ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Υλοποίηση συστημάτων RAG αξιοποιώντας τη νομοθεσία για τα τυχερά παίγνια

Από
Αγγέλου Αντώνιο

Υποβάλλεται
για την εκπλήρωση των προϋποθέσεων λήψης
Μεταπτυχιακού Διπλώματος
στην ειδίκευση «Προηγμένα Πληροφοριακά Συστήματα»
του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”
στο
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Φεβρουάριος 2026

Επιβλέπουσα: Ανδριάνα Πρέντζα
Ακαδημαϊκή θέση: Καθηγήτρια

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή: Αγγέλου Αντώνιος

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας:
Υλοποίηση συστημάτων RAG αξιοποιώντας τη νομοθεσία για τα τυχερά παίγνια

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις 25/2/2026 από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπουσα (Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς): Ανδριάνα Πρέντζα, Καθηγήτρια
Μέλος Εξεταστικής Επιτροπής: Μιχαήλ Φιλιππάκης, Καθηγητής
Μέλος Εξεταστικής Επιτροπής: Ανδρέας Μενύχτας, Επίκουρος Καθηγητής

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ
Ο Αντώνιος Αγγέλου, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Υλοποίηση συστημάτων RAG αξιοποιώντας τη νομοθεσία για τα τυχερά παίγνια», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Επιπλέον δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο ΔΗΛΩΝ

Ονοματεπώνυμο: Αγγέλου Αντώνιος
Αριθμός Μητρώου: ΜΕ2331

Υπογραφή:

Ευχαριστίες

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, θα ήθελα να εκφράσω τις θερμότερες ευχαριστίες μου στην επιβλέπουσα καθηγήτριά κυρία Ανδριάνα Πρέντζα, για τις πολύτιμες συμβουλές, την καθοδήγηση και τη συνεχή υποστήριξή της καθ' όλη τη διάρκεια της εκπόνησής της. Επιπλέον, θα ήθελα να ευχαριστήσω την Επιτροπή Εποπτείας και Ελέγχου Παιγνίων για την παροχή των δεδομένων που συνέβαλαν ουσιαστικά στην ολοκλήρωση της έρευνας. Θερμές ευχαριστίες οφείλω επίσης στον φίλο και συμφοιτητή μου, Περικλή Βαϊρακταρίδη, για τη συνεργασία και τη στήριξη κατά τη διάρκεια των σπουδών. Τέλος, ευχαριστώ τους φίλους μου και την οικογένειά μου για την ψυχολογική υποστήριξη και την ενθάρρυνση που μου προσέφεραν σε όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Λίστα Εικόνων

Συντμήσεις

AI	Artificial Intelligence
ANN	Approximate Nearest Neighbors
API	Application Programming Interface
ATT&CK	Adversarial Tactics, Techniques, and Common Knowledge
BERT	Bidirectional Encoder Representations from Transformers
CAPEC	Common Attack Pattern Enumeration and Classification
CUDA	Compute Unified Device Architecture
CVE	Common Vulnerabilities and Exposures
CWE	Common Weakness Enumeration
DNS	Domain Name System
DPO	Direct Preference Optimization
FAISS	Facebook AI Similarity Search
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HNSW	Hierarchical Navigable Small World
HTML	HyperText Markup Language
HyDE	Hypothetical Document Embeddings
k-NN	k-Nearest Neighbors
Labse	Language-agnostic BERT Sentence Embedding
Llama	Large Language Model Meta AI
LLM	Large Language Model
MIRACL	Multilingual Information Retrieval Across a Continuum of Languages
MLM	Masked Language Modeling
MTEB	Massive Text Embedding Benchmark
NLP	Natural Language Processing
OCR	Optical Character Recognition
PDF	Portable Document Format
RAG	Retrieval-Augmented Generation
SFT	Supervised Fine-Tuning
TLM	Translation Language Modeling
TPU	Tensor Processing Unit

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει την εφαρμογή της αρχιτεκτονικής Retrieval-Augmented Generation (RAG) σε κανονιστικά/νομικά κείμενα, με έμφαση στη νομοθεσία των τυχερών παιγνίων στην Ελλάδα. Στόχος είναι να διερευνηθεί πώς ένα σύστημα RAG μπορεί να υποστηρίξει τον εποπτικό ρόλο μιας αρμόδιας αρχής, παρέχοντας τεκμηριωμένες απαντήσεις που βασίζονται σε ανακτημένα αποσπάσματα νομοθεσίας. Για τον σκοπό αυτό υλοποιούνται τρεις εφαρμογές: (α) εντοπισμός άρθρων που ενδέχεται να παραβιάζει μια στοιχηματική εταιρεία με βάση το κείμενο μιας καταγγελίας, (β) ανάκτηση και σύγκριση παρόμοιων καταγγελιών από ιστορικό αρχείο, και (γ) chatbot ερωταποκρίσεων για θέματα αδειών καταλληλότητας. Κάθε εφαρμογή υλοποιείται σε δύο εκδόσεις, ώστε να συγκριθεί η επίδραση διαφορετικών πολυγλωσσικών μοντέλων ενσωμάτωσης (embedding models) (LaBSE και intfloat/multilingual-e5-base) στην ποιότητα της ανάκτησης. Η προσέγγιση βασίζεται σε προεπεξεργασία και τμηματοποίηση κειμένων, δημιουργία embeddings, αποθήκευση σε βάση διανυσμάτων (vector database) και παραγωγή απάντησης από μεγάλο γλωσσικό μοντέλο (LLM) με εμπλουτισμένο πλαίσιο. Τα αποτελέσματα δείχνουν ότι η χρήση RAG μειώνει τον κίνδυνο μη τεκμηριωμένων απαντήσεων και διευκολύνει την πρόσβαση σε εξειδικευμένη νομοθετική γνώση, ενώ αναδεικνύονται πρακτικές παράμετροι όπως η επιλογή embedding model, η ποιότητα των δεδομένων και οι περιορισμοί του context window.

Θεματική Περιοχή: Συστήματα RAG

Λέξεις-κλειδιά: Retrieval-Augmented Generation, νομοθεσία τυχερών παιγνίων, πολυγλωσσικά embedding models, vector databases, LLMs.

Abstract

This thesis investigates the use of Retrieval-Augmented Generation (RAG) in regulatory/legal texts, focusing on the Greek framework for gambling regulation. The goal is to explore how RAG-based systems can support supervisory activities by producing grounded answers that explicitly rely on retrieved legal excerpts. Three prototype applications are implemented: (a) identifying potentially violated legal provisions from the text of a complaint, (b) retrieving and comparing similar historical complaints, and (c) a question-answering chatbot for suitability licensing topics. Each application is developed in two variants to assess the impact of different multilingual embedding models (LaBSE and intfloat/multilingual-e5-base) on retrieval quality. The pipeline includes text preprocessing and chunking, embedding generation, storage in a vector database, and answer generation by a large language model using the retrieved context. The findings indicate that RAG improves access to domain-specific legal knowledge and helps mitigate ungrounded responses, while highlighting practical factors such as embedding selection, data quality, and context-window limitations.

Subject Area: RAG Systems

Keywords: Retrieval-Augmented Generation, gambling regulation, multilingual embedding models, vector databases, LLMs.

1. Εισαγωγή

1.1 Σκοπός και στόχοι της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία έχει ως σκοπό να δείξει την εφαρμογή των συστημάτων RAG στη νομοθεσία των τυχερών παιγνίων. Πιο συγκεκριμένα, υλοποιούνται τρεις εφαρμογές που αξιοποιούν κανονισμούς που αφορούν τη νομοθεσία για τα τυχερά παίγνια. Η πρώτη εφαρμογή αποτελεί ένα εργαλείο με το οποίο ο ελεγκτής έχει τη δυνατότητα να εισάγει την καταγγελία ενός πολίτη για μία στοιχηματική εταιρεία και το σύστημα να απαντήσει ποια αποσπάσματα της νομοθεσίας παραβιάζει η εταιρεία για την οποία γίνεται η καταγγελία. Η δεύτερη εφαρμογή αποτελεί ένα σύστημα στο οποίο ο ελεγκτής εισάγει μία καταγγελία και του εμφανίζει τις παρόμοιες καταγγελίες που έχουν υποβληθεί καθώς και τι ομοιότητες και διαφορές έχουν. Η τρίτη εφαρμογή αποτελεί ένα chatbot στην οποία ο ελεγκτής μπορεί να ρωτάει για θέματα που αφορούν τις άδειες καταλληλότητας και το σύστημα να απαντάει στα ερωτήματα με παραπομπές από τη νομοθεσία.

Οι ανωτέρω εφαρμογές έχουν ως στόχο να δείξουν χρήσεις που μπορεί να έχει ένα σύστημα RAG επάνω στη νομοθεσία για τα τυχερά παίγνια στην Ελλάδα, προτείνοντας εργαλεία που μπορούν να ενισχύσουν τον εποπτικό ρόλο της αρμόδιας αρχής.

Προκειμένου να δειχθούν τα πλεονεκτήματα των συστημάτων RAG, για κάθε εφαρμογή σχεδιάστηκαν δύο εκδόσεις. Η μία έκδοση χρησιμοποιεί ως embedding model το LaBSE και η άλλη το intfloat/multilingual-e5-base. Με αυτό τον τρόπο λαμβάνει χώρα ένα πείραμα που στόχο έχει να δείξει την ικανότητα που έχουν τα συγκεκριμένα μοντέλα στην ανάκτηση σχετικού περιεχομένου, σύμφωνα με το ερώτημα του χρήστη, από ελληνικά κείμενα και πως αυτές οι πληροφορίες μπορούν να εμπλουτίσουν την τελική απάντηση που θα δώσει το LLM στον χρήστη.

1.2 Δομή της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία αποτελείται από πέντε κύρια κεφάλαια. Το πρώτο κεφάλαιο εισάγει το θέμα της εργασίας, τον σκοπό και τη συμβολή της. Στο δεύτερο

κεφάλαιο περιλαμβάνεται η βιβλιογραφική ανασκόπηση στην οποία δίνεται έμφαση στον ορισμό της έννοιας RAG, στην αρχιτεκτονική, στα πλεονεκτήματα και μειονεκτήματα. Ακόμη παρουσιάζονται τα embedding models, οι vector databases και τα LLMs που δύναται να αξιοποιηθούν. Επιπροσθέτως, παρατίθενται γενικά παραδείγματα εφαρμογών συστημάτων RAG καθώς και παραδείγματα που σχετίζονται με τον δημόσιο τομέα και με νομικά κείμενα. Στο τρίτο κεφάλαιο παρουσιάζονται η ανάλυση απαιτήσεων των εφαρμογών που σχεδιάστηκαν, καθώς και η επιλογή embedding model, vector database, και LLM για την υλοποίησή τους. Επιπλέον, γίνεται παρουσίαση των τεχνολογιών και εργαλείων που χρησιμοποιήθηκαν. Στο τέταρτο κεφάλαιο γίνεται αναφορά στην εκτέλεση των εφαρμογών και ερμηνεύονται τα αποτελέσματα που παράγουν. Τέλος, στο πέμπτο κεφάλαιο παρατίθενται τα βασικά συμπεράσματα της εργασίας και προτάσεις για μελλοντικές προσθήκες.

2. Βιβλιογραφική Ανασκόπηση

2.1 Retrieval-Augmented Generation (RAG)

2.1.1 Ορισμός έννοιας, πλεονεκτήματα και περιορισμοί

Η έννοια Retrieval augmented generation (RAG) είναι μια αρχιτεκτονική για τη βελτιστοποίηση της απόδοσης ενός μοντέλου τεχνητής νοημοσύνης (AI) συνδέοντάς το με εξωτερικές βάσεις γνώσης. Το RAG βοηθά τα μεγάλα γλωσσικά μοντέλα (LLMs) να παρέχουν πιο σχετικές και υψηλότερης ποιότητας απαντήσεις. [1]

Προκειμένου να γίνει πιο κατανοητή η έννοια των συστημάτων που αξιοποιούν την αρχιτεκτονική RAG καθώς και για να αντιληφθούμε την συμβολή της στο generative AI, μπορούμε να την παρομοιάσουμε με την αίθουσα ενός δικαστηρίου. Οι δικαστές αποφασίζουν για τις υποθέσεις που τους ανατίθενται στηριζόμενοι στις γνώσεις που έχουν πάνω στην νομοθεσία. Μερικές φορές κάποιες υποθέσεις απαιτούν ειδική τεχνογνωσία, οπότε οι δικαστές στέλνουν τους δικαστικούς υπαλλήλους σε μια νομική βιβλιοθήκη, αναζητώντας δεδικασμένα και συγκεκριμένες υποθέσεις που μπορούν να επικαλεστούν. Όπως ένας καλός δικαστής, έτσι και τα μεγάλα γλωσσικά

μοντέλα (LLMs) μπορούν να απαντήσουν σε μια ευρεία γκάμα ερωτημάτων. Όμως, για να δώσουν έγκυρες απαντήσεις -για παράδειγμα βασισμένες σε συγκεκριμένα πρακτικά δικαστηρίου ή παρόμοιες πηγές- το μοντέλο πρέπει να τροφοδοτηθεί με αυτές τις πληροφορίες. Ο «δικαστικός υπάλληλος» της τεχνητής νοημοσύνης είναι μια διαδικασία που ονομάζεται retrieval-augmented generation ή εν συντομία RAG. [2]

Τον όρο RAG τον συναντάμε στην εργασία «Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks», η οποία δημοσιεύτηκε το 2020, όπου κύριος συγγραφέας ήταν ο Patrick Lewis. Στη συγκεκριμένη εργασία, οι συγγραφείς της πρότειναν μια αρχιτεκτονική όπου το μοντέλο ανακτά σχετικά έγγραφα από μια εξωτερική πηγή (για παράδειγμα Wikipedia) και στη συνέχεια τα χρησιμοποιεί για να παράγει την απάντηση. Πιο συγκεκριμένα, πρότειναν μία αρχιτεκτονική που συνδυάζει δύο είδη μνήμης: την parametric memory η οποία δηλώνει την γνώση που είναι ενσωματωμένη στα βάρη του μοντέλου κατά την εκπαίδευση, και την non-parametric memory, η οποία εκφράζει ένα εξωτερικό ευρετήριο, όπου το μοντέλο μπορεί να «διαβάσει» κείμενα σε πραγματικό χρόνο. Στα μοντέλα τους η parametric memory ήταν ένας pre-trained seq2seq transformer και η non-parametric memory ήταν ένας dense vector index της Wikipedia προσβάσιμος μέσω ενός προεκπαιδευμένου νευρωνικού συστήματος ανάκτησης. Ο retriever παρείχε τα έγγραφα που ήταν σχετικά με το ερώτημα του χρήστη και το seq2seq μοντέλο στη συνέχεια βασιζόταν στα συγκεκριμένα έγγραφα μαζί με το ερώτημα του χρήστη για να παραγάγει το αποτέλεσμα. [3]

Από τα ανωτέρω συμπεραίνουμε ότι ένα σύστημα που αξιοποιεί την αρχιτεκτονική RAG σε γενικές γραμμές λειτουργεί ως εξής: ο χρήστης κάνει μία ερώτηση. Στη συνέχεια το σύστημα κάνει ανάκτηση από μία εξωτερική πηγή τα πιο σχετικά αποσπάσματα με το ερώτημα του χρήστη (Retrieval). Τα αποσπάσματα αυτά προστίθενται στην ερώτηση του χρήστη (Augmentation). Το γλωσσικό μοντέλο διαβάζει τα αποσπάσματα και την ερώτηση του χρήστη και συνθέτει μια απάντηση βασισμένη σε αυτά (Generation).

Τα Generative AI μοντέλα εκπαιδεύονται σε τεράστια σύνολα δεδομένων και ανατρέχουν σε αυτές τις πληροφορίες για να παράγουν αποτελέσματα. Ωστόσο, τα

σύνολα δεδομένων εκπαίδευσης είναι πεπερασμένα και περιορίζονται στις πληροφορίες στις οποίες έχει πρόσβαση ο προγραμματιστής του μοντέλου τεχνητής νοημοσύνης. Τέτοια είδους σύνολα δεδομένων θα μπορούσαν να θεωρηθούν άρθρα στο διαδίκτυο, περιεχόμενο μέσων κοινωνικής δικτύωσης και άλλα δεδομένα που είναι δημόσια προσβάσιμα. Το RAG επιτρέπει στα Generative AI μοντέλα να έχουν πρόσβαση σε πρόσθετες εξωτερικές βάσεις γνώσεων, όπως εσωτερικά οργανωτικά δεδομένα, επιστημονικά περιοδικά και εξειδικευμένα σύνολα δεδομένων. Ενσωματώνοντας σχετικές πληροφορίες από εξωτερικές πηγές γνώσεων στη διαδικασία παραγωγής της απάντησης στο ερώτημα ενός χρήστη, τα chatbots και άλλα εργαλεία επεξεργασίας φυσικής γλώσσας (NLP) μπορούν να δημιουργήσουν πιο ακριβές περιεχόμενο για συγκεκριμένους τομείς, χωρίς να απαιτείται περαιτέρω εκπαίδευση. Με αυτόν τον τρόπο έχουμε εξοικονόμηση πόρων. Μια εταιρεία που αξιοποιεί μοντέλα τεχνητής νοημοσύνης δεν χρειάζεται να εκπαιδεύει από την αρχή ένα μοντέλο με τα δικά της δεδομένα. Μπορεί να χρησιμοποιήσει την αρχιτεκτονική RAG για να δείξει στο υπάρχον μοντέλο τα δεδομένα της.

Τα πλεονεκτήματα της αρχιτεκτονικής RAG είναι πολύ σημαντικά και η υιοθέτησή της βοηθά τους οργανισμούς που αξιοποιούν μοντέλα τεχνητής νοημοσύνης να παρέχουν καλύτερες απαντήσεις προς τους χρήστες. Ένα από τα πιο σημαντικά πλεονεκτήματα είναι η οικονομική και εύκολη στην επέκταση χρήση της τεχνητής νοημοσύνης. Πιο συγκεκριμένα, οι περισσότεροι οργανισμοί όταν θέλουν να εφαρμόσουν την τεχνητή νοημοσύνη, επιλέγουν πρώτα ένα θεμελιώδες μοντέλο (foundation model). Πρόκειται για τα μοντέλα βαθιάς μάθησης (deep-learning) που χρησιμεύουν ως θεμέλιο για την ανάπτυξη πιο εξελιγμένων εκδόσεων. Τα foundation models συνήθως διαθέτουν γενικευμένες βάσεις γνώσης που έχουν δημιουργηθεί από δημόσια διαθέσιμα δεδομένα εκπαίδευσης, όπως περιεχόμενο του διαδικτύου που ήταν προσβάσιμο τη χρονική στιγμή της εκπαίδευσής τους. Η επανεκπαίδευση, ή το fine tunning ενός foundation model αποτελεί μια δαπανηρή σε υπολογιστικούς πόρους διαδικασία. Με το RAG οι επιχειρήσεις μπορούν να αξιοποιήσουν τις δικές τους έγκυρες πηγές δεδομένων και να πετύχουν παρόμοια βελτίωση στην απόδοση του μοντέλου, χωρίς να χρειαστεί επανεκπαίδευση. Με αυτό τον τρόπο μπορούν να

υλοποιήσουν τις εφαρμογές τεχνητής νοημοσύνης ανάλογα με τις ανάγκες τους μειώνοντας το κόστος και την απαίτηση σε πόρους.[1]

Ένα ακόμα μεγάλο πλεονέκτημα της αρχιτεκτονικής RAG είναι η μείωση των «παραισθήσεων» (hallucinations). Τα hallucinations συμβαίνουν όταν τα μοντέλα τεχνητής νοημοσύνης κατά τη δημιουργία των απαντήσεων παρουσιάζουν λανθασμένες ή επινοημένες πληροφορίες σαν να ήταν πραγματικά γεγονότα. Το RAG είναι μια τεχνική που αξιοποιεί εξωτερικές πηγές για να ενισχύσει την αξιοπιστία των generative μοντέλων και αποτελεί μια δημοφιλή μέθοδο για τον μετριασμό των hallucinations. Για παράδειγμα, το Gemini ενσωματώνει τη μηχανή αναζήτησης για να παρέχει εξωτερικές αναφορές κατά τη διάρκεια της παραγωγής λόγου. Ωστόσο, το RAG αντιμετωπίζει αρκετές προκλήσεις Στο στάδιο της ανάκτησης της πληροφορίας από τις εξωτερικές πηγές, το μοντέλο απαιτείται να διακρίνει τον θόρυβο, τις άσχετες ή τις ψευδείς πληροφορίες. Στο στάδιο της ενίσχυσης, η ενσωμάτωση ετερογενών, ανεξάρτητων, ακόμη και αντικρουόμενων πληροφοριών παρουσιάζει δυσκολίες. Επιπλέον, το μοντέλο πρέπει να είναι σε θέση να αρνηθεί την παραγωγή απάντησης όταν οι παρεχόμενες πληροφορίες είναι ανεπαρκείς. Παρόλο που το RAG μπορεί να μειώσει τον κίνδυνο hallucinations, δεν μπορεί να καταστήσει ένα μοντέλο απόλυτα αλάνθαστο. [4]

Επιπροσθέτως, τα συστήματα παρέχουν πρόσβαση σε επίκαιρα και εξειδικευμένα δεδομένα. Τα Generative AI μοντέλα έχουν έναν περιορισμό στη γνώση που διαθέτουν, ο οποίος είναι το χρονικό σημείο στο οποίο ενημερώθηκαν τελευταία φορά τα δεδομένα εκπαίδευσής τους. Όσο περισσότερος χρόνος περνά από το συγκεκριμένο χρονικό σημείο και δεν έχει επανεκπαιδευτεί το μοντέλο, τόσο λιγότερο επίκαιρες και σχετικές απαντήσεις παρέχει. Τα συστήματα RAG συνδέουν τα μοντέλα με συμπληρωματικά εξωτερικά δεδομένα σε πραγματικό χρόνο και ενσωματώνουν ενημερωμένες πληροφορίες στις παραγόμενες απαντήσεις. Οι επιχειρήσεις χρησιμοποιούν το RAG για να εμπλουτίσουν τα μοντέλα με συγκεκριμένες πληροφορίες, όπως ιδιόκτητα δεδομένα της επιχείρησης τα οποία δεν είναι διαθέσιμα στο διαδίκτυο, έγκυρες έρευνες και άλλα σχετικά έγγραφα. Τα RAG συστήματα μπορούν επίσης να συνδεθούν στο διαδίκτυο μέσω APIs και να αποκτήσουν πρόσβαση σε ροές κοινωνικών δικτύων και αξιολογήσεις καταναλωτών

σε πραγματικό χρόνο. Παράλληλα, η πρόσβαση σε έκτακτες ειδήσεις και μηχανές αναζήτησης μπορεί να οδηγήσει σε πιο ακριβείς απαντήσεις, καθώς τα μοντέλα ενσωματώνουν τις ανακτημένες πληροφορίες στη διαδικασία παραγωγής κειμένου.

Τα μοντέλα που υιοθετούν την αρχιτεκτονική RAG, αυξάνουν την εμπιστοσύνη του χρήστη προς τα συγκεκριμένα μοντέλα. Για παράδειγμα τα chatbots, μια κοινή εφαρμογή της τεχνητής νοημοσύνης, απαντούν σε ερωτήματα που θέτουν οι χρήστες. Για να είναι επιτυχημένο ένα chatbot, οι χρήστες πρέπει να θεωρούν τα αποτελέσματά του αξιόπιστα. Τα συστήματα RAG έχουν τη δυνατότητα να περιλαμβάνουν στις απαντήσεις τους παραπομπές από τις πηγές γνώσης των εξωτερικών τους δεδομένων.

Τα μοντέλα που αξιοποιούν την αρχιτεκτονική RAG, έχουν πρόσβαση σε περισσότερα δεδομένα και μπορούν να απαντούν σε πιο σύνθετα ερωτήματα των χρηστών. Οι επιχειρήσεις μπορούν να βελτιστοποιήσουν τα μοντέλα και να αποκομίσουν μεγαλύτερη αξία από αυτά, διευρύνοντας τις βάσεις γνώσης τους και, κατά συνέπεια, επεκτείνοντας τα πλαίσια μέσα στα οποία τα μοντέλα παράγουν αξιόπιστα αποτελέσματα. Τα συστήματα RAG μπορούν να ανακτούν και να ενσωματώνουν πληροφορίες από πολλαπλές πηγές δεδομένων ως απόκριση σε σύνθετα ερωτήματα.

Δεν θα πρέπει να παραληφθεί πως τα μοντέλα που ακολουθούν την αρχιτεκτονική RAG παρέχουν μεγαλύτερη ασφάλεια δεδομένων. Επειδή η συγκεκριμένη αρχιτεκτονική συνδέει ένα μοντέλο με εξωτερικές πηγές γνώσης αντί να ενσωματώνει αυτή τη γνώση στα δεδομένα εκπαίδευσής του, διατηρείται ένας διαχωρισμός ανάμεσα στο μοντέλο και στις εξωτερικές πληροφορίες. Οι επιχειρήσεις μπορούν να χρησιμοποιούν RAG για να προστατεύουν τα πρωτογενή τους δεδομένα, ενώ ταυτόχρονα επιτρέπουν στα μοντέλα να έχουν πρόσβαση σε αυτά, η οποία μπορεί να ανακληθεί οποιαδήποτε στιγμή. Παρά τα οφέλη, οι επιχειρήσεις οφείλουν να μεριμνούν για την προστασία των εξωτερικών βάσεων δεδομένων τους. Το RAG βασίζεται σε διανυσματικές βάσεις (vector databases), οι οποίες χρησιμοποιούν διανυσματικές αναπαραστάσεις (embeddings) για τη μετατροπή της πληροφορίας σε αριθμητική μορφή. Σε περίπτωση παραβίασης, ένας επιτιθέμενος θα μπορούσε ενδεχομένως να ανασυνθέσει τα πρωτογενή δεδομένα από τα embeddings, κίνδυνος

που γίνεται ακόμα μεγαλύτερος εάν η βάση δεδομένων δεν είναι κρυπτογραφημένη.[1]

Αβίαστα συνάγεται το συμπέρασμα πως η αρχιτεκτονική RAG ανοίγει νέους δρόμους στην τεχνητή νοημοσύνη. Εκτός από τα πλεονεκτήματα που έχει η συγκεκριμένη αρχιτεκτονική καλό είναι να αναφερθούν και οι περιορισμοί που συναντώνται. Κατά τη σχεδίαση ενός συστήματος RAG μπορεί να εντοπιστούν σημεία αποτυχίας. Η πρώτη περίπτωση αποτυχίας είναι το ελλιπές περιεχόμενο. Όταν δηλαδή υποβάλλεται μια ερώτηση που δεν μπορεί να απαντηθεί από τα διαθέσιμα έγγραφα. Στην ιδανική περίπτωση, το σύστημα RAG θα απαντήσει κάτι όπως «Λυπάμαι, δεν γνωρίζω». Ωστόσο, για ερωτήσεις που σχετίζονται με το περιεχόμενο αλλά δεν υπάρχει απάντηση, το σύστημα θα μπορούσε να παραπλανηθεί στο να δώσει μια απάντηση.

Μία ακόμη περίπτωση αποτυχίας είναι η παράληψη των κορυφαίων σε κατάταξη εγγράφων. Όταν ένα σύστημα RAG ανακτά τα σχετικά έγγραφα από τη βάση δεδομένων τα οποία σχετίζονται με το ερώτημα του χρήστη, τα βαθμολογεί και επιστρέφει στον χρήστη τα πιο σχετικά με βάση το ερώτημά του. Υπάρχει περίπτωση η απάντηση στην ερώτηση να υπάρχει στο έγγραφο, αλλά αυτό δεν κατατάχθηκε αρκετά ψηλά ώστε να επιστραφεί στον χρήστη.

Στα ανωτέρω αξίζει να προστεθεί και η περίπτωση στην οποία τα έγγραφα που σχετίζονται με την απάντηση του χρήστη ανακτήθηκαν από τη βάση δεδομένων αλλά δεν εντάχθηκαν στο context που δόθηκε στο LLM για να παράγει την απάντηση στον χρήστη. Αυτό συμβαίνει συνήθως όταν επιστρέφονται πολλά έγγραφα από τη βάση δεδομένων και λαμβάνει χώρα μια διαδικασία ενοποίησης για την εξαγωγή της απάντησης. Πολλά LLM έχουν αυστηρά όρια στο πόσο κείμενο μπορεί να συμπεριληφθεί σε ένα prompt (token limit και rate limit). Αυτό περιορίζει τον αριθμό των τμημάτων που μπορούν να συμπεριληφθούν σε ένα prompt για την εξαγωγή μιας απάντησης, και απαιτείται μια στρατηγική μείωσης ώστε να αλυσοδεθούν πολλαπλά prompts για να προκύψει η τελική απάντηση.

Ακόμη υπάρχει περίπτωση το LLM να μην εξάγει τη σωστή απάντηση. Ενώ η απάντηση υπάρχει στο context που έχει δοθεί στο LLM εκείνο αποτυγχάνει να εξαγάγει τη

σωστή απάντηση. Συνήθως, αυτό συμβαίνει όταν υπάρχει υπερβολικός θόρυβος ή αντικρουόμενες πληροφορίες στο περιεχόμενο. Επιπλέον μπορεί να συμβεί ότι ενώ έχουν εξαχθεί τα σωστά αποσπάσματα από τη βάση δεδομένων και το LLM έχει καταλάβει το νόημα και έχει σχηματίσει μία σωστή απάντηση, να μην την δώσει στη μορφή που τη ζήτησε ο χρήστης. Για παράδειγμα η ερώτηση περιλάμβανε την εξαγωγή πληροφοριών σε μια συγκεκριμένη μορφή, όπως πίνακα ή λίστα, και το LLM αγνόησε την οδηγία.

Μία ακόμη αποτυχία ενός συστήματος RAG είναι η απάντηση στο ερώτημα του χρήστη να επιστρέφεται, αλλά δεν είναι αρκετά συγκεκριμένη ή είναι υπερβολικά συγκεκριμένη για να καλύψει τις ανάγκες του χρήστη. Αυτό συμβαίνει όταν οι σχεδιαστές του συστήματος RAG επιδιώκουν ένα συγκεκριμένο αποτέλεσμα για μια ερώτηση (για παράδειγμα οι εκπαιδευτικοί προς τους μαθητές). Σε αυτή την περίπτωση, θα πρέπει να παρέχεται συγκεκριμένο εκπαιδευτικό περιεχόμενο μαζί με τις απαντήσεις. Η λανθασμένη εξειδίκευση προκύπτει επίσης όταν οι χρήστες δεν είναι σίγουροι πώς να διατυπώσουν μια ερώτηση και τη διατυπώνουν πολύ γενικά.

Στις αποτυχίες ενός συστήματος RAG πρέπει να προστεθούν και οι ημιτελείς απαντήσεις. Οι συγκεκριμένες απαντήσεις δεν είναι λανθασμένες, αλλά υπολείπονται ορισμένων πληροφοριών, παρόλο που αυτές υπήρχαν στο πλαίσιο που δόθηκε στο LLM. Ένα παράδειγμα ερώτησης είναι: «Ποια είναι τα βασικά σημεία που καλύπτονται στα έγγραφα Α, Β και Γ;».[5]

Τα συστήματα RAG αποτελούν μια σημαντική πρόκληση για τη σύγχρονη τεχνητή νοημοσύνη, καθώς συνδυάζουν τεχνικές ανάκτησης πληροφοριών με LLMs. Πρόκειται για μια σύγχρονη προσέγγιση ανάκτησης γνώσης, η οποία υπερβαίνει τις παραδοσιακές μεθόδους, επιτρέποντας τη δυναμική ενσωμάτωση εξωτερικής πληροφορίας στη διαδικασία παραγωγής απαντήσεων. Οι μηχανικοί λογισμικού έχουν τη δυνατότητα να σχεδιάσουν και να υλοποιήσουν πολύπλοκα συστήματα RAG, λαμβάνοντας υπόψη τόσο τα πλεονεκτήματα όσο και τους περιορισμούς που προκύπτουν από την αρχιτεκτονική αυτή.

2.1.2 Μοντέλο λειτουργίας ενός RAG συστήματος

Τα συστήματα RAG λειτουργούν ως προσωπικός βοηθός βιβλιοθήκης για τα LLMs, παρέχοντάς τους πρόσβαση σε εξωτερικές βάσεις γνώσης ώστε να συμπληρώνουν την εσωτερική τους γνώση. [6] Αποτελούνται από τέσσερα βασικά components. Το πρώτο component το οποίο αποτελεί το πρώτο στάδιο στην ανάπτυξη ενός συστήματος RAG είναι το knowledge base, δηλαδή μία βάση γνώσης στην οποία μπορεί να γίνει αναζήτηση. Το εξωτερικό αποθετήριο δεδομένων μπορεί να περιλαμβάνει πληροφορίες από πολλές και διαφορετικές πηγές, όπως αρχεία PDF, έγγραφα, οδηγοί, ιστοσελίδες, αρχεία ήχου και άλλα. Το μεγαλύτερο μέρος αυτών των δεδομένων είναι μη δομημένο, δηλαδή δεν έχει ακόμη οργανωθεί ή επισημανθεί με συγκεκριμένες ετικέτες.

Τα συστήματα RAG αξιοποιούν μια διαδικασία που ονομάζεται embedding, μέσω της οποίας τα δεδομένα μετατρέπονται σε αριθμητικές αναπαραστάσεις, γνωστές ως διανύσματα. Το μοντέλο που αξιοποιείται για να γίνει η διαδικασία του embedding αποτυπώνει τις πληροφορίες σε έναν πολυδιάστατο μαθηματικό χώρο και τις οργανώνει με βάση τη σημασιολογική τους ομοιότητα· όσα δεδομένα θεωρούνται πιο σχετικά μεταξύ τους τοποθετούνται πιο κοντά. Επομένως, όλα τα δεδομένα των αρχείων από τα οποία θα ανακτηθούν τα πιο σχετικά για να εμπλουτίσουν την απάντηση που θα δώσει το LLM, διανυσματοποιούνται και αποθηκεύονται σε μία βάση διανυσμάτων (vector database). Για να διατηρείται η ποιότητα και η επικαιρότητα ενός συστήματος RAG, η βάση γνώσης του χρειάζεται συνεχή ενημέρωση.

