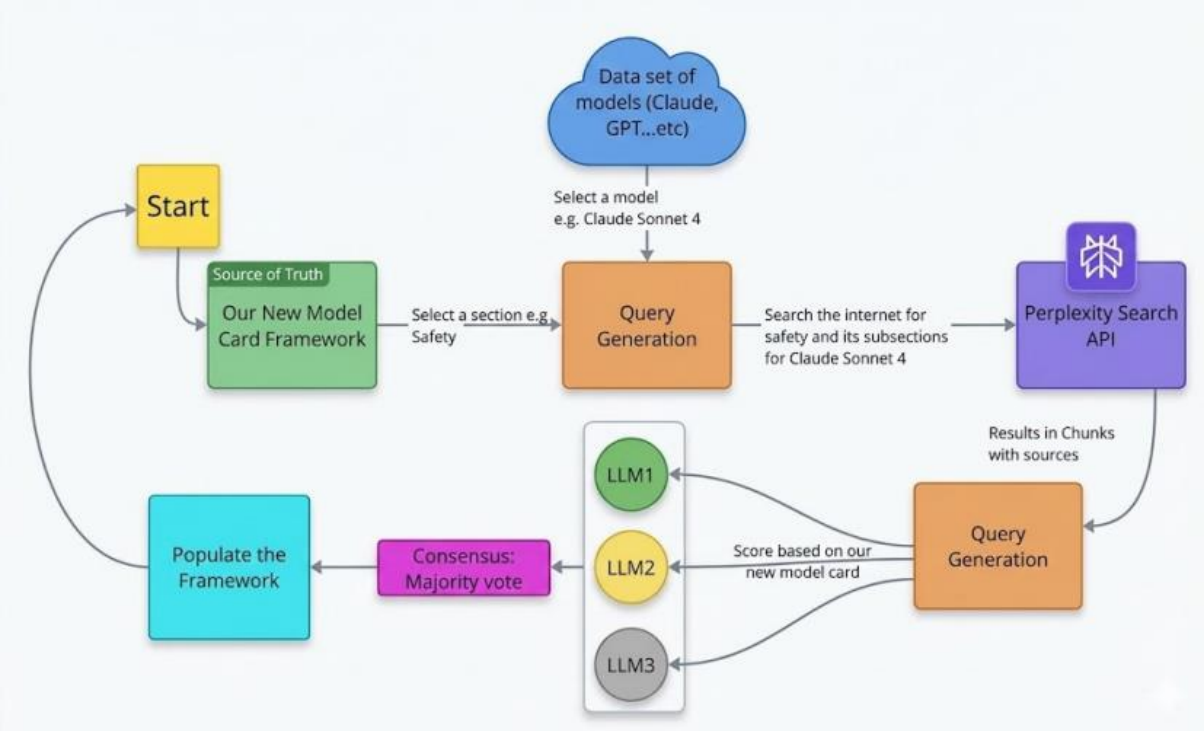
Ωστόσο, η πρακτική υλοποίηση των επιθεωρήσεων από τρίτα μέρη δημιουργεί μια φυσική ένταση ανάμεσα στην ανάγκη για διαφάνεια και στην προστασία εμπορικών μυστικών ή της πνευματικής ιδιοκτησίας των φορέων. Στις περιπτώσεις λευκού κουτιού, αν η πρόσβαση παρέχεται ανεξέλεγκτα, δημιουργείται κίνδυνος διαρροής του ίδιου του μοντέλου ή ευαίσθητων δεδομένων, κάτι που μπορεί να λειτουργήσει αποτρεπτικά για τους παρόχους. Για την αντιμετώπιση αυτών των απειλών, προτείνονται ειδικές τεχνικές, φυσικές, και νομικές δικλείδες ασφαλείας, όπως ελεγχόμενα περιβάλλοντα αξιολόγησης, περιορισμένη πρόσβαση στα κρίσιμα στοιχεία του συστήματος, και συμβατικές υποχρεώσεις εμπιστευτικότητας, ώστε να διασφαλίζεται η ουσιαστικότητα του ελέγχου χωρίς να κινδυνεύει η ασφάλεια και η πνευματική ιδιοκτησία.

Συμπερασματικά, οι μηχανισμοί επιθεώρησης από τρίτα μέρη αποτελούν ένα διακριτό και αναγκαίο επίπεδο διακυβέρνησης για τα μεγάλα γλωσσικά μοντέλα, το οποίο συμπληρώνει τις εσωτερικές διαδικασίες των οργανισμών. Η αποτελεσματικότητά τους συνδέεται άμεσα με το βάθος της πρόσβασης που παρέχεται στους επιθεωρητές, αλλά και από την ένταξή τους σε ένα θεσμοθετημένο πλαίσιο συμμόρφωσης και εποπτείας, όπου η ανεξαρτησία τους θα είναι προστατευμένη. Χωρίς αυτές τις προϋποθέσεις, οι έλεγχοι κινδυνεύουν να καταλήξουν επιφανειακοί, οδηγώντας σε φαινόμενα ξεπλύματος ασφαλείας, στα οποία δημιουργείται μια ψεύτικη αίσθηση ασφάλειας χωρίς πραγματική λογοδοσία.

13.2 Κάρτες Μοντέλων και Κάρτες Δεδομένων ως εργαλεία διαφάνειας

Η σύγχρονη πορεία των μεγάλων γλωσσικών μοντέλων χαρακτηρίζεται από τη σταδιακή υποχώρηση της ουσιαστικής διαφάνειας, αφού οι τεχνολογικοί κολοσσοί απομακρύνονται ολοένα και περισσότερο από τις πρακτικές ανοιχτής διάθεσης πληροφοριών. Μελέτες που αναλύουν διαχρονικά την τεκμηρίωση που παρέχουν τα μοντέλα GPT έχουν ανακαλύψει πως οι αρχικές εκδόσεις συνοδεύονταν από εκτενείς περιγραφές αρχιτεκτονικής, δεδομένων και μεθοδολογίας εκπαίδευσης, ενώ οι πιο πρόσφατες περιορίζουν εμφανώς το εύρος και το βάθος των δημοσιοποιούμενων πληροφοριών. Αυτή η αυξανόμενη αφαίρεση καίριων τεχνικών στοιχείων δυσχεραίνει την επιστημονική αξιολόγηση, περιορίζει τη δυνατότητα κατανόησης των πιθανών κινδύνων, και ενισχύει ασύμμετρες σχέσεις ισχύος μεταξύ χρηστών και παρόχων. Σε ένα τέτοιο περιβάλλον, υπάρχει ένα κενό που πρέπει να καλυφθεί από τυποποιημένα τεκμήρια διαφάνειας, τα οποία πρέπει να λειτουργούν ως σταθερά σημεία αναφοράς για τους ερευνητές, τις ρυθμιστικές αρχές, αλλά και τους κοινωνικούς φορείς.

Για την τεκμηρίωση των ίδιων των αλγοριθμικών συστημάτων έχουν αναπτυχθεί οι κάρτες μοντέλων, οι οποίες αποτυπώνουν την ταυτότητα ενός μοντέλου, τις προβλεπόμενες χρήσεις του, τους περιορισμούς του, αλλά και τις μεθόδους αξιολόγησής του. Ωστόσο, διαφορετικές εμπειρικές αναλύσεις της υφιστάμενης πρακτικής τεκμηρίωσης έχουν αποδείξει πως η εφαρμογή τους είναι αποσπασματική και ασυνεπής. Όταν αναλύεται μεγάλος αριθμός καρτών μοντέλων φαίνεται έντονα η ετερογένεια στη δομή, στην ονοματολογία, και στο περιεχόμενο, καθώς κρίσιμες πληροφορίες που αφορούν την ασφάλεια συχνά απουσιάζουν ή αναφέρονται επιφανειακά. Η ανάπτυξη πλαισίων αξιολόγησης και σύγχρονων συστημάτων βαθμολόγησης διαφάνειας, όπως προτείνεται στο AI Transparency Atlas, ένα ερευνητικό πλαίσιο και εργαλείο οπτικοποίησης για την αξιολόγηση της διαφάνειας των μοντέλων, αλλά και όπως αυτά που ευθυγραμμίζονται με τις κανονιστικές απαιτήσεις τεχνικής τεκμηρίωσης της Ευρωπαϊκής Ένωσης, επιτρέπει τη μετατροπή των καρτών από απλά αφηγηματικά έγγραφα σε μετρήσιμα εργαλεία λογοδοσίας και συγκρισιμότητας ανάμεσα στα συστήματα.



Εικόνα 12: Ροή εργασίας για τη δημιουργία και συμπλήρωση Καρτών Μοντέλου (Model Cards).

