

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Ευφυείς βοηθοί μελέτης: RAG και LLMs για την
αυτοματοποίηση της ακαδημαϊκής υποστήριξης από
πανεπιστημιακούς οδηγούς»



Των φοιτητριών:
Κουρτίδου Σοφία
Αρ. Μητρώου: 2019080

Κρυωνίδου Δήμητρα
Αρ. Μητρώου: 2019086

Επιβλέπων
Όνοματεπώνυμο:
Μπράτσας Χαράλαμπος

Βαθμίδα :
Επίκουρος Καθηγητής

Σεπτέμβριος 2025

Ευφυείς βοηθοί μελέτης: RAG και LLMs για την αυτοματοποίηση της ακαδημαϊκής υποστήριξης από πανεπιστημιακούς οδηγούς

24323

Κουρτίδου Σοφία
Α.Μ. : 2019080

Κρυωνίδου Δήμητρα
Α.Μ. : 2019086

Εισηγητής : Μπράτσας Χαράλαμπος

Ημερομηνία ανάληψης Δ.Ε. : 28-11-2024
Ημερομηνία περάτωσης Δ.Ε. : 28-11-2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία των φοιτητριών Κουρτίδου Σοφία και Κρυωνίδου Δήμητρα που την εκπόνησαν. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (downloading), ανάρτηση (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Αυτή η διπλωματική εργασία είναι αφιερωμένη στις οικογένειές μας. Είμαστε βαθιά ευγνώμων για την αγάπη, την στήριξη και την ενθάρρυνση σας. Σας ευχαριστούμε που σταθήκατε δίπλα μας και μας υποστηρίξατε σε κάθε στόχο που θέσαμε.

«Αφιέρωση»

Πρόλογος

Επιλέξαμε αυτό το συγκεκριμένο θέμα επειδή μας ενδιαφέρει η τεχνητή νοημοσύνη, ειδικά τα μεγάλα γλωσσικά μοντέλα, τα οποία έχουν γίνει όλο και πιο δημοφιλή τα τελευταία χρόνια και χρησιμοποιούνται πλέον σε ένα ευρύ φάσμα διαδικτυακών εφαρμογών. Θέλαμε να μάθουμε περισσότερα για το πώς λειτουργούν και πώς μπορούν να χρησιμοποιηθούν για την ανάλυση κειμένων, όπως οι σχολικοί κανονισμοί στους οποίους θα αναφερόμαστε σε αυτό το δοκίμιο.

Μέσα από την συνεργασία μας αποκτήσαμε εμπειρία στη χρήση διαφόρων εργαλείων, στη συγγραφή και διόρθωση κώδικα, και στην αναγνώριση των πλεονεκτημάτων και μειονεκτημάτων κάθε μοντέλου. Επιπλέον, βελτιώσαμε την ικανότητα μας να οργανώνουμε την έρευνα, να αντιπαραβάλουμε ευρήματα και να διατυπώνουμε πιο σαφή συμπεράσματα. Η εργασία αυτή συνέβαλε επίσης στην ενίσχυση των δεξιοτήτων μας στον προγραμματισμό.

Συνολικά, μάθαμε περισσότερα για τα LLMs και βελτιώσαμε τις δεξιότητες που θα μου φανούν χρήσιμες στο μελλοντικό μας επάγγελμα.

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με τη μελέτη και τη σύγκριση τριών διαφορετικών μεγάλων γλωσσικών μοντέλων (LLMs), με στόχο την κατανόηση του τρόπου με τον οποίο μπορούν να αξιοποιηθούν στην ανάλυση και στην επεξεργασία κειμένων. Η επιλογή του θέματος βασίστηκε στο πολύ μεγάλο ενδιαφέρον που υπάρχει σήμερα για την τεχνητή νοημοσύνη και τις δυνατότητες που προσφέρει στην εκπαίδευση, την έρευνα, τη διαχείριση δεδομένων κ.α.

Ως υλικό ανάλυσης χρησιμοποιήθηκαν περίπου 50 έγγραφα σε μορφή PDF, τα οποία αφορούν κανονισμούς σχολών. Η συγκεκριμένη επιλογή δεν έγινε τυχαία, καθώς τα κείμενα αυτού του τύπου χαρακτηρίζονται από επίσημο ύφος, μεγάλη έκταση και πολλές λεπτομέρειες, γεγονός που τα καθιστά ιδανικά για να δοκιμαστεί η αποτελεσματικότητα των LLMs. Η πρόκληση έγκειται στο αν τα μοντέλα μπορούν να αναγνωρίσουν σωστά την πληροφορία, να την οργανώσουν και να τη μετατρέψουν σε χρήσιμη γνώση.