Παράλληλα, τα LLMs μπορούν να επεξεργαστούν μόνο περιορισμένο όγκο πληροφοριών κάθε φορά, μέσα στο context window. Για τον λόγο αυτό, τα έγγραφα χωρίζονται σε μικρότερα τμήματα (chunking), ώστε τα παραγόμενα embeddings να μην υπερβαίνουν τα όρια του μοντέλου και να διατηρείται το νόημα.

Το μέγεθος των chunks αποτελεί κρίσιμη παράμετρο για τη σωστή λειτουργία του συστήματος RAG. Αν τα τμήματα είναι πολύ μεγάλα, οι πληροφορίες γίνονται υπερβολικά γενικές και δύσκολα συνδέονται με συγκεκριμένα ερωτήματα χρηστών.

Αντίθετα, όταν είναι υπερβολικά μικρά, υπάρχει κίνδυνος να χαθεί η σημασιολογική συνοχή του περιεχομένου.

Το επόμενο component για την υλοποίηση ενός συστήματος RAG είναι ο retriever, το οποίο αναζητά στη βάση γνώσης για δεδομένα που είναι σχετικά με το ερώτημα του χρήστη. Η μετατροπή των δεδομένων σε διανύσματα προετοιμάζει τη βάση γνώσης για σημασιολογική αναζήτηση, δηλαδή για μια μέθοδο που εντοπίζει πληροφορίες οι οποίες έχουν παρόμοιο νόημα με το ερώτημα του χρήστη. Οι αλγόριθμοι σημασιολογικής αναζήτησης μπορούν να εξετάζουν πολύ μεγάλες βάσεις δεδομένων και να βρίσκουν γρήγορα τις πιο σχετικές πληροφορίες, προσφέροντας σημαντικά μικρότερο χρόνο απόκρισης σε σύγκριση με τις κλασικές αναζητήσεις που βασίζονται αποκλειστικά σε λέξεις-κλειδιά. Στη συνέχεια, ο retriever αξιοποιεί ένα μοντέλο ενσωμάτωσης (embedding model) για τη μετατροπή του ερωτήματος του χρήστη σε διανυσματική αναπαράσταση και το συγκρίνει με τα διανύσματα που είναι αποθηκευμένα στη βάση γνώσης. Τα αποτελέσματα που παρουσιάζουν τη μεγαλύτερη ομοιότητα επιστρέφονται, ώστε να χρησιμοποιηθούν ως σχετικό περιεχόμενο για την παραγωγή της απάντησης.

Το τρίτο component είναι το integration layer το οποίο αποτελεί τον κεντρικό κόμβο της αρχιτεκτονικής RAG, καθώς συντονίζει τις επιμέρους διαδικασίες και διαχειρίζεται τη ροή των δεδομένων σε ολόκληρο το σύστημα. Αξιοποιώντας τις πληροφορίες που ανακτώνται από τη βάση γνώσης, το σύστημα RAG διαμορφώνει ένα νέο, εμπλουτισμένο ερώτημα προς το LLM που θα δημιουργήσει την απάντηση στον χρήστη. Το ερώτημα αυτό συνδυάζει την αρχική ερώτηση του χρήστη εμπλουτισμένη με την πληροφορία που παρείχε ο retriever. Για τη δημιουργία αποτελεσματικών προτροπών, τα συστήματα RAG χρησιμοποιούν τεχνικές prompt engineering, οι οποίες αυτοματοποιούν τη διαδικασία και ενισχύουν την ποιότητα της τελικής απάντησης του μοντέλου.

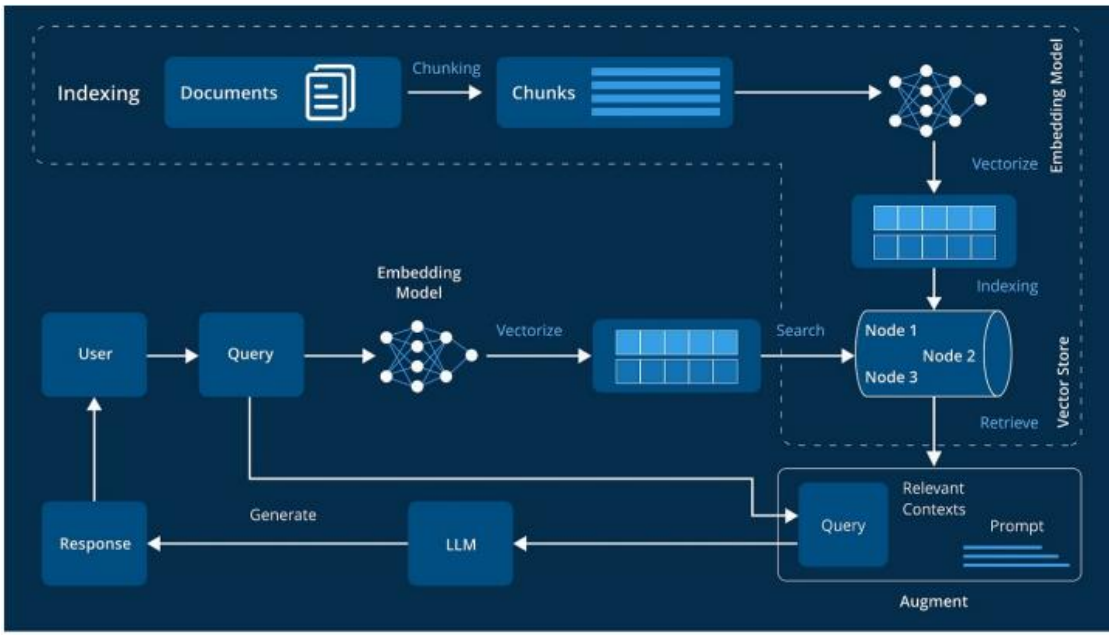
Το τέταρτο component είναι ο generator, οποίος δημιουργεί μία απάντηση στον χρήστη που βασίζεται στα δεδομένα που έλαβε από το integration layer. Το prompt που έλαβε ο generator αποτελεί μία σύνθεση του ερωτήματος του χρήστη και των πληροφοριών που ανακτήθηκαν από τη βάση γνώσης. Ο generator είναι συνήθως προ-εκπαιδευμένο γλωσσικό μοντέλο όπως το GPT, Gemini, Claude, Llama. Το

μοντέλο αυτό μαζί με την ερώτηση του χρήστη θα λάβει υπόψη του και όλα τα δεδομένα από τη βάση γνώσης, με αποτέλεσμα να παρέχει μία πιο ολοκληρωμένη απάντηση στον χρήστη.

Πρόσθετα components ενός συστήματος RAG μπορεί να περιλαμβάνουν έναν ranker, ο οποίος ιεραρχεί τα ανακτημένα δεδομένα με βάση τον βαθμό συνάφειάς τους, καθώς και έναν output handler, που αναλαμβάνει τη μορφοποίηση και παρουσίαση της παραγόμενης απάντησης με τρόπο κατανοητό και φιλικό προς τον χρήστη.[1]

Αφού παρουσιάστηκαν τα βασικά δομικά στοιχεία ενός συστήματος RAG, είναι απαραίτητο να αναλυθεί και η αρχιτεκτονική του, δηλαδή ο τρόπος με τον οποίο τα επιμέρους components αλληλεπιδρούν και οργανώνονται σε μια ενιαία ροή επεξεργασίας δεδομένων για την παραγωγή της τελικής απάντησης. Η αρχιτεκτονική ενός συστήματος RAG μπορεί να ιδωθεί ως μια ακολουθία διακριτών σταδίων, τα οποία αξιοποιούν τα components που περιγράφηκαν προηγουμένως.

Στο πλαίσιο της αρχιτεκτονικής η οποία παρουσιάζεται στην Εικόνα 1, το σύστημα RAG περιλαμβάνει ένα αρχικό στάδιο προεπεξεργασίας και ευρετηρίασης (indexing), κατά το οποίο τα δεδομένα της βάσης γνώσης διασπώνται σε τμήματα (chunking) και μετατρέπονται σε διανυσματικές αναπαραστάσεις (embedding), ώστε να καταστούν αναζητήσιμα. Κατά το στάδιο της ανάκτησης, ο retriever αξιοποιεί τα embeddings της βάσης γνώσης για να εντοπίσει τα πιο σχετικά αποσπάσματα σε σχέση με το ερώτημα του χρήστη. Σε σύγχρονες υλοποιήσεις, η αρχική ανάκτηση συχνά ακολουθείται από ένα επιπλέον στάδιο επανακατάταξης (re-ranking), το οποίο βελτιώνει την ακρίβεια των αποτελεσμάτων πριν αυτά προωθηθούν στο LLM. Στο τελικό στάδιο της αρχιτεκτονικής, ο generator παράγει την απάντηση λαμβάνοντας υπόψη τόσο το αρχικό ερώτημα όσο και τα επιλεγμένα αποσπάσματα γνώσης, διασφαλίζοντας ότι το παραγόμενο αποτέλεσμα είναι τεκμηριωμένο και βασισμένο σε εξωτερικές πηγές.[7]



Εικόνα 1: Παρουσίαση αρχιτεκτονικής ενός συστήματος RAG [7]

Η βιβλιογραφία διακρίνει διαφορετικές παραλλαγές αρχιτεκτονικής RAG (Εικόνα 2), όπως Naive, Advanced και Modular RAG, οι οποίες δεν διαφοροποιούνται ως προς τα βασικά components του συστήματος, αλλά ως προς τη ροή επεξεργασίας, τον αριθμό και το είδος των ενδιάμεσων σταδίων ανάκτησης και τον βαθμό ευελιξίας και επεκτασιμότητας της αρχιτεκτονικής.

Πιο συγκεκριμένα, το αρχιτεκτονικό πρότυπο του Naive RAG ακολουθεί την παραδοσιακή διαδικασία ενός RAG συστήματος η οποία περιλαμβάνει το indexing, το retrieval και το generation. Το indexing περιλαμβάνει τον καθαρισμό και την εξαγωγή ακατέργαστων δεδομένων από διαφορετικές μορφές, όπως PDF, HTML, Word τα οποία στη συνέχεια μετατρέπονται σε ενιαία μορφή απλού κειμένου. Για να αντιμετωπιστούν οι περιορισμοί του context window των γλωσσικών μοντέλων, το κείμενο διαχωρίζεται σε μικρότερα chunks. Στη συνέχεια, τα τμήματα αυτά κωδικοποιούνται σε διανυσματικές αναπαραστάσεις με τη χρήση ενός μοντέλου ενσωμάτωσης (embedding model) και αποθηκεύονται σε μια βάση διανυσμάτων. Το στάδιο αυτό είναι κρίσιμης σημασίας, καθώς επιτρέπει την αποδοτική εκτέλεση σημασιολογικών αναζητήσεων στη φάση της ανάκτησης που ακολουθεί. Στο στάδιο του retrieval το σύστημα RAG χρησιμοποιεί το ίδιο μοντέλο που αξιοποιήθηκε κατά

το στάδιο του indexing για να μετατρέψει το ερώτημα του χρήστη σε διανυσματική αναπαράσταση. Στη συνέχεια, υπολογίζει τον βαθμό ομοιότητας μεταξύ του διανύσματος του ερωτήματος και των διανυσμάτων των τμημάτων που περιλαμβάνονται στο ευρετηριασμένο σύνολο δεδομένων. Το σύστημα ιεραρχεί και ανακτά τα K τμήματα που εμφανίζουν τη μεγαλύτερη σημασιολογική συνάφεια με το ερώτημα. Τα τμήματα αυτά χρησιμοποιούνται στη συνέχεια ως εμπλουτισμένο πλαίσιο πληροφορίας στο prompt που παρέχεται στο γλωσσικό μοντέλο. Κατά το στάδιο του Generation το ερώτημα του χρήστη και τα επιλεγμένα έγγραφα συντίθενται σε ένα συνεκτικό prompt, στο οποίο ένα μεγάλο γλωσσικό μοντέλο καλείται να διαμορφώσει την απάντησή του.

Το αρχιτεκτονικό πρότυπο του Advanced RAG έρχεται να βελτιώσει περιορισμούς που συναντώνται στο Naive RAG, όπως είναι δυσκολίες στο retrieval, στο generation κατά τη δημιουργίας της τελικής απάντησης στον χρήστη ή προβλήματα στην ενσωμάτωση των πρόσθετων πληροφοριών από τις εξωτερικές πηγές που ανακτήθηκαν προκειμένου να διαμορφώσουν μία απάντηση στον χρήστη. Εστιάζοντας στη βελτίωση της ποιότητας της ανάκτησης, αξιοποιεί στρατηγικές τόσο πριν όσο και μετά το στάδιο της ανάκτησης (pre-retrieval και post-retrieval). Για την αντιμετώπιση ζητημάτων που σχετίζονται με την ευρετηρίαση, το Advanced RAG βελτιστοποιεί τις τεχνικές indexing μέσω της χρήσης μεθόδων όπως το sliding window, ο λεπτομερής διαχωρισμός του κειμένου σε τμήματα και η ενσωμάτωση μεταδεδομένων. Παράλληλα, ενσωματώνει διάφορες τεχνικές βελτιστοποίησης που αποσκοπούν στην αποδοτικότερη και πιο αποτελεσματική διαδικασία ανάκτησης.

Στο στάδιο του pre-retrieval, η κύρια έμφαση δίνεται στη βελτιστοποίηση τόσο της δομής ευρετηρίασης όσο και του αρχικού ερωτήματος του χρήστη. Στόχος της βελτιστοποίησης του indexing είναι η αναβάθμιση της ποιότητας του περιεχομένου που ευρετηριάζεται. Αυτό επιτυγχάνεται μέσω στρατηγικών όπως η αύξηση της λεπτομέρειας των δεδομένων (data granularity), η βελτιστοποίηση των δομών του ευρετηρίου, η προσθήκη μεταδεδομένων. Παράλληλα, η βελτιστοποίηση του ερωτήματος αποσκοπεί στο να καταστήσει την αρχική ερώτηση του χρήστη σαφέστερη και καταλληλότερη για τη διαδικασία ανάκτησης. Συνήθεις τεχνικές που

στοχεύουν στη βελτιστοποίηση του ερωτήματος αποτελούν οι query rewriting, query transformation, query expansion.

Στο στάδιο του post-retrieval, δηλαδή το στάδιο που έχει ανακτηθεί η απαραίτητη πληροφορία από τις εξωτερικές πηγές γνώσης, είναι κρίσιμη η αποτελεσματική ενσωμάτωσή της με το ερώτημα του χρήστη. Οι βασικές τεχνικές του σταδίου αυτού περιλαμβάνουν το re-ranking και το context compression. Το re-ranking της ανακτηθείσας πληροφορίας έχει ως αποτέλεσμα τα πιο σχετικά αποσπάσματα να ενσωματωθούν στο prompt στο γλωσσικό μοντέλο. Επιπλέον, η απευθείας εισαγωγή όλων των σχετικών εγγράφων στο γλωσσικό μοντέλο μπορεί να οδηγήσει σε υπερφόρτωση πληροφορίας, μειώνοντας την εστίαση στα ουσιώδη στοιχεία λόγω της παρουσίας άσχετου περιεχομένου. Για την αντιμετώπιση αυτού του προβλήματος, οι τεχνικές μετά την ανάκτηση επικεντρώνονται στην επιλογή των πιο κρίσιμων πληροφοριών και στην ανάδειξη των βασικών τμημάτων, προκειμένου το περιεχόμενο να μην υπερβαίνει τους περιορισμούς του context window του γλωσσικού μοντέλου και να υπάρχει διατήρηση της εστίασης στα πιο κρίσιμα τμήματα.

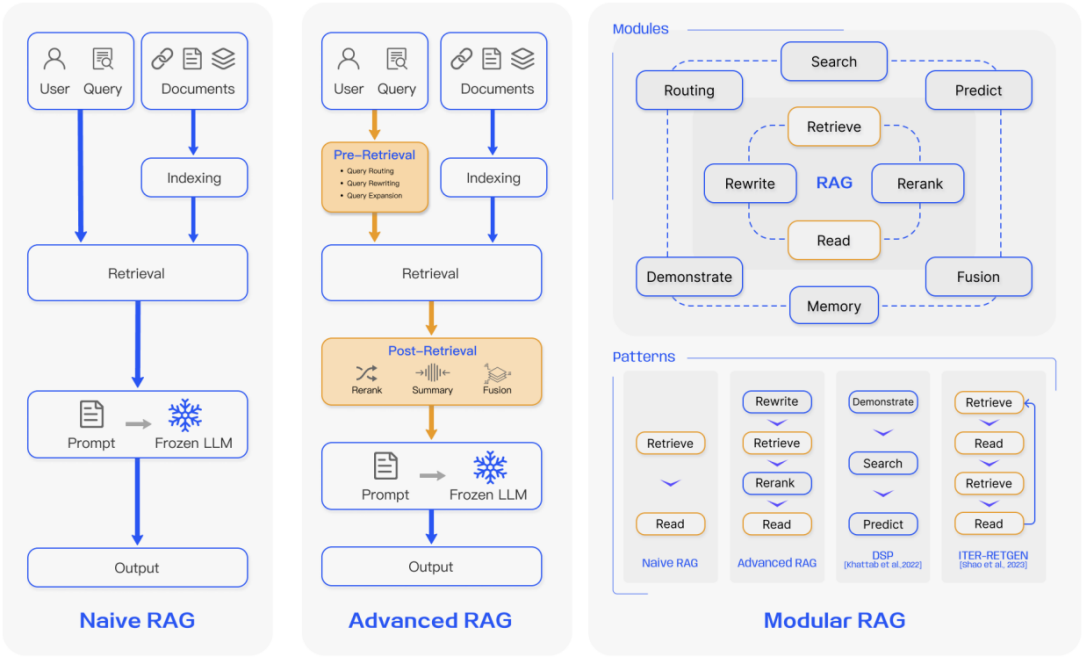
Το αρχιτεκτονικό πρότυπο του Modular RAG υπερέχει από τα δύο προηγούμενα παραδείγματα RAG, προσφέροντας αυξημένη προσαρμοστικότητα και ευελιξία. Ενσωματώνει ποικίλες στρατηγικές για τη βελτίωση των επιμέρους στοιχείων του, όπως η προσθήκη ενός search module για σημασιολογικές αναζητήσεις και η περαιτέρω βελτιστοποίηση του retriever μέσω fine-tuning. Το Search module προσαρμόζεται σε συγκεκριμένα σενάρια χρήσης, επιτρέποντας άμεσες αναζητήσεις σε ποικίλες πηγές δεδομένων, όπως μηχανές αναζήτησης, βάσεις δεδομένων και γράφους γνώσης, μέσω κώδικα και γλωσσών ερωτημάτων που παράγονται από LLMs. Το Memory module αξιοποιεί πληροφορίες που προκύπτουν από τη λειτουργία και τις προηγούμενες αλληλεπιδράσεις του γλωσσικού μοντέλου, ώστε να καθοδηγεί πιο αποτελεσματικά τη διαδικασία ανάκτησης. Με τον τρόπο αυτό, δημιουργείται ένα δυναμικό και επεκτάσιμο αποθετήριο μνήμης, το οποίο βελτιώνεται επαναληπτικά και συμβάλλει στην καλύτερη ευθυγράμμιση του ανακτημένου περιεχομένου με την κατανομή και τα χαρακτηριστικά των διαθέσιμων δεδομένων. Το Predict module στοχεύει στη μείωση της πλεονάζουσας πληροφορίας και του θορύβου, παράγοντας

απευθείας το κατάλληλο περιεχόμενο μέσω του LLM, διασφαλίζοντας συνάφεια και ακρίβεια. Τέλος, το Task Adapter module επιτρέπει την προσαρμογή του συστήματος RAG σε διαφορετικούς τύπους εργασιών, αυτοματοποιώντας τη διαμόρφωση των κατάλληλων προτροπών για σενάρια zero-shot και υποστηρίζοντας τη δημιουργία εξειδικευμένων μηχανισμών ανάκτησης μέσω παραγωγής few-shot ερωτημάτων, προσαρμοσμένων στις απαιτήσεις κάθε εργασίας.

Στο Modular RAG παρατηρούνται και νέα patterns τα οποία δεν υπάρχουν στα δύο προηγούμενα αρχιτεκτονικά παραδείγματα. Προσφέρει αξιοσημείωτη προσαρμοστικότητα, επιτρέποντας την αντικατάσταση ή αναδιάρθρωση modules ώστε να αντιμετωπίζονται συγκεκριμένες προκλήσεις. Αυτό υπερβαίνει τις σταθερές δομές των Naive και Advanced RAG, οι οποίες βασίζονται σε έναν απλό μηχανισμό «Retrieve» και «Read». Επιπλέον, το Modular RAG επεκτείνει αυτή την ευελιξία μέσω της ενσωμάτωσης νέων modules ή της προσαρμογής της αλληλεπίδρασης μεταξύ των υπαρχόντων, ενισχύοντας την εφαρμοσιμότητά της σε διαφορετικά είδη εργασιών.

Καινοτομίες όπως το μοντέλο Rewrite–Retrieve–Read αξιοποιούν τις δυνατότητες των LLMs για τη βελτίωση των ερωτημάτων ανάκτησης μέσω ενός module αναδιατύπωσης και ενός μηχανισμού ανάδρασης, βελτιώνοντας την απόδοση σε συγκεκριμένες εργασίες. Αντίστοιχα, προσεγγίσεις όπως Generate–Read αντικαθιστούν την παραδοσιακή ανάκτηση με περιεχόμενο που παράγεται από το LLM, ενώ το Recite–Read δίνει έμφαση στην ανάκτηση γνώσης από τα ίδια τα βάρη του μοντέλου, ενισχύοντας την ικανότητά του να χειρίζεται εργασίες υψηλής γνωσιακής απαίτησης.

Υβριδικές στρατηγικές ανάκτησης συνδυάζουν λέξεις-κλειδιά, σημασιολογική και διανυσματική αναζήτηση για την κάλυψη διαφορετικών τύπων ερωτημάτων. Επιπλέον, η χρήση υπο-ερωτημάτων και υποθετικών διανυσματικών αναπαραστάσεων εγγράφων (HyDE) αποσκοπεί στη βελτίωση της συνάφειας της ανάκτησης, εστιάζοντας στις ομοιότητες μεταξύ των embeddings των παραγόμενων απαντήσεων και των πραγματικών εγγράφων.[8]



Εικόνα 2: Τα αρχιτεκτονικά πρότυπα των Naive RAG, Advanced RAG, Modular RAG [8]

2.1.3 Μοντέλα Ενσωμάτωσης

Τα μοντέλα ενσωμάτωσης (embedding models) διαδραματίζουν σημαντικό ρόλο στην αρχιτεκτονική RAG. Τα συγκεκριμένα μοντέλα μετατρέπουν κείμενο, όπως μία πρόταση, μία παράγραφο σε ένα διάνυσμα αριθμών σταθερού μήκους, το οποίο αποδίδει τη σημασιολογική του έννοια. Αυτά τα διανύσματα επιτρέπουν στα υπολογιστικά συστήματα να συγκρίνουν και να αναζητούν κείμενο με βάση το νόημα και όχι με βάση τις ακριβείς λέξεις. Στην πράξη αυτό σημαίνει ότι κείμενα με παρόμοιο νόημα τοποθετούνται κοντά το ένα στο άλλο στον διανυσματικό χώρο. Για παράδειγμα, αντί να εντοπίζουν μόνο την ακριβή φράση «machine learning», τα μοντέλα ενσωμάτωσης μπορούν να φέρουν στην επιφάνεια έγγραφα που συζητούν σχετικές έννοιες, ακόμη και όταν χρησιμοποιείται διαφορετική διατύπωση.

Η λειτουργία του μοντέλου είναι να κωδικοποιεί κάθε input string ως ένα διάνυσμα πολλών διαστάσεων. Τα διανύσματα συγκρίνονται μεταξύ τους με τη χρήση μαθηματικών μετρικών, ώστε να μετρηθεί πόσο στενά σχετίζονται τα υποκείμενα κείμενα. Μετρικές σύγκρισης που χρησιμοποιούνται είναι το συνημίτονο ομοιότητας (cosine similarity) το οποίο μετρά τη γωνία μεταξύ δύο διανυσμάτων, η ευκλείδεια

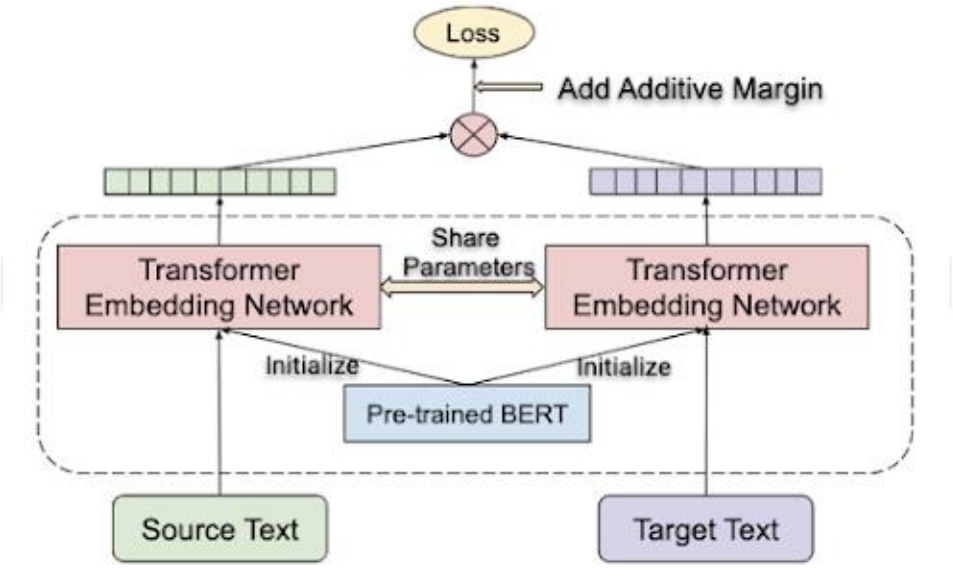
απόσταση (euclidean distance**)**, η οποία μετρά την απόσταση σε ευθεία γραμμή μεταξύ δύο σημείων, και το εσωτερικό γινόμενο το οποίο μετρά το μέγεθος της προβολής ενός διανύσματος πάνω σε ένα άλλο. Στο σημείο αυτό αξίζει να επισημανθεί πως υπάρχουν μοντέλα ενσωμάτωσης που χρησιμοποιούνται για την διανυσματοποίηση κειμένου (text-based embedding models) και μοντέλα που χρησιμοποιούνται για τη διανυσματοποίηση κειμένου, εικόνας και ήχου (multimodal embeddings models). [9]

Τα μοντέλα ενσωμάτωσης, όπως αναφέρθηκε σε προηγούμενη ενότητα, έχουν ως ρόλο τη διανυσματοποίηση των εξωτερικών πηγών. Τα διανύσματα που προκύπτουν αποθηκεύονται σε μία βάση διανυσμάτων. Το ερώτημα του χρήστη επίσης διανυσματοποιείται με το ίδιο μοντέλο που έγινε η διανυσματοποίηση στις εξωτερικές πηγές. Στη συνέχεια πραγματοποιείται semantic similarity προκειμένου να εντοπιστούν ποια έγγραφα από τις εξωτερικές πηγές δένουν νοηματικά στην ερώτηση του χρήστη, τα οποία θα χρησιμοποιηθούν για τον εμπλουτισμό της απάντησης που θα δώσει το LLM.

Υπάρχουν αρκετά μοντέλα ενσωμάτωσης και η χρησιμότητα του καθενός αποδεικνύεται από το περιβάλλον στο οποίο έχει εκπαιδευτεί καθώς και από τις εργασίες τις οποίες καλείται να επιτελέσει. Ένα πολυγλωσσικό μοντέλο ενσωμάτωσης (multilingual embedding model) είναι ένα ισχυρό εργαλείο που κωδικοποιεί κείμενο από διαφορετικές γλώσσες σε έναν κοινό διανυσματικό χώρο. Αυτό του επιτρέπει να εφαρμόζεται σε μια σειρά από εργασίες, όπως η ταξινόμηση κειμένου (text classification) και η ομαδοποίηση (clustering), αξιοποιώντας παράλληλα σημασιολογικές πληροφορίες για την κατανόηση της γλώσσας.

Το LaBSE (Language-agnostic BERT Sentence Embedding) αποτελεί ένα πολυγλωσσικό μοντέλο ενσωμάτωσης προτάσεων, βασισμένο σε αρχιτεκτονική BERT, το οποίο στοχεύει στην παραγωγή cross-lingual sentence embeddings. Το μοντέλο υποστηρίζει περισσότερες από 109 γλώσσες συμπεριλαμβανομένης της ελληνικής και έχει σχεδιαστεί για την ανάκτηση μεταφραστικά ισοδύναμων προτάσεων μεταξύ διαφορετικών γλωσσών. Η εκπαίδευσή του βασίζεται σε μεγάλο όγκο δεδομένων, αξιοποιώντας περίπου 17 δισεκατομμύρια μονόγλωσσες προτάσεις και 6 δισεκατομμύρια δίγλωσσα ζεύγη προτάσεων, σε συνδυασμό με τεχνικές προ-

εκπαίδευσης Masked Language Modeling (MLM) και Translation Language Modeling (TLM). Ως αποτέλεσμα, το LaBSE εμφανίζει ικανοποιητική απόδοση ακόμη και σε γλώσσες χαμηλών πόρων (low-resource languages), για τις οποίες δεν υπήρχαν εκτενή δεδομένα κατά την εκπαίδευση. Επιπλέον, το μοντέλο θέτει νέα δεδομένα σε πολλαπλές εργασίες ανάκτησης παράλληλων κειμένων (parallel text). Στο LaBSE χρησιμοποιήθηκε η αρχιτεκτονική του dual encoder (Εικόνα 3) για την τελική εκπαίδευσή του. Στην αρχιτεκτονική αυτή, υπάρχουν δύο πανομοιότυποι κωδικοποιητές (encoders) που μοιράζονται τα ίδια βάρη. Ο ένας κωδικοποιητής παίρνει μια πρόταση στη γλώσσα-πηγή (για παράδειγμα Ελληνικά). Ο άλλος κωδικοποιητής παίρνει μια πρόταση στη γλώσσα-στόχο (για παράδειγμα Αγγλικά). Το μοντέλο εκπαιδεύεται έτσι ώστε τα διανύσματα αυτών των δύο προτάσεων να καταλήγουν πολύ κοντά το ένα στο άλλο στον διανυσματικό χώρο αν η μία πρόταση είναι μετάφραση της άλλης. Χρησιμοποιήθηκε ένας transformer 12 επιπέδων με λεξιλόγιο 500 χιλιάδων tokens, ο οποίος έχει προ-εκπαιδευτεί με MLM και TLM σε 109 γλώσσες (Εικόνα 4), προκειμένου να αυξηθεί η κάλυψη του μοντέλου και του λεξιλογίου. Αυτό είχε ως αποτέλεσμα το LaBSE να παρέχει υποστήριξη μεγάλου αριθμού γλωσσών σε ένα ενιαίο μοντέλο.[10] [11]



Εικόνα 3: Dual encoder [10]

ISO	NAME	ISO	NAME	ISO	NAME
af	AFRIKAANS	ht	HAITIAN_CREOLE	pt	PORTUGUESE
am	AMHARIC	hu	HUNGARIAN	ro	ROMANIAN
ar	ARABIC	hy	ARMENIAN	ru	RUSSIAN
as	ASSAMESE	id	INDONESIAN	rw	KINYARWANDA
az	AZERBAIJANI	ig	IGBO	si	SINHALESE
be	BELARUSIAN	is	ICELANDIC	sk	SLOVAK
bg	BULGARIAN	it	ITALIAN	sl	SLOVENIAN
bn	BENGALI	ja	JAPANESE	sm	SAMOAN
bo	TIBETAN	jv	JAVANESE	sn	SHONA
bs	BOSNIAN	ka	GEORGIAN	so	SOMALI
ca	CATALAN	kk	KAZAKH	sq	ALBANIAN
ceb	CEBUANO	km	KHMER	sr	SERBIAN
co	CORSICAN	kn	KANNADA	st	SESOTHO
cs	CZECH	ko	KOREAN	su	SUNDANESE
cy	WELSH	ku	KURDISH	sv	SWEDISH
da	DANISH	ky	KYRGYZ	sw	SWAHILI
de	GERMAN	la	LATIN	ta	TAMIL
el	GREEK	lb	LUXEMBOURGISH	te	TELUGU
en	ENGLISH	lo	LAOTHIAN	tg	TAJIK
eo	ESPERANTO	lt	LITHUANIAN	th	THAI
es	SPANISH	lv	LATVIAN	tk	TURKMEN
et	ESTONIAN	mg	MALAGASY	tl	TAGALOG
eu	BASQUE	mi	MAORI	tr	TURKISH
fa	PERSIAN	mk	MACEDONIAN	tt	TATAR
fi	FINNISH	ml	MALAYALAM	ug	UIGHUR
fr	FRENCH	mn	MONGOLIAN	uk	UKRAINIAN
fy	FRISIAN	mr	MARATHI	ur	URDU
ga	IRISH	ms	MALAY	uz	UZBEK
gd	SCOTS_GAELIC	mt	MALTESE	vi	VIETNAMESE
gl	GALICIAN	my	BURMESE	wo	WOLOF
gu	GUJARATI	ne	NEPALI	xh	XHOSA
ha	HAUSA	nl	DUTCH	yi	YIDDISH
haw	HAWAIIAN	no	NORWEGIAN	yo	YORUBA
he	HEBREW	ny	NYANJA	zh	CHINESE
hi	HINDI	or	ORIYA	zu	ZULU
hmn	HMONG	pa	PUNJABI		
hr	CROATIAN	pl	POLISH		

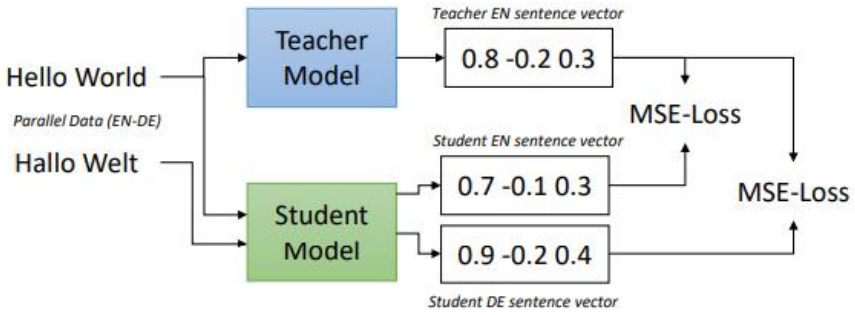
Εικόνα 4: Οι υποστηριζόμενες γλώσσες του LaBSE [10]

Το intfloat/multilingual-e5-base είναι ένα ακόμη πολυγλωσσικό text embedding μοντέλο, το οποίο έχει αρχικοποιηθεί από το xlm-roberta-base και έχει εκπαιδευτεί περαιτέρω σε ένα μείγμα πολυγλωσσικών συνόλων δεδομένων. Υποστηρίζει 100 γλώσσες από το xlm-roberta, ωστόσο το μοντέλο ενδέχεται να μην είναι τόσο

αποτελεσματικό στις γλώσσες που δεν έχουν πολλά διαθέσιμα δεδομένα εκπαίδευσης. Το μοντέλο είναι κατάλληλο για tasks που αφορούν ανάκτηση πληροφοριών, σημασιολογική ομοιότητα, ομαδοποίηση.