Εξίσου κρίσιμη με την τεκμηρίωση των μοντέλων είναι η διαφάνεια στα δεδομένα εκπαίδευσης, η οποία επιτυγχάνεται μέσω των καρτών δεδομένων ή φύλλων δεδομένων (Datasheets). Τα φύλλα δεδομένων οργανώνουν πληροφορίες που αφορούν το κίνητρο για τη δημιουργία του συνόλου δεδομένων, τη σύνθεσή του, τη διαδικασία συλλογής, καθώς και την προεπεξεργασία που έχει υποστεί. Ιδίως σε ευαίσθητους τομείς, η λεπτομερής καταγραφή της κατανομής των δεδομένων και των προτεινόμενων χρήσεων συνιστά ασπίδα απέναντι σε αλγοριθμικές μεροληψίες, και μειώνει την πιθανότητα δημιουργίας ψευδούς αίσθησης ουδετερότητας και αναπαραγωγής κοινωνικών ανισοτήτων. Με μια τέτοια τυποποιημένη καταγραφή καθίσταται εφικτός ο έγκαιρος εντοπισμός κενών και προκαταλήψεων στα δεδομένα, διασφαλίζοντας την σαφήνεια και ελεγξιμότητα των πηγών πληροφορίας στις οποίες βασίζεται η ανάπτυξη του συστήματος.

Τέλος, η αξία των καρτών μοντέλων και δεδομένων επεκτείνεται πέραν του επιπέδου εθνικής δεοντολογίας και διαφάνειας, στο θέμα της νομικής συμμόρφωσης, ιδίως υπό το πρίσμα του Ευρωπαϊκού Κανονισμού για την Τεχνητή Νοημοσύνη (AI Act). Οι απαιτήσεις για την τεχνική τεκμηρίωση μπορούν να υποστηριχθούν από σημασιολογικά πλαίσια, τα οποία χαρτογραφούν τις νομικές υποχρεώσεις, διευκολύνοντας την ενσωμάτωση των καρτών σε διαδικασίες διαχείρισης κινδύνου και ελέγχου συμμόρφωσης. Αξιοποιώντας οντολογίες και δομημένες αναπαραστάσεις, η τεκμηρίωση μετατρέπεται από στατικό σε δυναμικό μηχανισμό παρακολούθησης, το οποίο γεφυρώνει το χάσμα της τεχνικής ανάπτυξης και των κανονιστικών απαιτήσεων για την ασφαλή και αξιόπιστη χρήση της τεχνητής νοημοσύνης.

13.3 Προδιαγραφές για πιστοποίηση ασφάλειας και ιδιωτικότητας

Η πιστοποίηση της ασφάλειας και της ιδιωτικότητας στα μεγάλα γλωσσικά μοντέλα δεν πρέπει να περιορίζεται στον τεχνικό έλεγχο του κώδικα, αλλά να προϋποθέτει την υιοθέτηση ενός ολοκληρωμένου σύστημα διαχείρισης που διέπει ολόκληρη τη λειτουργία του συστήματος. Σύγχρονες μελέτες αναδεικνύουν ότι η συμμόρφωση δεν σταματά στην απλή αξιολόγηση αλγορίθμων και στην ανάλυση του πηγαίου κώδικα, καθώς προϋποθέτει την ύπαρξη θεσμοθετημένων μηχανισμών διακυβέρνησης που διατρέχουν ολόκληρο τον οργανισμό. Με αυτή τη λογική, η πιστοποίηση λειτουργεί σαν ένα εργαλείο ενσωμάτωσης της ασφάλειας και της προστασίας δεδομένων στον κύκλο ζωής των συστημάτων τεχνητής νοημοσύνης, από την αρχή του σχεδιασμού μέχρι τη λειτουργία και την απόσυρσή τους.

Βασικό θεμέλιο για αυτή την προσέγγιση αποτελεί το διεθνές πρότυπο ISO/IEC 42001, το οποίο εισάγει τις κύριες απαιτήσεις για την εγκαθίδρυση, την υλοποίηση, και τη συνεχή βελτίωση ενός συστήματος διαχείρισης τεχνητής νοημοσύνης (AIMS), παρέχοντας το απαραίτητο πλαίσιο για την πιστοποίηση των οργανωτικών διαδικασιών. Το πρότυπο εκτός των τεχνικών χαρακτηριστικών, επικεντρώνεται στη θεσμική οργάνωση της γενικότερης υπεύθυνης χρήσης τέτοιου είδους συστημάτων, καθώς βασίζεται στη λογική του συνεχούς κύκλου βελτίωσης Plan-Do-Check-Act, όπως στο ISO 27001 που αφορά την πιστοποίηση ενός ώριμου συστήματος διαχείρισης ασφάλειας πληροφοριών. Μέσω μιας τέτοιας δομής, ζητήματα όπως η λογοδοσία, η διαφάνεια, η τεκμηρίωση των αποφάσεων, και η διαχείριση των κινδύνων ενσωματώνονται αυτόματα στις επιχειρησιακές διαδικασίες, καθιστώντας την πιστοποίηση ένα αποτέλεσμα μιας ώριμης οργανωτικής πρακτικής και όχι απλής τυπικής συμμόρφωσης.

Παράλληλα με τα πρότυπα διαχείρισης συστημάτων, η διαδικασία πιστοποίησης ολοκληρώνεται σε στιβαρά πλαίσια διαχείρισης κινδύνου, όπως το NIST AI Risk Management Framework (AI RMF), το οποίο αναδεικνύεται ως χαρακτηριστικό παράδειγμα προσέγγισης που γεφυρώνει τις παραδοσιακές πρακτικές κυβερνοασφάλειας με τις σύγχρονες ιδιαιτερότητες των συστημάτων τεχνητής νοημοσύνης. Χρησιμοποιώντας τέσσερις βασικές λειτουργίες: τη διακυβέρνηση, τη χαρτογράφηση, τη μέτρηση, και τη

διαχείριση, το πλαίσιο αυτό παρέχει μια δομημένη μεθοδολογία για τον εντοπισμό και τον έλεγχο της σοβαρότητας των κινδύνων που μπορούν να απορρέουν από την πολυπλοκότητα, τη δυναμική συμπεριφορά, αλλά και την αλληλεπίδραση των μοντέλων με ανθρώπους και δεδομένα.

Σε καθαρά τεχνικό επίπεδο, οι προδιαγραφές ασφάλειας για τα μεγάλα γλωσσικά μοντέλα καθορίζονται από πιο εξειδικευμένα πρότυπα, τα οποία χαρτογραφούν τις πιο κρίσιμες ευπάθειες, όπως το OWASP Top 10 for LLM Applications. Για να λάβει ένα σύστημα πιστοποίηση ασφάλειας, χρειάζεται να αποδείξει την ανθεκτικότητά του απέναντι σε συγκεκριμένες απειλές, με κυριότερη την επίθεση τύπου ενέσιμης εντολής (Prompt Injection), μέσω της οποίας ο επιτιθέμενος μπορεί να χειραγωγήσει το μοντέλο μέσω κακόβουλων εισόδων. Επίσης, οι προδιαγραφές ελέγχου δείχνουν ιδιαίτερο ενδιαφέρον στην προστασία από τη διαρροή ευαίσθητων πληροφοριών και στην ασφάλεια της εφοδιαστικής αλυσίδας, καθώς η χρήση ευπαθών στοιχείων ή δεδομένων τρίτων μερών μπορεί να υπονομεύσει τόσο την ασφάλεια του μοντέλου όσο και την εμπιστοσύνη των χρηστών.

Συνολικά, οι προδιαγραφές πιστοποίησης της ιδιωτικότητας θεμελιώνονται άμεσα στον Γενικό Κανονισμό Προστασίας Δεδομένων (GDPR) και συγκεκριμένα από το άρθρο 42, το οποίο εισάγει τη δυνατότητα επίσημων μηχανισμών πιστοποίησης, που θα λειτουργούν ως αποδεικτικά συμμόρφωσης με τις αρχές προστασίας δεδομένων εκ σχεδιασμού και εξ ορισμού. Χαρακτηριστικό παράδειγμα αποτελεί το σχήμα Europrivacy, το οποίο είναι εγκεκριμένο ως επίσημη Ευρωπαϊκή Σφραγίδα Προστασίας Δεδομένων για να αξιολογεί συστηματικά τη νομιμότητα της επεξεργασίας, τη διαφάνεια, τα τεχνικά και οργανωτικά αντίμετρα ασφαλείας, και τον σεβασμό των δικαιωμάτων των υποκειμένων των δεδομένων. Έτσι, το εργαλείο αυτό καθιστά την πιστοποίηση ένα ουσιαστικό εργαλείο εμπιστοσύνης και κανονιστικής ωριμότητας.