Για τον σκοπό αυτό, αναπτύχθηκε ένας κώδικας, μέσω του οποίου ήταν δυνατή η αξιοποίηση και η σύγκριση των τριών διαφορετικών γλωσσικών μοντέλων. Η μεθοδολογία περιλαμβάνει βήματα όπως η μετατροπή των PDF σε επεξεργάσιμα δεδομένα, η προετοιμασία των κειμένων, η εισαγωγή τους στα μοντέλα και η παρατήρηση των αποτελεσμάτων που παράγαν.

Στόχος της εργασίας είναι να δοθεί μια ολοκληρωμένη εικόνα των δυνατοτήτων και των περιορισμών που εμφανίζουν τα LLMs όταν εφαρμόζονται σε τέτοιου είδους κείμενα. Η ανάλυση αυτή μπορεί να αποτελέσει χρήσιμη βάση για μελλοντικές εφαρμογές και να προσφέρει κατευθύνσεις για περαιτέρω έρευνα στον τομέα της εκπαίδευσης και της τεχνητής νοημοσύνης.

«Intelligent study assistants: RAG and LLMs for automating academic support from university guides»

«Sophia Kourtidou and Dimitra Krionidou»

Abstract

This thesis deals with the study and comparison of three different large language models (LLMs), with the aim of understanding how they can be utilized in text analysis and processing. The choice of the topic was based on the great interest that exists today in artificial intelligence and the possibilities it offers in education, research, data management, etc.

Approximately 50 documents in PDF format were used as analysis material, which concern school regulations. This specific choice was not made by chance, as texts of this type are characterized by a formal style, large extent and many details, which makes them ideal for testing the effectiveness of LLMs. The challenge lies in whether the models can correctly recognize information, organize it and transform it into useful knowledge.

For this purpose, a code was developed, through which it was possible to utilize and compare the three different language models. The methodology includes steps such as converting PDFs into editable data, preparing the texts, introducing them into the models and observing the results they produced.

The aim of the work is to provide a comprehensive picture of the capabilities and limitations of LLMs when applied to such texts. This analysis can be a useful basis for future applications and offer directions for further research in the field of education and artificial intelligence.

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μας εργασίας, θα θέλαμε να εκφράσουμε την ειλικρινή μας ευγνωμοσύνη σε όλους όσους μας στήριξαν καθ' όλη τη διάρκεια των σπουδών μας και συνέλαβαν με οποιονδήποτε τρόπο στην επιτυχή ολοκλήρωση αυτής της προσπάθειας.

Πρώτα απ' όλα, ευχαριστούμε όλους τους καθηγητές του Τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε για τις πολύτιμες γνώσεις που μας προσέφεραν μέσα από τα μαθήματα τους.

Ιδιαίτερα, θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα καθηγητή μας, κ. Χαράλαμπο Μπράτσα, για την εμπιστοσύνη, την υποστήριξη και τις καθοριστικές συμβουλές του κατά τη διάρκεια αυτής της πτυχιακής.

Τέλος, ευχαριστούμε θερμά τις οικογένειές μας, που στέκονται πάντα δίπλα μας και μας στηρίζουν σε κάθε βήμα. Η βοήθειά τους ήταν καταλυτική για την επίτευξη των στόχων μας.

Περιεχόμενα

Πρόλογος.....	v
Περίληψη	vi
Abstract.....	vii
Ευχαριστίες	viii
Περιεχόμενα	ix
Κατάλογος Σχημάτων	xii
Κατάλογος Πινάκων	xii
Συνομογραφίες.....	xiii
Κεφάλαιο 1ο: Εισαγωγή	1
1.1 Τεχνητή Νοημοσύνη και Μεγάλα Γλωσσικά Μοντέλα.....	1
1.2 Ιστορική Αναδρομή της Τεχνητής Νοημοσύνης.....	2
1.3 Από τη Μηχανική Μάθηση στη Βαθιά Μάθηση.....	3
1.4 Τα Μεγάλα Γλωσσικά Μοντέλα και ο Ρόλος τους στην Επεξεργασία Κειμένου	4
1.5 Chatbots με LLMs ως Υποστηρικτικά Εργαλεία στην Εκπαίδευση	5
1.6 Σκοπός και Στόχοι	6
Κεφάλαιο 2ο: Θεωρητικό Πλαίσιο	7
2.1 Εισαγωγή στο Θεωρητικό Πλαίσιο	7
2.2 Τεχνητή Νοημοσύνη: Ορισμοί και Βασικές Αρχές.....	7
2.3 Μηχανική Μάθηση και Βαθιά Μάθηση	8
2.3.1 Μηχανική Μάθηση	8
2.3.2 Βαθιά Μάθηση	10
2.4 Επεξεργασία Φυσικής Γλώσσας (NLP).....	13
2.4.1 Ορισμός και Σημασία.....	13
2.4.2 Βασικά Στάδια Επεξεργασίας.....	13
2.4.3 Εφαρμογές του NLP	14
2.5 Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs)	16
2.5.1 Ορισμός και Σημασία.....	16
2.5.2 Αρχιτεκτονική Transformers	17
2.6 Retrieval-Augmented Generation (RAG)	18
2.6.1 Ορισμός και Σημασία.....	18
2.6.2 Πως Λειτουργεί το RAG	18
2.7 Αξιολόγηση Μεγάλων Γλωσσικών Μοντέλων	19