Η εκπαίδευσή του ακολουθεί μια διφασική διαδικασία, η οποία περιλαμβάνει αρχικά το στάδιο του weakly-supervised contrastive pre-training σε περίπου ένα δισεκατομμύριο πολυγλωσσικά ζεύγη κειμένων και στη συνέχεια fine-tuning σε μικρή ποσότητα labeled δεδομένων υψηλής ποιότητας. Σύμφωνα με τα πειραματικά αποτελέσματα, το multilingual-E5 επιτυγχάνει ανταγωνιστική απόδοση τόσο σε αγγλόφωνα όσο και σε πολυγλωσσικά benchmarks, όπως τα MTEB και MIRACL, γεγονός που καταδεικνύει την καταλληλότητά του για εφαρμογές πολυγλωσσικής σημασιολογικής ανάκτησης. [12]

Το paraphrase-multilingual-mpnet-base-v2 αποτελεί ένα πολυγλωσσικό μοντέλο που είναι αξιόλογο για tasks που αφορούν σημασιολογική ομοιότητα. Αποτελεί την πολυγλωσσική έκδοση του paraphrase-mpnet-base-v2 και είναι εκπαιδευμένο σε παράλληλα δεδομένα σε περισσότερες από 50 γλώσσες συμπεριλαμβανομένης και της ελληνικής. Το συγκεκριμένο μοντέλο είναι κατάλληλο να βρίσκει σημασιολογικά όμοιες προτάσεις εντός μίας γλώσσας ή σε διάφορες γλώσσες. Το paraphrase-multilingual-mpnet-base-v2 έχει εκπαιδευτεί με τη μέθοδο Knowledge Distillation, όπου ένα πολυγλωσσικό Student μοντέλο μαθαίνει να προσεγγίζει τη συμπεριφορά ισχυρότερων μοντέλων σημασιολογικής ομοιότητας, με στόχο την αποτελεσματική αναπαράσταση παραφραστικά ισοδύναμων προτάσεων σε πολλαπλές γλώσσες. Με αυτή τη μέθοδο μπορούν να δημιουργηθούν εκδόσεις πολυγλωσσικών μοντέλων τα οποία προηγουμένως ήταν μονογλωσσικά. Η εκπαίδευση βασίζεται στην ιδέα ότι μια μεταφρασμένη πρόταση θα πρέπει να αντιστοιχίζεται στην ίδια θέση στον διανυσματικό χώρο με την αρχική πρόταση. Στη συγκεκριμένη μέθοδο υπάρχει ένα μονογλωσσικό Teacher Model το οποίο γνωρίζει πολύ καλά μία συγκεκριμένη γλώσσα. Στη συνέχεια εκπαιδεύεται ένα Student Model σε μεταφρασμένες προτάσεις του Teacher Model προκειμένου να το μιμηθεί. Όπως παρουσιάζεται στην Εικόνα 5, τα δεδομένα που δίνονται στο Student μοντέλο το εκπαιδεύουν ώστε τα παραγόμενα διανύσματα για τις αγγλικές και γερμανικές προτάσεις να είναι κοντά στο διάνυσμα της αγγλικής πρότασης του Teacher Model. [13] [14]



Εικόνα 5: Knowledge Distillation [13]

Πέρα από τα μοντέλα ενσωμάτωσης που προτείνονται κυρίως στο πλαίσιο της ακαδημαϊκής έρευνας, στη σύγχρονη βιβλιογραφία και στην πράξη αξιοποιούνται επίσης ευρέως εμπορικά διαθέσιμες λύσεις που προσφέρονται ως υπηρεσία (as-a-service), όπως τα μοντέλα ενσωμάτωσης κειμένου της OpenAI. Τα μοντέλα αυτά μετατρέπουν κειμενικές εισόδους σε πυκνά διανύσματα σταθερού μήκους, τα οποία κωδικοποιούν τη σημασιολογική πληροφορία του κειμένου και μπορούν να χρησιμοποιηθούν σε εφαρμογές σημασιολογικής αναζήτησης, ανάκτησης πληροφορίας και ομαδοποίησης δεδομένων.

Τα embeddings της OpenAI αποτελούν μοντέλα γενικού σκοπού για αναπαράσταση κειμένου, σχεδιασμένα ώστε να αποδίδουν σε κείμενα σε πολλές γλώσσες και να είναι αξιοποιήσιμα σε ευρύ φάσμα θεματικών πεδίων και εφαρμογών. Η σχεδίασή τους αποσκοπεί στην ευρεία εφαρμοσιμότητα σε διαφορετικά πληροφοριακά συστήματα, χωρίς εξειδίκευση σε συγκεκριμένο γνωστικό πεδίο. Επιπλέον, οι νεότερες εκδόσεις των μοντέλων επιτρέπουν τη ρύθμιση της διάστασης των παραγόμενων διανυσμάτων, διευκολύνοντας τη βελτιστοποίηση μεταξύ υπολογιστικού κόστους και ακρίβειας ανάλογα με τις ανάγκες της εφαρμογής. Ως εκ τούτου, παρουσιάζουν σταθερή και ικανοποιητική απόδοση σε ένα ευρύ φάσμα εργασιών σημασιολογικής αναζήτησης και σε αρχιτεκτονικές RAG. [15]

Αντίστοιχη προσέγγιση ακολουθείται και από την Google μέσω της οικογένειας μοντέλων Gemini, η οποία περιλαμβάνει και μοντέλα ενσωμάτωσης κειμένου. Τα embeddings του Gemini στοχεύουν στη δημιουργία διανυσματικών αναπαραστάσεων γενικού σκοπού, οι οποίες μπορούν να αξιοποιηθούν σε

εφαρμογές σημασιολογικής αναζήτησης, ανάκτησης πληροφορίας και σε αρχιτεκτονικές Retrieval-Augmented Generation (RAG).

Η σχεδίαση των μοντέλων αυτών δίνει έμφαση στην ευελιξία και στη δυνατότητα ενσωμάτωσής τους σε σύνθετα πληροφοριακά συστήματα. Η οικογένεια Gemini, ως σύνολο, υποστηρίζει πολυτροπική επεξεργασία δεδομένων. Το Gemini API προσφέρει μοντέλα ενσωμάτωσης κειμένου (text embedding models) για τη δημιουργία embeddings για λέξεις, φράσεις, προτάσεις και κώδικα. Τα embeddings εκτελούν εργασίες παρέχοντας αποτελέσματα με μεγαλύτερη ακρίβεια και επίγνωση του πλαισίου (context) σε σύγκριση με τις προσεγγίσεις που βασίζονται σε λέξεις-κλειδιά. Ως αποτέλεσμα, τα embeddings του Gemini αποτελούν μία σύγχρονη λύση ενσωμάτωσης κειμένου γενικού σκοπού, η οποία χρησιμοποιείται τόσο σε ερευνητικές όσο και σε εφαρμοστικές προσεγγίσεις. [16] [17]

2.1.4 Vector Databases

Μία βάση διανυσμάτων (vector database) αποθηκεύει, διαχειρίζεται και ευρετηριάζει δεδομένα διανυσμάτων υψηλών διαστάσεων. Τα σημεία των δεδομένων αποθηκεύονται ως πίνακες αριθμών που ονομάζονται διανύσματα (vectors), οι οποίοι ομαδοποιούνται με βάση την ομοιότητά τους (similarity). Αυτός ο μηχανισμός είναι ιδανικός για ερωτήματα γρήγορης απόκρισης και είναι κατάλληλος για εφαρμογές τεχνητής νοημοσύνης.

Σε αντίθεση με τις παραδοσιακές σχεσιακές βάσεις δεδομένων, τα σημεία δεδομένων σε μία διανυσματική βάση αναπαρίστανται από διανύσματα με συγκεκριμένο αριθμό διαστάσεων. Οι διανυσματικές βάσεις είναι καλύτερες στο να διαχειρίζονται αδόμητα σύνολα δεδομένων (unstructured datasets). Η διανυσματική αναζήτηση αναπαριστά τα δεδομένα ως dense vectors, δηλαδή διανύσματα στα οποία τα περισσότερα ή όλα τα στοιχεία τους είναι μη μηδενικά. Τα διανύσματα αναπαρίστανται σε έναν συνεχή διανυσματικό χώρο (vector space), ο οποίος είναι το μαθηματικό πλαίσιο μέσα στο οποίο τα δεδομένα εκφράζονται ως διανύσματα.

Οι διανυσματικές αναπαραστάσεις καθιστούν δυνατή την αναζήτηση βάσει ομοιότητας (similarity search). Κάθε διάσταση ενός dense vectors αντιστοιχεί σε ένα λανθάνον χαρακτηριστικό (latent feature) ή μια πτυχή των δεδομένων. Ένα λανθάνον χαρακτηριστικό είναι μια υποκείμενη ιδιότητα ή γνώρισμα που δεν παρατηρείται άμεσα, αλλά συνάγεται από τα δεδομένα μέσω μαθηματικών μοντέλων ή αλγορίθμων. Τα λανθάνοντα χαρακτηριστικά αποτυπώνουν τα κρυμμένα μοτίβα και τις σχέσεις στα δεδομένα, επιτρέποντας πιο ουσιαστικές και ακριβείς αναπαραστάσεις των αντικειμένων ως διανύσματα σε έναν χώρο υψηλών διαστάσεων.

Όπως ειπώθηκε και σε προηγούμενη ενότητα, τα μοντέλα ενσωμάτωσης έχουν εκπαιδευτεί να μετατρέπουν τα δεδομένα σε διανύσματα. Οι βάσεις διανυσμάτων αποθηκεύουν και ευρετηριάζουν τα αποτελέσματα των μοντέλων ενσωμάτωσης. Μέσα στη βάση, τα διανύσματα μπορούν να ομαδοποιηθούν ή να αναγνωριστούν ως αντίθετα, με βάση τη σημασιολογική τους έννοια ή τα χαρακτηριστικά τους. Το μοντέλο ενσωμάτωσης θα μπορούσε να παρομοιαστεί με έναν μεταφραστή που μετατρέπει την ανθρώπινη γλώσσα, εικόνες ή ήχο σε μία μαθηματική αναπαράσταση. Για παράδειγμα εάν έχουμε τις λέξεις «Βασιλιάς» και «Βασίλισσα» το μοντέλο θα τις τοποθετήσει πολύ κοντά στον διανυσματικό χώρο γιατί έχουν παρόμοιο σημασιολογικό πλαίσιο. Η συγκεκριμένη τεχνολογία δεν περιορίζεται μόνο σε κείμενο. Υπάρχει η δυνατότητα να αποθηκευτεί το διάνυσμα μιας φωτογραφίας ενός σκύλου και η βάση να επιστρέψει άλλες φωτογραφίες σκύλων.

Οι βάσεις διανυσμάτων είναι απαραίτητες για την υλοποίηση συστημάτων RAG. Το RAG είναι ένα AI framework που επιτρέπει στα LLMs να αντλούν επίκαιρη πληροφορία από εξωτερικές πηγές γνώσης. Κάθε ερώτημα του χρήστη μετατρέπεται σε διανύσματα και συγκρίνεται σημασιολογικά με τα ήδη αποθηκευμένα διανύσματα της βάσης, τα οποία προέρχονται από την προ-επεξεργασία των εξωτερικών πηγών. Η σύγκριση αυτή εντοπίζει τα πιο σχετικά αποσπάσματα πληροφορίας, τα οποία στη συνέχεια παρέχονται στο LLM μαζί με το ερώτημα του χρήστη, επιτρέποντάς του να παράγει μια τεκμηριωμένη και ακριβή απάντηση.[18]

Μία από τις σύγχρονες υλοποιήσεις βάσεων διανυσμάτων που χρησιμοποιούνται ευρέως σε εφαρμογές RAG είναι η ChromaDB. Η Chroma αποτελεί μία βάση

διανυσμάτων ανοιχτού κώδικα, σχεδιασμένη με έμφαση στην απλότητα χρήσης και στην εύκολη ενσωμάτωσή της σε συστήματα που αξιοποιούν μοντέλα ενσωμάτωσης κειμένου. Παρέχει δυνατότητες αποθήκευσης και ευρετηρίασης dense vector αναπαραστάσεων, καθώς και μηχανισμούς αναζήτησης βασισμένους στη σημασιολογική ομοιότητα, επιτρέποντας την αποδοτική ανάκτηση των πιο σχετικών εγγραφών σε σχέση με ένα ερώτημα.

Η ChromaDB υποστηρίζει τη συσχέτιση διανυσμάτων με μεταδεδομένα (metadata), γεγονός που διευκολύνει το φιλτράρισμά τους και την οργάνωση των αποθηκευμένων δεδομένων σε σύνθετα πληροφοριακά συστήματα. Επιπλέον, έχει σχεδιαστεί ώστε να λειτουργεί αποδοτικά σε τοπικά περιβάλλοντα ανάπτυξης, καθιστώντας την ιδιαίτερα κατάλληλη για πειραματισμό, έρευνα και ανάπτυξη πρωτοτύπων συστημάτων RAG. Λόγω αυτών των χαρακτηριστικών, η ChromaDB έχει υιοθετηθεί ευρέως σε εφαρμογές που συνδυάζουν μοντέλα ενσωμάτωσης και γλωσσικά μοντέλα, προσφέροντας μία ευέλικτη και πρακτική λύση για τη διαχείριση διανυσματικών δεδομένων. [19]

Το FAISS (Facebook AI Similarity Search) αποτελεί μία εξειδικευμένη βιβλιοθήκη για την αποδοτική αναζήτηση πλησιέστερων γειτόνων σε σύνολα διανυσματικών δεδομένων υψηλών διαστάσεων. Σε αντίθεση με τις παραδοσιακές σχεσιακές βάσεις δεδομένων, οι οποίες έχουν σχεδιαστεί για αναζητήσεις βασισμένες σε συγκεκριμένες τιμές και κλειδιά, το FAISS εστιάζει στην αναζήτηση ομοιότητας μεταξύ διανυσματικών αναπαραστάσεων. Με τον τρόπο αυτό, υποστηρίζει αποδοτικά την αναζήτηση σε πολύ μεγάλης κλίμακας σύνολα δεδομένων, τα οποία μπορεί να περιλαμβάνουν από εκατομμύρια έως και δισεκατομμύρια διανύσματα. Η βιβλιοθήκη περιλαμβάνει πληθώρα αλγορίθμων αναζήτησης πλησιέστερου γείτονα (nearest neighbor search), επιτρέποντας την προσαρμογή μεταξύ ταχύτητας, ακρίβειας και απαιτήσεων μνήμης, ανάλογα με το εκάστοτε σενάριο χρήσης. Επιπλέον, το FAISS έχει υλοποιηθεί σε γλώσσα C++ και παρέχει διεπαφή για Python, ενώ υποστηρίζει και επιτάχυνση μέσω GPU με τη χρήση CUDA, γεγονός που το καθιστά κατάλληλο για εφαρμογές μεγάλης υπολογιστικής κλίμακας. [20]

Στα ανωτέρω δεν θα έπρεπε να παραληφθεί το Pinecone, το οποίο αποτελεί μία πλήρως διαχειριζόμενη βάση διανυσμάτων, η οποία προσφέρεται ως υπηρεσία και

έχει σχεδιαστεί για την αποδοτική αποθήκευση και αναζήτηση διανυσματικών αναπαραστάσεων σε μεγάλης κλίμακας εφαρμογές. Σε αντίθεση με αλγοριθμικές βιβλιοθήκες αναζήτησης διανυσμάτων, όπως το FAISS, το Pinecone παρέχει ένα ολοκληρωμένο περιβάλλον διαχείρισης διανυσματικών δεδομένων, υποστηρίζοντας λειτουργίες διαχείρισης υποδομής και αποδοτικής αναζήτησης διανυσμάτων σε περιβάλλοντα μεγάλης κλίμακας. Η πλατφόρμα υποστηρίζει αναζητήσεις πλησιέστερων γειτόνων με χρήση προσεγγιστικών τεχνικών (Approximate Nearest Neighbor), καθώς και τη συσχέτιση διανυσμάτων με μεταδεδομένα, επιτρέποντας σύνθετα φίλτρα και πιο ελεγχόμενη ανάκτηση πληροφορίας. Λόγω της σχεδίασής του ως υπηρεσία, το Pinecone χρησιμοποιείται ευρέως σε παραγωγικά συστήματα και αρχιτεκτονικές RAG, όπου απαιτείται αξιοπιστία, επεκτασιμότητα και ελαχιστοποίηση της πολυπλοκότητας διαχείρισης της υποδομής. [21] [22]

Στις ανωτέρω βάσεις διανυσμάτων θα πρέπει να προστεθεί και το Weaviate. Αποτελεί μία σύγχρονη βάση διανυσμάτων ανοιχτού κώδικα, σχεδιασμένη για την αποθήκευση, ευρετηρίαση και αναζήτηση διανυσματικών αναπαραστάσεων σε εφαρμογές σημασιολογικής ανάκτησης πληροφορίας. Σε αντίθεση με αλγοριθμικές βιβλιοθήκες αναζήτησης διανυσμάτων, όπως το FAISS, το Weaviate παρέχει ένα ολοκληρωμένο σύστημα διαχείρισης διανυσματικών δεδομένων, υποστηρίζοντας μόνιμη αποθήκευση, μεταδεδομένα και αναζήτηση βάσει ομοιότητας. Το Weaviate υποστηρίζει αναζητήσεις πλησιέστερων γειτόνων με χρήση προσεγγιστικών τεχνικών (Approximate Nearest Neighbor), ενώ επιτρέπει τον συνδυασμό διανυσματικής αναζήτησης με φίλτρα βασισμένα σε μεταδεδομένα, προσφέροντας μεγαλύτερο έλεγχο στη διαδικασία ανάκτησης πληροφορίας. Επιπλέον, έχει σχεδιαστεί ώστε να μπορεί να λειτουργήσει τόσο ως αυτόνομη εγκατάσταση όσο και ως διαχειριζόμενη υπηρεσία, γεγονός που το καθιστά κατάλληλο για χρήση σε ερευνητικά περιβάλλοντα, αλλά και σε παραγωγικά συστήματα και αρχιτεκτονικές RAG. [23]

2.1.5 Μεγάλα Γλωσσικά Μοντέλα

Τα μεγάλα γλωσσικά μοντέλα (LLMs) αποτελούν αλγόριθμους βαθιάς μάθησης και έχουν τη δυνατότητα να αναγνωρίζουν, να συνοψίζουν, να μεταφράζουν, να

προβλέπουν και να δημιουργούν περιεχόμενο, αξιοποιώντας πολύ μεγάλα σύνολα δεδομένων. Τα LLMs αντιπροσωπεύουν σε μεγάλο βαθμό μια κατηγορία αρχιτεκτονικών βαθιάς μάθησης που ονομάζονται δίκτυα μετασχηματιστών (transformer networks). Ένα μοντέλο transformer είναι ένα νευρωνικό δίκτυο που μαθαίνει το πλαίσιο και το νόημα παρακολουθώντας τις σχέσεις σε διαδοχικά δεδομένα, όπως οι λέξεις σε μία πρόταση Ένας transformer αποτελείται από πολλαπλά τμήματα (transformer blocks), γνωστά και ως επίπεδα (layers). Για παράδειγμα, ένας transformer διαθέτει self-attention layers, feed-forward layers και normalization layers, τα οποία συνεργάζονται για να αποκωδικοποιήσουν την είσοδο και να προβλέψουν ροές εξόδου κατά τη φάση της εξαγωγής συμπερασμάτων. Τα επίπεδα αυτά μπορούν να στοιβάζονται το ένα πάνω στο άλλο για τη δημιουργία βαθύτερων transformers και ισχυρών γλωσσικών μοντέλων.

Υπάρχουν δύο βασικές καινοτομίες που καθιστούν τους transformers ιδιαίτερα κατάλληλους για μεγάλα γλωσσικά μοντέλα: το positional encoding και το self-attention. Μέσω των positional encodings, η πληροφορία της θέσης κάθε token ενσωματώνεται στην αναπαράστασή του, επιτρέποντας στο μοντέλο να επεξεργάζεται ολόκληρη την ακολουθία παράλληλα, χωρίς σειριακή επεξεργασία. Χωρίς τη χρήση positional encodings, ο transformer δεν διαθέτει εγγενή πληροφορία σχετικά με τη σειρά των tokens και αντιμετωπίζει την είσοδο ως σύνολο μη διατεταγμένων στοιχείων. Η συγκεκριμένη τεχνική επιτρέπει την ενσωμάτωση της έννοιας της σχετικής θέσης των tokens μέσα στην ακολουθία.

Το self-attention αναθέτει ένα βάρος (weight) σε κάθε μέρος των δεδομένων εισόδου κατά την επεξεργασία τους. Το συγκεκριμένο βάρος υποδηλώνει τη σημασία του συγκεκριμένου μέρους των δεδομένων εισόδου σε σχέση με το υπόλοιπο περιεχόμενο. Με την συγκεκριμένη τεχνική τα μοντέλα δεν χρειάζεται να αφιερώνουν την ίδια προσοχή σε όλες τις εισόδους και μπορούν να εστιάσουν στα μέρη που πραγματικά έχουν σημασία. Η αναπαράσταση των σημείων στα οποία πρέπει να δώσει προσοχή το νευρωνικό δίκτυο μαθαίνεται με την πάροδο του χρόνου, καθώς το μοντέλο επεξεργάζεται και αναλύει τεράστιους όγκους δεδομένων. Ο συνδυασμός των positional encodings και του μηχανισμού self-attention επιτρέπει στο μοντέλο να συλλαμβάνει και να αναπαριστά σχέσεις μεταξύ στοιχείων της ακολουθίας, ακόμη και

όταν αυτά απέχουν σημαντικά μεταξύ τους, χωρίς την ανάγκη σειριακής επεξεργασίας. Η ικανότητα επεξεργασίας δεδομένων μη διαδοχικά επιτρέπει την αποδόμηση του σύνθετου προβλήματος σε πολλαπλούς, μικρότερους, ταυτόχρονους υπολογισμούς. Τέτοιου είδους υπολογισμοί μπορούν να εκτελεστούν παράλληλα σε μία κάρτα γραφικών (GPU). Οι κάρτες γραφικών είναι κατάλληλες για την εκτέλεση αυτών των υπολογισμών, επιτρέποντας την επεξεργασία μεγάλης κλίμακας μη επισημασμένων συνόλων δεδομένων (unlabelled datasets) και τεράστιων δικτύων transformers.

Στη βιβλιογραφία, τα μοντέλα που βασίζονται στην αρχιτεκτονική transformer μπορούν να κατηγοριοποιηθούν σε τρεις βασικές κλάσεις, ανάλογα με τη δομή τους και τον ρόλο τους στην επεξεργασία της γλώσσας. Η πρώτη κατηγορία είναι τα encoder only μοντέλα. Τα συγκεκριμένα μοντέλα είναι κατάλληλα για εργασίες που απαιτούν την κατανόηση της γλώσσας, όπως η ταξινόμηση (classification) και η ανάλυση συναισθήματος (sentiment analysis). Ένα παράδειγμα που εντάσσεται στην κατηγορία encoder only μοντέλων αποτελεί το BERT (Bidirectional Encoder Representations from Transformers). Η δεύτερη κατηγορία είναι τα decoder only μοντέλα. Αυτή η κατηγορία μοντέλων είναι εξαιρετικά ικανή στην παραγωγή γλώσσας και περιεχομένου. Ορισμένες περιπτώσεις χρήσης περιλαμβάνουν τη συγγραφή ιστοριών και τη δημιουργία αναρτήσεων σε ιστολόγια (blogs). Στη συγκεκριμένη κατηγορία μοντέλων εντάσσεται το GPT-3 (Generative Pretrained Transformer 3). Η τρίτη κατηγορία είναι τα encoder-decoder μοντέλα. Τα μοντέλα που ανήκουν στη συγκεκριμένη κατηγορία συνδυάζουν τα components encoder και decoder της αρχιτεκτονικής transformer, ώστε να μπορούν τόσο να κατανοούν όσο και να παράγουν περιεχόμενο. Ορισμένες περιπτώσεις χρήσης όπου αυτή η αρχιτεκτονική διαπρέπει περιλαμβάνουν τη μετάφραση και τη σύνοψη κειμένου. Ένα παραδείγματα αρχιτεκτονικής encoder-decoder αποτελεί το T5 (Text-to-Text Transformer).[24]

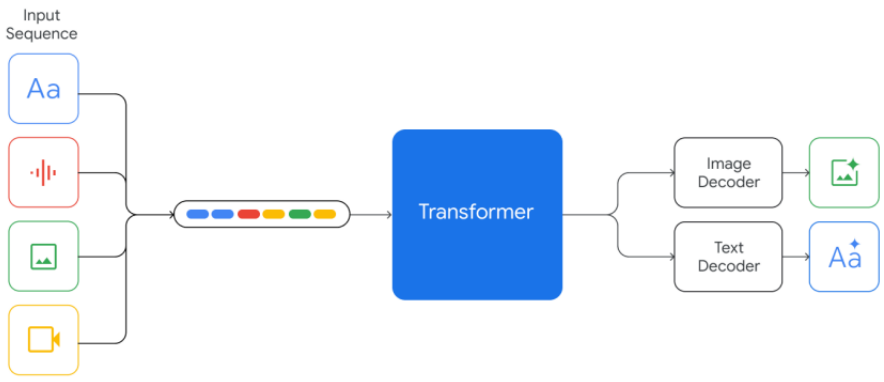
Όπως έχει ειπωθεί και σε προηγούμενη ενότητα, τα LLMs διαδραματίζουν έναν σημαντικό ρόλο. Είναι τα μοντέλα που αναλαμβάνουν το έργο να δώσουν την τελική απάντηση στον χρήστη. Δέχονται το ερώτημα του χρήστη και τις πληροφορίες που ανακτώνται από τις εξωτερικές πηγές και δημιουργούν μια απάντηση προς αυτόν. Με

αυτή την τεχνική τα LLMs παρέχουν απαντήσεις με μεγάλο ποσοστό ακρίβειας χωρίς να χρειαστεί επανεκπαίδευση του μοντέλου. Για τον σκοπό αυτό, χρησιμοποιούνται συνήθως ισχυρά decoder-only LLMs, μοντέλα (GPT, Gemini) τα οποία είναι κατάλληλα για τη σύνθεση συνεκτικών και τεκμηριωμένων απαντήσεων.

Τα μοντέλα GPT βασίζονται σε νευρωνικά δίκτυα που χρησιμοποιούνται για εργασίες επεξεργασίας φυσικής γλώσσας (NLP), όπως η γλωσσική μοντελοποίηση (language modelling), η ταξινόμηση κειμένου και η παραγωγή κειμένου. Η αρχιτεκτονική του μοντέλου GPT βασίζεται στην αρχιτεκτονική των μοντέλων transformer. Το μοντέλο transformer χρησιμοποιεί μηχανισμούς self-attention για να επεξεργάζεται αλληλουχίες εισόδου μεταβλητού μήκους, γεγονός που το καθιστά ιδιαίτερα κατάλληλο για εργασίες NLP. Το GPT απλοποιεί την αρχιτεκτονική καθώς δεν ακολουθεί την αρχιτεκτονική encoder-decoder αλλά αποτελεί ένα decoder only μοντέλο. Το GPT αξιοποιεί την αρχιτεκτονική transformer και προ-εκπαιδεύεται σε μεγάλες ποσότητες δεδομένων κειμένου, χρησιμοποιώντας τεχνικές μη επιβλεπόμενης μάθησης (unsupervised learning). Η διαδικασία της προ-εκπαίδευσης περιλαμβάνει την πρόβλεψη της επόμενης λέξης σε μια αλληλουχία, δεδομένων των προηγούμενων λέξεων, μια εργασία γνωστή ως γλωσσική μοντελοποίηση. Αυτή η διαδικασία προ-εκπαίδευσης επιτρέπει στο μοντέλο να μάθει αναπαραστάσεις της φυσικής γλώσσας που μπορούν στη συνέχεια να υποστούν fine-tuning για συγκεκριμένες δευτερεύουσες εργασίες. [25]

Στο πλαίσιο αρχιτεκτονικών RAG, τα μοντέλα GPT αξιοποιούνται ως ο μηχανισμός παραγωγής της τελικής απάντησης. Το γλωσσικό μοντέλο λαμβάνει ως είσοδο τόσο το ερώτημα του χρήστη όσο και τα ανακτημένα αποσπάσματα πληροφορίας από εξωτερικές πηγές γνώσης, τα οποία έχουν επιλεγεί μέσω σημασιολογικής ανάκτησης. Με βάση αυτό το εμπλουτισμένο συμφραζόμενο, το GPT συνθέτει μια συνεκτική και τεκμηριωμένη απάντηση, χωρίς να απαιτείται η ενσωμάτωση της εξωτερικής γνώσης στις παραμέτρους του μοντέλου. Η ικανότητα των GPT να αξιοποιούν αποτελεσματικά μεγάλο κειμενικό context καθιστά τα decoder-only μοντέλα ιδιαίτερα κατάλληλα για χρήση σε συστήματα RAG, ειδικά σε εφαρμογές που απαιτούν ευελιξία, γενίκευση και δυναμική αξιοποίηση επικαιροποιημένων πληροφοριών.

Στα ανωτέρω αξίζει να γίνει αναφορά στα μοντέλα Gemini τα οποία ακολουθούν την αρχιτεκτονική των transformer decoders ενισχυμένη με βελτιώσεις που επιτρέπουν σταθερή εκπαίδευση σε μεγάλη κλίμακα και βελτιστοποιημένη εξαγωγή συμπερασμάτων. Από το ξεκίνημά τους, εκπαιδεύτηκαν να διαχειρίζονται μεγάλο όγκο πληροφοριών, με το μήκος του πλαισίου τους (context length) να φτάνει τα 32.000 tokens. Τα μοντέλα Gemini είναι εκπαιδευμένα να δέχονται εισαγωγή κειμένου που εναλλάσσεται με μια μεγάλη ποικιλία ακουστικών και οπτικών δεδομένων, όπως φυσικές εικόνες, διαγράμματα, στιγμιότυπα οθόνης (screenshots), αρχεία PDF και βίντεο, ενώ μπορούν να παράγουν αποτελέσματα σε μορφή κειμένου και εικόνας. Αυτό τα καθιστά εγγενώς πολυτροπικά (multimodal), δίνοντάς τους τη δυνατότητα να κατανοούν, να επεξεργάζονται και να συνδυάζουν διαφορετικούς τύπους δεδομένων ταυτόχρονα (Εικόνα 6). [26] Στο πλαίσιο υλοποίησης συστημάτων RAG, το Gemini μπορεί να αξιοποιηθεί ως ο μηχανισμός παραγωγής της τελικής απάντησης, λαμβάνοντας ως είσοδο τόσο το ερώτημα του χρήστη όσο και τα ανακτημένα αποσπάσματα πληροφορίας από εξωτερικές πηγές γνώσης. Η δυνατότητα επεξεργασίας μεγάλου κειμενικού context, σε συνδυασμό με τη γενικού σκοπού σχεδίασή του, καθιστά το Gemini κατάλληλο για χρήση σε συστήματα RAG, ιδίως σε εφαρμογές που απαιτούν ευελιξία, δυναμική ενημέρωση της γνώσης και αξιοποίηση ετερογενών πηγών πληροφορίας.



Εικόνα 6: Τα μοντέλα Gemini μπορούν να υποστηρίζουν διαφορετικούς τύπους δεδομένων ταυτόχρονα [26]

Ένα ακόμη γλωσσικό μοντέλο που δεν θα πρέπει να παραληφθεί είναι το Llama (Large Language Model Meta AI). Τα μοντέλα Llama αποτελούν μια οικογένεια προηγμένων γλωσσικών μοντέλων τα οποία αναπτύχθηκαν από την Meta. Τα μοντέλα αυτά έχουν τη δυνατότητα να κατανοούν και να παράγουν κείμενο που προσομοιάζει τον ανθρώπινο λόγο, καθιστώντας τα εξαιρετικά πολύτιμα για εργασίες στην επεξεργασία φυσικής γλώσσας (NLP), στην τεχνητή νοημοσύνη συνομιλίας (conversational AI), στην παραγωγή κειμένου και σε άλλες γλωσσικές εφαρμογές που βασίζονται στην τεχνητή νοημοσύνη.

Προκειμένου να γίνουν κατανοητές οι δυνατότητες του Llama θα γίνει αναφορά στο μοντέλο Llama 3. Τα μοντέλα που ανήκουν στο Llama 3 (Εικόνα 7) αποτελούν ένα σύνολο foundation γλωσσικών μοντέλων τα οποία υποστηρίζουν την πολυγλωσσία, τον προγραμματισμό, την επίλυση προβλημάτων και είναι κατάλληλα σε πολλές γλωσσικές εφαρμογές. Το μεγαλύτερο μοντέλο αποτελεί έναν dense transformer με 405 δισεκατομμύρια παραμέτρους, ο οποίος επεξεργάζεται πληροφορίες σε ένα παράθυρο πλαισίου (context window) έως και 128 χιλιάδων tokens.