14 ΣΥΜΠΕΡΑΣΜΑΤΑ, ΠΕΡΙΟΡΙΣΜΟΙ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

14.1 Συνοπτικά συμπεράσματα

Η παρούσα εργασία ανέδειξε ότι τα μεγάλα γλωσσικά μοντέλα εισάγουν μια διαφορετική κατηγορία κινδύνων για την ασφάλεια και την ιδιωτικότητα, οι οποίοι δεν μπορούν να εντοπιστούν και να καταπολεμηθούν επαρκώς από τους παραδοσιακούς τρόπους αξιολόγησης λογισμικού. Παράγοντες όπως ο τεράστιος όγκος των δεδομένων εκπαίδευσης, η πολυπλοκότητα της εσωτερικής λειτουργίας και η ικανότητα παραγωγής νέου περιεχομένου, μεταβάλλουν ριζικά τη φύση των κινδύνων διαρροής και ακούσιας αποκάλυψης πληροφοριών. Καθίσταται, συνεπώς, σαφές πως το ζήτημα προστασίας προσωπικών δεδομένων στα μεγάλα γλωσσικά μοντέλα δεν αποτελεί αποσπασματικό πρόβλημα, αλλά συστημική και διαχρονική πρόκληση.

Η ανάλυση των τεχνικών μηχανισμών επιβεβαίωσε ότι ορισμένα φαινόμενα, όπως η απομνημόνευση, η αναπαραγωγή δεδομένων εκπαίδευσης, και οι επιθέσεις μέσω χειραγώγησης εισόδων αποτελούν καίριες εστίες κινδύνου, ακόμη και σε περιβάλλοντα όπου εφαρμόζονται βασικά μέτρα ασφάλειας. Παράλληλα, η διερεύνηση τεχνικών προστασίας της ιδιωτικότητας ανέδειξε την ύπαρξη εγγενών αντισταθμισμάτων ανάμεσα στο επίπεδο προστασίας και στην λειτουργική απόδοση, υπογραμμίζοντας ότι οι τεχνικές λύσεις, αν και είναι απαραίτητες, δεν φτάνουν από μόνες τους για την εξασφάλιση της συμμόρφωσης και της εμπιστοσύνης των χρηστών.

Σε αυτό το πλαίσιο, η εργασία ανέδειξε τον κομβικό ρόλο της τεκμηρίωσης, της διαφάνειας, και των ελεγκτικών μηχανισμών για την υπεύθυνη διακυβέρνηση των μεγάλων γλωσσικών μοντέλων. Σύγχρονα εργαλεία, όπως οι κάρτες μοντέλων και δεδομένων, οι διαδικασίες

ελέγχου τρίτων μερών, και οι προδιαγραφές πιστοποίησης δεν λειτουργούν απλώς επικουρικά, αλλά γεφυρώνουν ουσιαστικά το χάσμα μεταξύ της τεχνικής πολυπλοκότητας και των κανονιστικών απαιτήσεων. Με την συστηματική ενσωμάτωσή τους καθ'όλη τη διάρκεια του κύκλου ζωής των συστημάτων διευκολύνεται η λογοδοσία και η ουσιαστική αξιολόγηση των απειλών.

Συμπερασματικά, η σύνθεση της τεχνικής και ρυθμιστικής ανάλυσης καταλήγει στη διαπίστωση ότι η ουσιαστική αντιμετώπιση των προκλήσεων ιδιωτικότητας στα μεγάλα γλωσσικά μοντέλα προϋποθέτει μια ολιστική προσέγγιση διακυβέρνησης. Η συμμόρφωση με το νομικό πλαίσιο και τα πρότυπα πιστοποίησης οφείλει να αντιμετωπίζεται ως ένας δυναμικός μηχανισμός συνεχούς αξιολόγησης, και όχι ως μια τυπική διαδικασία. Υπό αυτή την οπτική, η εργασία συνεισφέρει στη διαμόρφωση ενός συνεκτικού πλαισίου κατανόησης και αντιμετώπισης των κινδύνων, θέτοντας τα θεμέλια για περαιτέρω επιστημονική και θεσμική έρευνα.

14.2 Περιορισμοί της θεωρητικής προσέγγισης

Ένας βασικός περιορισμός της παρούσας εργασίας αφορά τη μεθοδολογική της φύση, η οποία βασίστηκε στη συστηματική βιβλιογραφική ανασκόπηση και στη συνθετική κανονιστική ανάλυση, χωρίς την εκτέλεση κάποιου πρωτογενούς εμπειρικού ελέγχου. Μολονότι η διερεύνηση των κινδύνων στηρίχθηκε σε τεκμηριωμένα περιστατικά και δημοσιευμένες μελέτες, δεν πραγματοποιήθηκαν πρακτικές δοκιμές, όπως ασκήσεις ελεγχόμενης επίθεσης ή τεχνικοί έλεγχοι διείσδυσης σε πραγματικά περιβάλλοντα λειτουργίας μεγάλων γλωσσικών μοντέλων. Συνεπώς, τα συμπεράσματα αφορούν την αποτελεσματικότητα των προτεινόμενων μηχανισμών ελέγχου και διακυβέρνησης σε θεωρητικό επίπεδο και η πρακτική τους εφαρμογή ενδέχεται να αναδείξει πρόσθετες τεχνικές δυσκολίες που δεν ήταν δυνατόν να αποτυπωθούν στο πλαίσιο της παρούσας θεωρητικής διερεύνησης.

Επιπρόσθετα, η ανάλυση επηρεάζεται αναπόφευκτα από τη ραγδαία και συνεχή εξέλιξη του πεδίου της τεχνητής νοημοσύνης. Οι ταχύτατες αλλαγές στις αρχιτεκτονικές των μεγάλων γλωσσικών μοντέλων, αλλά και η παράλληλη διαμόρφωση του ρυθμιστικού τοπίου, όπως η σταδιακή εφαρμογή του Ευρωπαϊκού Κανονισμού για την Τεχνητή Νοημοσύνη (AI Act), καθιστούν τα ευρήματα της εργασίας χρονικά προσδιορισμένα. Δεν είναι απίθανο ορισμένες από τις τεχνικές προδιαγραφές ή τις ρυθμιστικές ερμηνείες, που παρουσιάζονται σήμερα, να απαιτήσουν αναθεώρηση στο άμεσο μέλλον, εφόσον νέα δεδομένα και νομολογιακές πρακτικές έρχονται στο φως, μεταβάλλοντας τις βέλτιστες πρακτικές. Υπό αυτή την έννοια, η παρούσα μελέτη αποτελεί μια φωτογραφική αποτύπωση της τρέχουσας συγκυρίας, χωρίς να αποκλείει την ανάγκη μελλοντικής αναθεώρησης ορισμένων συμπερασμάτων.

Τέλος, το προτεινόμενο πλαίσιο διακυβέρνησης και συμμόρφωσης υιοθετεί σκόπιμα έναν γενικευμένο χαρακτήρα, προκειμένου να παραμείνει εφαρμόσιμο σε ευρύ φάσμα περιπτώσεων χρήσης μεγάλων γλωσσικών μοντέλων. Η οριζόντια αυτή προσέγγιση, δεν λαμβάνει υπόψη τις εξειδικευμένες απαιτήσεις που ενδέχεται να προκύψουν σε τομείς αυξημένης ευαισθησίας, όπως η υγεία, η εθνική άμυνα, και ο χρηματοοικονομικός τομέας. Σε τέτοιες περιπτώσεις, οι απαιτήσεις ιδιωτικότητας είναι πιθανό να επιβάλλουν αυστηρότερες και περισσότερο εξειδικευμένες ρυθμίσεις, οι οποίες υπερβαίνουν το πεδίο της παρούσας γενικής ανάλυσης.

14.3 Μελλοντικές ερευνητικές κατευθύνσεις

Με βάση τα ευρήματα και τους μεθοδολογικούς περιορισμούς της παρούσας εργασίας, αναδεικνύονται συγκεκριμένες προοπτικές για την περαιτέρω εξέλιξη της έρευνας τόσο σε

τεχνικό όσο και σε θεσμικό επίπεδο. Ιδιαίτερο ενδιαφέρον συγκεντρώνει η διερεύνηση μεθόδων μηχανικής επιλεκτικής απομάθησης από μεγάλα γλωσσικά μοντέλα, καθώς η απλή αφαίρεση δεδομένων εκπαίδευσης δεν διασφαλίζει την πλήρη εξάλειψη της αντίστοιχης επιρροής της πληροφορίας από τη συμπεριφορά του μοντέλου. Έτσι, η ανάπτυξη αξιόπιστων και επαληθεύσιμων αλγορίθμων που βοηθούν το σύστημα να ξεχνά επιλεκτικά, χωρίς να απαιτείται η πλήρης επανεκπαίδευση, θα μπορούσε να συμβάλει ουσιαστικά στη λειτουργική υλοποίηση των αρχών προστασίας δεδομένων, και ειδικότερα στις περιπτώσεις που απαιτείται η αναδρομική συμμόρφωση.