	Finetuned	Multilingual	Long context	Tool use	Release
Llama 3 8B	✗	✗[1]	✗	✗	April 2024
Llama 3 8B Instruct	✓	✗	✗	✗	April 2024
Llama 3 70B	✗	✗[1]	✗	✗	April 2024
Llama 3 70B Instruct	✓	✗	✗	✗	April 2024
Llama 3.1 8B	✗	✓	✓	✗	July 2024
Llama 3.1 8B Instruct	✓	✓	✓	✓	July 2024
Llama 3.1 70B	✗	✓	✓	✗	July 2024
Llama 3.1 70B Instruct	✓	✓	✓	✓	July 2024
Llama 3.1 405B	✗	✓	✓	✗	July 2024
Llama 3.1 405B Instruct	✓	✓	✓	✓	July 2024

Εικόνα 7: Μοντέλα που ανήκουν στην οικογένεια των μοντέλων Llama 3. Περιλαμβάνονται τα μοντέλα που ανήκουν μέχρι την έκδοση 3.1 [27]

Με βάση τη μελέτη «The Llama 3 Herd of Models», η ανάπτυξη των μοντέλων Llama 3 περιλαμβάνει δύο κύρια στάδια. Το στάδιο της προ-εκπαίδευσης (Language model pre-training) και το στάδιο της μετά-εκπαίδευσης (Language model post-training). Κατά το στάδιο της προ-εκπαίδευσης πραγματοποιείται η μετατροπή ενός μεγάλου, πολυγλωσσικού σώματος κειμένων σε tokens και η εκπαίδευση ενός μεγάλου γλωσσικού μοντέλου (LLM) στα δεδομένα αυτά, με στόχο την πρόβλεψη του

επόμενου συμβόλου (next-token prediction). Στο συγκεκριμένο στάδιο, το μοντέλο μαθαίνει τη δομή της γλώσσας και αποκτά τεράστιο όγκο γνώσεων για τον κόσμο από τα κείμενα που επεξεργάζεται. Για την αποτελεσματική επίτευξη αυτού του στόχου, η προ-εκπαίδευση πραγματοποιήθηκε σε μαζική κλίμακα. Εκπαιδεύτηκε ένα μοντέλο 405 δισεκατομμυρίων παραμέτρων σε 15,6 τρισεκατομμύρια tokens, χρησιμοποιώντας παράθυρο πλαισίου (context window) 8 χιλιάδων tokens. Μετά το τυπικό αυτό στάδιο, ακολουθήθηκε μία φάση συνεχούς προ-εκπαίδευσης (continued pre-training), η οποία αύξησε το υποστηριζόμενο παράθυρο πλαισίου στα 128 χιλιάδες tokens.

Στο στάδιο της μετα-εκπαίδευσης (θεωρείται κάθε στάδιο εκπαίδευσης που συμβαίνει εκτός του σταδίου της προ-εκπαίδευσης) γίνεται ευθυγράμμιση του μοντέλου με ανθρώπινη ανατροφοδότηση σε διάφορους γύρους εκπαίδευσης καθένας από τους οποίους περιλαμβάνει τεχνικές supervised finetuning (SFT) και Direct Preference Optimization (DPO). Αυτό συμβαίνει καθώς ενώ το προ-εκπαιδευμένο μοντέλο διαθέτει βαθιά κατανόηση της γλώσσας, δεν είναι ακόμη σε θέση να ακολουθεί οδηγίες. Τα μοντέλα που δημιουργήθηκαν με τη συγκεκριμένη διαδικασία διακρίνονταν από αυξημένες δυνατότητες. Είχαν τη δυνατότητα να απαντήσουν σε μία ερώτηση σε τουλάχιστον 8 διαφορετικές γλώσσες, να συνθέσουν κώδικα υψηλής ποιότητας και να επιλύσουν πολύπλοκα προβλήματα.[27]

Σε ένα τυπικό RAG σύστημα, τα Llama μπορούν να αναλάβουν τον ρόλο του generator, δηλαδή να συνθέτουν την τελική απάντηση αξιοποιώντας τα έγγραφα που ανακτήθηκαν από την εξωτερική πηγή γνώσης. Η αύξηση του context window (έως 128 χιλιάδες tokens που παρατηρήθηκε σε νεότερες εκδόσεις) διευκολύνει την ένταξη περισσότερων αποσπασμάτων στο prompt, βελτιώνοντας τη δυνατότητα τεκμηριωμένων απαντήσεων όταν ο χρήστης απαιτεί παραπομπές ή εκτενές αποδεικτικό υλικό.

2.1.6 Παραδείγματα χρήσης των συστημάτων RAG

Τα τελευταία χρόνια παρατηρείται μία ραγδαία αύξηση στις επενδύσεις των εταιρειών στην τεχνητή νοημοσύνη. Πολλοί οργανισμοί έχουν εξετάσει την υλοποίηση συστημάτων RAG για να αξιοποιήσουν αποτελεσματικά τα δεδομένα που

ανήκουν αποκλειστικά στον οργανισμό και δεν είναι διαθέσιμα στο κοινό (proprietary data).

Ένα παράδειγμα αποτελεί η εταιρεία Bayer η οποία αξιοποίησε την αρχιτεκτονική RAG για τη διαχείριση και ανάκτηση ιδιόκτητων αγροτικών δεδομένων, με στόχο να παρέχει στους αγρότες ακριβείς και έγκαιρες πληροφορίες. Ενσωματώνοντας τα εκτεταμένα σύνολα δεδομένων στο RAG σύστημά της, η Bayer βελτίωσε την προσβασιμότητα και τη χρηστικότητα κρίσιμων αγροτικών πληροφοριών. Αυτή η ενσωμάτωση διευκόλυνε καλύτερες διαδικασίες λήψης αποφάσεων για τους αγρότες, οδηγώντας σε βελτιωμένη διαχείριση καλλιεργειών και αυξημένη παραγωγικότητα.

Η Cohere ανέπτυξε περαιτέρω την τεχνολογία RAG ώστε τα συστήματα τεχνητής νοημοσύνης να αναφέρουν τις πηγές τους όταν δημιουργούν μία απάντηση στον χρήστη, διευκολύνοντας την ανθρώπινη επαλήθευση των παραγόμενων πληροφοριών. Με την ενσωμάτωση εξωτερικών κειμένων, όπως εταιρικών εγγράφων ή ειδησεογραφικών ιστοσελίδων, το σύστημα RAG της Cohere μείωσε τις παραισθήσεις (hallucinations) που μπορεί να παρουσιάσει ένα γλωσσικό μοντέλο και παρείχε πρόσβαση σε επίκαιρη πληροφορία. Αυτή η εξέλιξη ανέδειξε το δυναμικό του RAG στη βελτίωση της αξιοπιστίας και της διαφάνειας του περιεχομένου που παράγεται από την τεχνητή νοημοσύνη.

Ένα ακόμη παράδειγμα είναι η NVIDIA η οποία ενσωμάτωσε την αρχιτεκτονική RAG, ώστε να συνδέει μεγάλα γλωσσικά μοντέλα με εξειδικευμένες πληροφορίες, όπως ιδιόκτητα δεδομένα πελατών και έγκυρη επιστημονική έρευνα. Αυτή η ενσωμάτωση επέτρεψε την παροχή πιο ακριβών και σχετικών απαντήσεων στα ερωτήματα των χρηστών, ενισχύοντας την παραγωγικότητα και μειώνοντας τον κίνδυνο εμφάνισης παραισθήσεων. Η υλοποίηση της NVIDIA ανέδειξε τη χρησιμότητα του RAG σε διάφορους κλάδους, όπως η εξυπηρέτηση πελατών και η διαχείριση δεδομένων.

Η IBM χρησιμοποίησε το RAG για να εμπλουτίσει τα μοντέλα τεχνητής νοημοσύνης με εξειδικευμένες πληροφορίες, όπως ιδιόκτητα δεδομένα πελατών και έγκυρα ερευνητικά έγγραφα. Αυτή η προσέγγιση επέτρεψε στα συστήματα AI της εταιρείας να ενσωματώνουν επίκαιρη πληροφορία στις παραγόμενες απαντήσεις,

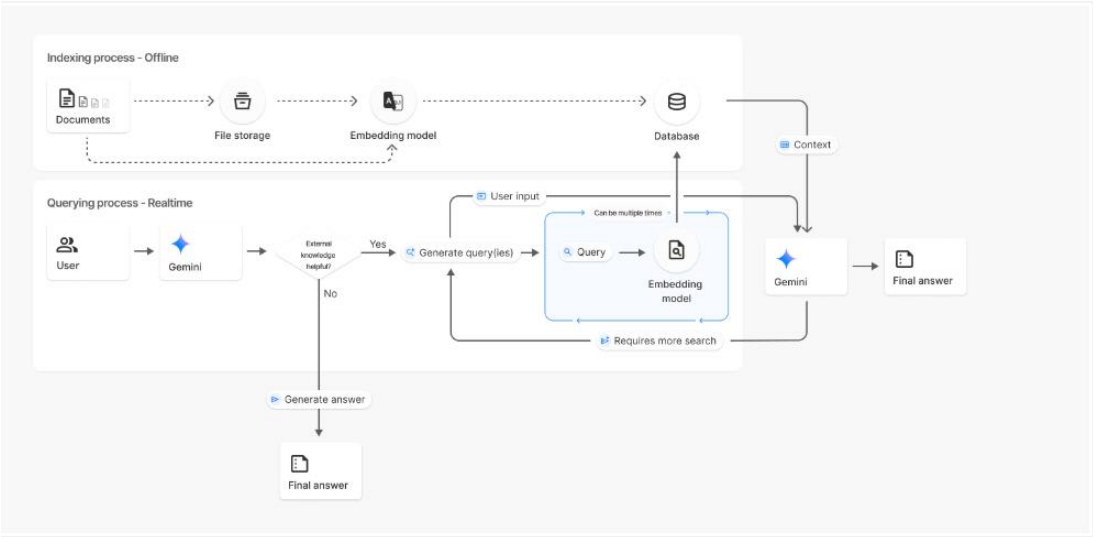
βελτιώνοντας την ακρίβεια και τη συνάφεια σε εφαρμογές συγκεκριμένων τομέων. Η αξιοποίηση του RAG από την IBM ανέδειξε την αποτελεσματικότητά του στην ενίσχυση των δυνατοτήτων της τεχνητής νοημοσύνης σε διάφορους κλάδους.[7]

Αξίζει να αναφερθεί και το case study της Samsung SDS στο οποίο παρουσιάζει την ανάπτυξη του SKE-GPT, ενός εργαλείου υποστήριξης troubleshooting για Kubernetes clusters στο πλαίσιο του SCP Kubernetes Engine. Το SKE-GPT αποτελείται από δύο κύρια components. Το διαγνωστικό τμήμα (diagnostic area) το οποίο ελέγχει την κατάσταση των workloads με βάση προκαθορισμένους κανόνες και, όταν απαιτείται, επεκτείνει τον έλεγχο σε συναφείς υπηρεσίες υποδομής (load balancing, DNS, storage), και το τμήμα ανάλυσης (analysis area) που συνθέτει προτάσεις ενεργειών για την αντιμετώπιση του περιστατικού. Για την ενίσχυση της απάντησης με τεκμηρίωση πέρα από την εγγενής γνώση του μοντέλου, υιοθετείται η αρχιτεκτονική RAG. Τεχνικά έγγραφα και οδηγοί τεκμηρίωσης προ-επεξεργάζονται (πραγματοποιείται chunking με overlap), μετατρέπονται σε embeddings και αποθηκεύονται σε vector database, ώστε κατά την εκτέλεση να ανακτώνται τα πιο σχετικά αποσπάσματα και να ενσωματώνονται στο prompt του LLM, με στόχο πιο στοχευμένη καθοδήγηση επίλυσης. [28]

Παραδείγματα εφαρμογής της αρχιτεκτονικής RAG συναντάμε και στον τομέα της υγείας. Η εφαρμογή AlzheimerRAG, μια εφαρμογή RAG για κλινικές χρήσεις, με κύρια εστίαση σε μελέτες περίπτωσης της νόσου Alzheimer από άρθρα του PubMed. Η εφαρμογή αυτή ενσωματώνει τεχνικές για την ενοποίηση της επεξεργασίας κειμενικών και οπτικών δεδομένων, επιτρέποντας την αποτελεσματική ευρετηρίαση και πρόσβαση σε τεράστιους όγκους βιοϊατρικής βιβλιογραφίας. [29]

Από όσα παραδείγματα χρήσης των συστημάτων RAG αναφέρθηκαν, δεν θα πρέπει να παραλειφθεί η αναφορά του εργαλείου File Search Tool. Το Gemini API αξιοποιεί την αρχιτεκτονική RAG μέσω του εργαλείου File Search. Το File Search εισάγει, τεμαχίζει και δημιουργεί ευρετήριο για τα δεδομένα του χρήστη, επιτρέποντας ταχεία ανάκτηση σχετικών πληροφοριών με βάση το εκάστοτε prompt. Αυτές οι πληροφορίες χρησιμοποιούνται στη συνέχεια ως πλαίσιο για το μοντέλο, επιτρέποντάς του να παρέχει πιο ακριβείς και σχετικές απαντήσεις. Το File Search χρησιμοποιεί σημασιολογική αναζήτηση (semantic search) προκειμένου να εντοπίσει

πληροφορία σχετική με το prompt του χρήστη. Όταν ο χρήστης ανεβάσει ένα αρχείο, αυτό μετατρέπεται σε embeddings, τα οποία αποτυπώνουν το σημασιολογικό περιεχόμενο του κειμένου. Τα embeddings αποθηκεύονται σε μια ειδική βάση δεδομένων File Search. Όταν ο χρήστης κάνει ένα query, μετατρέπεται και αυτό σε embedding. Στη συνέχεια, το σύστημα εκτελεί File Search για να βρει τα πιο παρόμοια και σχετικά τμήματα εγγράφων από το File Search store. [30] Η αρχιτεκτονική του File Search απεικονίζεται στην Εικόνα 8.



Εικόνα 8: Η αρχιτεκτονική του File Search [30]

2.2 Συστήματα RAG και δημόσιες υπηρεσίες

2.2.1 Ανάγκη εφαρμογής των συστημάτων RAG στον δημόσιο τομέα

Στον δημόσιο τομέα, η καθημερινή εξυπηρέτηση βασίζεται σε πληροφορία που είναι διασκορπισμένη σε νόμους, εγκυκλίους, οδηγούς διαδικασιών και εσωτερική τεκμηρίωση, ενώ συχνά αλλάζει και απαιτείται να εφαρμόζεται ομοιόμορφα από διαφορετικές υπηρεσίες. Αυτή η πραγματικότητα αυξάνει τον χρόνο αναζήτησης, δημιουργεί ασυνέπειες στις απαντήσεις και δυσκολεύει τη λογοδοσία όταν ο πολίτης ζητά να ενημερωθεί με ποιο κανονισμό προκύπτει η απάντηση την οποία λαμβάνει.

Ο OECD [31] επισημαίνει ότι η τεχνητή νοημοσύνη μπορεί να βελτιώσει τις παρεχόμενες προς τον πολίτη υπηρεσίες και να υποστηρίξει το έργο των δημοσίων υπαλλήλων. Τα οφέλη όμως εξαρτώνται από το πώς αντιμετωπίζονται ζητήματα κινδύνου και διαφάνειας καθώς η έλλειψη λογοδοσίας ή η διάδοση λαθών από την εφαρμογή τεχνητής μπορεί να πλήξει την εμπιστοσύνη των πολιτών προς το κράτος. Για παράδειγμα εάν μία εφαρμογή τεχνητής νοημοσύνης όπως είναι ένας προσωπικός βοηθός παράγει αποτελέσματα που δεν είναι ορθά, αυτό έχει ως άμεση συνέπεια την έλλειψη εμπιστοσύνη των πολιτών προς τις δημόσιες υπηρεσίες.

Σε αυτό το πλαίσιο, τα συστήματα RAG απαντούν σε μια πολύ πρακτική ανάγκη: να μπορεί ένα LLM να δίνει απαντήσεις δεμένες με τις επίσημες πηγές του οργανισμού, και όχι μόνο να δημιουργεί περιεχόμενο βασιζόμενο στην γνώση που έχει αποκτήσει από την εκπαίδευσή του. Το RAG ακολουθεί την αρχιτεκτονική πρώτα να γίνεται ανάκτηση των πιο σχετικών αποσπασμάτων από ευρετήρια, βάσεις εγγράφων που σχετίζονται με το ερώτημα του πολίτη και έπειτα το LLM αποκρίνεται με βάση τα δεδομένα που έχουν ανακτηθεί και το ερώτημα του χρήστη. Η απάντηση η οποία παράγεται είναι τεκμηριωμένη, είναι ευκολότερο να δοθούν παραπομπές ώστε να διασταυρωθεί το περιεχόμενό της. Το γλωσσικό μοντέλο έχει λιγότερες πιθανότητες να παρουσιάσει hallucinations. Η λογική αυτή συνδέεται και με τη βασική ερευνητική ιδέα του RAG ως συνδυασμού γνώσης που είναι ενσωματωμένη στα βάρη του μοντέλου με τη γνώση από ανακτημένα τεκμήρια (που ενημερώνονται χωρίς να εκπαιδευτεί από την αρχή το μοντέλο). Τέλος, το NIST (NIST AI RMF 1.0 – Generative AI Profile (NIST AI 600-1)) [32] δίνει έμφαση σε πρακτικές που ενισχύουν την διαφάνεια, την τεκμηρίωση και τη διαχείριση κινδύνων όταν χρησιμοποιούνται LLMs σε οργανισμούς υψηλής ευθύνης. Αυτές οι κατευθύνσεις ευθυγραμμίζονται με αρχιτεκτονικές όπως το RAG, οι οποίες μπορούν να υποστηρίξουν απαντήσεις με παραπομπές και να περιορίσουν τον κίνδυνο εμφάνισης hallucinations.

2.2.2 Χρήση συστημάτων RAG στον δημόσιο τομέα

Προκειμένου ο δημόσιος τομέας να γίνει πιο ανταγωνιστικός και να ανταποκριθεί στις προκλήσεις για εξυπηρέτηση των πολιτών, διαφάνεια, και έγκυρη πληροφόρηση

αξιοποιεί τα συστήματα RAG. Σε πολλές χώρες έχουν δημιουργηθεί εφαρμογές που ακολουθούν τη συγκεκριμένη αρχιτεκτονική. Ένα παράδειγμα είναι το Ηνωμένο Βασίλειο που παρέχει την εφαρμογή GOV.UK Chat, η οποία αποτελεί ένα σύστημα συνομιλίας βασισμένο στην αρχιτεκτονική RAG, το οποίο βοηθά τους χρήστες του GOV.UK να περιηγούνται και να αξιοποιούν το περιεχόμενο του GOV.UK με τρόπο εύκολο και κατανοητό. Το GOV.UK Chat έχει σχεδιαστεί ώστε να χρησιμοποιεί αποκλειστικά περιεχόμενο του GOV.UK για οποιαδήποτε απάντηση παράγει και ζητά ρητά από το LLM να αγνοεί οποιαδήποτε προηγούμενη γνώση είχε κατά την εκπαίδευσή του. Το GOV.UK χρησιμοποιεί ένα LLM, το οποίο παρέχεται από την AWS, για να επεξεργάζεται ερωτήσεις χρηστών σχετικά με το περιεχόμενό του και τα ζητήματα που αντιμετωπίζουν. [33]

Ένα ακόμη παράδειγμα είναι η πλατφόρμα AIBots της Government Technology Agency of Singapore (GovTech Singapore) που παρουσιάζεται ως εργαλείο που επιτρέπει σε δημόσιους υπαλλήλους στη Singapore να δημιουργούν, να παραμετροποιούν και να αναπτύσσουν ασφαλή chatbots πολύ γρήγορα, αξιοποιώντας αρχιτεκτονική RAG ώστε οι απαντήσεις να ενισχύονται με εσωτερικές βάσεις γνώσης (internal agency knowledge bases) και να γεφυρώνεται η δημόσια πληροφορία με την εσωτερική τεκμηρίωση κάθε φορέα. Παρέχει δυνατότητες όπως ανέβασμα εσωτερικών εγγράφων (ακόμα και με διαβαθμίσεις ως «Restricted / Sensitive (Normal)»), ρύθμιση system prompts και διαμοιρασμό των bots σε ομάδες, ενώ αναφέρεται ότι η πρόσβαση είναι διαθέσιμη μόνο εντός κυβερνητικού δικτύου. [34]

Από τα ανωτέρω δεν θα πρέπει να παραληφθεί και η υλοποίηση RAG του Pacific Northwest National Laboratory. Ο συγκεκριμένος φορέας προχώρησε στην ανάπτυξη συστημάτων RAG για κυβερνοάμυνα με την ονομασία CyRAG και GraphCyRAG. Το CyRAG στοχεύει να βοηθήσει αναλυτές κυβερνοασφάλειας να αντλούν γρήγορα αξιόπιστη πληροφορία από μεγάλης κλίμακας, δομημένες πηγές. Έχει σχεδιαστεί για να χειρίζεται δομημένα δεδομένα από σχεσιακές βάσεις δεδομένων, εστιάζοντας σε οντότητες CVE (Common Vulnerabilities and Exposures) και CWE (Common Weakness Enumeration), ώστε να παράγει πιο ακριβείς και πλούσιες σε περιεχόμενο απαντήσεις. Όταν ο χρήστης υποβάλει ένα ερώτημα (για παράδειγμα ένα

συγκεκριμένο CVE), το σύστημα ανακτά σχετικά στοιχεία από τη βάση (όπως συσχετισμένα CWE) και τα χρησιμοποιεί ως τεκμήριο για να παραχθεί μια πιο ολοκληρωμένη, περιγραφική απάντηση από το LLM, κάτι ιδιαίτερα χρήσιμο όταν ο όγκος και η πολυπλοκότητα των δεδομένων κάνουν τις κλασικές αναζητήσεις ανεπαρκείς.

Το GraphCyRAG αξιοποιεί γράφους γνώσης Neo4j για την ανάκτηση διασυνδεδεμένων πληροφοριών από τα σύνολα δεδομένων CVE, CWE, CAPEC (Common Attack Pattern Enumeration and Classification) και ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge). Χρησιμοποιώντας το framework του Neo4j, το GraphCyRAG επιτρέπει βαθύτερη διερεύνηση των σχέσεων μεταξύ ευπαθειών και μοτίβων επίθεσης, προσφέροντας στους αναλυτές κυβερνοασφάλειας πιο ολοκληρωμένες γνώσεις σχετικά με πιθανούς διαύλους επίθεσης και στρατηγικές μετριασμού ενδεχόμενων επιθέσεων.[35]

Ιδιαίτερη μνεία αξίζει η υλοποίηση RAG του δήμου Borough of Barking and Dagenham στο Λονδίνο. Το AI Search and Webchat παρουσιάζεται ως ένα εργαλείο που στοχεύει να απλοποιήσει την πρόσβαση των πολιτών σε πληροφορίες και υπηρεσίες του δήμου, συνδυάζοντας την «έξυπνη» αναζήτηση, την συνομιλία με chat σε ένα ενιαίο σημείο επαφής. Το πρόβλημα επικεντρωνόταν στο ότι οι παραδοσιακές λειτουργίες αναζήτησης εμφάνιζαν υπερβολικά μεγάλο αριθμό αποτελεσμάτων, γεγονός που δυσκολεύει τους χρήστες να εντοπίσουν τις πληροφορίες που χρειάζονται. Αυτό οδηγεί σε απογοήτευση, με αποτέλεσμα πολλοί κάτοικοι να εγκαταλείπουν την επιλογή αυτοεξυπηρέτησης και να προτιμούν είτε να καλέσουν το κέντρο εξυπηρέτησης είτε να εγκαταλείψουν εντελώς την προσπάθεια. Με τη δημιουργία του συγκεκριμένου εργαλείου η βασική ιδέα είναι ότι ο χρήστης δεν χρειάζεται να γνωρίζει τη δομή του ιστότοπου ή τους ορθούς όρους αναζήτησης. Μπορεί να διατυπώσει ένα ερώτημα σε καθημερινή γλώσσα και να λάβει μια συνοπτική απάντηση μαζί με σχετικούς συνδέσμους προς τις αντίστοιχες σελίδες και υπηρεσίες, ώστε να ολοκληρώσει τη διαδικασία. Σε επίπεδο τεχνικής προσέγγισης, το εργαλείο αποτελεί ένα LLM που έχει δημιουργηθεί με το Azure AI Search. Εμπεριέχει το βασικό περιεχόμενο της ιστοσελίδας του δήμου. Χρησιμοποιεί την αρχιτεκτονική RAG για να απαντά σε ερωτήματα σε απλή γλώσσα με συνοπτικές περιλήψεις και άμεσους

συνδέσμους προς τις σχετικές υπηρεσίες. Οι χρήστες μπορούν να αλληλοεπιδρούν μέσω διεπαφών αναζήτησης ή συνομιλίας. [36]

Μία ακόμη σημαντική υλοποίηση είναι το Genesys ChatBot του Cafcass (Children and Family Court Advisory and Support Service). Αποτελεί έναν πολύ σημαντικό δημόσιο οργανισμό στο Ηνωμένο Βασίλειο, και διαχειρίζεται δικαστικές υποθέσεις που αφορούν παιδιά. Στη συγκεκριμένη εφαρμογή οι χρήστες εισάγουν μία ερώτηση στο chat και εκείνο αποκρίνεται στο ερώτημά τους με την κατάλληλη απάντηση. Απευθύνεται σε χρήστες που θέλουν να λάβουν απευθείας απάντηση στα ερωτήματά τους μέσω της εφαρμογής. Σε περίπτωση που δεν είναι εφικτή η απάντηση, η συζήτηση ανακατευθύνεται σε agent του τηλεφωνικού κέντρου. Το Genesys ChatBot αξιοποιεί μια αρχιτεκτονική RAG, η οποία έχει αναπτυχθεί με ασφάλεια σε περιβάλλον Amazon Web Services (AWS). Το σύστημα έχει σχεδιαστεί για να διαχειρίζεται ερωτήματα χρηστών που σχετίζονται με πληροφορίες του Cafcass. Η εισαγωγή των δεδομένων που βρίσκονται στον δημόσιο ιστότοπο του Cafcass στο περιβάλλον AWS πραγματοποιείται μέσω του εργαλείου Genesys Architect. Τα δεδομένα ενσωματώνονται και μετατρέπονται σε διανυσματικές αναπαραστάσεις (vector embeddings) χρησιμοποιώντας το Amazon Lex. Τα ανακτημένα έγγραφα επεξεργάζονται μέσω των Digital Engine Bot Flows της Genesys, τα οποία είναι ενσωματωμένα μέσω του AWS Lex. Το Genesys Digital Engine Bot συνοψίζει τις πληροφορίες που ανακτήθηκαν και δημιουργεί μια φιλική προς τον χρήστη απάντηση, η οποία στη συνέχεια παρουσιάζεται στον χρήστη. [37]

Στο σημείο αυτό αξίζουν να αναφερθούν παραδείγματα εφαρμογών RAG τα οποία έχουν ως περιεχόμενο την νομοθεσία και προσπαθούν να συμβάλουν στην επίλυση διάφορών νομικών ζητημάτων που αντιμετωπίζουν οι χρήστες. Οι εφαρμογές αυτές δεν αποτελούν όλες παραγωγικά συστήματα μίας δημόσιας υπηρεσίας. Το περιεχόμενό τους όμως έχει άμεση συσχέτιση με θέματα του δημόσιου τομέα.

Αρχικά αξίζει να γίνει αναφορά στην εργασία «Chat-Based Access to Legislation», η οποία αναφέρει ένα ερευνητικό project στο οποίο παρουσιάζεται ένα chatbot που στόχο έχει να βελτιώσει την πρόσβαση του κοινού στο νομοθετικό περιεχόμενο της Νέας Ζηλανδίας. Αναπτύχθηκε ένα πειραματικό chatbot χρησιμοποιώντας την αρχιτεκτονική RAG για την ανάκτηση νομοθετικού περιεχομένου και τη δημιουργία

απαντήσεων που βασίζονται στο πρωτογενές κείμενο, αποφεύγοντας την νομοθετική ερμηνεία. Το έργο αναπτύχθηκε για τις ανάγκες του Parliamentary Counsel Office (PCO). Το μοντέλο λάμβανε το ερώτημα του χρήστη, επέλεγε τα κατάλληλα μέρη της νομοθεσίας σύμφωνα με το ερώτημα και όλα τα δεδομένα εισέρχονταν στο LLM προκειμένου να δημιουργήσει μία εμπεριστατωμένη απάντηση. [38]

Ένα ακόμη project που πρέπει επισημανθεί είναι η δημιουργία ενός Legal Assistant για governance εφαρμογές πάνω στο δίκαιο της ευρωπαϊκής ένωσης. Η εφαρμογή ενσωματώνει τα μοντέλα GPT-3.5 και GPT-4 ως LLMs, ένα σαφώς καθορισμένο νομικό σώμα που περιλαμβάνει νομοθεσία της Ευρωπαϊκής Ένωσης (ΕΕ) και νομολογία σχετική με τον Γενικό Κανονισμό Προστασίας Δεδομένων (GDPR), καθώς και μια σειρά από ενδεικτικά νομικά ερωτήματα ποικίλης πολυπλοκότητας. Επιπλέον, χρησιμοποιήθηκε η αρχιτεκτονική RAG καθώς και μεθοδολογίες agents, ώστε να ενσωματωθούν απρόσκοπτα οι δυνατότητες των LLMs στο προσαρμοσμένο σύνολο δεδομένων. [39]

Αξιοσημείωτη είναι η εφαρμογή LawPal, η οποία αποτελεί ένα RAG σύστημα που βοηθά στη βελτιωμένη πρόσβαση σε νομικές πληροφορίες στην Ινδία. Πιο συγκεκριμένα, το LawPal είναι ένα chatbot για αποδοτική και ακριβή ανάκτηση νομικών πληροφοριών. Έχει εκπαιδευτεί με ένα εκτενές σύνολο δεδομένων που περιλαμβάνει νομικά βιβλία, επίσημη τεκμηρίωση και το Σύνταγμα της Ινδίας. Επιπλέον, έχει σχεδιαστεί με τεχνικές prompt engineering ώστε να χειρίζεται στρεβλά ή αμφίσημα νομικά ερωτήματα, μειώνοντας τις πιθανότητες λανθασμένων ερμηνειών. Επροσθέτως, για να παραμείνει το σύνολο δεδομένων ενημερωμένο, αυτοματοποιημένα εργαλεία web scraping εξάγουν τις πιο πρόσφατες τροποποιήσεις και δικαστικές αποφάσεις. Σχετικά με την αρχιτεκτονική του, τα δεδομένα υποβάλλονται σε αυστηρή προεπεξεργασία, η οποία περιλαμβάνει καθαρισμό, ψηφιοποίηση μέσω OCR και κανονικοποίηση κειμένου. Στη συνέχεια, τμηματοποιούνται σε μικρότερα κομμάτια χρησιμοποιώντας το RecursiveCharacterTextSplitter του LangChain, ώστε να διατηρείται η λογική συνοχή. Κάθε τμήμα του κειμένου μετατρέπεται σε embeddings αξιοποιώντας το μοντέλο DeepSeek, τα οποία αποθηκεύονται στην vector database FAISS. Όταν ο χρήστης πραγματοποιήσει ένα ερώτημα, οι απαραίτητες νομικές πληροφορίες ανακτώνται

από το FAISS και μαζί με το ερώτημα του χρήστη παρέχονται στο DeepSeek για να δημιουργήσει την απάντηση στον χρήστη. [40]

Από τα ανωτέρω προκύπτει ότι αρχιτεκτονική RAG έχει εφαρμογή σε πολλούς τομείς του δημοσίου τομέα που σχετίζονται με την εξυπηρέτηση του πολίτη και τη βελτίωση της καθημερινότητας του. Επιπλέον, η συγκεκριμένη αρχιτεκτονική αξιοποιείται και στο πλαίσιο της νομοθεσίας με στόχο την εύκολή πρόσβαση στους νόμους και τη δημιουργία έγκυρων απαντήσεων σε νομικά ζητήματα.

3. Ανάλυση και Σχεδίαση

3.1 Ανάλυση Απαιτήσεων

Προκειμένου να παρουσιαστούν οι δυνατότητες των συστημάτων RAG, όπως αναφέρθηκαν στη βιβλιογραφική ανασκόπηση, επιλέχθηκε η υλοποίηση τριών εφαρμογών οι οποίες δείχνουν την εφαρμογή της αρχιτεκτονικής RAG πάνω σε νομικά κείμενα και δεδομένα καταγγελιών πολιτών.

Πιο συγκεκριμένα, το πλαίσιο στο οποίο βασίζονται οι εφαρμογές είναι η νομοθεσία των τυχερών παιγνίων στην Ελλάδα. Αρμόδιος ρυθμιστής για την διεξαγωγή τυχερών παιγνίων στην Ελλάδα είναι η Επιτροπή Εποπτείας και Ελέγχου Παιγνίων (Ε.Ε.ΕΠ.). Η Ε.Ε.Ε.Π. είναι ανεξάρτητη αρχή επιφορτισμένη με τον έλεγχο, τη ρύθμιση και την εποπτεία των παιγνίων τα οποία λαμβάνουν χώρα στην Ελλάδα. Αναλαμβάνει δράσεις για τη διασφάλιση του υπεύθυνου παιγνιδιού καθώς και για θέματα σχετικά με την προστασία των ανηλίκων και των ευαίσθητα κοινωνικών ομάδων του πληθυσμού από τον εθισμό στα παίγνια.

Στο ανωτέρω πλαίσιο σχεδιάστηκαν τρεις εφαρμογές. Η πρώτη στοχεύει στην ανεύρεση των νόμων που ενδεχομένως να παραβιάζει η στοιχηματική εταιρεία για την οποία κάνει καταγγελία ένας πολίτης. Ουσιαστικά πρόκειται για ένα εργαλείο ελέγχου με βάση το οποίο ο ελεγκτής μπορεί να εισάγει μία καταγγελία και το σύστημα να του εμφανίσει ποιους νόμους ενδέχεται να παραβιάζει η στοιχηματική εταιρεία σε σχέση με τα όσα αναφέρονται στην καταγγελία. Αποτελεί εργαλείο που

βοηθά στην επίσπευση του ελέγχου των καταγγελιών, καθώς βοηθά τον ελεγκτή να επικεντρώνεται σε συγκεκριμένα άρθρα της νομοθεσίας χωρίς να αφιερώνει αρκετό χρόνο στον εντοπισμό των άρθρων που απαιτούνται αναζητώντας τα σε ολόκληρη την νομοθεσία. Με αυτό τον τρόπο οι διαδικασίες επισπεύδονται και αυξάνεται η εμπιστοσύνη του πολίτη προς τις αρμόδιες ελεγκτικές αρχές. Το μοντέλο στη συγκεκριμένη εφαρμογή ακολουθεί την αρχιτεκτονική RAG με τον ακόλουθο τρόπο: ο χρήστης εισάγει την καταγγελία. Η καταγγελία μετατρέπεται σε embeddings με βάση το embedding model. Πραγματοποιείται semantic search μεταξύ της δοθείσας καταγγελίας και των άρθρων της νομοθεσίας, τα οποία και αυτά έχουν μετατραπεί σε embeddings και είναι αποθηκευμένα στην vector database. Στη συνέχεια ανακτώνται τα πιο σχετικά άρθρα σύμφωνα με την καταγγελία. Το περιεχόμενο της καταγγελίας και τα άρθρα της νομοθεσίας δίνονται στο LLM προκειμένου να δοθεί μία τεκμηριωμένη απάντηση σχετικά με το ποια άρθρα παραβιάζονται σύμφωνα με την καταγγελία του πολίτη, παραθέτοντας και τις αντίστοιχες πηγές.