Παράλληλα, η σταδιακή στροφή προς τα πολυτροπικά μοντέλα, τα οποία μπορούν να επεξεργάζονται συνδυαστικά κείμενο, εικόνα, και ήχο, δημιουργεί καινούργιες προκλήσεις για την προστασία της ασφάλειας και της ιδιωτικότητας. Η μελλοντική έρευνα καλείται να χαρτογραφήσει τον τρόπο που διαφοροποιείται το προφίλ του κινδύνου στις περιπτώσεις που εισάγονται ετερογενή δεδομένα, καθώς και να αναπτύξει σενάρια ελέγχου για την αντιμετώπιση καταχρήσεων, οι οποίες ενδέχεται να διαφεύγουν των υφιστάμενων μηχανισμών φιλτραρίσματος κειμένου. Η κατανόηση τέτοιου είδους φαινομένων μπορεί να οδηγήσει στη διαμόρφωση νέων πλαισίων αξιολόγησης που ανταποκρίνονται στη σύνθετη φύση των σύγχρονων συστημάτων.

Συνολικά, σε επίπεδο διακυβέρνησης και συμμόρφωσης, πρέπει να επιβάλλεται η μετάβαση από στατικές διαδικασίες ελέγχου, σε πιο δυναμικές και διαρκούς επιτήρησης μορφές παρακολούθησης. Η περαιτέρω έρευνα θα μπορούσε να εστιάσει στην τυποποίηση αυτοματοποιημένων μηχανισμών που θα ευθυγραμμίζουν τις κανονιστικές απαιτήσεις με τις βέλτιστες πρακτικές ανάπτυξης και λειτουργίας των μοντέλων, καθώς και στην παραμετροποίηση του προτεινόμενου πλαισίου σε τομείς αυξημένης ευαισθησίας. Μια τέτοια εξειδίκευση σε τομείς, όπως η υγεία ή η εθνική άμυνα, θα υποστήριζε την ουσιαστικότερη ενσωμάτωση αρχών ασφαλείας και ιδιωτικότητας σε περιβάλλοντα με αυστηρές επιχειρησιακές και ηθικές απαιτήσεις, μεγιστοποιώντας την πραγματική αξία των συμπερασμάτων της παρούσας μελέτης.

15 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Cheng, X. (2025). *Approaching Memorization in Large Language Models* (Master's thesis). University of Waterloo, Ontario, Canada.
2. Huang, Y., Xu, J., Lai, J., Jiang, Z., Chen, T., Li, Z., ... & Ma, X. (2024). Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv:2311.12351*.
3. Jindal, I., Badrinath, C., Bharti, P., Vinay, L., & Sharma, S. D. (2024). Balancing Continuous Pre-Training and Instruction Fine-Tuning: Optimizing Instruction-Following in LLMs. *arXiv preprint arXiv:2410.10739*.
4. Liu, B. (n.d.). Comparative Analysis of Encoder-Only, Decoder-Only, and Encoder-Decoder Language Models. *College of Liberal Arts & Sciences, University of Illinois Urbana-Champaign*.
5. Liu, Y. (2025). Attention is All Large Language Model Need. *ITM Web of Conferences, 73*, 02025. https://doi.org/10.1051/itmconf/20257302025

6. Shukla, S. M., & Magoo, C. (2024). Comparing Fine Tuned-LMs for Detecting LLM-Generated Text. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*. IEEE.
7. Sun, K., & Dredze, M. (n.d.). Amuro & Char: Analyzing the Relationship between Pre-Training and Fine-Tuning of Large Language Models. *Johns Hopkins University*.
8. Zheng, Z., Wang, Y., Huang, Y., Song, S., Yang, M., Tang, B., ... & Li, Z. (2025). Attention heads of large language models. *Patterns*. https://doi.org/10.1016/j.patter.2025.101176
9. Mienye, I. D., Jere, N., Obaido, G., Ogunruku, O. O., Esenogho, E., & Modisane, C. (2025). Large language models: an overview of foundational architectures, recent trends, and a new taxonomy. *Discover Applied Sciences, 7*, 1027. https://doi.org/10.1007/s42452-025-07668-w
10. Neel, S., & Chang, P. W. (2024). Privacy Issues in Large Language Models: A Survey. *arXiv preprint arXiv:2312.06717*.
11. Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., ... & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine, 154*, 102900.
12. Thakkar, O., Ramaswamy, S., Mathews, R., & Beaufays, F. (n.d.). Understanding Unintended Memorization in Language Models Under Federated Learning. *Google LLC*.
13. Wang, Z., Chu, Z., Doan, T. V., Ni, S., Yang, M., & Zhang, W. (2024). History, development, and principles of large language models: an introductory survey. *AI and Ethics, 5*, 1955-1971. https://doi.org/10.1007/s43681-024-00583-7
14. Bai, Y., Pei, G., Gu, J., Yang, Y., & Ma, X. (2024). Special Characters Attack: Toward Scalable Training Data Extraction From Large Language Models. *arXiv preprint arXiv:2405.05990*.
15. Li, Z., Wu, Y., Chen, Y., Tonin, F., Rocamora, E. A., & Cevher, V. (n.d.). Membership Inference Attacks against Large Vision-Language Models. *LIONS, EPFL*.
16. Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., & Berg-Kirkpatrick, T. (2023). Membership Inference Attacks against Language Models via Neighbourhood Comparison. *arXiv preprint arXiv:2305.18462*.
17. Meeus, M., Shilov, I., Jain, S., Faysse, M., Rei, M., & de Montjoye, Y. A. (2025). SoK: Membership Inference Attacks on LLMs are Rushing Nowhere (and How to Fix It). *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*.
18. Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., ... & Lee, K. (2023). Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.
19. Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., ... & Lee, K. (2025). Scalable Extraction of Training Data from Aligned, Production Language Models. *International Conference on Learning Representations (ICLR)*.
20. Zhao, K., Gong, N. Z., Li, L., & Zhao, Y. (n.d.). A Survey on Model Extraction Attacks and Defenses for Large Language Models. *University of Notre Dame & Florida State University*.
21. Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., ... & Xu, K. (2024). Privacy Leakage on DNNs: A Survey of Model Inversion Attacks and Defenses. *arXiv preprint arXiv:2402.04013*.
22. Hao, G., & Wu, J. (2025). Privacy-Preserving Prompt Injection Detection for Smart Cloud-Deployed Large Language Models. *2025 IEEE 10th International Conference on Smart Cloud (SmartCloud)*.
23. Li, Y., Anciaux, N., Ghozzi, S., Bensamoun, A., Eichler, C., & Manzano, L. G. (2025). Data Provenance Auditing of Fine-Tuned Large Language Models with a Text-Preserving Technique. *arXiv preprint arXiv:2510.09655*.

24. Luo, X., & Yu, T. (n.d.). Prompt Inference Attack on Distributed Large Language Model. *Mohamed bin Zayed University of Artificial Intelligence*.
25. Mukherjee, K. (n.d.). LLM-driven Provenance Forensics for Threat Intelligence and Detection. *Virginia Tech*.
26. Qu, W., Zhou, Y., Wu, Y., Xiao, T., Yuan, B., Li, Y., & Zhang, J. (2025). Prompt Inversion Attack against Collaborative Inference of Large Language Models. *2025 IEEE Symposium on Security and Privacy (SP)*.
27. Thomas, R., Zahran, L., Choi, E., Potti, A., Goldblum, M., & Pal, A. (n.d.). Hidden No More: Attacking and Defending Private Third-Party LLM Inference. *arXiv preprint*.
28. Yang, W., Wang, S., Wu, D., Cai, T., Zhu, Y., Wei, S., ... & Li, Y. (2025). Deep learning model inversion attacks and defenses: a comprehensive survey. *Artificial Intelligence Review, 58*, 242.
29. Zeng, Z., Wang, J., Yang, J., Lu, Z., Li, H., Zhuang, H., & Chen, C. (n.d.). Privacy Restore: Privacy-Preserving Inference in Large Language Models via Privacy Removal and Restoration. *South China University of Technology*.
30. Zhan, J., Zhang, W., Zhang, Z., Xue, H., Zhang, Y., & Wu, Y. (2025). Portcullis: A Scalable and Verifiable Privacy Gateway for Third-Party LLM Inference. *The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*.
31. Bowen, D., Murphy, B., Cai, W., Khachaturov, D., Gleave, A., & Pelrine, K. (2025). Scaling Trends for Data Poisoning in LLMs. *The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*.
32. Carlini, N., Ippolito, D., Lee, K., Tramèr, F., Jagielski, M., & Zhang, C. (2023). Quantifying Memorization Across Neural Language Models. *International Conference on Learning Representations (ICLR)*.
33. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting Training Data from Large Language Models. *30th USENIX Security Symposium (USENIX Security 21)*, 2633-2650.
34. Duan, S., Khona, M., Iyer, A., Schaeffer, R., & Fiete, I. (n.d.). Uncovering Latent Memories: Assessing Data Leakage and Memorization Patterns in Large Language Models.
35. He, P., Xing, Y., Xu, H., Xiang, Z., & Tang, J. (2025). Multi-Faceted Studies on Data Poisoning can Advance LLM Development. *arXiv preprint arXiv:2502.14182*.
36. Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing Leakage of Personally Identifiable Information in Language Models. *2023 IEEE Symposium on Security and Privacy (SP)*, 346-363.
37. Anonymous. (2023). Benchmark Probing: Investigating Data Leakage in Large Language Models. *In NeurIPS 2023 Workshop on Backdoors in Deep Learning*.
38. Armeni, K., Honey, C., & Linzen, T. (2022). Characterizing Verbatim Short-Term Memory in Neural Language Models. *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 405–424.
39. Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., ... & Carlini, N. (2023). Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy. *Proceedings of the 16th International Natural Language Generation Conference (INLG)*, 28–53.
40. Kandpal, N., Jagielski, M., Tramèr, F., & Carlini, N. (2023). Backdoor Attacks for In-Context Learning with Language Models. *arXiv preprint arXiv:2307.14692*.
41. Keum, S., Shin, D., Marchyok, L., Hong, S., & Son, S. (2025). Private Investigator: Extracting Personally Identifiable Information from Large Language Models Using Optimized Prompts. *34th USENIX Security Symposium*.
42. Lin, Y., Yang, R., Mao, Y., Zhang, Q., Hong, J., Cai, Q., ... & Zhong, S. (2025). ObfusLM: Privacy-preserving Language Model Service against Embedding Inversion

Attacks. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1160–1174.
43. Shao, H., Huang, J., Zheng, S., & Chang, K. C.-C. (2024). Quantifying Association Capabilities of Large Language Models and Its Implications on Privacy Leakage. *Findings of the Association for Computational Linguistics: EACL 2024*, 814–825.
44. Al Harbi, S. H., Tidjon, L. N., & Khomh, F. (2023). Responsible Design Patterns for Machine Learning Pipelines. *arXiv preprint arXiv:2306.01788*.
45. de Hert, P., & Lazcoz, G. (2022). When GDPR-Principles Blind Each Other: Accountability, Not Transparency, at the Heart of Algorithmic Governance. *European Data Protection Law Review (EDPL), 8*(1), 31-40.
46. Gruschka, N., Mavroeidis, V., Vishi, K., & Jensen, M. (2018). Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR. *2018 IEEE International Conference on Big Data (Big Data)*.
47. Hoel, T., Griffiths, D., & Chen, W. (2016). The Influence of Data Protection and Privacy Frameworks on the Design of Learning Analytics Systems. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*.
48. Kassem, A. M., Mahmoud, O., & Saad, S. (n.d.). Preserving Privacy Through DeMemorization: An Unlearning Technique For Mitigating Memorization Risks In Language Models. *University of Windsor & Deakin University*.
49. Oladosu, S. A., Ike, C. C., Adepoju, P. A., Afolabi, A. I., Ige, A. B., & Amoo, O. O. (2024). Frameworks for ethical data governance in machine learning: Privacy, fairness, and business optimization. *Magna Scientia Advanced Research and Reviews, 7*(02), 096-106.
50. Pothukuchi, S. N. M. (2025). LLMOps: A Comprehensive Guide to Deploying Large Language Models in Production. *International Journal on Science and Technology (IJSAT), 16*(1).
51. Vaezi, A. (2025). *Legal Challenges in the Deployment of Large Language Models: A Comparative Analysis under the GDPR and EU AI Act* (Master's thesis). Politecnico di Torino.
52. Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., & Xu, X. (2025). Right to be forgotten in the Era of large language models: implications, challenges, and solutions. *AI and Ethics, 5*, 2445-2454.
53. European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.
54. European Parliament & Council of the European Union. (2016). *Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. Official Journal of the European Union, L119.
55. Fabiano, N. (2024). AI Act and Large Language Models (LLMs): When critical issues and privacy impact require human and ethical oversight. *arXiv preprint arXiv:2404.00600*.
56. Feretzakis, G., Vagena, E., Kalodanis, K., Peristera, P., Kalles, D., & Anastasiou, A. (2025). GDPR and Large Language Models: Technical and Legal Obstacles. *Future Internet, 17*(4), 151. https://doi.org/10.3390/fi17040151
57. Huwyler, H. (2025). A Unified Framework for Operationalizing EU AI Act Compliance: Integrating Risk Management, Technical Documentation, and Human Oversight for High-Risk Systems. *IE Law School Executive Education*. https://doi.org/10.5281/zenodo.17703640
58. Kingston, J. K. (2017). Using artificial intelligence to support compliance with the general data protection regulation. *Artificial Intelligence and Law, 25*, 429-443.

59. Minssen, T., Seitz, C., Aboy, M., & Corrales Compagnucci, M. (2020). The EU-US Privacy Shield Regime for Cross-Border Transfers of Personal Data under the GDPR: What Are the Legal Challenges and How Might These Affect Cloud-Based Technologies, Big Data, and AI in the Medical Sector? *European Pharmaceutical Law Review, 4*(1), 34-50.
60. Moradi, N. (2025). *Legal and Ethical Challenges of Large Language Models: An Analysis Under the GDPR and the AI Act* (Master's thesis). Politecnico di Torino.
61. Nolte, H., Finck, M., & Meding, K. (2025). Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data. *arXiv preprint arXiv:2503.01630*.
62. Poullet, Y. (2023). The data protection impact assessment or rather the privacy impact assessment, a revolution with a future in the age of artificial intelligence? In *Artificial intelligence law: between sectoral rules and comprehensive regime: comparative law* (pp. 627-649). Bruylant.
63. Voss, W. G. (2020). Cross-Border Data Flows, the GDPR, and Data Governance. *Washington International Law Journal, 29*(3), 485-532.
64. Wachter, S., & Mittelstadt, B. (2019). A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review, 2019*(2), 494-620.
65. Allen, J. W., Schaefer, O., Mann, S. P., Earp, B. D., & Wilkinson, D. (2025). Augmenting research consent: should large language models (LLMs) be used for informed consent to clinical research?. *Research Ethics, 21*(4), 644-670. https://doi.org/10.1177/17470161241298726
66. Bouhouita-Guermech, S., Gogognon, P., & Bélisle-Pipon, J. C. (2023). Specific challenges posed by artificial intelligence in research ethics. *Frontiers in Artificial Intelligence, 6*, 1149082. https://doi.org/10.3389/frai.2023.1149082
67. Bygrave, L. A. (2020). Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions. *University of Oslo Faculty of Law Research Paper No. 2020-35*. https://ssrn.com/abstract=3721118
68. Hutson, J., Whitney, C., & Conrad, J. T. (2025). Forget Me Not? Machine Unlearning's Implications for Privacy Law.
69. Mitrou, L. (2019). Data Protection, Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR) 'Artificial Intelligence-Proof'?. *SSRN Electronic Journal*. https://ssrn.com/abstract=3386914
70. Pruski, M. (2024). AI-Enhanced Healthcare: Not a new Paradigm for Informed Consent. *Bioethical Inquiry, 21*, 475–489. https://doi.org/10.1007/s11673-023-10320-0
71. Zhang, M., Sankaranarayanapillai, M., Du, J., Xiang, Y., Manion, F. J., Harris, M. R., ... & Tao, C. (2023). Machine learning-based donor permission extraction from informed consent documents. *BMC Bioinformatics, 24*, 477. https://doi.org/10.1186/s12859-023-05568-7
72. Lima Junior, A. G., Coimbra, P. P. A., Silveira, F. M., & RezayatiZoj, M. (2025). Safeguarding Privacy in the Age of AI: Ethical and Technical Challenges in Healthcare Data Anonymization. *Revista Científica COGNITIONIS, 8*(2), e679. https://doi.org/10.38087/2595.8801.679
73. Lundell, B., Gamalielsson, J., Katz, K., Persson, B., Mattsson, M., Fischer, G., & Olén, J. (2025). Exploring the Viability of ChatGPT for Personal Data Anonymization in Government: A Comprehensive Analysis of Possibilities, Risks, and Ethical Implications. *Digital Government: Research and Practice, 6*(2), Article 25.
74. Nyffenegger, A., Stürmer, M., & Niklaus, J. (2024). Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models in Court Decisions. *arXiv preprint arXiv:2308.11103*.