Η δεύτερη εφαρμογή έχει ως στόχο τον εντοπισμό παρόμοιων καταγγελιών. Πιο αναλυτικά, ο χρήστης μπορεί να εισάγει μία καταγγελία και η εφαρμογή επιστρέφει καταγγελίες που μοιάζουν ως προς το περιεχόμενο σε σχέση με την υποβληθείσα καταγγελία. Η συγκεκριμένη εφαρμογή είναι ιδιαίτερα χρήσιμη καθώς βοηθά τον ελεγκτή να αναζητά παρόμοιες υποθέσεις. Επιπλέον, δίνεται στο LLM η καταγγελία του χρήστη και οι παρόμοιες καταγγελίες και ζητείται να παρουσιάσει μέχρι πέντε ομοιότητες και μέχρι τρεις διαφορές. Με αυτό τον τρόπο δείχνει τις διαφορές και τις ομοιότητες που υπάρχουν με τις υπόλοιπες καταγγελίες, ενέργεια που βοηθά στον έλεγχο παρόμοιων υποθέσεων. Στη συγκεκριμένη εφαρμογή υλοποιείται η αρχιτεκτονική RAG. Η αρχική καταγγελία μετατρέπεται σε embeddings, και συγκρίνεται με τις καταγγελίες που είναι στην vector database, η οποία περιλαμβάνει το σύνολο των καταγγελιών που έχουν μετατραπεί και αυτές σε embeddings. Στη συνέχεια η αρχική καταγγελία και οι όμοιες της εισάγονται στο LLM προκειμένου να εντοπιστούν ομοιότητες και διαφορές.

Η τρίτη εφαρμογή αποτελεί ένα chatbot στο οποίο ο χρήστης μπορεί να θέτει ερωτήματα που αφορούν τον κανονισμό για την χορήγηση αδειών καταλληλόλητας. Ο χρήστης υποβάλει το ερώτημά του και το σύστημα αποκρίνεται με βάση την

κείμενη νομοθεσία και εμφανίζει τις πηγές στις οποίες βασίστηκε. Η συγκεκριμένη εφαρμογή έχει την χρησιμότητα ενός προσωπικού βοηθού. Ο χρήστης δεν χρειάζεται να μελετά εντατικά τα τεράστια νομοθετικά κείμενα καθώς μπορεί να λάβει εμπεριστατωμένη απάντηση μέσα από το chatbot. Και εδώ εφαρμόζεται η μεθοδολογία της αρχιτεκτονικής RAG. Η νομοθεσία έχει μετατραπεί σε embeddings και έχει αποθηκευτεί στην vector database. Όταν ο χρήστης εισάγει το ερώτημά του, γίνεται μετατροπή σε embeddings και με τη βοήθεια των metadata ανακτώνται από την βάση οι πιο σχετικές παράγραφοι νόμων. Όλα τα δεδομένα (ερώτημα χρήστη και νόμοι) προωθούνται στο LLM το οποίο απαντά το ερώτημα του χρήστη παραθέτοντας πηγές.

3.2 Προετοιμασία Δεδομένων

Όπως αναφέρθηκε και στη βιβλιογραφική ανασκόπηση απαραίτητη προϋπόθεση για την υλοποίηση της αρχιτεκτονικής RAG είναι η ύπαρξη εξωτερικής πηγής γνώσης από την οποία το σύστημα ανακτά σχετικά αποσπάσματα τα οποία αξιοποιούνται στη συνέχεια για την παραγωγή τεκμηριωμένης απάντησης. Στην παρούσα διπλωματική εργασία δημιουργήθηκε ένα ξεχωριστό σύνολο δεδομένων για κάθε εφαρμογή, από το οποίο ανακτάται η απαραίτητη πληροφορία που θα χρησιμοποιηθεί για την δημιουργία της απάντησης από το LLM.

Στην πρώτη εφαρμογή παρουσιάζεται ένα εργαλείο, μέσω του οποίου ο χρήστης δύναται να εισάγει μία καταγγελία παίκτη που αφορά στοιχηματική εταιρεία. Με βάση το περιεχόμενο της καταγγελίας, το σύστημα εντοπίζει και παρουσιάζει διατάξεις της ισχύουσας νομοθεσίας περί τυχερών παιγνίων που ενδέχεται να παραβιάζονται. Για την υλοποίηση της λειτουργικότητας αυτής αξιοποιήθηκαν επιλεγμένα αποσπάσματα της σχετικής νομοθεσίας. Πιο συγκεκριμένα, χρησιμοποιήθηκε η 79835 ΕΞ 2020/Β’ 3265/05.08.2020 απόφαση «Θέσπιση Κανονισμού Παιγνίων για τη Διοργάνωση και Διεξαγωγή Τυχερών Παιγνίων μέσω Διαδικτύου», η οποία καθορίζει κανόνες συμμόρφωσης για τους Κατόχους Άδειας που προσφέρουν τυχερά παίγνια μέσω διαδικτύου. Επιπλέον, στο σύνολο των δεδομένων εντάχθηκαν άρθρα από τον «Κανονισμό Εφαρμογής Μέτρων για την

Καταπολέμηση της Νομιμοποίησης Εσόδων από Εγκληματικές Δραστηριότητες και της Χρηματοδότησης της Τρομοκρατίας από τα Υπόχρεα Πρόσωπα στην Αγορά Υπηρεσιών Τυχερών Παιγνίων (554/5/15.04.2021 απόφαση της Ε.Ε.Ε.Π.)».

Το σύνολο δεδομένων οργανώθηκε σε μορφή csv και περιλαμβάνει τα πεδία id (μοναδικό αναγνωριστικό εγγραφής), law_source (πληροφορία σε ποιον νόμο αναφέρεται η συγκεκριμένη εγγραφή), article_title (το άρθρο του νόμου της συγκεκριμένης εγγραφής), article_text (το απόσπασμα της νομοθεσίας). Τα δεδομένα του συνόλου αποτελούν την εξωτερική βάση γνώσης από την οποία ανακτώνται τα πλέον συναφή αποσπάσματα με βάση την καταγγελία. Για την αποτελεσματική ανάκτηση εφαρμόστηκε τμηματοποίηση του κειμένου σε μικρότερα τμήματα (chunks). Το μήκος κάθε chunk υπολογίστηκε με κώδικα σε Python, ο οποίος εκτιμούσε το μήκος της κάθε παραγράφου σε tokens. Ως ανώτατο όριο ορίστηκαν 250 tokens ανά chunk, ώστε τα τμήματα να παραμένουν εντός των περιορισμών του embedding model και να διατηρείται υψηλή ακρίβεια κατά την αναζήτηση. Σε περίπτωση που ένα άρθρο του νόμου έπρεπε να διασπαστεί σε περισσότερα τμήματα, εφαρμόστηκε επικάλυψη συμφραζομένων: η θεματική πρόταση του πρώτου τμήματος ενσωματωνόταν και στα επόμενα, ώστε να διατηρείται η συνοχή του νοήματος και να μειώνεται ο κίνδυνος απώλειας κρίσιμου πλαισίου.

Η δεύτερη εφαρμογή αποτελεί ένα σύστημα εντοπισμού παρόμοιων καταγγελιών, στο οποίο το LLM εντοπίζει τις ομοιότητες και τις διαφορές ανάμεσα στην αρχική καταγγελία που δίνεται και των παρόμοιων καταγγελιών που ανακτώνται από την εξωτερική βάση γνώσης. Για την υλοποίηση της συγκεκριμένης βάσης γνώσης δημιουργήθηκε ένα σύνολο δεδομένων σε μορφή csv, το οποίο περιλαμβάνει τα πεδία ID (μοναδικό ID της καταγγελίας) και SUMMARY (περίληψη της καταγγελίας μέχρι 50 λέξεις). Ουσιαστικά το συγκεκριμένο σύνολο δεδομένων εμπεριέχει το ιστορικό καταγγελιών των πολιτών και έχει παρασχεθεί από την Ε.Ε.ΕΠ. στο πλαίσιο της παρούσας διπλωματικής. Στο σημείο αυτό πρέπει να διευκρινιστεί πως επειδή η περίληψη της κάθε καταγγελίας είναι μέχρι 50 λέξεις δεν χρειάζεται η εφαρμογή κάποιας τεχνικής τμηματοποίησης του κειμένου σε μικρότερα τμήματα.

Για τρίτη εφαρμογή η οποία αποτελεί ένα chatbot όπου χρήστης λαμβάνει πληροφορίες σε ερωτήματα που θέτει, οργανώθηκε ένα σύνολο δεδομένων σε

μορφή csv που περιλαμβάνει δεδομένα από την απόφαση «Τροποποίηση και κωδικοποίηση της υπ' αρ. 79305/27.7.2020 απόφασης του Υπουργού Οικονομικών Θέσπιση Κανονισμού Παιγνίων περί Καταλληλότητας Προσώπων» (56580 ΕΞ 2022 Β' 2166/04-05-2022) και την «Χορήγηση Άδειας Καταλληλότητας Συνεργατών Προωθητικών Ενεργειών Τυχερών Παιγνίων μέσω Διαδικτύου και εγγραφή στο Μητρώο Συνεργατών (Affiliates)» (509/1/11.09.2020 (Β 4140)). Τα πεδία που περιλαμβάνει είναι id (μοναδικό αναγνωριστικό εγγραφής), law_source (πληροφορία σε ποιον νόμο αναφέρεται η συγκεκριμένη εγγραφή), article_title (το άρθρο του νόμου της συγκεκριμένης εγγραφής), article_context_name (ο τίτλος του άρθρου), subject (το θέμα το οποίο πραγματεύεται το συγκεκριμένο άρθρο), paragraph_content (ένας πλαγιότιτλος της παραγράφου του άρθρου), article_text (το απόσπασμα της νομοθεσίας), topic_norm (το ειδικότερο θέμα που πραγματεύεται η παράγραφος του άρθρου), process_norm (την διαδικασία στην οποία εντάσσεται η συγκεκριμένη παράγραφος), area_norm (την περιοχή της νομοθεσίας που αφορά η συγκεκριμένη παράγραφος), role_norm (το πρόσωπο ή επιχείρηση που περιγράφει η συγκεκριμένη παράγραφος το οποίο πρόκειται να λάβει άδεια καταλληλότητας), entity_type_norm (εάν το πρόσωπο που αναφέρεται είναι φυσικό ή νομικό), subject_norm (επιπρόσθετη πληροφορία που αφορά το θέμα της παραγράφου).

Στο συγκεκριμένο σύνολο δεδομένων εφαρμόστηκε επίσης τμηματοποίηση του κειμένου της νομοθεσίας σε μικρότερα τμήματα (chunks). Το μήκος κάθε chunk υπολογίστηκε με κώδικα σε Python, ο οποίος εκτιμούσε το μήκος της κάθε παραγράφου σε tokens. Το ανώτατο όριο που ορίστηκε ήταν τα 250 tokens ανά chunk. Πρέπει να διευκρινιστεί πως στην περίπτωση που ένα άρθρο της νομοθεσίας έπρεπε να διαιρεθεί σε μικρότερα τμήματα, εφαρμόστηκε και εδώ επικάλυψη συμφραζομένων, δηλαδή η θεματική πρόταση του πρώτου τμήματος ενσωματωνόταν στα επόμενα.

Εν κατακλείδι και οι τρεις εφαρμογές ανακτούν δεδομένα από εξωτερική πηγή γνώσης η οποία αποτελείται από νόμους και καταγγελίες στην ελληνική γλώσσα σχετικά με τα τυχερά παίγνια. Τα δεδομένα που ανακτώνται θα αποτελέσουν το υλικό στο οποίο θα βασιστεί το LLM για την δημιουργία μία εμπεριστατωμένης απάντησης στον χρήστη.

3.3 Επιλογή Μοντέλου Ενσωμάτωσης, Βάση Διανυσμάτων και Μεγάλου Γλωσσικού Μοντέλου

Το επόμενο βήμα για τη δημιουργία ενός συστήματος RAG είναι η επιλογή κατάλληλου embedding model. Δεδομένου ότι τόσο τα κείμενα της εξωτερικής βάσης γνώσης όσο και τα ερωτήματα και οι καταγγελίες των χρηστών είναι στην ελληνική γλώσσα, επιλέχθηκαν πολυγλωσσικά μοντέλα που υποστηρίζουν την ελληνική και είναι κατάλληλα για ανάκτηση σημασιολογικά σχετικών αποσπασμάτων. Επιπροσθέτως, δόθηκε προτεραιότητα σε μοντέλα η χρήση των οποίων να μην οδηγεί σε χρέωση του τελικού χρήστη. Γι' αυτό τον λόγο δεν επιλέχθηκαν embedding models της Open AI και του Gemini, καθώς υπάρχει χρέωση ανάλογα με τα tokens από τα οποία αποτελείται το κείμενο που πρέπει να μετατραπεί σε embeddings. Επιπλέον, προτιμήθηκαν μοντέλα που είναι ελεύθερα διαθέσιμα. Τέτοιου είδους μοντέλα είναι το LaBSE, το intfloat/multilingual-e5-base και το paraphrase-multilingual-mpnet-base-v2, τα οποιά έχουν εκπαιδευτεί σε ελληνικά κείμενα και διατίθενται ελεύθερα μέσω αδειών Apache-2.0 (LaBSE και paraphrase-multilingual-mpnet-base-v2) και MIT (intfloat/multilingual-e5-base). Τα ανωτέρω μοντέλα είναι διαθέσιμα και αξιοποιήσιμα μέσω της βιβλιοθήκης Sentence-Transformers.

Για την παρούσα διπλωματική τα μοντέλα που επιλέχθηκαν για τη δημιουργία διανυσμάτων και στις τρεις εφαρμογές είναι τα LaBSE και intfloat/multilingual-e5-base. Το μοντέλο paraphrase-multilingual-mpnet-base-v2 δεν επιλέχθηκε καθώς υποστηρίζει πολύ μικρό αριθμό tokens (128) σε αντίθεση με τα LaBSE (256 tokens) και intfloat/multilingual-e5-base (512 tokens). Ουσιαστικά για κάθε μία εφαρμογή δημιουργήθηκε μία εκδοχή που χρησιμοποιεί το LaBSE ως embedding model και μία εκδοχή που χρησιμοποιεί το intfloat/multilingual-e5-base.

Για την υλοποίηση του σταδίου ανάκτησης δηλαδή του σταδίου κατά το οποίο ανακτώνται από την vector database τα πιο κοντινά διανύσματα σε σχέση με το ερώτημα του χρήστη, τα οποία αντιστοιχούν σε εξωτερική πηγή γνώσης, χρησιμοποιήθηκαν δύο διαφορετικές τεχνολογίες, ανάλογα με τις απαιτήσεις της εκάστοτε εφαρμογής. Στις εφαρμογές που υλοποιούν RAG πάνω σε νομοθετικά

κείμενα (στην πρώτη και την τρίτη εφαρμογή) επιλέχθηκε η ChromaDB, καθώς αποτελεί open-source vector database, οργανώνει τη γνώση σε συλλογές (collections) και υποστηρίζει αποθήκευση κειμένων μαζί με αναγνωριστικά (ids) και προαιρετικά μεταδεδομένα (metadata), διευκολύνοντας παραπομπές στις ανακτημένες πηγές. Επιπλέον, υποστηρίζει αποδοτική αναζήτηση πλησιέστερων γειτόνων μέσω HNSW (Hierarchical Navigable Small World) indexing και ρυθμιζόμενο distance metric (για παράδειγμα cosine) σε επίπεδο συλλογής. [41]
Αντίθετα, στην εφαρμογή εύρεσης όμοιων καταγγελιών (στη δεύτερη εφαρμογή) επιλέχθηκε το FAISS, το οποίο είναι βιβλιοθήκη υψηλής απόδοσης για similarity search σε dense embeddings και παρέχει αποδοτικούς μηχανισμούς indexing και k-NN (k-Nearest Neighbors) αναζήτησης ακόμη και σε πολύ μεγάλα σύνολα διανυσμάτων. Η παραπάνω επιλογή επέτρεψε την υλοποίηση κάθε εφαρμογής με το κατάλληλο επίπεδο «βαρύτητας» υποδομής: διαχείριση κειμένων-πηγών (και μεταδεδομένων όπου απαιτείται) για RAG με ChromaDB, και ελαφρύς, ταχύς vector index για similarity search με FAISS.

Το LLM που χρησιμοποιήθηκε για την παραγωγή της απάντησης του χρήστη και στις τρεις εφαρμογές είναι το Gemini. Πιο συγκεκριμένα το LLM λαμβάνει το ερώτημα του χρήστη και τα δεδομένα που έχει ανακτήσει από το vector database και δημιουργεί μία εμπεριστατωμένη απάντηση. Το Gemini επιλέχθηκε καθώς έχει προηγμένες δυνατότητες στην παραγωγή κειμένου. Το API του επιτρέπει χωρίς επιπλέον χρέωση μεγάλα όρια στην χρήση tokens. Το input token limit φτάνει μέχρι τα 1.048.576 tokens ενώ το output στα 65.536 tokens. Δέχεται διαφορετικούς τύπους δεδομένων όπως κείμενο, PDF αρχεία, εικόνες, βίντεο, ήχους. Αυτό είναι πολύ σημαντικό καθώς αποτελεί πλεονέκτημα σε μεταγενέστερες εξελίξεις των εφαρμογών.

Με βάση όσα αναφέρθηκαν ανωτέρω, η αρχιτεκτονική της κάθε εφαρμογής έχει ως εξής: Στην πρώτη εφαρμογή ο χρήστης εισάγει μία καταγγελία. Τα αποσπάσματα της νομοθεσίας μετατρέπονται σε embeddings (με χρήση του embedding model LaBSE ή intfloat/multilingual-e5-base). Δημιουργείται ένας in-memory client της ChromaDB και μία συλλογή (collection), στην οποία αποθηκεύονται για κάθε απόσπασμα το μοναδικό αναγνωριστικό του (id), το κείμενο και το αντίστοιχο διάνυσμα. Στη συνέχεια, η καταγγελία του χρήστη μετατρέπεται και εκείνη σε embeddings με το ίδιο

μοντέλο που μετατράπηκαν τα αποσπάσματα της νομοθεσίας. Το επόμενο βήμα είναι να εντοπιστούν τα αποσπάσματα της νομοθεσίας που ταιριάζουν σημασιολογικά με την καταγγελία. Αξιοποιείται η μετρική του cosine similarity, η οποία υπολογίζει πόσο «κοντά» δείχνουν δύο διανύσματα στον χώρο. Μετρά μόνο τη γωνία μεταξύ των διανυσμάτων (αγνοώντας το μέγεθος), γεγονός που την καθιστά ιδανική για text embeddings ή γενικότερες περιπτώσεις όπου μας ενδιαφέρει η κατεύθυνση του διανύσματος και όχι το μέγεθός του. Με την συγκεκριμένη μετρική εάν δύο διανύσματα δείχνουν προς την ίδια κατεύθυνση, σημαίνει ότι έχουν παρόμοια αναλογία χαρακτηριστικών, άρα παρόμοιο νόημα. Το σύστημα ανακτά από την ChromaDB τα έξι πιο σχετικά αποσπάσματα που σχετίζονται νοηματικά με την καταγγελία του χρήστη. Τα συγκεκριμένα αποσπάσματα τα παρουσιάζει στον χρήστη μαζί με το cosine similarity score καθώς και με το μοναδικό id που είχαν στο σύνολο δεδομένων προκειμένου να μπορέσει να επαληθεύσει ο χρήστης την ορθότητά τους. Η καταγγελία του χρήστη μαζί με τα αποσπάσματα εισάγονται ως prompt στο Gemini, το οποίο δημιουργεί μια τεκμηριωμένη απάντηση στον χρήστη σχετικά με τις διατάξεις που ενδέχεται να παραβιάζει η συγκεκριμένη καταγγελία, βασιζόμενο αποκλειστικά στο παρεχόμενο νομοθετικό περιεχόμενο. Για την ανάκτηση των έξι πιο συναφών αποσπασμάτων, η ChromaDB αξιοποιεί δείκτη HNSW για προσεγγιστική αναζήτηση πλησιέστερων γειτόνων (Approximate Nearest Neighbors – ANN) στον χώρο των embeddings.

Στη δεύτερη εφαρμογή ο χρήστης εισάγει μία καταγγελία. Το σύστημα χρησιμοποιώντας το Gemini δημιουργεί μία περίληψη της συγκεκριμένης καταγγελίας μέχρι 50 λέξεις. Στη συνέχεια οι περιλήψεις του ιστορικού καταγγελιών μετατρέπονται σε embeddings με χρήση του ανάλογου embedding model (LaBSE ή intfloat/multilingual-e5-base) και κατασκευάζεται ένας διανυσματικός δείκτης (FAISS index), ο οποίος επιτρέπει ταχύτατη αναζήτηση στον χώρο των embeddings. Το επόμενο βήμα είναι να μετατραπεί σε embeddings και η περίληψη που δημιούργησε το Gemini με το ίδιο embedding model που χρησιμοποιήθηκε και στο ιστορικό καταγγελιών. Πραγματοποιείται semantic similarity search προκειμένου να εντοπιστούν οι πέντε κορυφαίες περιλήψεις των καταγγελιών που έχουν μεγαλύτερη νοηματική συνάφεια με την περίληψη της καταγγελίας του χρήστη. Τα αποτελέσματα

επιστρέφονται με τα αντίστοιχα IDs και cosine similarity scores. Έπειτα δίνονται στο Gemini η περίληψη της καταγγελίας που εισήγαγε ο χρήστης και οι περιλήψεις των πιο σχετικών καταγγελιών που ανακτήθηκαν από το FAISS. Το LLM προσπαθεί να βρει ομοιότητες και διαφορές τις οποίες παρουσιάζει στον χρήστη.

Στην τρίτη εφαρμογή ο χρήστης κάνει μία ερώτηση στο chatbot. Γίνεται προσπάθεια να εξαχθεί η πληροφορία για ποιο role_norm (το πρόσωπο ή επιχείρηση το οποίο πρόκειται να λάβει άδεια καταλληλότητας) αφορά το ερώτημά του. Αυτό γίνεται με τη βοήθεια ενός λεξικού. Η ερώτηση του χρήστη αρχικά περνά από ένα στάδιο καθαρισμού. Όλα τα γράμματα μετατρέπονται σε πεζά και αφαιρούνται τα κενά από την αρχή και το τέλος. Αφαιρούνται οι τόνοι. Το τελικό «s» μετατρέπεται σε «σ». Διατηρούνται μόνο λατινικοί και ελληνικοί χαρακτήρες, αριθμοί, κενά και παύλες. Οτιδήποτε άλλο αντικαθίσταται με το κενό. Αν μετά τον καθαρισμό μείνουν πολλά συνεχόμενα κενά, μετατρέπονται σε ένα μόνο κενό. Το τελικό κείμενο που δημιουργείται, συγκρίνεται με τη βοήθεια της βιβλιοθήκης rapidfuzz, με το λεξικό το οποίο περιλαμβάνει διαφορετικές μορφές ορθογραφίας των τιμών που βρίσκονται στο πεδίο role_norm του συνόλου δεδομένων. Ουσιαστικά γίνεται μία προσπάθεια να γίνει αντιληπτό ποιο role_norm αφορά η ερώτηση του χρήστη προκειμένου να χρησιμοποιηθεί ως φίλτρο μεταδεδομένων (metadata filter) στο στάδιο ανάκτησης.

Πιο συγκεκριμένα, από το τελικό κείμενο αφαιρούνται stopwords και δημιουργούνται ακολουθίες n διαδοχικών λέξεων (n-grams), για n = 1 έως 3. Μέσω της βιβλιοθήκης rapidfuzz υπολογίζεται ένα score ομοιότητας για κάθε n-gram που έχει δημιουργηθεί με κάθε role_norm που υπάρχει στο σύνολο δεδομένων, δίνοντας ένα βάρος στα n-grams με τις περισσότερες λέξεις. Τα role_norms που θα προκύψουν μέσα από τη συγκεκριμένη διαδικασία, τα οποία είναι πιο κοντά στο ερώτημα του χρήστη, ανάλογα με το score ομοιότητάς τους κατατάσσονται σε confident και ambiguous. Σε περίπτωση που υπάρξει ambiguous role_norm, δίνεται η δυνατότητα στον χρήστη εάν θέλει να το συμπεριλάβει ως φίλτρο μεταδεδομένων στο στάδιο της ανάκτησης.

Το επόμενο βήμα είναι τα αποσπάσματα της νομοθεσίας να μετατραπούν σε embeddings (με χρήση του embedding model LaBSE ή intfloat/multilingual-e5-base). Δημιουργείται ένας in-memory client της ChromaDB και μία συλλογή (collection), στην οποία αποθηκεύονται για κάθε απόσπασμα το κείμενο, το αντίστοιχο διάνυσμα

και τα πεδία id, law_source, article_title, article_context_name, subject, paragraph_content, role_norm ως metadata. Στη συνέχεια, το αρχικό input του χρήστη μετατρέπεται σε embeddings με το ίδιο μοντέλο που μετατράπηκαν τα αποσπάσματα της νομοθεσίας. Στόχος είναι να εντοπιστούν αποσπάσματα της νομοθεσίας που είναι πιο κοντά στο ερώτημα που πληκτρολόγησε ο χρήστης στην εφαρμογή. Για να επιτευχθεί αυτό αποδοτικά και με μεγαλύτερη ακρίβεια, η αναζήτηση γίνεται μόνο στις εγγραφές της ChromaDB που έχουν ως metadata ένα από τα role_norms που προέκυψαν στην προηγουμένη διαδικασία. Σε περίπτωση που δεν έχουν προκύψει role_norms που να σχετίζονται με το ερώτημα του χρήστη, τότε η αναζήτηση γίνεται σε όλο το περιεχόμενο της ChromaDB. Η ανάκτηση των πιο συναφών αποσπασμάτων της νομοθεσίας βάσει σημασιολογικής ομοιότητας, πραγματοποιείται με αναζήτηση σε cosine space και η ChromaDB αξιοποιεί δείκτη HNSW για προσεγγιστική αναζήτηση πλησιέστερων γειτόνων (Approximate Nearest Neighbors – ANN) στον χώρο των embeddings, επιστρέφοντας τα δέκα πιο σχετικά αποσπάσματα. Στη συνέχεια το αρχικό ερώτημα του χρήστη, τα αποσπάσματα που ανακτήθηκαν δίνονται ως input στο Gemini, το οποίο δημιουργεί μία απάντηση βασισμένη αποκλειστικά στα αποσπάσματα της νομοθεσίας παρέχοντας παραπομπές.

3.4 Παρουσίαση Τεχνολογιών και Εργαλείων που Χρησιμοποιήθηκαν

Για την υλοποίηση των ανωτέρω εφαρμογών που ακολουθούν την αρχιτεκτονική RAG, χρησιμοποιήθηκαν τεχνολογίες και εργαλεία που ως στόχο είχαν την δημιουργία εφαρμογών που να είναι όσο τον δυνατόν ταχύτερες και να έχουν μειωμένο υπολογιστικό κόστος. Ως περιβάλλον υλοποίησης επιλέχθηκε το Google Colab προκειμένου οι κώδικες να εκτελούνται στο cloud και μην καταναλώνουν πόρους της υποδομής του χρήστη. Το Google Colab αποτελεί ένα υπολογιστικό περιβάλλον βασισμένο στο υπολογιστικό νέφος (cloud-based), το οποίο υποστηρίζει τη διαδραστική εκτέλεση κώδικα Python μέσω διεπαφής Jupyter Notebook. Το κύριο πλεονέκτημα της πλατφόρμας είναι η παροχή υπολογιστικών πόρων υψηλών

επιδόσεων, όπως Μονάδες Επεξεργασίας Γραφικών (GPUs) και Tensor Processing Units (TPUs), διευκολύνοντας εργασίες Μηχανικής Μάθησης και επεξεργασίας κειμένου χωρίς απαίτηση τοπικής υποδομής. Επιπλέον, η ενσωμάτωση του Colab με το οικοσύστημα της Google διευκολύνει τη διαχείριση δεδομένων και τη συνεργατική ανάπτυξη κώδικα. [42]

Εφόσον αξιοποιήθηκε ως εργαλείο εκτέλεσης του κώδικα το Google Colab, η γλώσσα προγραμματισμού που επιλέχθηκε για την συγγραφή του κώδικα είναι η Python. Αποτελεί μια γλώσσα προγραμματισμού υψηλού επιπέδου, που εκτελείται μέσω διερμηνευτή και υποστηρίζει τον αντικειμενοστραφή προγραμματισμό. Διαθέτει δυναμική τυποποίηση και «ευέλικτη» σημασιολογία κατά την εκτέλεση, με αποτέλεσμα να διευκολύνει τη γρήγορη ανάπτυξη εφαρμογών. Παράλληλα, χάρη στην καθαρή και λιτή σύνταξή της, χρησιμοποιείται συχνά τόσο για αυτοματοποίηση εργασιών (scripting) όσο και ως «glue language», δηλαδή για τη διασύνδεση διαφορετικών εργαλείων και βιβλιοθηκών σε ένα ενιαίο σύστημα. [43]

Για την υλοποίηση των εφαρμογών χρησιμοποιούνται βιβλιοθήκες οι οποίες διασφαλίζουν την ορθή λειτουργία τους. Η βιβλιοθήκη sentence-transformers είναι ένα framework σε Python για την παραγωγή διανυσματικών αναπαραστάσεων (embeddings) προτάσεων ή μικρών κειμένων, ώστε κείμενα με παρόμοιο νόημα να βρίσκονται «κοντά» στον ίδιο διανυσματικό χώρο και να μπορούν να συγκριθούν εύκολα (για παράδειγμα με την χρήση της μετρικής cosine similarity) για εργασίες όπως σημασιολογική αναζήτηση, ομαδοποίηση και εντοπισμό ομοιότητας. Η κλάση SentenceTransformer (from sentence_transformers import SentenceTransformer) αποτελεί το βασικό σημείο χρήσης της βιβλιοθήκης, καθώς επιτρέπει τη φόρτωση ενός προεκπαιδευμένου μοντέλου και την κωδικοποίηση κειμένων σε embeddings μέσω της μεθόδου encode. Με την συγκεκριμένη βιβλιοθήκη φορτώνουμε τα μοντέλα LaBSE και intfloat/multilingual-e5-base. Ακόμη πρέπει να γίνει αναφορά και στη βιβλιοθήκη chromadb η οποία δίνει τη δυνατότητα πρόσβασης στις λειτουργίες της ChromaDB, όπως είναι η δημιουργία collections, αποθήκευση embeddings, και metadata. Επιπλέον, η βιβλιοθήκη faiss αποτελεί βιβλιοθήκη υψηλής απόδοσης για similarity search σε dense embeddings και παρέχει αποδοτικούς μηχανισμούς indexing ακόμη και σε πολύ μεγάλα σύνολα διανυσμάτων. Η βιβλιοθήκη rapidfuzz

είναι μια εξαιρετικά γρήγορη βιβλιοθήκη της Python που χρησιμοποιείται για τη σύγκριση συμβολοσειρών (string matching). Στην ουσία, επιτρέπει να βρεθεί πόσο μοιάζουν δύο λέξεις ή προτάσεις, ακόμα και αν υπάρχουν ορθογραφικά λάθη, διαφορετική σύνταξη ή ελλιπή στοιχεία. Χρησιμοποιείται στην εφαρμογή στην οποία έχει υλοποιηθεί το chatbot προκειμένου να εντοπιστεί σε ποιο role_norm αναφέρεται ο χρήστης στο ερώτημά του, αναζητώντας ομοιότητα ανάμεσα στο ερώτημα του χρήστη και στα role_norm που υπάρχουν στο λεξικό. Τέλος, η βιβλιοθήκη google-genai επιτρέπει την αλληλεπίδραση με LLMs και συγκεκριμένα με την οικογένεια μοντέλων Gemini. Με τη συγκεκριμένη βιβλιοθήκη δίνεται η δυνατότητα να χρησιμοποιούνται τα γλωσσικά μοντέλα Gemini και αποτελεί απαραίτητο συστατικό στην αρχιτεκτονική RAG, καθώς μέσω της συγκεκριμένης βιβλιοθήκης το LLM παράγει την απάντηση στον χρήστη.

4. Παρουσίαση Εφαρμογών και Αποτελεσμάτων

4.1 Παρουσίαση Εφαρμογής Εύρεσης Νόμων που Παραβιάζει μία Στοιχηματική Εταιρεία με Βάση την Καταγγελία

Στη συγκεκριμένη εφαρμογή ο χρήστης ο οποίος είναι ένας ελεγκτής, εισάγει μία καταγγελία ενός πολίτη που αφορά μία στοιχηματική εταιρεία, στο σύστημα. Η εφαρμογή χρησιμοποιώντας τα δεδομένα από την νομοθεσία για τα τυχερά παίγνια απαντά στον ελεγκτή ποια αποσπάσματα από την νομοθεσία ενδεχομένως παραβιάζει η στοιχηματική.

Στο πείραμα που ακολουθεί εισάγεται μία καταγγελία ενός χρήστη και το embedding model που χρησιμοποιείται είναι το LaBSE. Το σύνολο δεδομένων καθώς και η καταγγελία δεν ξεπερνούν τα 250 tokens. Η καταγγελία που εισάγεται είναι η ακόλουθη:

«Καλησπέρα. Διατηρώ ενεργό λογαριασμό στην εταιρεία Testbet και παίζω λίγους μήνες. Κατά τη διάρκεια αυτού του διαστήματος, κάποιες καταθέσεις έγιναν από

τραπεζικό λογαριασμό που ανήκει στον σύζυγό μου. Η συμμετοχή μου στο παιχνίδι εξελισσόταν κανονικά και κατά καιρούς είχα κέρδη, από τα οποία πραγματοποίησα και μικρές αναλήψεις (περίπου 100–200€), χωρίς πρόβλημα. Πρόσφατα, για πρώτη φορά προέκυψε κέρδος μεγαλύτερου ύψους, συγκεκριμένα 900€. Ωστόσο, η εταιρεία δεν προχώρησε στην πλήρη ανάληψη, αλλά επέστρεψε μόνο 300€ και στη συνέχεια ακύρωσε τη σχετική δραστηριότητα, αναφέροντας ότι το κέρδος δεν ισχύει επειδή είχε προηγηθεί κατάθεση από διαφορετικό λογαριασμό. Θα ήθελα τη συνδρομή σας σχετικά με αυτή τη μονομερή αντιμετώπιση των κερδών από την εταιρεία. Από την πλευρά μου θεωρώ ότι είτε (α) θα πρέπει να καταβληθεί το κέρδος κανονικά, είτε (β) αν κρίνεται ότι υπάρχει λόγος ακύρωσης λόγω του τρόπου κατάθεσης, τότε θα πρέπει να ακυρωθούν και να επιστραφούν όλες οι καταθέσεις μου, και όχι μόνο το ποσό που υπερβαίνει την τελευταία κατάθεση».

Το σύστημα αξιοποιώντας τη μετρική cosine similarity ανακτά από την chroma τα πιο σχετικά αποσπάσματα νομοθεσίας, τα οποία δείχνουν τις πιθανές παραβάσεις της εταιρείας. Στον χρήστη εμφανίζονται τα έξι πιο σχετικά αποσπάσματα μαζί με το μοναδικό id και στο similarity score. Στη συνέχεια η καταγγελία του χρήστη και τα αποσπάσματα δίνονται στο Gemini το οποίο δημιουργεί μία τεκμηριωμένη απάντηση στον χρήστη βασιζόμενο μόνο στα αποσπάσματα που του δίνονται. Στο συγκεκριμένο πείραμα χρησιμοποιούμε το μοντέλο gemini-3-flash-preview.

Προκειμένου το Gemini να απαντήσει βασιζόμενο μόνο στην πληροφορία που λαμβάνει, του έχει δοθεί αναλυτικό prompt το οποίο παρουσιάζεται στην ακόλουθη εικόνα.

```
# 🔵 7. Σύνθεση prompt για το Gemini 3 Flash
prompt = f"""
Ενεργείς ως νομικός σύμβουλος εξειδικευμένος στη νομοθεσία τυχερών παιγνίων
στην Ελλάδα.

Σου παρέχω:

1 Μία καταγγελία παίκτη εναντίον στοιχηματικής εταιρείας.
2 Αποσπάσματα από τη σχετική νομοθεσία.

Η αποστολή σου είναι:

- Να εξετάσεις αν υπάρχει παραβίαση νομοθεσίας από την εταιρεία, βασισμένος
ΜΟΝΟ στα δεδομένα που σου παρέχονται.
- Αν εντοπίζεις παραβίαση, να εξηγήσεις ΠΟΙΟ άρθρο παραβιάζεται και ΓΙΑΤΙ.
- Αν δεν υπάρχει παραβίαση, να το δηλώσεις ρητά και να εξηγήσεις.
- ΜΗΝ χρησιμοποιείς καθόλου εξωτερικές γενικές γνώσεις ή υποθέσεις.
- Να βασιστείς ΑΠΟΚΛΕΙΣΤΙΚΑ στα αποσπάσματα της νομοθεσίας που δίνονται.

### Καταγγελία πελάτη:
{kataggelia}

### Απόσπασματα Νομοθεσίας:
{nomothesia_context}

Παρακαλώ ξεκίνα την ανάλυσή σου με ξεκάθαρη και λογική νομική αιτιολόγηση.
"""
"""
```

Εικόνα 9: Prompt στο Gemini προκειμένου να βασιστεί μόνο στα δεδομένα που του παρέχονται για την δημιουργία της απάντησης στον χρήστη

Από τα αποτελέσματα της εκτέλεσης της συγκεκριμένης εφαρμογής, τα οποία παρουσιάζονται στις Εικονές 10 και 11, μπορούμε να συμπεράνουμε πως το LaBSE μπόρεσε να εντοπίσει αποσπάσματα νομοθεσίας τα οποία σχετίζονται με την καταγγελία του πολίτη. Το απόσπασμα με ID: 554_5_13.7d αναφέρεται στην υποχρέωση των Κατόχων Άδειας σε αναλήψεις ποσών άνω των 800 ευρώ να προβαίνουν σε ταυτοποίηση του αιτούντος ανάληψη, δηλαδή έχουν την υποχρέωση να επιβεβαιώνουν ότι ο λογαριασμός πληρωμών που έχει δηλώσει ο παίκτης ανήκει στον ίδιο. Το απόσπασμα με ID: 554_5_13.7a-c διευκρινίζει ότι Μεταφορές Χρημάτων από και προς τον Ηλεκτρονικό Λογαριασμό Παίκτη διενεργούνται υποχρεωτικά μέσω των Παρόχων Υπηρεσιών Πληρωμών και ότι καταθέσεις (deposits) στον Ηλεκτρονικό Λογαριασμό Παίκτη, ποσού πέντε χιλιάδων (5.000) ευρώ και άνω ανά πράξη, γίνονται

αποδεκτές και χρησιμοποιούνται για τη διενέργεια Συμμετοχών, μόνο εφόσον επιβεβαιωθεί ότι ο λογαριασμός πληρωμών που έχει δηλώσει ο Παίκτης ανήκει πραγματικά σε αυτόν. Επιπλέον, το απόσπασμα με ID: 554_5_13.7e επισημαίνει ότι οι Κάτοχοι Άδειας διασφαλίζουν ότι λαμβάνουν κάθε πρόσφορο, οργανωτικό και τεχνικό μέτρο, ώστε οι Μεταφορές Χρημάτων από και προς τον Ηλεκτρονικό Λογαριασμό Παίκτη να διενεργούνται με Μέσα Πληρωμής που ανήκουν στον Παίκτη. Σε περίπτωση που διαπιστωθεί ότι το μέσο πληρωμής δεν ανήκει στο άτομο που διατηρεί Ηλεκτρονικό Λογαριασμό Παίκτη, απαγορεύουν το μέσο και επιστρέφουν το ποσό της συναλλαγής στον δικαιούχο του μέσου πληρωμής.

Τα ανωτέρω αποσπάσματα είναι αρκετά κοντά στο νόημα της καταγγελίας, καθώς αναφέρονται σε διαδικασίες που ακολουθούνται κατά τη διαδικασία της ανάληψης και κατάθεσης ποσού. Μέσα στην ίδια την καταγγελία γίνεται λόγος για καταθέσεις και αναλήψεις του χρήστη και μάλιστα με μέσο πληρωμής που ανήκε σε τρίτο πρόσωπο. Επιπλέον, παρατηρείται ότι τα πρώτα τρία αποσπάσματα έχουν κοντινό cosine similarity score.

Στο σημείο αυτό πρέπει να διευκρινιστεί πως το LaBSE δεν είναι ένα generative μοντέλο. Επομένως δεν μπορεί εκτός από την παράθεση των αποσπασμάτων, να δημιουργήσει μία απάντηση στον χρήστη. Τον συγκεκριμένο ρόλο αναλαμβάνει το LLM. Το Gemini δέχεται την καταγγελία του χρήστη και τα αποσπάσματα της νομοθεσίας προκειμένου να δημιουργήσει μία απάντηση στον χρήστη.

Από την απάντηση που παρήγαγε το Gemini, η οποία απεικονίζεται στις Εικόνες 12 και 13, μπορεί να εξαχθεί το συμπέρασμα πως το LLM θεώρησε ποιο σχετικό απόσπασμα που βασίζεται στην καταγγελία, ένα από τα τρία πρώτα αποσπάσματα που έφερε το LaBSE. Πιο συγκεκριμένα, στην απάντηση διευκρινίζεται ότι με βάση το απόσπασμα με ID: 554_5_13.7e η εταιρεία ορθώς δεν επέτρεψε την ανάληψη των 900 ευρώ καθώς διαπιστώθηκε πως το μέσο πληρωμής ανήκε σε τρίτο πρόσωπο. Αφήνει όμως ανοικτή μία ενδεχόμενη παράβαση, να μην έχει επιστρέψει η εταιρεία όλο το ποσό που διακυβεύτηκε στα τυχερά παίγνια στον πραγματικό δικαιούχο του μέσου πληρωμής.

🔍 Σχετικά Άρθρα: 📄 ID: 554_5_13.7d 🔗 Cosine similarity: 0.572 Καθ' όλη τη διάρκεια της Επιχειρηματικής Σχέσης, τα Υπόχρεα Πρόσωπα που διεξάγουν Παίγνια μέσω Ηλεκτρονικού Λογαριασμού Παίκτη, διασφαλίζουν ότι: δ. Καταβολές (withdrawals) προς τους Παίκτες, ποσού οκτακοσίων (800) ευρώ και άνω ανά πράξη, διενεργούνται μόνο εφόσον επιβεβαιωθεί, πριν τη συναλλαγή, ότι ο λογαριασμός πληρωμών που έχει δηλώσει ο Παίκτης ανήκει πραγματικά σε αυτόν. Για τις πληρωμές των Παικτών στην περίπτωση των Τυχερών Παιγνίων που διεξάγονται μέσω παιγνιομηχανημάτων τύπου VLT, εφαρμόζονται τα όρια και οι διαδικασίες της παρ. 15.1.

📄 ID: 554_5_13.7a-c 🔗 Cosine similarity: 0.569 Καθ' όλη τη διάρκεια της Επιχειρηματικής Σχέσης, τα Υπόχρεα Πρόσωπα που διεξάγουν Παίγνια μέσω Ηλεκτρονικού Λογαριασμού Παίκτη, διασφαλίζουν ότι: α. Δεν μεταφέρονται πιστωτικές μονάδες μεταξύ Ηλεκτρονικών Λογαριασμών Παικτών. β. Μεταφορές Χρημάτων από και προς τον Ηλεκτρονικό Λογαριασμό Παίκτη διενεργούνται υποχρεωτικά μέσω των Παρόχων Υπηρεσιών Πληρωμών. γ. Καταθέσεις (deposits) στον Ηλεκτρονικό Λογαριασμό Παίκτη, ποσού πέντε χιλιάδων (5.000) ευρώ και άνω ανά πράξη, γίνονται αποδεκτές και χρησιμοποιούνται για τη διενέργεια Συμμετοχών, μόνο εφόσον επιβεβαιωθεί ότι ο λογαριασμός πληρωμών που έχει δηλώσει ο Παίκτης ανήκει πραγματικά σε αυτόν.

📄 ID: 554_5_13.7e 🔗 Cosine similarity: 0.566 Καθ' όλη τη διάρκεια της Επιχειρηματικής Σχέσης, τα Υπόχρεα Πρόσωπα που διεξάγουν Παίγνια μέσω Ηλεκτρονικού Λογαριασμού Παίκτη, διασφαλίζουν ότι: ε. Λαμβάνουν κάθε πρόσφορο, οργανωτικό και τεχνικό μέτρο, ώστε οι Μεταφορές Χρημάτων από και προς τον Ηλεκτρονικό Λογαριασμό Παίκτη να διενεργούνται με Μέσα Πληρωμής που ανήκουν στον Παίκτη. Σε περίπτωση που εκ των υστέρων επιβεβαιωθεί Μεταφορά Χρημάτων με την χρήση Μέσων Πληρωμής που ανήκουν σε τρίτους, απαγορεύουν την χρήση αυτών, επιστρέφουν το κεφάλαιο που έχει κατατεθεί στον δικαιούχο του Μέσου Πληρωμής, δεν αποδίδουν τυχόν κέρδη που προέκυψαν από την χρήση των Μέσων αυτών και εξετάζουν εάν συντρέχει λόγος αναφοράς στην Αρχή.

📄 ID: 79835_25.2 🔗 Cosine similarity: 0.552 Για να ξεκινήσει τη διεξαγωγή τυχερών παιγνίων, ο Κάτοχος Άδειας πρέπει να διατηρεί σε λογαριασμό Παικτών ποσό ίσο με το μισό της εγγυητικής επιστολής, ανά Τύπο Άδειας. Προκειμένου να λάβουν Άδεια, οι ενδιαφερόμενοι οφείλουν να προσκομίσουν: Αναλυτική κατάσταση με το συνολικό πιστωμένο ποσό των Ηλεκτρονικών Λογαριασμών των Παικτών. Αντίγραφο του τραπεζικού λογαριασμού Παικτών που δείχνει το υπόλοιπο. Το ποσό στον λογαριασμό Παικτών πρέπει να είναι τουλάχιστον ίσο με το συνολικό πιστωμένο ποσό των Ηλεκτρονικών Λογαριασμών των Παικτών. Εάν υπάρχει έλλειμμα, ο Κάτοχος Άδειας οφείλει να το αναπληρώσει εντός τριών (3) ημερών.

Εικόνα 10: LaBSE: Αποσπάσματα νομοθεσίας που είναι πιο κοντά στο περιεχόμενο της καταγγελίας (1)

📄 ID: 554_5_13.4a-b 🔗 Cosine similarity: 0.544 Για το χρονικό διάστημα που ο Ηλεκτρονικός Λογαριασμός Παίκτη έχει προσωρινό χαρακτήρα σύμφωνα με την παρ. 13.2, τα Υπόχρεα Πρόσωπα: α. Επιτρέπουν καταθέσεις από τον Παίκτη στον Ηλεκτρονικό του Λογαριασμό ποσού έως και οκτακόσια (800) ευρώ, είτε το ποσό αυτό καλύπτεται με μία πράξη είτε με περισσότερες πράξεις. β. Επιτρέπουν τη Συμμετοχή του Παίκτη σε στοιχήματα επί γεγονότων τα οποία προγραμματίζεται να λάβουν χώρα έως και την λήξη της 30ης ημέρας από την ημέρα δημιουργίας του Ηλεκτρονικού του Λογαριασμού ή σε διαδικτυακά παίγνια που διενεργούνται εντός του ίδιου χρονικού διαστήματος. Στοιχήματα επί γεγονότων που αναβλήθηκαν για μετά την λήξη της 30ης ημέρας \από την ημέρα δημιουργίας του Ηλεκτρονικού Λογαριασμού, διευθετούνται με απόδοση τη μονάδα.

📄 ID: 79835_24.5c-bb 🔗 Cosine similarity: 0.534 Ο Κάτοχος Άδειας ενημερώνει τον Παίκτη μέσω μηνύματος στον Ιστότοπο ότι, μόλις καλυφθεί το χρονικό όριο, η συμμετοχή του σε τουρνουά πόκερ (peer to peer) θα διακοπεί υποχρεωτικά. Η διακοπή γίνεται με την ολοκλήρωση του τουρνουά για το οποίο ο Παίκτης έχει καταβάλει συμμετοχή. Εάν έχει πληρώσει για περισσότερα τουρνουά που έχουν ξεκινήσει, η διακοπή γίνεται με την ολοκλήρωση του τουρνουά που τελειώνει τελευταίο. Σε περίπτωση που ο Παίκτης έχει καταβάλει συμμετοχή σε τουρνουά που δεν έχουν ακόμη ξεκινήσει (registration open), οι συμμετοχές του ακυρώνονται και το καταβληθέν ποσό επιστρέφεται χωρίς προμήθεια (rake, fee).

Εικόνα 11: LaBSE: Αποσπάσματα νομοθεσίας που είναι πιο κοντά στο περιεχόμενο της καταγγελίας (2)

🎯 Απάντηση Gemini:

Βάσει των δεδομένων που μου παρείχατε και των συγκεκριμένων αποσπασμάτων της νομοθεσίας, η νομική αξιολόγηση της υπόθεσης έχει ως εξής:

1. Διαπίστωση Παραβίασης ή μη

Από τα στοιχεία της καταγγελίας και τα σχετικά άρθρα, προκύπτει ότι:

- **Η εταιρεία ορθώς αρνήθηκε την καταβολή των κερδών (900€).**
- **Η εταιρεία ενδέχεται να παραβιάζει τη νομοθεσία ως προς το ύψος του ποσού που επέστρεψε**, εάν το ποσό των 300€ δεν καλύπτει το σύνολο του κεφαλαίου που κατατέθηκε από τον λογαριασμό του συζύγου.

2. Νομική Αιτιολόγηση & Ανάλυση

Α. Η χρήση Μέσου Πληρωμής τρίτου (Σύζυγος)

Σύμφωνα με το **ID: 554_5_13.7e**, οι εταιρείες οφείλουν να διασφαλίζουν ότι οι μεταφορές χρημάτων διενεργούνται με Μέσα Πληρωμής που ανήκουν αποκλειστικά στον Παίκτη. Στην περίπτωση που επιβεβαιωθεί εκ των υστέρων ότι έγινε χρήση μέσου τρίτου (στην προκειμένη περίπτωση του συζύγου), ο νόμος είναι ρητός:

1. **Απαγορεύεται** η χρήση του συγκεκριμένου μέσου για το μέλλον.
2. **Δεν αποδίδονται τυχόν κέρδη** που προέκυψαν από τη χρήση αυτών των μέσων.
3. **Επιστρέφεται το κεφάλαιο** που έχει κατατεθεί στον δικαιούχο του μέσου πληρωμής.

Συνεπώς, η αξίωση του παίκτη για καταβολή του κέρδους των 900€ (αίτημα α) **αντιτίθεται ρητά** στη διάταξη 554_5_13.7e και η εταιρεία ορθώς ακύρωσε τα κέρδη αυτά.

Εικόνα 12: LaBSE: Απάντηση Gemini (1)

Β. Η διαδικασία ελέγχου κατά την ανάληψη

Ο παίκτης αναφέρει ότι προηγούμενες μικρές αναλήψεις (100-200€) εκτελέστηκαν κανονικά. Αυτό εξηγείται από το **ID: 554_5_13.7d**, το οποίο ορίζει ότι ο υποχρεωτικός έλεγχος ταυτοποίησης του λογαριασμού πληρωμών (αν ο λογαριασμός ανήκει πραγματικά στον παίκτη) είναι επιβεβλημένος **πριν από συναλλαγή ανάληψης ποσού οκτακοσίων (800) ευρώ και άνω**. Καθώς το τρέχον αίτημα ήταν 900€, η εταιρεία ήταν υποχρεωμένη από τον νόμο να προβεί στον έλεγχο που οδήγησε στην αποκάλυψη της χρήσης λογαριασμού τρίτου.

Γ. Το ύψος της επιστροφής (Κεφάλαιο)

Το κρίσιμο σημείο της καταγγελίας αφορά το αίτημα (β) του παίκτη: την επιστροφή όλων των καταθέσεων και όχι μόνο μέρους αυτών. Το άρθρο **ID: 554_5_13.7e** ορίζει ότι η εταιρεία «**επιστρέφει το κεφάλαιο που έχει κατατεθεί** στον δικαιούχο του Μέσου Πληρωμής».

- Αν ο παίκτης κατά το διάστημα των «λίγων μηνών» κατέθεσε συνολικά από τον λογαριασμό του συζύγου ποσό μεγαλύτερο των 300€, η εταιρεία **παραβιάζει το άρθρο 554_5_13.7e**, καθώς οφείλει να επιστρέψει το **σύνολο του κεφαλαίου** που εισέρρευσε από το παράτυπο μέσο και όχι μόνο το ποσό της τελευταίας κατάθεσης ή μέρος αυτού.
- Η νομοθεσία δεν προβλέπει επιλεκτική επιστροφή κεφαλαίου, αλλά επιστροφή του κεφαλαίου που προήλθε από τη χρήση του μέσου τρίτου.

Συμπέρασμα

1. **Δεν υπάρχει παραβίαση** ως προς την άρνηση καταβολής των κερδών (βάσει του ID: 554_5_13.7e).
2. **Υπάρχει πιθανή παραβίαση του άρθρου ID: 554_5_13.7e** ως προς την επιστροφή των χρημάτων: Η Testbet υποχρεούται να επιστρέψει στον σύζυγο (δικαιούχο) το **συνολικό κεφάλαιο** των καταθέσεων που πραγματοποιήθηκαν από τον λογαριασμό του και όχι μόνο το ποσό των 300€, εφόσον οι συνολικές καταθέσεις του ήταν υψηλότερες.

Εικόνα 13: Απάντηση Gemini (2)

Στη συνέχεια επιχειρείται η χρήση της ίδιας καταγγελίας και του ίδιου συνόλου δεδομένων στο intfloat/multilingual-e5-base. Η εφαρμογή φέρνει και πάλι τα έξι αποσπάσματα που είναι πιο σχετικά με την καταγγελία. Με τη χρήση του intfloat/multilingual-e5-base παρατηρείται και πάλι ότι το σύστημα εμφανίζει

αποσπάσματα νομοθεσίας που σχετίζονται με την καταγγελία του χρήστη (Εικόνες 14, 15). Το απόσπασμα με ID: 554_5_13.7d εντοπίστηκε και στα αποσπάσματα που εμφανίστηκαν με χρήση του μοντέλου LaBSE. Τονίζει ότι η καταβολή ποσού 800 ευρώ και άνω προς τον παίκτη, πρέπει να γίνεται αφού επιβεβαιωθεί πως ο λογαριασμός πληρωμών ανήκει πραγματικά σε αυτόν. Επιπλέον, το απόσπασμα με ID: 79835_25.4 επισημαίνει πως σε περίπτωση που επιβεβαιωθεί χρήση μέσων πληρωμής που ανήκουν σε τρίτους, απαγορεύεται η χρήση αυτών, το κεφάλαιο που έχει κατατεθεί επιστρέφεται στον δικαιούχο του μέσου πληρωμής και δεν αποδίδονται τυχόν κέρδη.

Το Gemini στην απάντηση που δίνει στον χρήστη καταλήγει ότι δεν δικαιούται ο καταγγέλλων το ποσό των 900 ευρώ (Εικόνα 16). Η συγκεκριμένη απάντηση είναι κοντά στην απάντηση που έδωσε το Gemini όταν χρησιμοποιήθηκε ως embedding model το LaBSE. Στηριζόμενο σε διαφορετικά αποσπάσματα, το LLM αναφέρει πως με βάση το απόσπασμα με ID: 79835_25.4 δεν οφείλει η εταιρεία να επιτρέψει ανάληψη ποσού 900 ευρώ καθώς το μέσο πληρωμής ανήκε σε τρίτο πρόσωπο. Βέβαια, αναγνωρίζει και σε αυτή την περίπτωση την ενδεχόμενη παράλειψη καταβολής ολόκληρου του κατατεθέντος κεφαλαίου.

🔍 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΚΤΗΣΗΣ (E5 + cosine similarity)

1

ID: 554_5_13.7d Cosine similarity: 0.8417 Άρθρο (απόσπασμα): Καθ' όλη τη διάρκεια της Επιχειρηματικής Σχέσης, τα Υπόχρεα Πρόσωπα που διεξάγουν Παίγνια μέσω Ηλεκτρονικού Λογαριασμού Παίκτη, διασφαλίζουν ότι: δ. Καταβολές (withdrawals) προς τους Παίκτες, ποσού οκτακοσίων (800) ευρώ και άνω ανά πράξη, διενεργούνται μόνο εφόσον επιβεβαιωθεί, πριν τη συναλλαγή, ότι ο λογαριασμός πληρωμών που έχει δηλώσει ο Παίκτης ανήκει πραγματικά σε αυτόν. Για τις πληρωμές των Παικτών στην περίπτωση των Τυχερών Παιγνίων που διεξάγονται μέσω παιγνιομηχανημάτων τύπου VLT, εφαρμόζονται τα όρια και οι διαδικασίες της παρ. 15.1.

2

ID: 554_5_13.6 Cosine similarity: 0.8401 Άρθρο (απόσπασμα): Εφόσον ο Ηλεκτρονικός Λογαριασμός Παίκτη κλείσει σύμφωνα με την παρ. 13.3, τα Υπόχρεα Πρόσωπα διευθετούν εκ νέου τις Συμμετοχές που απέδωσαν κέρδη στον Παίκτη, καθ' όλο το χρονικό διάστημα που ο Λογαριασμός είχε προσωρινό χαρακτήρα, με απόδοση τη μονάδα, καταβάλλουν στον Παίκτη, με την επιφύλαξη της περ. δ' της παρ. 13.7, το υπόλοιπο του κεφαλαίου που έχει πιστωθεί στον Ηλεκτρονικό Λογαριασμό σύμφωνα με τα προβλεπόμενα στην παρ. 2 του άρθρου 21 της υπό στοιχεία 79835 ΕΞ 2020/24.7.2020 (Β' 3265) απόφασης του Υπουργού Οικονομικών, και εξετάζουν εάν συντρέχει η λόγος αναφοράς στην Αρχή. Συμμετοχές που δεν απέδωσαν κέρδη δεν επιστρέφονται.

3

ID: 79835_25.4 Cosine similarity: 0.8387 Άρθρο (απόσπασμα): Η κατάθεση (deposit) ποσών για τη Συμμετοχή καθώς και η εκταμίευση πιστωτικού υπολοίπου (withdrawal) του Παίκτη, πραγματοποιούνται υποχρεωτικά μεταξύ του Παίκτη και του Κατόχου της Άδειας, χωρίς τη μεσολάβηση τρίτου, πλην των αναφερόμενων στην παρ. 1 του παρόντος άρθρου Παρόχων Υπηρεσιών Πληρωμών και σύμφωνα με τα προβλεπόμενα στον Κανονισμό Καταπολέμησης Νομιμοποίησης Εσόδων. Σε περίπτωση που εκ των υστέρων επιβεβαιωθεί χρήση μέσων πληρωμής που ανήκουν σε τρίτους, απαγορεύεται η χρήση αυτών, το κεφάλαιο που έχει κατατεθεί επιστρέφεται στον δικαιούχο του μέσου πληρωμής και δεν αποδίδονται τυχόν κέρδη.

Εικόνα 14: intfloat/multilingual-e5-base: Αποσπάσματα νομοθεσίας που είναι πιο κοντά στο περιεχόμενο της καταγγελίας (1)

4

ID: 554_5_13.4a-b Cosine similarity: 0.8377 Άρθρο (απόσπασμα): Για το χρονικό διάστημα που ο Ηλεκτρονικός Λογαριασμός Παίκτη έχει προσωρινό χαρακτήρα σύμφωνα με την παρ. 13.2, τα Υπόχρεα Πρόσωπα: α. Επιτρέπουν καταθέσεις από τον Παίκτη στον Ηλεκτρονικό του Λογαριασμό ποσού έως και οκτακόσια (800) ευρώ, είτε το ποσό αυτό καλύπτεται με μία πράξη είτε με περισσότερες πράξεις. β. Επιτρέπουν τη Συμμετοχή του Παίκτη σε στοιχήματα επί γεγονότων τα οποία προγραμματίζεται να λάβουν χώρα έως και την λήξη της 30ης ημέρας από την ημέρα δημιουργίας του Ηλεκτρονικού του Λογαριασμού ή σε διαδικτυακά παίγνια που διενεργούνται εντός του ίδιου χρονικού διαστήματος. Στοιχήματα επί γεγονότων που αναβλήθηκαν για μετά την λήξη της 30ης ημέρας από την ημέρα δημιουργίας του Ηλεκτρονικού Λογαριασμού, διευθετούνται με απόδοση τη μονάδα.

5

ID: 79835_23.4 Cosine similarity: 0.8360 Άρθρο (απόσπασμα): Δεν επιτρέπεται η επιβολή τελών σε Ηλεκτρονικούς Λογαριασμούς που έχουν τεθεί σε κατάσταση «Φραγής».

6

ID: 79835_7.2 Cosine similarity: 0.8338 Άρθρο (απόσπασμα): Δεν επιτρέπεται η με οποιονδήποτε τρόπο αξιοποίηση των στοιχείων του Ηλεκτρονικού Λογαριασμού Παίκτη για λόγους εμπορικής επικοινωνίας, χωρίς τη συναίνεση του Παίκτη.

Εικόνα 15: intfloat/multilingual-e5-base: Αποσπάσματα νομοθεσίας που είναι πιο κοντά στο περιεχόμενο της καταγγελίας (2)

Ως νομικός σύμβουλος εξειδικευμένος στη νομοθεσία τυχερών παιγνίων, κατόπιν εξέτασης της καταγγελίας σας και των διατάξεων που παρατέθηκαν, παραθέτω την εξής αξιολόγηση:

1. Διαπίστωση Παραβιάσεων

Με βάση τα δεδομένα που παρείχατε, προκύπτουν τα εξής συμπεράσματα:

Α. Σχετικά με την ακύρωση των κερδών (900€): Η εταιρεία **δεν παραβιάζει** τη νομοθεσία στο σκέλος της άρνησης καταβολής των κερδών. Σύμφωνα με το άρθρο **ID: 79835_25.4**, η χρήση μέσων πληρωμής που ανήκουν σε τρίτους απαγορεύεται ρητά. Η ίδια διάταξη ορίζει κατηγορηματικά ότι σε περίπτωση που επιβεβαιωθεί (έστω και εκ των υστέρων) τέτοια χρήση, **«δεν αποδίδονται τυχόν κέρδη»**. Συνεπώς, η απαίτησή σας για καταβολή των κερδών (σενάριο α) δεν εδράζεται στο νόμο.

Β. Σχετικά με την επιστροφή των καταθέσεων (κεφαλαίου): Εδώ εντοπίζεται **πιθανή παραβίαση** των υποχρεώσεων της εταιρείας. Το άρθρο **ID: 79835_25.4** ορίζει ότι σε περίπτωση χρήσης μέσου πληρωμής τρίτου, **«το κεφάλαιο που έχει κατατεθεί επιστρέφεται στον δικαιούχο του μέσου πληρωμής»**.

- Η εταιρεία επέστρεψε μόνο 300€, ενώ παραδέχεστε ότι «κάποιες καταθέσεις» (πληθυντικός) έγιναν από τον λογαριασμό του συζύγου σας.
- Ο νόμος δεν διακρίνει μεταξύ της «τελευταίας» κατάθεσης και των προηγούμενων. Εφόσον διαπιστωθεί η χρήση τρίτου προσώπου, **όλο το κεφάλαιο** που προήλθε από αυτόν τον λογαριασμό πρέπει να επιστραφεί στον δικαιούχο (τον σύζυγό σας) και όχι να παρακρατηθεί από την εταιρεία, εφόσον τα αντίστοιχα κέρδη ακυρώνονται.

Γ. Σχετικά με τον έλεγχο κατά την ανάληψη: Σύμφωνα με το άρθρο **ID: 554_5_13.7d**, για αναλήψεις άνω των **800€**, η εταιρεία οφείλει υποχρεωτικά να επιβεβαιώσει ότι ο λογαριασμός πληρωμών ανήκει στον παίκτη πριν τη συναλλαγή. Η εταιρεία συμμορφώθηκε με αυτόν τον έλεγχο τη στιγμή που ζητήσατε τα 900€, γεγονός που οδήγησε στην αποκάλυψη της παράβασης.

2. Σχετικά Άρθρα Νομοθεσίας

- **ID: 79835_25.4:** Είναι το βασικό άρθρο που εφαρμόζεται στην περίπτωσή σας. Προβλέπει ότι η συναλλαγή πρέπει να γίνεται χωρίς μεσολάβηση τρίτου. Η χρήση μέσων τρίτου επιφέρει: (1) απαγόρευση χρήσης τους, (2) επιστροφή του κεφαλαίου στον δικαιούχο, (3) μη απόδοση κερδών.
- **ID: 554_5_13.6:** Αν και αναφέρεται σε κλείσιμο λογαριασμού, ενισχύει τη λογική ότι σε περιπτώσεις μη κανονικότητας, οι συμμετοχές διευθετούνται με απόδοση τη μονάδα (δηλαδή επιστροφή πονταρίσματος/κεφαλαίου) και τα κέρδη δεν αποδίδονται.
- **ID: 554_5_13.7d:** Καθορίζει το όριο των 800€ ως το σημείο υποχρεωτικού ελέγχου ταυτοποίησης του λογαριασμού πληρωμών.

3. Σύνοψη - Συμπέρασμα

Η εταιρεία Testbet:

1. **Ορθώς ακύρωσε τα κέρδη των 900€**, καθώς η χρήση του λογαριασμού του συζύγου σας αποτελεί παράβαση του άρθρου **79835_25.4**.
2. **Οφείλει να επιστρέψει το σύνολο του κεφαλαίου** που κατατέθηκε από τον λογαριασμό του συζύγου σας και όχι μόνο τα τελευταία 300€, βάσει του ιδίου άρθρου (**79835_25.4**), εφόσον ακυρώνει τη δραστηριότητα που παρήχθη από αυτά τα χρήματα.

Σύσταση: Μπορείτε να διεκδικήσετε την επιστροφή όλων των κατατεθειμένων ποσών που προήλθαν από τον λογαριασμό του συζύγου σας (τα οποία δεν έχουν ήδη αναληφθεί), επικαλούμενη ότι η διάταξη 79835_25.4 επιβάλλει την επιστροφή του κεφαλαίου στον δικαιούχο χωρίς εξαιρέσεις για προγενέστερες καταθέσεις.

Εικόνα 16: intfloat/multilingual-e5-base: Απάντηση Gemini

Συγκρίνοντας τα δύο embeddings models εξάγεται το συμπέρασμα πως είναι κατάλληλα για την ανάκτηση σχετικών αποσπασμάτων νομοθεσίας που αφορούν καταγγελία για τα τυχερά παίγνια. Και τα δύο μοντέλα ανέκτησαν ως κορυφαίο αποτέλεσμα το απόσπασμα με ID: 554_5_13.7d, το οποίο συνδέεται άμεσα με αναλήψεις άνω των 800 ευρώ και την υποχρέωση επιβεβαίωσης ότι ο λογαριασμός πληρωμών ανήκει στον παίκτη. Το LaBSE επέστρεψε επιπλέον πολύ στοχευμένες παραγράφους για χρήση μέσων πληρωμής τρίτων (ID: 554_5_13.7e) ενώ το intfloat/multilingual-e5-base ανέκτησε αντίστοιχες διατάξεις μέσω πιο γενικών άρθρων (ID: 79835_25.4).