75. Senavirathne, N., & Torra, V. (2020). On the Role of Data Anonymization in Machine Learning Privacy. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 355-370.
76. Staab, R., Vero, M., Balunović, M., & Vechev, M. (2024). Beyond Memorization: Violating Privacy via Inference with Large Language Models. *International Conference on Learning Representations (ICLR)*.
77. Štarchoň, P., & Pikulík, T. (2019). GDPR principles in Data protection encourage pseudonymization through most popular and full-personalized devices – mobile phones. *Procedia Computer Science, 151*, 303–312.
78. Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2024). On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. *arXiv preprint arXiv:2403.05156*.
79. Al-Qahtani, F. (2025). Human-AI Collaboration in Decision Making: Balancing Automation and Ethical Responsibility. *International Journal of Innovative Science and Technology Studies (IJISTS), 1*(1), 79-94.
80. Gerdon, F., Bach, R. L., Kern, C., & Kreuter, F. (2022). Social impacts of algorithmic decision-making: A research agenda for the social sciences. *Big Data & Society, 9*(1), 1-13. https://doi.org/10.1177/20539517221089305
81. Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & Society, 38*, 549–563. https://doi.org/10.1007/s00146-022-01455-6
82. Li, Y., Du, M., Song, R., Wang, X., & Wang, Y. (2024). A Survey on Fairness in Large Language Models. *arXiv preprint arXiv:2308.10149*.
83. Liao, Y., & Naghizadeh, P. (2023). Social Bias Meets Data Bias: The Impacts of Labeling and Measurement Errors on Fairness Criteria. *The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*.
84. McGraw, D. K. (2024). Ethical Responsibility in the Design of Artificial Intelligence (AI) Systems. *International Journal on Responsibility, 7*(1), Article 4. https://doi.org/10.62365/2576-0955.1114
85. Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: what it is and how it works. *AI & Society, 39*, 1871–1882. https://doi.org/10.1007/s00146-023-01635-y
86. Brenneis, A. (2025). Assessing dual use risks in AI research: necessity, challenges and mitigation strategies. *Research Ethics, 21*(2), 302-330. https://doi.org/10.1177/17470161241267782
87. Das, B. C., Amini, M. H., & Wu, Y. (2024). Security and Privacy Challenges of Large Language Models: A Survey. *arXiv preprint arXiv:2402.00888*.
88. Prokopowicz, D. (2025). The Role of Artificial Intelligence in Social Media and Internet Portals: From Combating Disinformation to the Rise of Cyber Threats. *ResearchGate*. https://doi.org/10.13140/RG.2.2.20968.74244
89. Sammouri, K. (2025). *An LLM-Based Threat Modeling Tool with CAPEC Semantic Retrieval* (Master's thesis). Miami University, Oxford, Ohio.
90. Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., ... & Thompson, N. (2025). The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence. *MIT FutureTech*.
91. Soice, E. H., Rocha, R., Cordova, K., Specter, M., & Esvelt, K. M. (n.d.). Can large language models democratize access to dual-use biotechnology?. *Massachusetts Institute of Technology & SecureBio*.
92. Veluru, C. S. (n.d.). Responsible Artificial Intelligence on Large Scale Data to Prevent Misuse, Unethical Challenges and Security Breaches. *Journal of Artificial Intelligence & Cloud Computing*.

93. Wu, T., Yang, S., Liu, S., Nguyen, D., Jang, S., & Abuadbba, A. (2025). ThreatModeling-LLM: Automating Threat Modeling using Large Language Models for Banking System. *arXiv preprint arXiv:2411.17058*.
94. Zhao, H. (2023). Artificial intelligence-based public safety data resource management in smart cities. *Open Computer Science, 13*(1), 20220271. https://doi.org/10.1515/comp-2022-0271
95. Bodeau, D. J., McCollum, C. D., & Fox, D. B. (2018). *Cyber Threat Modeling: Survey, Assessment, and Representative Framework*. Homeland Security Systems Engineering and Development Institute (HSSEDI) & The MITRE Corporation.
96. Cantini, R., Orsino, A., Ruggiero, M., & Talia, D. (2025). Benchmarking adversarial robustness to bias elicitation in large language models: scalable automated assessment with LLM-as-a-judge. *Machine Learning, 114*, 249. https://doi.org/10.1007/s10994-025-06862-6
97. Clusmann, J., Ferber, D., Wiest, I. C., Schneider, C. V., Brinker, T. J., Foersch, S., Truhn, D., & Kather, J. N. (2024). Prompt Injection Attacks on Large Language Models in Oncology. *arXiv preprint arXiv:2407.18981*.
98. Desai, A., Abdelhamid, M., & Padalkar, N. R. (2025). What is reproducibility in artificial intelligence and machine learning research? *AI Magazine*. https://doi.org/10.1002/aaai.70004
99. Gittens, A., Yener, B., & Yung, M. (2022). An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML. *IEEE Access*. https://doi.org/10.1109/ACCESS.2022.3218715
100. Raina, V., Liusie, A., & Gales, M. (2024). Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment. *arXiv preprint arXiv:2402.14016*.
101. Shaha, H. (2025). *Understanding Prompt Injection Attacks in Web Based LLM Applications and Basic Mitigation Strategies* (Bachelor's thesis). Centria University of Applied Sciences.
102. Sun, J., Yin, Z., Zhang, H., Chen, X., & Zheng, W. (2025). Adversarial generation method for smart contract fuzz testing seeds guided by chain-based LLM. *Automated Software Engineering, 32*(12). https://doi.org/10.1007/s10515-024-00483-4
103. Sun, S. (2024). *Investigating Adversarial Robustness in Language Models: Adversarial Attacks, Certification, and Defense* (Doctoral dissertation). University of Liverpool.
104. Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., & Chen, K. (2025). Large Language Models for Cyber Security: A Systematic Literature Review. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3769676
105. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*.
106. Alexander, X. F. (2020). Federated Learning and Secure Data Exchange Mechanisms for Scalable Cloud–Edge–IoT Ecosystems in Intelligent Computing Environments. *American International Journal of Computer Science and Technology, 2*(2), 1-9. https://doi.org/10.63282/3117-5481/AIJCST-V212P101
107. Blanco-Justicia, A., Sánchez, D., Domingo-Ferrer, J., & Muralidhar, K. (2022). A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning. *ACM Computing Surveys (CSUR), 55*(8), 1-32.
108. Charles, Z., McMahan, H. B., Ganesh, A., Pillutla, K., McKenna, R., Mitchell, N., & Rush, K. (2024). Fine-Tuning Large Language Models with User-Level Differential Privacy. *arXiv preprint arXiv:2407.07737*.

109. Gupta, S., Huang, Y., Zhong, Z., Gao, T., Li, K., & Chen, D. (2022). Recovering Private Text in Federated Learning of Language Models. *Advances in Neural Information Processing Systems (NeurIPS), 35*, 8130-8143.
110. Ha, T., Dang, T. K., Dang, T. T., & Truong, T. A. (2019). Differential Privacy in Deep Learning: An Overview. *2019 International Conference on Advanced Computing and Applications (ACOMP)* (pp. 1-8). IEEE.
111. Harder, F. (2023). *Methods for Generative Modeling and Interpretable Classification with Strong Differential Privacy* (Doctoral dissertation). Eberhard Karls Universität Tübingen.
112. Mawela, C., Issaid, C. B., & Bennis, M. (2025). A Web-Based Solution for Federated Learning With LLM-Based Automation. *IEEE Internet of Things Journal, 12*(12), 19488-19500.
113. Shi, W., Shea, R., Chen, S., Zhang, C., Jia, R., & Yu, Z. (2022). Just Fine-tune Twice: Selective Differential Privacy for Large Language Models. *arXiv preprint arXiv:2204.07667*.
114. Wang, H., Yin, Z., Chen, B., Zeng, Y., Yan, X., Zhou, C., & Li, A. (2026). ROFED-LLM: Robust Federated Learning for Large Language Models in Adversarial Wireless Environments. *IEEE Transactions on Network Science and Engineering, 13*, 1084.
115. Wu, Y., Tian, C., Li, J., Sun, H., Tam, K., Zhou, Z., ... & Xu, C. (2025). A Survey on Federated Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2503.12016*.
116. Boemer, F., Costache, A., Cammarota, R., & Wierzynski, C. (2019). nGraph-HE2: A High-Throughput Framework for Neural Network Inference on Encrypted Data. *Proceedings of the 7th ACM Workshop on Encrypted Computing & Applied Homomorphic Cryptography*, 53-62.
117. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021). Machine Unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, 141-159.
118. Brophy, J., & Lowd, D. (2021). Machine Unlearning for Random Forests. *International Conference on Machine Learning (ICML)*, 1092-1104.
119. de Castro, L., Escudero, D., & Polychroniadou, A. (n.d.). Privacy-Preserving Large Language Model Inference via GPU-Accelerated Fully Homomorphic Encryption. *J.P. Morgan AI Research*.
120. Huang, Y., & Canonne, C. L. (2023). Tight Bounds for Machine Unlearning via Differential Privacy. *arXiv preprint arXiv:2309.00886*.
121. Jandali, Y., Zhang, R., Sheybani, N., & Koushanfar, F. (2025). Optimizing Privacy-Preserving Primitives to Support LLM-Scale Applications. *arXiv preprint arXiv:2509.25072*.
122. Lin, Y., Gao, Z., Du, H., Niyato, D., Gui, G., Cui, S., & Ren, J. (2024). Scalable Federated Unlearning via Isolated and Coded Sharding. *arXiv preprint arXiv:2401.15957*.
123. Xu, P., Chen, M., & Xu, W. (2025). Privacy-Preserving Large Language Model in Terms of Secure Computing: A Survey. *2025 2nd International Conference on Algorithms, Software Engineering and Network Security (ASENS)*.
124. Zhang, H., Nakamura, T., Isohara, T., & Sakurai, K. (2023). A Review on Machine Unlearning. *SN Computer Science, 4*, 337.
125. Zhou, I., Tofigh, F., Piccardi, M., Abolhasan, M., Franklin, D., & Lipman, J. (2024). Secure Multi-Party Computation for Machine Learning: A Survey. *IEEE Access, 12*, 53899-53926.
126. Dasgupta, A., Tanvir, A. A., & Zhong, X. (2025). Watermarking Language Models through Language Models. *arXiv preprint arXiv:2411.05091*.