4.2 Παρουσίαση Εφαρμογής Εύρεσης Παρόμοιων Καταγγελιών.

Στη συγκεκριμένη εφαρμογή ο χρήστης εισάγει μία καταγγελία και το σύστημα αντλώντας δεδομένα από περιλήψεις καταγγελιών που έχουν υποβληθεί στο παρελθόν, εμφανίζει πέντε περιλήψεις καταγγελιών που είναι νοηματικά πιο κοντά στην καταγγελία που εισήγαγε ο χρήστης.

Στο πείραμα που παρουσιάζεται στη συνέχεια, αξιοποιείται το μοντέλο LaBSE και ένα σύνολο δεδομένων που περιλαμβάνει περιλήψεις καταγγελιών, που η κάθε περίληψη δεν ξεπερνά τις πενήντα λέξεις. Όταν ο χρήστης εισάγει την καταγγελία, το σύστημα δημιουργεί μία περίληψη με χρήση του Gemini μέχρι 50 λέξεις. Το prompt που δίνεται για τη δημιουργία της περίληψης διευκρινίζει το όριο των λέξεων, να είναι σε γ' ενικό πρόσωπο, να μην περιλαμβάνονται προσωπικά στοιχεία και να βασίζεται μόνο στην καταγγελία (Εικόνα 17).

Το επόμενο βήμα είναι η εφαρμογή να δημιουργήσει την περίληψη. Στη συνέχεια ανακτώνται οι πέντε πιο σχετικές περιλήψεις καταγγελιών από το FAISS, με βάση τη σημασιολογική ομοιότητα των embeddings της περίληψης που παρήγαγε το Gemini και των αποθηκευμένων περιλήψεων. Η αρχική περίληψη και οι παρόμοιες περιλήψεις εισάγονται στο Gemini το οποίο θα επιστρέψει τις ομοιότητες και τις διαφορές.

Από τα αποτελέσματα εξάγεται το συμπέρασμα πως με τη χρήση του μοντέλου LaBSE ανακτήθηκαν παρόμοιες περιλήψεις καταγγελιών σε σύγκριση με την περίληψη που δημιούργησε το LLM της εφαρμογής (Εικόνα 18). Οι παρόμοιες καταγγελίες είναι νοηματικά κοντά με την αρχική περίληψη καθώς οι περισσότερες περιγράφουν θέματα όπως παρακράτηση κερδών, την άρνηση των παικτών ότι παραβίασαν τους όρους που ισχυρίζεται η εταιρεία και ότι οι πράξεις της εταιρείας δεν συνοδεύονται από ικανοποιητικές εξηγήσεις.

Επιπλέον, η παρουσίαση των ομοιοτήτων και των διαφορών ανάμεσα στην αρχική περίληψη και τις παρόμοιες περιλήψεις βασίζεται στο prompt που δίνεται στο LLM, το οποίο παρουσιάζεται στην Εικόνα 19. Πιο συγκεκριμένα, ζητείται από το LLM να εντοπίσει πέντε κοινά μοτίβα που εμφανίζονται σε τουλάχιστον δύο από τις παρόμοιες καταγγελίες. Για κάθε μοτίβο να δώσει μία σύντομη περιγραφή, λίστα με

IDs, λέξεις κλειδιά. Επίσης, ζητείται να εντοπίσει τρεις διαφορές που κάνουν τη νέα καταγγελία να ξεχωρίζει σε σχέση με το σύνολο των παρόμοιων. Για κάθε διαφορά να δώσει σύντομη περιγραφή και λίστα με IDs. Εάν δεν μπορέσει να κάνει κάτι από τα παραπάνω να εμφανίσει ανάλογο μήνυμα. Επιπλέον, να μην συμπεριλάβει προσωπικά δεδομένα και να βασιστεί μόνο στα δοθέντα δεδομένα. Αν βρει λιγότερα από πέντε μοτίβα ή λιγότερες από τρεις διαφορές, να επιστρέψει όσα βρήκε. Με αυτό τον τρόπο το LLM παρουσιάζει τις διαφορές και τις ομοιότητες με τρόπο απλό και κατανοητό, παρέχοντας στον ελεγκτή χρήσιμη πληροφόρηση για την εξέταση παρόμοιων υποθέσεων. Οι ομοιότητες και διαφορές του συγκεκριμένου πειράματος, παρουσιάζονται στην Εικόνα 20.

```
prompt = (
    "Δώσε περίληψη του παρακάτω κειμένου στα ελληνικά, μέγιστο 50 λέξεις, σε γ' πρόσωπο, "
    "και ξεκίνα με τη φράση: 'Η καταγγελία αναφέρεται'. "
    "Μην αναφέρεις προσωπικά στοιχεία ούτε του καταγγέλλοντος ούτε του καταγγελλόμενου. "
    "Να βασιστείς μόνο στα δεδομένα του κειμένου που σου στέλνω. "
    "Ο καταγγέλων να αναφέρεται με την λέξη χρήστης.\n\n"
    f"Κείμενο:\n{complaint_text}"
)
```

Εικόνα 17: Prompt για δημιουργία περίληψης στο Gemini

Περίληψη νέας καταγγελίας: Η καταγγελία αναφέρεται στην ακύρωση ανάληψης κερδών ύψους 840 ευρώ από αδειοδοτημένη εταιρεία. Ο χρήστης κέρδισε το ποσό παίζοντας καζίνο, αλλά η εταιρεία αρνείται την καταβολή, επικαλούμενη αβάσιμα παραβίαση όρων χρήσης (άρθρο 6.4.1). Ο χρήστης διαψεύδει τις κατηγορίες και ζητά τη δίκαιη διερεύνηση της υπόθεσης. (42 λέξεις)

Top 5 πιο σχετικές καταγγελίες με τη νέα:

1. ID: 40423 | Similarity: 0.7396 Κείμενο: Η καταγγελία αναφέρεται σε αυθαίρετη παρακράτηση κερδών από στοιχηματική εταιρεία, η οποία κλείνει τον λογαριασμό χρήστη χωρίς εξηγήσεις, επικαλούμενη παραβιάσεις όρων. Ο χρήστης δηλώνει ότι η δραστηριότητά του ήταν νόμιμη και ζητά την επιστροφή των κερδών και του κεφαλαίου του. Περιμένει ενέργειες από τις αρχές.
2. ID: 40975 | Similarity: 0.7298 Κείμενο: Η καταγγελία αναφέρεται σε περιστατικό όπου παίκτης κέρδισε από φρουτάκι σε καζίνο, αλλά η εταιρεία αρνήθηκε να πιστώσει τα κέρδη του. Παρά την αίτηση ανάληψης, τα χρήματα δεν επιστράφηκαν και η εταιρεία χρησιμοποιεί αθέμιτες πρακτικές, παραβιάζοντας τους κανόνες.
3. ID: 41047 | Similarity: 0.7285 Κείμενο: Η καταγγελία αναφέρεται σε παράνομη και καταχρηστική κατακράτηση υπολοίπου λογαριασμού από εταιρεία στοιχημάτων. Ο χρήστης υποστηρίζει ότι δεν παραβίασε τους όρους χρήσης και ότι τα κέρδη του προήλθαν από νόμιμες δραστηριότητες. Ζητά την επιστροφή του κατακρατηθέντος ποσού.
4. ID: 111881 | Similarity: 0.7182 Κείμενο: Η καταγγελία αναφέρεται σε παράνομη απενεργοποίηση λογαριασμού από στοιχηματική εταιρεία, μετά από αίτημα για επιπλέον έγγραφα ταυτοποίησης. Ο χρήστης δεν έλαβε πειστική εξήγηση για την παραβίαση όρων και ζητά την αποκατάσταση του ποσού που αφαιρέθηκε, θεωρώντας την ενέργεια της εταιρείας αντισυμβατική.
5. ID: 71302 | Similarity: 0.7160 Κείμενο: Η καταγγελία αναφέρεται στην απενεργοποίηση λογαριασμού από στοιχηματική εταιρεία λόγω φερόμενης παραβίασης όρων. Ο καταγγέλλων ζητά την επιστροφή του υπολοίπου, υποστηρίζοντας ότι η παρακράτηση κερδών είναι παράνομη. Η εταιρεία αρνείται την επιστροφή, προκαλώντας περιουσιακή ζημία στον καταγγέλλοντα και μη παρέχοντας απαντήσεις στις αιτιάσεις του.

Εικόνα 18: Η περίληψη που δημιούργησε το Gemini και οι παρόμοιες καταγγελίες που ανακτήθηκαν

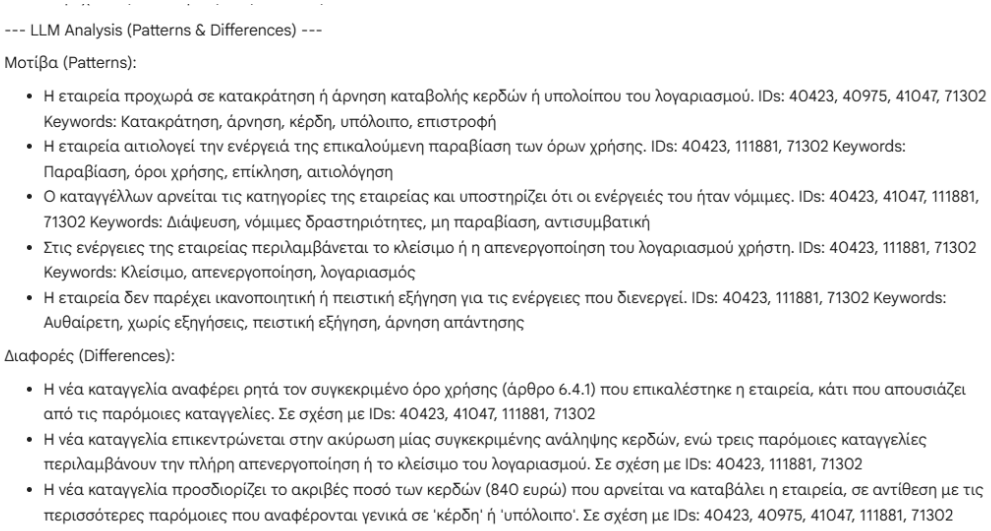
```
 prompt = f"""
Είσαι αναλυτής καταγγελιών. Θα σου δώσω:
(α) Περίληψη νέας καταγγελίας
(β) Περιλήψεις από 5 παρόμοιες καταγγελίες με IDs

Στόχος:
1) Εντόπισε ΕΩΣ 5 κοινά μοτίβα (patterns) που εμφανίζονται σε τουλάχιστον 2 από
τις παρόμοιες καταγγελίες.
   - Για κάθε μοτίβο, δώσε:
     * "pattern": σύντομη περιγραφή (1 πρόταση)
     * "supporting_ids": λίστα με IDs (τουλάχιστον 2 IDs)
     * "keywords": 3-6 λέξεις-κλειδιά (προαιρετικά αλλά επιθυμητά)
   - Αν δεν μπορείς να βρεις κανένα μοτίβο με evidence (>=2 IDs), τότε το
   "patterns" να είναι ακριβώς: "Δεν εντοπίστηκε".

2) Εντόπισε ΕΩΣ 3 διαφορές που κάνουν τη νέα καταγγελία να ξεχωρίζει σε σχέση
με το σύνολο των παρόμοιων.
   - Για κάθε διαφορά, δώσε:
     * "difference": σύντομη περιγραφή (1 πρόταση)
     * "reference_ids": λίστα IDs (1 ή περισσότερα) σε σχέση με τα οποία
     προκύπτει η διαφορά
   - Αν δεν εντοπίζονται διαφορές, τότε το "differences" να είναι ακριβώς: "Δεν
   εντοπίστηκε".

Κανόνες:
- ΜΗΝ συμπεριλάβεις προσωπικά δεδομένα (ονόματα, emails, τηλέφωνα, διευθύνσεις,
λογαριασμούς).
- ΜΗΝ κάνεις υποθέσεις. Μόνο ό,τι τεκμηριώνεται από τα κείμενα που σου δίνω.
- Κάθε μοτίβο ΠΡΕΠΕΙ να έχει τουλάχιστον 2 IDs που το στηρίζουν. Αλλιώς μην το
συμπεριλάβεις.
- Αν βρεις λιγότερα από 5 μοτίβα ή λιγότερες από 3 διαφορές, επέστρεψε όσα
βρήκες.
```

Εικόνα 19: Prompt για την εύρεση ομοιοτήτων και διαφορών



Εικόνα 20: Αποτελέσματα ομοιοτήτων και διαφορών που επέστρεψε το LLM

Προκειμένου να δούμε τις δυνατότητες του μοντέλου intfloat/multilingual-e5-base, χρησιμοποιήθηκε η ίδια περίληψη, το ίδιο σύνολο δεδομένων και τα ίδια prompts. Παρατηρείται ότι το συγκεκριμένο μοντέλο εντοπίζει παρόμοιες περιλήψεις καταγγελιών. Οι περιλήψεις καταγγελιών που ανακτώνται περιγράφουν θέματα ακύρωσης ανάληψης κερδών, παραβίαση όρων που ο παίκτης αρνείται, κέρδη που προέρχονται από δραστηριότητα καζίνο (Εικόνες 21, 22).

Περίληψη νέας καταγγελίας: Η καταγγελία αναφέρεται στην ακύρωση ανάληψης κερδών ύψους 840 ευρώ από αδειοδοτημένη εταιρεία. Ο χρήστης κέρδισε το ποσό παίζοντας καζίνο, αλλά η εταιρεία αρνείται την καταβολή, επικαλούμενη αβάσιμα παραβίαση όρων χρήσης (άρθρο 6.4.1). Ο χρήστης διαψεύδει τις κατηγορίες και ζητά τη δίκαιη διερεύνηση της υπόθεσης. (42 λέξεις)

Top 5 πιο σχετικές καταγγελίες με τη νέα (χρήση intfloat/multilingual-e5-base):

1. ID: 91673 | Similarity: 0.9053 Κείμενο: Η καταγγελία αναφέρεται σε φραγή λογαριασμού και ακύρωση αιτήματος ανάληψης κερδών από παιχνίδι καζίνο. Ο χρήστης επικοινώνησε επανειλημμένα με την εταιρία, ζητώντας διευκρινίσεις, αλλά δεν έλαβε ικανοποιητική απάντηση. Δηλώνει δυσαρεστημένος και προγραμματίζει να ενημερώσει τις αρμόδιες αρχές για την κατάσταση.
2. ID: 40246 | Similarity: 0.9033 Κείμενο: Η καταγγελία αναφέρεται σε καθυστέρηση και ακύρωση αναλήψεων κερδών από καζίνο, μετά από κατάθεση και λήψη μπόνους. Ο χρήστης υποστηρίζει ότι δεν παραβίασε τους όρους, ενώ η εταιρεία επικαλείται κατάχρηση της προσφοράς. Ο χρήστης εκφράζει αγανάκτηση και ζητά βοήθεια για την αποκατάσταση των κερδών του.
3. ID: 40147 | Similarity: 0.8967 Κείμενο: Η καταγγελία αναφέρεται σε προβλήματα που προέκυψαν κατά τη διαδικασία ανάληψης κερδών από καζίνο. Ο χρήστης αντιμετώπισε καθυστερήσεις και απαιτήσεις για ταυτοποίηση, ενώ το καζίνο απέρριψε την ανάληψη λόγω χρήσης κοινών λογαριασμών. Ο χρήστης εκφράζει την απογοήτευσή του για την αδιαφορία του καζίνο και ζητάει διερεύνηση της υπόθεσης.
4. ID: 40134 | Similarity: 0.8954 Κείμενο: Η καταγγελία αναφέρεται σε προβλήματα που προέκυψαν κατά τη διάρκεια της εγγραφής σε διαδικτυακό καζίνο. Ο χρήστης αντιμετώπισε αφαίρεση κερδών και απενεργοποίηση λογαριασμού, παρά την αποδοχή καταθέσεων από τρίτο πρόσωπο. Ζητήθηκε επανεξέταση της καταγγελίας, αλλά η άρνηση επιστροφής των κερδών συνεχίζεται, οδηγώντας σε καταγγελία στην αρμόδια αρχή.
5. ID: 40768 | Similarity: 0.8937 Κείμενο: Η καταγγελία αναφέρεται σε προβλήματα που προέκυψαν κατά τη διαδικασία ανάληψης κερδών από στοιχηματική εταιρεία. Ο χρήστης ισχυρίζεται ότι η ανάληψη ακυρώθηκε και τα κέρδη του αφαιρέθηκαν, ενώ η εταιρεία επικαλέστηκε παραβίαση κανόνων. Ο χρήστης διαφωνεί με τους ισχυρισμούς και καταγγέλλει αδικαιολόγητες πρακτικές.

Εικόνα 21: Η περίληψη που είχε δημιουργηθεί στην LaBSE έκδοση της εφαρμογής και οι παρόμοιες καταγγελίες που ανακτήθηκαν με χρήση του μοντέλου intfloat/multilingual-e5-base

--- LLM Analysis (Patterns & Differences) ---

Μοτίβα (Patterns):

- Οι καταγγελίες αφορούν την ακύρωση αιτημάτων ανάληψης ή την αφαίρεση κερδών από τον πάροχο τυχερών παιγνίων. IDs: 91673, 40246, 40134, 40768 Keywords: ακύρωση ανάληψης, αφαίρεση κερδών, άρνηση πληρωμής
- Η εταιρεία δικαιολογεί την ενέργεια της επικαλούμενη παραβίαση συγκεκριμένων όρων χρήσης ή κανονισμών λειτουργίας. IDs: 40246, 40147, 40134, 40768 Keywords: παραβίαση όρων, κανόνες, κατάχρηση, κοινοί λογαριασμοί
- Οι χρήστες δηλώνουν την πρόθεση ή έχουν ήδη προβεί σε καταγγελία στις αρμόδιες ρυθμιστικές αρχές. IDs: 91673, 40134 Keywords: αρμόδιες αρχές, καταγγελία, ενημέρωση
- Τα κέρδη που αποτέλεσαν αντικείμενο διαφωνίας προέρχονταν από δραστηριότητα σε διαδικτυακό καζίνο. IDs: 91673, 40246, 40147, 40134 Keywords: καζίνο, διαδικτυακό, παιχνίδι
- Σε ορισμένες περιπτώσεις, αναφέρονται καθυστερήσεις ή επιπλέον απαιτήσεις κατά τη διαδικασία ανάληψης των κερδών. IDs: 40246, 40147 Keywords: καθυστέρηση, ανάληψη, ταυτοποίηση

Διαφορές (Differences):

- Η νέα καταγγελία προσδιορίζει το ακριβές ποσό των κερδών (840 ευρώ) που ακυρώθηκαν. Σε σχέση με IDs: 91673, 40246, 40147, 40134, 40768
- Η εταιρεία επικαλείται ρητά ένα συγκεκριμένο αριθμημένο άρθρο (6.4.1) των όρων χρήσης για την ακύρωση. Σε σχέση με IDs: 40246, 40147, 40768
- Η νέα καταγγελία αναφέρει ρητά ότι η εμπλεκόμενη εταιρεία είναι «αδειοδοτημένη» (υπονοώντας νομική λειτουργία). Σε σχέση με IDs: 91673, 40246, 40147, 40134, 40768

Εικόνα 22: Οι ομοιότητες και διαφορές ανάμεσα στην αρχική περίληψη και στις παρόμοιες περιλήψεις καταγγελιών που ανακτήθηκαν με χρήση του μοντέλου intfloat/multilingual-e5-base

Συμπερασματικά, και τα δύο μοντέλα κατάφεραν να εντοπίσουν παρόμοιες περιλήψεις καταγγελιών. Το LaBSE έδωσε περιλήψεις που μοιάζουν στη θεματική με το περιεχόμενο που πραγματεύεται η αρχική περίληψη (παρακράτηση κερδών, παραβίαση όρων) ενώ το intfloat/multilingual-e5-base έδωσε περιπτώσεις που είναι πολύ κοντά στο ακριβές σενάριο που περιγράφει η αρχική περίληψη (ακύρωση ανάληψης κερδών, παικτική δραστηριότητα σε καζίνο).

4.3 Παρουσίαση Εφαρμογής Chatbot

Η υλοποίηση ενός chatbot αποτελεί την τρίτη εφαρμογή της παρούσας διπλωματικής. Μέσα από το chatbot ο χρήστης έχει τη δυνατότητα να λαμβάνει πληροφορίες για θέματα που αφορούν τις άδειες καταλληλότητας. Πιο συγκεκριμένα, μπορεί να λαμβάνει απαντήσεις σε ερωτήματα που αφορούν διοικητικό και λοιπό προσωπικό καζίνο, εισαγωγείς, προμηθευτές, κατασκευαστές, junket operators, μετόχους, πρόσωπα που επιτελούν κρίσιμες λειτουργίες, οργανισμούς πιστοποίησης, τεχνικούς, συνεργάτες-affiliates.

Όταν εισάγει ένα ερώτημα στην εφαρμογή γίνεται έλεγχος μέσα στο ερώτημα, εάν υπάρχει λέξη που να ταιριάζει με κάποια από τις λέξεις του λεξικού. Οι λέξεις του λεξικού παραπέμπουν σε κατηγορίες προσώπων που δύναται να λάβουν άδεια καταλληλότητας (για παράδειγμα τεχνικοί, κατασκευαστές). Εφόσον εντοπιστεί μία λέξη στο ερώτημα που να ταιριάζει με λέξη που υπάρχει στο λεξικό, η κατηγορία στην οποία παραπέμπει δύναται να χρησιμοποιηθεί ως metadata filter στην ChromaDB. Η αναζήτηση ομοιοτήτων (string similarity) ανάμεσα σε λέξεις του ερωτήματος και σε λέξεις του λεξικού γίνεται όπως αναφέρθηκε και στην ενότητα 3.4 με χρήση της βιβλιοθήκης rapidfuzz, προκειμένου να προβλεφθούν περιπτώσεις που ο χρήστης μπορεί να έχει κάνει κάποιο ορθογραφικό λάθος.

Αφού γίνει ο «καθαρισμός» του ερωτήματος όπως περιγράφεται στην ενότητα 3.4, αφαιρούνται stopwords, και δημιουργούνται ακολουθίες n διαδοχικών λέξεων (n-grams), για n = 1 έως 3. Για παράδειγμα το ερώτημα «Τί ισχύει για παράβολα συνεργατών» μπορεί να διαιρεθεί στις φράσεις-λέξεις: «ισχυει παραβολα συνεργατων», «ισχυει παραβολα», «παραβολα συνεργατων» «παραβολα»,

«συνεργατων», «ισχυει». Κάθε φράση που προκύπτει ελέγχεται πόσο κοντά είναι με κάθε λέξη του λεξικού και υπολογίζεται ένα score 0-100 (χρήση βιβλιοθήκης rapidfuzz). Όσο πιο υψηλό είναι το score, τόσο μεγαλύτερη ομοιότητα υπάρχει στα συγκρινόμενα μέρη. Η εφαρμογή κρατάει μόνο όσα συγκρινόμενα μέρη έχουν σκορ μεγαλύτερο ή ίσο του 85. Στη συνέχεια δίνει βάρος στις φράσεις με τις περισσότερες λέξεις (1-gram: x 1.00, 2-gram: x 1.08, 3-gram: x 1.15) προκειμένου να αποφευχθούν περιπτώσεις ταύτισης με κατηγορίες οι οποίες δεν σχετίζονται με το ερώτημα του χρήστη.

Έπειτα παίρνει την κατηγορία της οποίας η λέξη συγκρινόμενη με μία φράση απέκτησαν το μεγαλύτερο score. Μετά εφαρμόζεται περιορισμός σύμφωνα με τον οποίο γίνονται δεκτές άλλες κατηγορίες αρκεί η λέξη την οποία εμπεριέχει η συγκεκριμένη κατηγορία, συγκρινόμενη με μία φράση να έχει συγκεντρώσει σε βαθμολογία τουλάχιστον το 82% της κατηγορίας με το υψηλότερο score. Τέλος, οι κατηγορίες που περνούν και τον συγκεκριμένο περιορισμό κατατάσσονται σε Confident και Ambiguous. Confident σημαίνει ότι η συγκεκριμένη κατηγορία θα χρησιμοποιηθεί σίγουρα ως metadata filter στην ChromaDB και ισχύει για όσες κατηγορίες επιλέχθηκαν μετά από σύγκριση λέξης που περιείχαν, με φράση η οποία αποτελείτο από δύο λέξεις και πάνω. Όλες οι υπόλοιπες κατηγορίες θεωρούνται Ambiguous, δηλαδή θα χρειαστεί η επιβεβαίωση του χρήστη για να χρησιμοποιηθούν ως metadata filter στην ChromaDB. Εξαίρεση αποτελεί όταν μία λέξη από το ερώτημα του χρήστη συγκριθεί με λέξη από το λεξικό και συγκεντρώσει score μεγαλύτερο από 92, τότε η κατηγορία της συγκεκριμένης λέξης θεωρείται Confident.

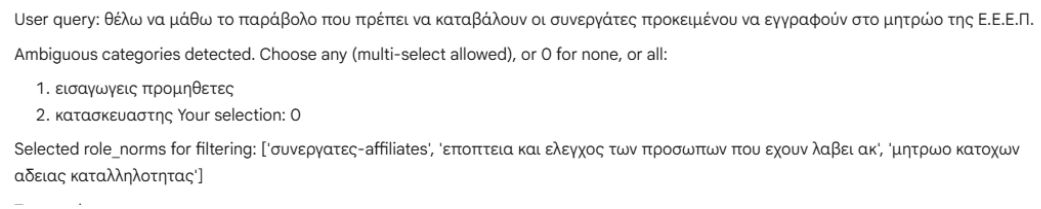
Στο πείραμα που ακολουθεί, ο χρήστης χρησιμοποιεί το chatbot αξιοποιώντας το μοντέλο LaBSE. Το σύνολο δεδομένων που χρησιμοποιείται περιέχει chunks που αποτελούνται το πολύ από 250 tokens. Το LLM μοντέλο που παράγει την τελική απάντηση είναι το gemini-3-flash-preview. Ο χρήστης εισάγει το ερώτημα όπως παρουσιάζεται στην Εικόνα 23, «θέλω να μάθω το παράβολο που πρέπει να καταβάλουν οι συνεργάτες προκειμένου να εγγραφούν στο μητρώο της Ε.Ε.Ε.Π.». Από το mapping που γίνεται του ερωτήματος με το λεξικό προκύπτουν οι ακόλουθες κατηγορίες οι οποίες θα χρησιμοποιηθούν ως metadata filters: «συνεργατες-affiliates», «εποπτεια και ελεγχος των προσωπων που εχουν λαβει ακ», «μητρωο

κατοχων αδειας καταλληλοτητας», «εισαγωγεις προμηθετες», «κατασκευαστης». Οι τρεις πρώτες κατηγορίες θεωρούνται Confident και οι υπόλοιπες Ambiguous. Στον χρήστη εμφανίζεται ανάλογο prompt στο οποίο πρέπει να επιλέξει εάν θέλει κάποια από τις Ambiguous κατηγορίες να συμπεριληφθεί ως metadata filter μαζί με τις Confident (έχει τη δυνατότητα να επιλέξει όλες τις κατηγορίες ή μία κατηγορία ή καμία). Στο πείραμα που διεξάγεται ο χρήστης δεν επιλέγει κάποια από τις Ambiguous κατηγορίες καθώς δεν έχουν άμεση σχέση με το ερώτημά του. Το σύστημα χρησιμοποιεί τις Confident κατηγορίες ως metadata filters στην ChromaDB και απαντά φέρνοντας τα δέκα πιο σχετικά αποσπάσματα της νομοθεσίας με βάση το ερώτημά του.

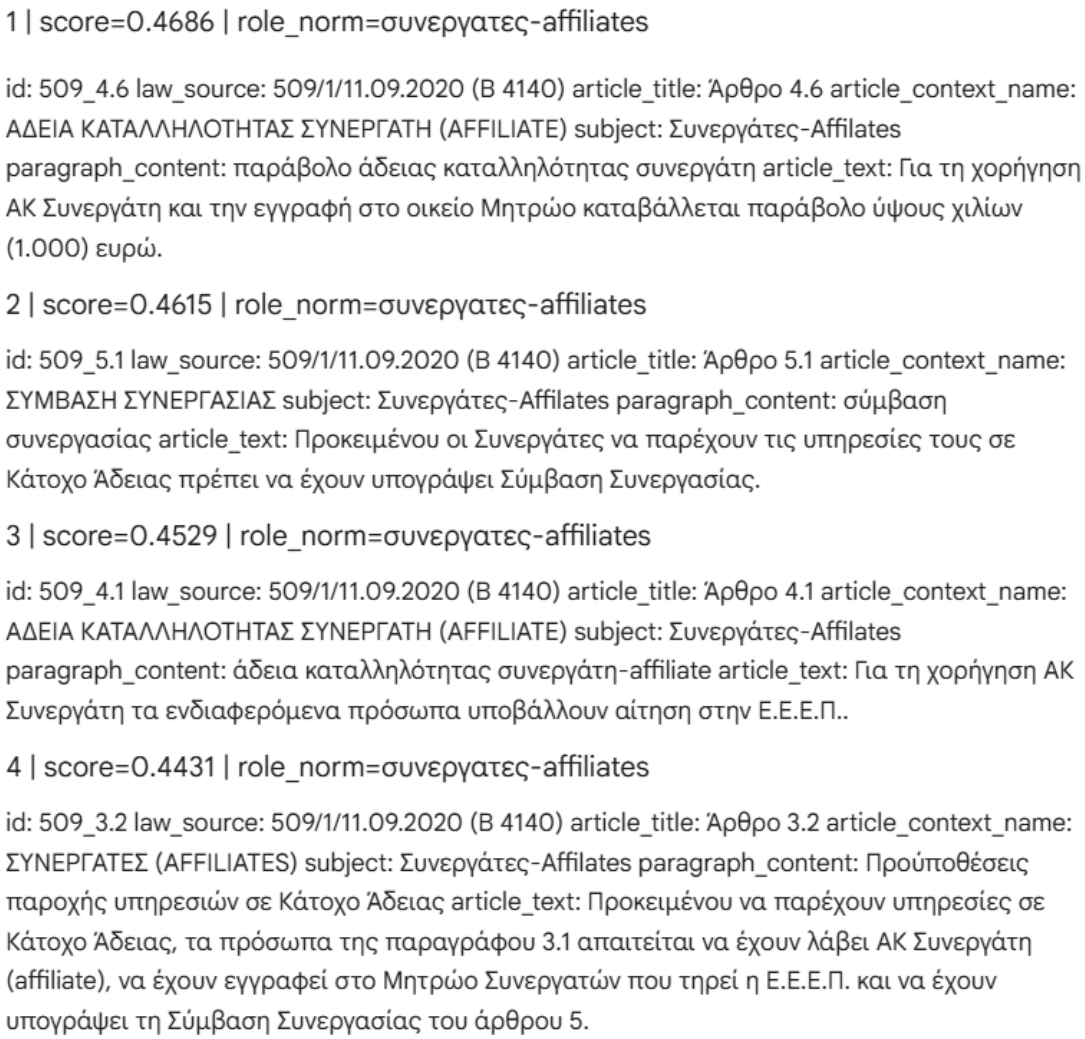
Αξίζει να σημειωθεί πως τα αποσπάσματα που παρουσιάζονται στον χρήστη (Εικόνες 24, 25, 26) είναι πλήρως αναλυτικά καθώς απεικονίζουν πληροφορίες για το similarity score (δηλαδή πόσο κοντά νοηματικά είναι με το ερώτημα που έθεσε), το role_norm στο οποίο έχουν ταξινομηθεί στο σύνολο δεδομένων, το id, το law_source, το article_title, το article_context_name, το subject και το paragraph_content. Με αυτό τον τρόπο ο χρήστης έχει τη δυνατότητα να έχει αναλυτική πληροφόρηση για το κάθε απόσπασμα που ανακτάται από την ChromaDB. Στο πείραμα στο οποίο γίνεται αναφορά, γίνεται αντιληπτό ότι το σχετικό απόσπασμα που απαντά στο ερώτημα του χρήστη εμφανίζεται πρώτο (Για τη χορήγηση ΑΚ Συνεργάτη και την εγγραφή στο οικείο Μητρώο καταβάλλεται παράβολο ύψους χιλίων (1.000) ευρώ), πράγμα το οποίο αποδεικνύει ότι το LaBSE κατάφερε να εντοπίσει το κατάλληλο απόσπασμα με μεγάλη ακρίβεια. Επιπλέον, τα υπόλοιπα αποσπάσματα που παρουσιάζονται, εμφανίζουν λέξεις που συνδέονται νοηματικά με το ερώτημα του χρήστη.

Στη συνέχεια τα αποσπάσματα της νομοθεσίας και το ερώτημα του χρήστη προωθούνται στο LLM και δημιουργείται μία τεκμηριωμένη απάντηση στον χρήστη με αναφορές στην νομοθεσία. Προκειμένου το LLM να απαντήσει βασιζόμενο στην νομοθεσία με παραπομπές, του έχει δοθεί σχετικό prompt οδηγιών. Επιπλέον, έχει ρυθμιστεί η παράμετρος temperature να έχει τιμή 0.2, η οποία ελέγχει το βαθμό τυχαιότητας ή δημιουργικότητας στις απαντήσεις του Gemini. Αυτό οδηγεί σε απαντήσεις που είναι πιο ακριβείς, τυπικές και σταθερές. Η απάντηση του Gemini επιβεβαιώνει (Εικόνα 27) ότι το παράβολο για να λάβει άδεια καταλληλότητας ένας

συνεργάτης είναι 1000 ευρώ. Η συγκεκριμένη απάντηση συνοδεύεται από παραπομπές στα αποσπάσματα της νομοθεσίας.



Εικόνα 23: Ερώτημα του χρήστη και επιλογή Ambiguous κατηγοριών



Εικόνα 24: Εμφάνιση αποσπασμάτων νομοθεσίας με χρήση του LaBSE (1)

5 | score=0.4136 | role_norm=συνεργατες-affiliates

id: 509_6.1 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 6.1 article_context_name: ΕΜΠΟΡΙΚΗ ΕΠΙΚΟΙΝΩΝΙΑ subject: Συνεργάτες-Affilates paragraph_content: εμπορική επικοινωνία για τον συνεργάτη article_text: Ο Συνεργάτης οφείλει να συμμορφώνεται με τις διατάξεις περί Εμπορικής Επικοινωνίας του ν. 4002/2011 (Α' 180) και του Κανονισμού Παιγνίων.