127. Henderson, J., & Milton, T. (2025). Auditing and Logging Systems for Privacy Assurance in Medical AI Pipelines. *Preprints.org*. https://doi.org/10.20944/preprints202506.1209.v1
128. Peng, Y. (2024). Semantic Context Modeling for Fine-Grained Access Control Using Large Language Models. *Journal of Computer Technology and Software, 3*(7).
129. Rastogi, S., & Pruthi, D. (2024). Revisiting the Robustness of Watermarking to Paraphrasing Attacks. *arXiv preprint arXiv:2411.05277*.
130. Shanmugarasa, Y., Ding, M., Arachchige, C. M., & Rakotoarivelo, T. (2025). SoK: The Privacy Paradox of Large Language Models: Advancements, Privacy Risks, and Mitigation. *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security (ASIA CCS '25)*. ACM.
131. Singh, S. (n.d.). Enhancing Privacy and Security in Large-Language Models: A Zero-Knowledge Proof Approach. *University of KwaZulu-Natal*.
132. Yi, J., Xie, Y., Zhu, B. B., Kiciman, E., Sun, G., & Xie, X. (2025). Benchmarking and Defending against Indirect Prompt Injection Attacks on Large Language Models. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*. ACM.
133. Zhang, R., & Koushanfar, F. (2024). Watermarking Large Language Models and the Generated Content: Opportunities and Challenges. *2024 58th Asilomar Conference on Signals, Systems, and Computers*. IEEE.
134. Zhou, T., Xu, X., Zhao, X., & Ren, S. (2024). Bileve: Securing Text Provenance in Large Language Models Against Spoofing with Bi-level Signature. *Advances in Neural Information Processing Systems (NeurIPS)*.
135. Bezzi, M. (2010). An information theoretic approach for privacy metrics. *Transactions on Data Privacy, 3*(3), 199-215.
136. Duddu, V., Szyller, S., & Asokan, N. (2022). SHAPR: An Efficient and Versatile Membership Privacy Risk Metric for Machine Learning. *arXiv preprint arXiv:2112.02230*.
137. Mendes, R., & Vilela, J. P. (2017). Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access, 5*, 10562-10582. https://doi.org/10.1109/ACCESS.2017.2706947
138. Negoescu, D. M., Gonzalez, H., Al Orjany, S. E., Yang, J., Lut, Y., Tandra, R., ... & Samorodnitsky, G. (2024). Epsilon*: Privacy Metric for Machine Learning Models. *Accepted at The 5th AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-24)*. arXiv:2307.11280.
139. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and Privacy in Machine Learning. *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 399-414. https://doi.org/10.1109/EuroSP.2018.00035
140. Semantha, F. H., Azam, S., Shanmugam, B., Yeo, K. C., & Beeravolu, A. R. (2021). A Conceptual Framework to Ensure Privacy in Patient Record Management System. *IEEE Access, 9*, 165667-165689. https://doi.org/10.1109/ACCESS.2021.3134873
141. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP)*, 3-18.
142. Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268-282.
143. Arous, A., Guesmi, A., Hanif, M. A., Alouani, I., & Shafique, M. (2023). Exploring Machine Learning Privacy/Utility Trade-Off from a Hyperparameters Lens. *2023 International Joint Conference on Neural Networks (IJCNN)*, 1-8. https://doi.org/10.1109/IJCNN54540.2023.10191743

144. Joshi, S. (2025). Evaluation of Large Language Models: Review of Metrics, Applications, and Methodologies. *Preprints.org*. https://doi.org/10.20944/preprints202504.0369.v2
145. Mauger, C., Le Mahec, G., & Dequen, G. (2023). Optimizing Privacy and Data Utility: Metrics and Strategies. *Transactions on Data Privacy, 16*(2), 153-189.
146. Miaschi, A., Brunato, D., Dell'Orletta, F., & Venturi, G. (2020). What Makes My Model Perplexed? A Linguistic Investigation on Neural Language Models Perplexity. *Proceedings of the First Workshop on Insights from Negative Results in NLP*, 117–122.
147. Pan, X., Zhang, M., Ji, S., & Yang, M. (2020). Privacy Risks of General-Purpose Language Models. *2020 IEEE Symposium on Security and Privacy (SP)*, 1314-1331.
148. Talukdar, W., & Biswas, A. (2023). Improving Large Language Model (LLM) fidelity through context-aware grounding: A systematic approach to reliability and veracity. *World Journal of Advanced Engineering Technology and Sciences, 10*(02), 283-296. https://doi.org/10.30574/wjaets.2023.10.2.0317
149. Xu, Y., Zhang, K., Dong, H., Sun, Y., Zhao, W., & Tu, Z. (2020). Rethinking Exposure Bias in Adversarial Language Modeling. *arXiv preprint arXiv:1910.11235*.
150. Xu, Z., Gupta, A., Li, T., Bentham, O., & Srikumar, V. (2024). Beyond Perplexity: Multi-dimensional Safety Evaluation of LLM Compression. *arXiv preprint arXiv:2407.04965*.
151. Benraouane, S. A. (2024). *AI Management System Certification According to the ISO/IEC 42001 Standard: How to Audit, Certify, and Build Responsible AI Systems*. Routledge.
152. Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., & Shokri, R. (2022). Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. *arXiv preprint arXiv:2203.03929*.
153. Ongwae, B. (2025). An Industrial Application of a Large Language Model to Enhancing Asset Integrity and Process Safety Management. *Preprints.org*. https://doi.org/10.20944/preprints202503.1172.v1
154. Pathade, C. (2025). Red Teaming the Mind of the Machine: A Systematic Evaluation of Prompt Injection and Jailbreak Vulnerabilities in LLMs. *arXiv preprint arXiv:2505.04806*.
155. Tantithamthavorn, C., Palomba, F., Khomh, F., & Chua, J. (2025). MLOps, LLMOps, FMOps, and Beyond. *IEEE Software*. https://doi.org/10.1109/MS.2024.3477014
156. Werthwein, M., Annighoefer, B., & Daw, Z. (2025). Towards Continuous Security Assessment: Integrating Model-Based Risk Assessment and Large Language Models. *2025 Integrated Communications, Navigation and Surveillance Conference (ICNS)*. IEEE.
157. Xu, R., & Ding, K. (2024). Large Language Models for Anomaly and Out-of-Distribution Detection: A Survey. *arXiv preprint*.
158. National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
159. Agarwal, A., & Nene, M. J. (2024). Standardised Schema and Taxonomy for AI Incident Databases in Critical Digital Infrastructure. *2024 IEEE Pune Section International Conference (PuneCon)*. IEEE.
160. Fredrikson, G. (2024). *Secure Interactions with Large Language Models in Financial Services: A Study on Implementing Safeguards for Large Language Models* (Master's thesis). Uppsala Universitet.