6 | score=0.4015 | role_norm=συνεργατες-affiliates

id: 509_5.5 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 5.5 article_context_name: ΣΥΜΒΑΣΗ ΣΥΝΕΡΓΑΣΙΑΣ subject: Συνεργάτες-Affilates paragraph_content: υποχρεώσεις κατόχου άδειας που έχει συνάψει σύμβαση με συνεργάτη article_text: Ο Κάτοχος Άδειας γνωστοποιεί, αμελλητί, στην Ε.Ε.Ε.Π. τη Σύμβαση Συνεργασίας και κάθε τροποποίηση αυτής, με την υπογραφή της, καθώς και τον λογαριασμό που τηρεί ο Συνεργάτης ως μοναδικός δικαιούχος σε Πάροχο Υπηρεσιών Πληρωμών και στον οποίο εκκαθαρίζονται οι μεταξύ τους συναλλαγές στο πλαίσιο της Σύμβασης.

7 | score=0.3793 | role_norm=συνεργατες-affiliates

id: 509_7.3 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 7.3 article_context_name: ΤΗΡΗΣΗ ΜΗΤΡΩΟΥ ΚΑΤΟΧΩΝ ΑΔΕΙΑΣ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ ΑΠΟ ΤΟΝ ΚΑΤΟΧΟ ΑΔΕΙΑΣ subject: Συνεργάτες-Affilates paragraph_content: έλεγχος τήρησης προϋποθέσεων εκδοσης άδειας καταλληλότητς Συνεργάτη από τον κάτοχο article_text: Ο Κάτοχος Άδειας υποχρεούται στον έλεγχο τήρησης των προϋποθέσεων έκδοσης ΑΚ Συνεργάτη καθόλη τη διάρκεια ισχύος της συμβατικής σχέσης.

8 | score=0.3786 | role_norm=συνεργατες-affiliates

id: 509_5.6 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 5.6 article_context_name: ΣΥΜΒΑΣΗ ΣΥΝΕΡΓΑΣΙΑΣ subject: Συνεργάτες-Affilates paragraph_content: υποχρεώσεις κατόχου άδειας σε περίπτωση άρσης άδειας καταλληλότητας συνεργάτη article_text: Στην περίπτωση άρσης της ΑΚ Συνεργάτη, ο Κάτοχος Άδειας υποχρεούται άμεσα να παύσει κάθε σχετική με τη Σύμβαση Συνεργασίας ενέργεια, να καταγγείλει τη Σύμβαση και να γνωστοποιήσει τις παραπάνω ενέργειες στην Ε.Ε.Ε.Π.

Εικόνα 25: Εμφάνιση αποσπασμάτων νομοθεσίας με χρήση του LaBSE (2)

9 | score=0.3663 | role_norm=συνεργατες-affiliates

id: 509_5.2 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 5.2 article_context_name: ΣΥΜΒΑΣΗ ΣΥΝΕΡΓΑΣΙΑΣ subject: Συνεργάτες-Affilates paragraph_content: σύμβαση συνεργασίας article_text: Η Σύμβαση Συνεργασίας περιλαμβάνει, κατ΄ ελάχιστον, τη διάρκεια και το αντικείμενο της Συνεργασίας, τα Σημεία των Συνεργατών, τη διαδικασία τροποποίησης της Σύμβασης, τους λόγους καταγγελίας της Σύμβασης, τον τρόπο υπολογισμού και τον τρόπο και χρόνο καταβολής των αμοιβών του Συνεργάτη, τον τρόπο επίλυσης των διαφορών, τη ρητή δέσμευση του Συνεργάτη για την εκ μέρους του πλήρη, απόλυτη και ανεπιφύλακτη αποδοχή των διατάξεων της κείμενης νομοθεσίας, του Κανονισμού Παιγνίων και των Οδηγιών της Ε.Ε.Ε.Π. για τη Διοργάνωση και Διεξαγωγή των Παιγνίων.

10 | score=0.3572 | role_norm=συνεργατες-affiliates

id: 509_5.4 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 5.4 article_context_name: ΣΥΜΒΑΣΗ ΣΥΝΕΡΓΑΣΙΑΣ subject: Συνεργάτες-Affilates paragraph_content: υποχρεώσεις συνεργάτη που έχει συνάψει σύμβαση με κάτοχο άδειας article_text: Ο Συνεργάτης στο πλαίσιο της συμβατικής του σχέσης με τον Κάτοχο Άδειας: α. Χρησιμοποιεί για την προώθηση του Κατόχου Άδειας, αποκλειστικά, Σημεία Συνεργάτη. β. Ενημερώνει, αμελλητί, την Ε.Ε.Ε.Π., για οιαδήποτε μεταβολή, μεταβίβαση, τροποποίηση, διακοπή ή προσθήκη Σημείων Συνεργάτη. γ. Εκκαθαρίζει τις συναλλαγές του με τον Κάτοχο Άδειας, σε λογαριασμό που τηρεί ο ίδιος ως μοναδικός δικαιούχος σε Πάροχο Υπηρεσιών Πληρωμών, σύμφωνα με τις κείμενες διατάξεις.

Εικόνα 26: Εμφάνιση αποσπασμάτων νομοθεσίας με χρήση του LaBSE (3)

LLM Answer:

1) Σύντομη απάντηση

Για τη χορήγηση Άδειας Καταλληλότητας (ΑΚ) Συνεργάτη και την εγγραφή στο οικείο Μητρώο της Ε.Ε.Ε.Π., καταβάλλεται παράβολο ύψους **χιλίων (1.000) ευρώ**.

2) Νομική τεκμηρίωση

- Για την εγγραφή στο Μητρώο και τη χορήγηση της σχετικής άδειας καταλληλότητας, ορίζεται η καταβολή παραβόλου ύψους 1.000 ευρώ (509/1/11.09.2020 (Β 4140), Άρθρο 4.6, ID: 509_4.6).
- Η εγγραφή στο Μητρώο Συνεργατών που τηρεί η Ε.Ε.Ε.Π. και η λήψη της Άδειας Καταλληλότητας αποτελούν απαραίτητες προϋποθέσεις προκειμένου ένας συνεργάτης (affiliate) να παρέχει υπηρεσίες σε Κάτοχο Άδειας (509/1/11.09.2020 (Β 4140), Άρθρο 3.2, ID: 509_3.2).

3) Παραπομπές

- **id: 509_4.6**, law_source: 509/1/11.09.2020 (Β 4140), article_title: Άρθρο 4.6
- **id: 509_3.2**, law_source: 509/1/11.09.2020 (Β 4140), article_title: Άρθρο 3.2 User query: exit Bye.

Εικόνα 27: Απάντηση του Gemini βασιζόμενο στο ερώτημα του χρήστη και στα αποσπάσματα που ανακτήθηκαν με χρήση του LaBSE

Στη συνέχεια χρησιμοποιώντας το ίδιο σύνολο δεδομένων και LLM, δίνεται το ίδιο ερώτημα στο chatbot αλλά ως embedding model χρησιμοποιείται το intfloat/multilingual-e5-base. Οι κατηγορίες που εμφανίζονται για επιλογή metadata

filters είναι ίδιες με το προηγούμενο πείραμα καθώς η αλλαγή embedding model δεν σχετίζεται με την συγκεκριμένη λειτουργικότητα.

Στα αποτελέσματα (Εικόνες 28, 29, 30) παρατηρείται πως το σύστημα εμφάνισε πάλι ως πρώτο το απόσπασμα με id: 509_4.6, πράγμα που σημαίνει ότι και το δεύτερο embedding model κατάφερε να ανακτήσει το πιο σχετικό απόσπασμα με μεγάλη ακρίβεια. Επιπλέον, παρατηρούνται αποσπάσματα νομοθεσίας που εντοπίστηκαν και με τη χρήση του μοντέλου LaBSE (id: 509_4.1, id: 509_3.2). Πρέπει να σημειωθεί πως εκτός του πρώτου, τα αποσπάσματα που ανακτήθηκαν με χρήση του μοντέλου intfloat/multilingual-e5-base είναι πιο κοντά στην ευρύτερη θεματική του ερωτήματος σε σχέση με του LaBSE. Πιο συγκεκριμένα, το ερώτημα του χρήστη αναφέρεται στο παράβολο που πρέπει να καταβάλει ο συνεργάτης για να εγγραφεί στο μητρώο. Αυτό αποτελεί ένα μέρος της διαδικασίας της αδειοδότησής του. Πολλά από τα αποσπάσματα που ανακτήθηκαν αναφέρονται στα δικαιολογητικά που πρέπει να προσκομίσει ένας συνεργάτης προκειμένου να αδειοδοτηθεί και να εγγραφεί στο μητρώο (id: 509_4.5, id: 509_4.1, id: 509_4.2i, id: 509_4.2h_dd, id: 509_4.7, id: 509_4.2fpar2). Επιπροσθέτως, πρέπει να αναφερθεί πως το LLM (Εικόνα 31) επιβεβαίωσε βασιζόμενο στην ερώτηση του χρήστη και στα αποσπάσματα της νομοθεσίας παραθέτοντας παραπομπές, ότι το παράβολο που πρέπει να καταβάλει ένας συνεργάτης είναι 1000 ευρώ.

1 | score=0.8585 | role_norm=συνεργατες-affiliates

id: 509_4.6 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 4.6 article_context_name: ΑΔΕΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ (AFFILIATE) subject: Συνεργάτες-Affilates paragraph_content: παράβολο άδειας καταλληλότητας συνεργάτη article_text: Για τη χορήγηση ΑΚ Συνεργάτη και την εγγραφή στο οικείο Μητρώο καταβάλλεται παράβολο ύψους χιλίων (1.000) ευρώ.

2 | score=0.8569 | role_norm=συνεργατες-affiliates

id: 509_4.5 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 4.5 article_context_name: ΑΔΕΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ (AFFILIATE) subject: Συνεργάτες-Affilates paragraph_content: εγγραφή του συνεργάτη-affiliate στο οικείο μητρώο article_text: Εφόσον πληρούνται οι νόμιμες προϋποθέσεις, η Ε.Ε.Ε.Π. χορηγεί την ΑΚ και εγγράφει τον Συνεργάτη στο οικείο Μητρώο. Η εγγραφή αποτελεί επαρκή γνωστοποίηση και απόδειξη για την έκδοση της ΑΚ και κατά συνέπεια, τη νόμιμη παροχή των υπηρεσιών.

3 | score=0.8525 | role_norm=συνεργατες-affiliates

id: 509_4.1 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 4.1 article_context_name: ΑΔΕΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ (AFFILIATE) subject: Συνεργάτες-Affilates paragraph_content: άδεια καταλληλότητας συνεργάτη-affiliate article_text: Για τη χορήγηση ΑΚ Συνεργάτη τα ενδιαφερόμενα πρόσωπα υποβάλλουν αίτηση στην Ε.Ε.Ε.Π..

4 | score=0.8508 | role_norm=συνεργατες-affiliates

id: 509_4.2i law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 4.2θ article_context_name: ΑΔΕΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ (AFFILIATE) subject: Συνεργάτες-Affilates paragraph_content: δικαιολογητικά άιτησης χορήγησης άδειας καταλληλότητας συνεργάτη article_text: Η αίτηση συνοδεύεται από φάκελο που περιλαμβάνει τα παρακάτω στοιχεία και δικαιολογητικά: θ. Αποδεικτικό καταβολής του προβλεπόμενου παραβόλου υπέρ της Ε.Ε.Ε.Π.

Εικόνα 28: Εμφάνιση αποσπασμάτων νομοθεσίας με χρήση του intfloat/multilingual-e5-base (1)

5 | score=0.8440 | role_norm=συνεργατες-affiliates

id: 509_3.2 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 3.2 article_context_name: ΣΥΝΕΡΓΑΤΕΣ (AFFILIATES) subject: Συνεργάτες-Affilates paragraph_content: Προϋποθέσεις παροχής υπηρεσιών σε Κάτοχο Άδειας article_text: Προκειμένου να παρέχουν υπηρεσίες σε Κάτοχο Άδειας, τα πρόσωπα της παραγράφου 3.1 απαιτείται να έχουν λάβει ΑΚ Συνεργάτη (affiliate), να έχουν εγγραφεί στο Μητρώο Συνεργατών που τηρεί η Ε.Ε.Ε.Π. και να έχουν υπογράψει τη Σύμβαση Συνεργασίας του άρθρου 5.

6 | score=0.8407 | role_norm=συνεργατες-affiliates

id: 509_4.2h_dd law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 4.2η-δδ article_context_name: ΑΔΕΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ (AFFILIATE) subject: Συνεργάτες-Affilates paragraph_content: δικαιολογητικά άιτησης χορήγησης άδειας καταλληλότητας συνεργάτη article_text: Η αίτηση συνοδεύεται από φάκελο που περιλαμβάνει τα παρακάτω στοιχεία και δικαιολογητικά: η. Υπεύθυνη δήλωση του υποψήφιου Συνεργάτη με την οποία δηλώνει ότι δδ) Παρέχει, χωρίς επιφύλαξη, στην Ε.Ε.Ε.Π. την πλήρη συναίνεσή του καθώς και εντολή για την επεξεργασία των προσωπικών στοιχείων και των δεδομένων που τον αφορούν, συμπεριλαμβανομένης της διάθεσης στην Αρχή ή/και της επαλήθευσης από αυτήν, των ανωτέρω στοιχείων και δεδομένων από τρίτα μέρη. Στην περίπτωση που στο πρόσωπο του αιτούντος την ΑΚ του άρθρου αυτού συντρέχει κατάσταση των περιπτώσεων ββ' ή/και γγ' ανωτέρω, ο αιτών δηλώνει υποχρεωτικά την κατάσταση στην οποία τελεί, προσκομίζει τυχόν στοιχεία που διαθέτει και αιτείται από την Ε.Ε.Ε.Π. να εξετάσει εάν, παρά τη συνδρομή της κατάστασης αυτής, δύναται να εγκρίνει την αίτηση εφόσ...

7 | score=0.8362 | role_norm=συνεργατες-affiliates

id: 509_4.7 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 4.7 article_context_name: ΑΔΕΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ (AFFILIATE) subject: Συνεργάτες-Affilates paragraph_content: μεταβολή άδειας καταλληλότητας συνεργάτη article_text: Οι εγγεγραμμένοι στο Μητρώο γνωστοποιούν στην Ε.Ε.Ε.Π. κάθε μεταβολή, η οποία αφορά, άμεσα ή έμμεσα, στην Καταλληλότητά τους, εντός είκοσι (20) ημερών από την ημερομηνία που επήλθαν, συνυποβάλλοντας με την γνωστοποίηση τα απαραίτητα για την τεκμηρίωση της μεταβολής στοιχεία και δικαιολογητικά.

Εικόνα 29: Εμφάνιση αποσπασμάτων νομοθεσίας με χρήση του intfloat/multilingual-e5-base (2)

8 | score=0.8352 | role_norm=συνεργατες-affiliates

id: 509_4.9 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 4.9 article_context_name: ΑΔΕΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ (AFFILIATE) subject: Συνεργάτες-Affilates paragraph_content: βεβαίωση ισχύος Άδειας Καταλληλότητας Συνεργάτη article_text: Η Ε.Ε.Ε.Π. χορηγεί στους κατόχους ΑΚ Συνεργάτη, κατόπιν αίτησής τους, βεβαίωση ισχύος της ΑΚ, για κάθε νόμιμη χρήση.

9 | score=0.8330 | role_norm=συνεργατες-affiliates

id: 509_7.4 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 7.4 article_context_name: ΤΗΡΗΣΗ ΜΗΤΡΩΟΥ ΚΑΤΟΧΩΝ ΑΔΕΙΑΣ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ ΑΠΟ ΤΟΝ ΚΑΤΟΧΟ ΑΔΕΙΑΣ subject: Συνεργάτες-Affilates paragraph_content: πρόσβαση της Ε.Ε.Ε.Π. στα μητρώα κατόχων άδειας καταλληλότητς Συνεργάτη article_text: Ανεξάρτητα από την υποχρέωση της παραγράφου 7.2, ο Κάτοχος Άδειας παρέχει στην Ε.Ε.Ε.Π. πρόσβαση στα μητρώα κατόχων ΑΚ Συνεργάτη, στον χρόνο και με τον τρόπο που η Αρχή ορίζει. Με Οδηγίες της Ε.Ε.Ε.Π. δύναται να προστίθενται στοιχεία και δεδομένα που τηρούνται στο σχετικό μητρώο.

10 | score=0.8325 | role_norm=συνεργατες-affiliates

id: 509_4.2fpar2 law_source: 509/1/11.09.2020 (Β 4140) article_title: Άρθρο 4.2στπαρ2 article_context_name: ΑΔΕΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΣΥΝΕΡΓΑΤΗ (AFFILIATE) subject: Συνεργάτες-Affilates paragraph_content: δικαιολογητικά άιτησης χορήγησης άδειας καταλληλότητας συνεργάτη article_text: Ειδικότερα εάν ο υποψήφιος Συνεργάτης είναι εγκατεστημένος στην Ελλάδα, τα πιστοποιητικά ότι δεν τελεί υπό πτώχευση, υπό πτωχευτικό συμβιβασμό ή υπό αναγκαστική διαχείριση ή ότι δεν έχει υπαχθεί σε διαδικασία εξυγίανσης, εκδίδονται από το εκάστοτε αρμόδιο δικαστήριο της έδρας αυτών. Το πιστοποιητικό ότι το νομικό πρόσωπο δεν έχει τεθεί υπό εκκαθάριση με δικαστική απόφαση εκδίδεται από το αρμόδιο δικαστήριο της έδρας του προσώπου, το δε πιστοποιητικό ότι δεν έχει τεθεί υπό εκκαθάριση με απόφαση των εταίρων εκδίδεται από το Γ.Ε.Μ.Η., σύμφωνα με τις κείμενες διατάξεις, ως κάθε φορά ισχύουν. Τα φυσικά πρόσωπα ή οι ατομικές επιχειρήσεις δεν προσκομίζουν πιστοποιητικό περί μη θέσεως σε εκκαθάριση.

Εικόνα 30: Εμφάνιση αποσπασμάτων νομοθεσίας με χρήση του intfloat/multilingual-e5-base (3)

LLM Answer:

1. **Σύντομη απάντηση** Το παράβολο που απαιτείται για τη χορήγηση Άδειας Καταλληλότητας (ΑΚ) Συνεργάτη και την εγγραφή στο οικείο Μητρώο της Ε.Ε.Ε.Π. ανέρχεται στο ποσό των **χιλίων (1.000) ευρώ**.
2. **Νομική τεκμηρίωση**
   - Για τη χορήγηση της ΑΚ Συνεργάτη και την εγγραφή στο οικείο Μητρώο καταβάλλεται παράβολο ύψους 1.000 ευρώ (509_4.6).
   - Το αποδεικτικό καταβολής του προβλεπόμενου παραβόλου υπέρ της Ε.Ε.Ε.Π. αποτελεί υποχρεωτικό δικαιολογητικό που συνοδεύει τον φάκελο της αίτησης για τη χορήγηση της άδειας (509_4.2i).
3. **Παραπομπές**
   - **id: 509_4.6**, law_source: 509/1/11.09.2020 (Β 4140), article_title: Άρθρο 4.6
   - **id: 509_4.2i**, law_source: 509/1/11.09.2020 (Β 4140), article_title: Άρθρο 4.2θ User query: exit Bye.

Εικόνα 31: Απάντηση του Gemini βασιζόμενο στο ερώτημα του χρήστη και στα αποσπάσματα που ανακτήθηκαν με χρήση του intfloat/multilingual-e5-base

Από όσα παρουσιάστηκαν και στα δύο πειράματα εξάγεται το συμπέρασμα πως και τα δύο embedding models κατάφεραν να εντοπίσουν με ακρίβεια το απόσπασμα που περιέχει την απάντηση στο ερώτημα του χρήστη. Τα αποσπάσματα που ανακτήθηκαν με τη χρήση του μοντέλου intfloat/multilingual-e5-base ήταν πιο κοντά στην ευρύτερη θεματική. Λόγω του ότι και τα δύο μοντέλα ανέκτησαν το απόσπασμα που περιείχε την απάντηση στο ερώτημα του χρήστη, το LLM κατέληξε στο ίδιο συμπέρασμα και στα δύο πειράματα. Ικανότητα των συγκεκριμένων μοντέλων να ανακτήσουν σχετικά αποσπάσματα ενισχύθηκε και από τα metadata filters τα οποία μείωσαν το εύρος αναζήτησης στην ChromaDB επιτρέποντας την ανάκτηση μόνο από τις κατηγορίες που ήταν πιο σχετικές με τη φύση του ερωτήματος.

5. Συμπεράσματα και Μελλοντικές Προεκτάσεις

Στην παρούσα εργασία αναδείχθηκε το θέμα της αρχιτεκτονικής RAG. Η συγκεκριμένη αρχιτεκτονική έχει εφαρμογές σε αρκετούς τομείς και στον ιδιωτικό αλλά και στον δημόσιο τομέα. Η εφαρμογή της επεκτείνει την ικανότητα των LLMs να απαντούν σε θέματα στα οποία δεν έχουν εκπαιδευτεί. Με την συγκεκριμένη τεχνική μία επιχείρηση που αξιοποιεί LLMs, δεν χρειάζεται να τα εκπαιδεύει από την αρχή στο νέο γνωστικό αντικείμενο. Μπορεί να τα τροφοδοτήσει με περιεχόμενο από εξωτερικές πηγές γνώσεις προκειμένου να έχουν την δυνατότητα να απαντούν σε ερωτήματα στα οποία δεν είχαν εκπαιδευτεί, μειώνοντας το κόστος και τον κίνδυνο εμφάνισης hallucinations.

Το πεδίο εφαρμογής της αρχιτεκτονικής RAG που πραγματεύεται η παρούσα εργασία είναι η ελληνική νομοθεσία στην αγορά τυχερών παιγνίων. Αναπτύχθηκαν τρεις εφαρμογές οι οποίες αποτελούν ένα χρήσιμο εργαλείο για την εποπτεία των τυχερών παιγνίων. Η πρώτη υλοποίηση αφορούσε μία εφαρμογή, η οποία δέχεται μία καταγγελία ενός πολίτη ενάντια σε μία στοιχηματική εταιρεία και εμφανίζει ποια αποσπάσματα από την νομοθεσία ενδέχεται να παραβιάζει. Η δεύτερη αφορούσε την ανάπτυξη μίας εφαρμογής που δεχόταν μία καταγγελία και εντόπιζε άλλες παρόμοιες καταγγελίες. Το LLM εμφάνιζε στον χρήστη ομοιότητες και διαφορές ανάμεσα στην

δοθείσα και στις ανακτηθείσες καταγγελίες. Η τρίτη εφαρμογή αποτελούσε μία υλοποίηση ενός chatbot, στο οποίο ο χρήστης είχε την δυνατότητα να κάνει ερωτήσεις πάνω στον κανονισμό για την καταλληλότητα προσώπων και να λάβει απάντηση στηριζόμενη σε αναφορές στην νομοθεσία. Όλες οι εφαρμογές στηρίχθηκαν σε σύνολα δεδομένων που ήταν στην ελληνική γλώσσα. Γι' αυτό τον λόγο επιλέχθηκαν και συγκρίθηκαν τα embedding models LaBSE και intfloat/multilingual-e5-base ως προς την ικανότητά τους να ανακτούν την απαραίτητη πληροφορία στα ερωτήματα του χρήστη από κείμενα της ελληνικής νομοθεσίας.

Τα αποτελέσματα του πειράματος ήταν αρκετά ενθαρρυντικά, καθώς και τα δύο μοντέλα κατάφεραν να ανακτήσουν σχετικό περιεχόμενο ανάλογα με το task που απαιτούσε η κάθε εφαρμογή. Στην πρώτη εφαρμογή το LaBSE επέστρεψε πολύ στοχευμένες παραγράφους ως προς το θέμα, ενώ το intfloat/multilingual-e5-base ανέκτησε αντίστοιχες διατάξεις μέσω πιο γενικών άρθρων. Στη δεύτερη εφαρμογή γίνεται αντιληπτό πως το μοντέλο το intfloat/multilingual-e5-base έδωσε περιλήψεις παλαιότερων καταγγελιών που ήταν πολύ κοντά στο ακριβές σενάριο που περιέγραφε η περίληψη που δημιούργησε το σύστημα δεχόμενο την καταγγελία του χρήστη. Στην τρίτη εφαρμογή είναι εμφανές ότι η αξιοποίηση των metatada αποτελεί σημαντική τεχνική για την ανάκτηση αποσπασμάτων που είναι πιο κοντά στο περιεχόμενο του ερωτήματος που θέτει ο χρήστης. Στη συγκεκριμένη εφαρμογή και τα δύο μοντέλα κατάφεραν να έχουν ως πρώτο στα αποτελέσματά τους το απόσπασμα που περιείχε την απάντηση στο ερώτημα του χρήστη, με το intfloat/multilingual-e5-base να παρέχει αποσπάσματα που ήταν πιο κοντά σημασιολογικά στην ευρύτερη θεματική.

Οι συγκεκριμένες εφαρμογές κατάφεραν με αποτελεσματικό τρόπο να διεκπεραιώσουν τα tasks τους. Η χρήση τους μπορεί να αποτελέσει ένα σημαντικό εργαλείο στον έλεγχο τήρησης της νομοθεσίας στα τυχερά παίγνια. Τα μοντέλα που χρησιμοποιήθηκαν καθώς και η αξιοποίηση των βιβλιοθηκών ChromaDB και Faiss, έδειξαν πως η υλοποίηση μίας αρχιτεκτονικής RAG πάνω σε νομοθετικά κείμενα θα ήταν πολύ χρήσιμη, καθώς θα διευκόλυνε το έλεγχο των αρχών και θα επέτρεπε τις

απαντήσεις σε νομοθετικά ζητήματα χωρίς να εκπαιδεύονται από την αρχή οι εφαρμογές.

Μελλοντικές αναβαθμίσεις στις υπάρχουσες εφαρμογές θα μπορούσε να είναι η προσθήκη μεγαλύτερου μέρους από την νομοθεσία για τα τυχερά παίγνια, όπως είναι το κανονιστικό πλαίσιο για τα επίγεια καζίνο και οι κανονισμοί για την διεξαγωγή τυχερών παιγνίων μέσω παιγνιομηχανημάτων τύπου Video Lottery Terminal (VLT). Οι συγκεκριμένες προσθήκες θα μπορούσαν να γίνουν χωρίς ιδιαίτερο υπολογιστικό κόστος καθώς δεν απαιτείται εκπαίδευση του μοντέλου, απλά εμπλουτισμός της εξωτερικής πηγής γνώσης με νέα δεδομένα.

Βιβλιογραφία

[1] IBM, «What is retrieval augmented generation (RAG)?,» [Ηλεκτρονικό]. Available: https://www.ibm.com/think/topics/retrieval-augmented-generation. [Πρόσβαση 15 01 2026].

[2] NVIDIA, «What Is Retrieval-Augmented Generation, aka RAG?,» [Ηλεκτρονικό]. Available: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/. [Πρόσβαση 15 01 2026].

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, «Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,» 2021.

[4] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, Pascale Fung, «Survey of Hallucination in Natural Language Generation,» 2024.

[5] Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, Mohamed Abdelrazek, «Seven Failure Points When Engineering a Retrieval Augmented,» 2024.

[6] European Data Protection Supervisor, «Retrieval-augmented generation (RAG),» [Ηλεκτρονικό]. Available: https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/retrieval-augmented-generation-rag_en#_ednref2. [Πρόσβαση 18 01 2026].

[7] Agada Joseph Oche, Ademola Glory Folashade, Tirthankar Ghosal, Arpan Biswas, «A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions,» 2025.

[8] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang, «Retrieval-Augmented Generation for Large Language Models: A Survey,» 2024.

[9] LangChain, «Embedding model integrations,» [Ηλεκτρονικό]. Available: https://docs.langchain.com/oss/python/integrations/text_embedding. [Πρόσβαση 18 01 2026].

[10] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, Wei Wang, «Language-agnostic BERT Sentence Embedding,» 2022.

[11] Yinfei Yang, Fangxiaoyu Feng, «Language-Agnostic BERT Sentence Embedding,» Google Research, [Ηλεκτρονικό]. Available: https://research.google/blog/language-agnostic-bert-sentence-embedding/. [Πρόσβαση 18 01 2026].

[12] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, Furu Wei, «Multilingual E5 Text Embeddings: A Technical Report,» 2024.

[13] Nils Reimers, Iryna Gurevych, «Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation,» 2020.

[14] sbert.net, «Pretrained Models,» [Ηλεκτρονικό]. Available: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html. [Πρόσβαση 18 01 2026].

[15] OpenAI Developers, «Vector embeddings,» [Ηλεκτρονικό]. Available: https://developers.openai.com/api/docs/guides/embeddings. [Πρόσβαση 18 01 2026].

[16] Gemini API, «Embeddings,» [Ηλεκτρονικό]. Available: https://ai.google.dev/gemini-api/docs/embeddings. [Πρόσβαση 18 01 2026].

[17] J. Lee, F. Chen, S. Dua, D. Cer, M. Shanbhogue, I. Naim, et al., «Gemini Embedding: Generalizable Embeddings from Gemini,» Google, 2025.

[18] Jim Holdsworth, Matthew Kosinski, «What is a vector database?,» IBM, [Ηλεκτρονικό]. Available: https://www.ibm.com/think/topics/vector-database. [Πρόσβαση 20 01 2026].

[19] Chroma, «Getting Started,» [Ηλεκτρονικό]. Available: https://docs.trychroma.com/docs/overview/getting-started. [Πρόσβαση 04 02 2026].

[20] Meta, «Faiss,» [Ηλεκτρονικό]. Available: https://ai.meta.com/tools/faiss/. [Πρόσβαση 20 01 2026].

[21] Pinecone, «Pinecone documentation,» [Ηλεκτρονικό]. Available: https://docs.pinecone.io/guides/get-started/overview. [Πρόσβαση 04 02 2026].

[22] Amir Ingber, Edo Liberty, «Accurate and Efficient Metadata Filtering,» 2025.

[23] Weaviate, «Weaviate Database,» [Ηλεκτρονικό]. Available: https://docs.weaviate.io/weaviate. [Πρόσβαση 04 02 2026].

[24] NVIDIA, «Large Language Models Explained,» [Ηλεκτρονικό]. Available: https://www.nvidia.com/en-us/glossary/large-language-models/. [Πρόσβαση 20 01 2026].

[25] Gokul Yenduri, M. Ramalingam, G. Chemmalar Selvi, Y. Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G. Deepti Raj, Rutvij H. Jhaveri, B. Prabadevi, Weizheng Wang, Athanasios V. Vasilakos, Thippa Reddy Gadekallu, «GPT (Generative Pre-Trained Transformer)-A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions,» 2024.

[26] Gemini Team, «Gemini: A Family of Highly Capable Multimodal Models,» Google, 2025.

[27] Llama Team, Meta AI, «The Llama 3 Herd of Models,» 2024.

[28] Suhui Gang, «Case Study: Samsung SDS Uses Retrieval Augmented Generation for Kubernetes Troubleshooting,» Samsung SDS, 04 11 2024. [Ηλεκτρονικό]. Available: https://www.samsungsds.com/en/insights/rag-customization.html. [Πρόσβαση 01 02 2026].

[29] Aritra Kumar Lahir, Qinmin Vivian Hu, «AlzheimerRAG: Multimodal Retrieval-Augmented Generation for Clinical Use Cases,» 2025.

[30] Gemini API, «File Search,» [Ηλεκτρονικό]. Available: https://ai.google.dev/gemini-api/docs/file-search. [Πρόσβαση 05 02 2026].

[31] OECD, «Building an AI-ready public workforce: Implications and strategies,» OECD Publishing, 2026. [Ηλεκτρονικό]. Available: https://doi.org/10.1787/b89244c7-en. [Πρόσβαση 08 02 2026].

[32] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, K. Roberts, «Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,» NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, 2024. [Ηλεκτρονικό]. Available: https://doi.org/10.6028/NIST.AI.600-1. [Πρόσβαση 08 02 2026].

[33] Gov.UK, «DSIT: GOV.UK Chat,» [Ηλεκτρονικό]. Available: https://www.gov.uk/algorithmic-transparency-records/dsit-gov-dot-uk-chat. [Πρόσβαση 05 02 2026].

[34] GovTech Singapore, «AIBots,» [Ηλεκτρονικό]. Available: https://www.tech.gov.sg/products-and-services/for-government-agencies/productivity-and-marketing/aibots/. [Πρόσβαση 06 02 2026].

[35] Pacific Northwest National Laboratory, «Retrieval Augmented Generation for Robust Cyber Defense,» 2024.

[36] AI.Gov.UK Knowledge Hub, «Barking and Dagenham AI Search and Webchat,» [Ηλεκτρονικό]. Available: https://ai.gov.uk/knowledge-hub/tools/barking-and-dagenham-ai-search-and-webchat/. [Πρόσβαση 06 02 2026].

[37] Gov.UK, «Cafcass: Genesys ChatBot,» [Ηλεκτρονικό]. Available: https://www.gov.uk/algorithmic-transparency-records/cafcass-genesys-chatbot. [Πρόσβαση 06 02 2026].

[38] Parliamentary Counsel Office Research & Development Findings, «Chat-Based Access to Legislation,» 2025.

[39] Marios Evangelos Mamalis, Evangelos Kalampokis, Fotios Fitsilis, Georgios Theodorakopoulos, Konstantinos Tarabanis, «A Large Language Model Agent Based Legal Assistant for Governance Applications,» 2024.

[40] Dnyanesh Panchal, Aaryan Gole,Vaibhav Narute, Raunak Joshi, «LawPal : A Retrieval Augmented Generation Based System for Enhanced Legal Accessibility in India,» 2025.

[41] Chroma, «Configure Collections,» [Ηλεκτρονικό]. Available: https://docs.trychroma.com/docs/collections/configure#what-is-an-hnsw-index. [Πρόσβαση 08 02 2026].

[42] Google Colab, «Colaboratory,» [Ηλεκτρονικό]. Available: https://research.google.com/colaboratory/faq.html. [Πρόσβαση 08 02 2026].

[43] Python, «What is Python? Executive Summary,» [Ηλεκτρονικό]. Available: https://www.python.org/doc/essays/blurb/. [Πρόσβαση 08 02 2026].