161. Liu, Z. (2024). Multi-Agent Collaboration in Incident Response with Large Language Models. *arXiv preprint arXiv:2412.00652*.
162. Mahar, M. A., Raza, A., Larik, R. S. A., Burdi, A., Shabbir, M., & Iftikhar, M. (2025). Transformative Role of LLMs in Digital Forensic Investigation: Exploring Tools, Challenges, and Emerging Opportunities. *VFAST Transactions on Computer Sciences, 13*(1), 217-229.
163. Rathod, V., Nabavirazavi, S., Zad, S., & Iyengar, S. S. (2025). Privacy and Security Challenges in Large Language Models. *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE.
164. Yigit, Y., Ferrag, M. A., Sarker, I. H., Maglaras, L. A., Chrysoulas, C., Moradpoor, N., & Janicke, H. (2024). Critical Infrastructure Protection: Generative AI, Challenges, and Opportunities. *arXiv preprint arXiv:2405.04874*.
165. Zhu, X., Zhou, W., Han, Q. L., Ma, W., Wen, S., & Xiang, Y. (2025). When Software Security Meets Large Language Models: A Survey. *IEEE/CAA Journal of Automatica Sinica, 12*(2), 317.
166. Aboukadri, S., Ouaddah, A., Mezrioui, A., & El Asri, I. (2025). Leveraging RAG and LLMs for Access Control Policy Extraction From User Stories in Agile Software Development. *IEEE Access, 13*, 116472. https://doi.org/10.1109/ACCESS.2025.3586203
167. Belmoukadam, O., De Jonghe, J., Ajridi, S., Krifa, A., Van Damme, J., Mkadem, M., & Latinne, P. (2024). AdversLLM: A Practical Guide to Governance, Maturity and Risk Assessment for LLM-based Applications. *International Journal on Cybernetics & Informatics (IJCI), 13*(6).
168. Gurram, A. (2025). Generative AI for enhanced cybersecurity: building a zero-trust architecture with agentic AI. *World Journal of Advanced Engineering Technology and Sciences, 15*(01), 2380-2396. https://doi.org/10.30574/wjaets.2025.15.1.0504
169. Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C., ... & Mustafa, M. A. (2024). A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review, 57*, 175. https://doi.org/10.1007/s10462-024-10824-0
170. Kaya, Y., Landerer, A., Pletinckx, S., Zimmermann, M., Kruegel, C., & Vigna, G. (2025). When AI Meets the Web: Prompt Injection Risks in Third-Party AI Chatbot Plugins. *arXiv preprint arXiv:2511.05797*. (Accepted to 2026 IEEE Symposium on Security and Privacy).
171. Li, E., Mallick, T., Rose, E., Robertson, W., Oprea, A., & Nita-Rotaru, C. (2025). ACE: A Security Architecture for LLM-Integrated App Systems. *arXiv preprint arXiv:2504.20984*.
172. Nathanson, S., Lee, A., Kieffer, C. C., Junkin, J., Ye, J., Saeed, A., ... & Peterson, E. (2025). AI Bill of Materials and Beyond: Systematizing Security Assurance through the AI Risk Scanning (AIRS) Framework. *arXiv preprint arXiv:2511.12668*.
173. Zakkiya, S. A., Selviandro, N., & Utomo, R. G. (2025). A Guideline for the Adoption of Generative AI to Support Secure Software Development Life Cycle (SSDLC): A Case Study of ChatGPT. *2025 IEEE International Conference on Artificial Intelligence for Learning and Optimization (ICoAILO)*. IEEE.
174. Korkmaz, B. S., & Daly, E. (2025). Paying Alignment Tax with Contrastive Learning. *arXiv preprint arXiv:2505.19327*.
175. Li, J., & Kim, J. E. (2025). Superficial Safety Alignment Hypothesis. *arXiv preprint arXiv:2410.10862*.

176. Lu, H., Fang, L., Zhang, R., Li, X., Cai, J., Cheng, H., ... & Ma, P. (2025). Alignment and Safety in Large Language Models: Safety Mechanisms, Training Paradigms, and Emerging Challenges. *arXiv preprint arXiv:2507.19672*.
177. Reddy, S., Khan, K., Patil, R., Chakraborty, A., Khan, F. A., Kulkarni, S., ... & Singh, N. (2025). Computational Economics in Large Language Models: Exploring Model Behavior and Incentive Design under Resource Constraints. *arXiv preprint arXiv:2508.10426*.
178. Wu, F. (2025). *Differential Privacy in the Era of Generative AI: Promises and Challenges* (Doctoral dissertation). University of Illinois Urbana-Champaign.
179. Xiao, Y., Jin, Y., Bai, Y., Wu, Y., Yang, X., Luo, X., ... & Cheng, W. (n.d.). Large Language Models Can Be Contextual Privacy Protection Learners. *University of California, Los Angeles & NEC Laboratories America*.
180. Zhang, J., Chen, R., Zhou, Q., Deng, X., & Jiang, W. (2025). Understanding and Mitigating Over-refusal for Large Language Models via Safety Representation. *arXiv preprint arXiv:2511.19009*.
181. Zhou, Z., Ning, X., Hong, K., Fu, T., Xu, J., Li, S., ... & Wang, Y. (2024). A Survey on Efficient Inference for Large Language Models. *arXiv preprint arXiv:2404.14294*.
182. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., & Bucknall, B. S. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 2254–2272. https://doi.org/10.1145/3630106.3659037
183. Chua, J., Li, Y., Yang, S., Wang, C., & Yao, L. (2024). AI Safety in Generative AI Large Language Models: A Survey. *arXiv preprint arXiv:2407.18369*.
184. Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, 1112–1123. https://doi.org/10.1145/3593013.3594067
185. Narayanan, S., & Vishwakarma, S. (n.d.). GUARD-D-LLM: An LLM-Based Risk Assessment Engine for the Downstream uses of LLMs. *University of Bordeaux & Symbiosis Institute of Technology*.
186. Oliva, F. L., Bution, J. L., Motta, F. G., Fenner, G., Randolph-Seng, B., Papa, M., & Naqshbandi, M. M. (2026). Appetite for risk: theoretical framework and practical application in a technology-based environment. *Journal of Intellectual Capital, 26*(1), 71-103.
187. Ottun, A. R. O., & Flores, H. (2025). Trustworthy AI in Practice: A Comprehensive Review of Human Oversight and Human-in-the-Loop Approaches. *Preprints*.
188. Thelisson, E., & Verma, H. (2024). Conformity assessment under the EU AI act general approach. *AI and Ethics, 4*, 113-121. https://doi.org/10.1007/s43681-023-00402-5
189. Wei, Z., Wei, Z.-Y., Liang, C.-W., Lang, S.-N., Yan, H., Han, Z.-M., ... & Wang, M.-J.-S. (n.d.). Learning-Based Automated Adversarial Red-Teaming for Robustness Evaluation of Large Language Models.
190. Golpayegani, D. (2024). *Semantic Frameworks to Support the EU AI Act's Risk Management and Documentation* (Doctoral dissertation). Trinity College Dublin.
191. Khatik, A. A., & Sheikh, Y. M. (n.d.). Artificial Intelligence for Cyber Risk Management: Frameworks, Innovations, and Challenges. *Quinnipiac University & Binghamton University*.
192. Koulierakis, E. (2024). Certification as guidance for data protection by design. *International Review of Law, Computers & Technology, 38*(2), 245-263. https://doi.org/10.1080/13600869.2023.2269498

193. Kucharavy, A., Plancherel, O., Mulder, V., Mermoud, A., & Lenders, V. (Eds.). (2024). *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*. Springer Nature. https://doi.org/10.1007/978-3-031-54827-7
194. Mamirov, A., Azmain, F., & Wang, H. (2025). AI Transparency Atlas: Framework, Scoring, and Real-Time Model Card Evaluation Pipeline. *arXiv preprint arXiv:2512.12443*.
195. Mazzinghy, A. O. C., Silva, R. M. S., Fernandes, R. M., Batista, E. D., Picanço, A. R. S., ... & Martins, V. W. B. (2024). Assessing Benefits of ISO/IEC 42001 Artificial Intelligence Management System: Insights from Brazilian Logistics. *Preprints.org*. https://doi.org/10.20944/preprints202410.0699.v1
196. Siddik, M., & Pandit, H. J. (2025). Datasheets for Healthcare AI: A Framework for Transparency and Bias Mitigation. *arXiv preprint arXiv:2501.05617*.
197. Xu, Z. (2024). *The Mysteries of Large Language Models: Tracing the Evolution of Transparency for OpenAI's GPT Models* (Honors thesis). Wellesley College.
198. Bilal, A., Ebert, D., & Lin, B. (2025). LLMs for Explainable AI: A Comprehensive Survey. *arXiv preprint arXiv:2504.00125*.
199. Chen, K., Zhou, X., Lin, Y., Su, J., Yu, Y., Shen, L., & Lin, F. (2025). A Survey on Data Security in Large Language Models. *arXiv preprint arXiv:2508.02312*.
200. Chen, Z. Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., ... & Wang, W. Y. (2024). A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. *arXiv preprint arXiv:2405.01769*.
201. Kumar, B. V. P., & Ahmed, M. D. S. (2024). Beyond Clouds: Locally Runnable LLMs as a Secure Solution for AI Applications. *Digital Society, 3*, 49. https://doi.org/10.1007/s44206-024-00141-y
202. Li, A., Zhou, Y., Raghuram, V. C., Goldstein, T., & Goldblum, M. (2025). Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks. *arXiv preprint arXiv:2502.08586*.
203. Resnik, P. (2025). Large Language Models Are Biased Because They Are Large Language Models. *Computational Linguistics*. https://doi.org/10.48550/arXiv.2406.13138
204. House of Lords Communications and Digital Committee. (2024). *Large language models and generative AI* (1st Report of Session 2023-24, HL Paper 54). UK Parliament.
205. OWASP Top 10 for LLM Applications Team. (2024). *LLM AI Cybersecurity & Governance Checklist* (Version 1.1). The OWASP Foundation.
206. UNESCO & International Research Centre on Artificial Intelligence (IRCAI). (2024). *Challenging Systematic Prejudices: An Investigation into Bias Against Women and Girls in Large Language Models*. UNESCO.