Κεφάλαιο 3ο: Μεθοδολογία.....	21
3.1 Εισαγωγή	21
3.2 Δεδομένα.....	21
3.3 Επεξεργασία και Οργάνωση Δεδομένων	21
3.4 Αρχιτεκτονική Συστήματος	24
3.4.1 Επίπεδο Δεδομένων	24
3.4.2 Επίπεδο Ανάκτησης	25
3.4.3 Επίπεδο Επεξεργασίας	26
3.4.4 Επίπεδο Διεπαφής Χρήστη (UI)	26
3.4.5 Παράδειγμα Χρήσης της Εφαρμογής.....	27
3.5 Εργαλεία και Τεχνολογίες.....	30
3.5.1 Γλώσσα Προγραμματισμού (Python)	30
3.5.2 Βιβλιοθήκες.....	31
3.5.3 Embeddings και Βάση Διανυσμάτων.....	32
3.5.4 Διαχείριση API Tokens	35
3.5.5 Υλοποίηση με Πολλαπλά LLMs	36
3.6 Μετρικές Αξιολόγησης και Μαθηματικοί Ορισμοί.....	38
3.6.1 Θεωρητικό Υπόβαθρο των Μετρικών	38
3.6.2 BLEU	38
3.6.3 ROUGE	39
3.6.4 BERTScore.....	39
3.6.5 Συγκριτική Ερμηνεία των Μετρικών	41
3.6.6 Εφαρμογή των Μετρικών στην Παρούσα Εργασία	41
Κεφάλαιο 4ο: Αποτελέσματα	42
4.1 Εισαγωγή	42
4.2 BERTScore	42
4.3 BLEU.....	44
4.4 ROUGE.....	46
4.5 Συνολική Σύγκριση	48
Κεφάλαιο 5ο: Συμπεράσματα και Μελλοντική Εργασία	51
5.1 Συμπεράσματα	51
5.2 Μελλοντική Εργασία	52
ΒΙΒΛΙΟΓΡΑΦΙΑ.....	53
ΠΑΡΑΡΤΗΜΑ Α : Βασικός Κώδικας Python	55

Κατάλογος Σχημάτων

Σχήμα 1.1: Σχέση μεταξύ Τεχνητής Νοημοσύνης, Μηχανικής Μάθησης, Βαθιάς Μάθησης και Επεξεργασίας Φυσικής Γλώσσας.	1
Σχήμα 1.2: Ο Alan Turing, θεμελιωτής της ιδέας της Τεχνητής Νοημοσύνης, δίπλα στη μηχανή του...2	2
Σχήμα 1.3: Απλό διάγραμμα νευρωνικού δικτύου.	4
Σχήμα 1.4: Ενδεικτικοί τομείς εφαρμογής των Μεγάλων Γλωσσικών Μοντέλων.....	5
Σχήμα 2.1: Σύγκριση ανάμεσα σε StrongAI και WeakAI.....	8
Σχήμα 2.2: Παράδειγμα υπολογισμού του MSE. Η ευθεία προσπαθεί να προσεγγίσει τη σχέση εισόδων-εξόδων, ενώ το MSE μετράει τη διαφορά ανάμεσα στις πραγματικές τιμές (Y) και στις προβλέψεις του μοντέλου (Y').	9
Σχήμα 2.3: Βασικές κατηγορίες της Μηχανικής Μάθησης.....	10
Σχήμα 2.4: Συνάρτηση ενεργοποίησης Sigmoid.	11
Σχήμα 2.5: Συνάρτηση ενεργοποίησης ReLU.	12
Σχήμα 2.6: Συνάρτηση ενεργοποίησης Tanh.	12
Σχήμα 2.7: Νευρωνικό δίκτυο με επίπεδο εισόδου, κρυφά επίπεδα και επίπεδο εξόδου. Η εκπαίδευση γίνεται μέσω οπισθοδιάδοσης (backpropagation).....	13
Σχήμα 2.8: Τα βασικά στάδια προεπεξεργασίας κειμένου στο NLP, από τον καθαρισμό μέχρι τη διανυσματοποίηση.	14
Σχήμα 2.9: Ενδεικτικές εφαρμογές της Επεξεργασίας Φυσικής Γλώσσας (NLP), όπως τα chatbots, η σύνοψη κειμένων, η ανάλυση κοινωνικών δικτύων και η στοχευμένη διαφήμιση.	15
Σχήμα 2.10: Ο μηχανισμός Προσοχής (Attention) σε ένα Transformer.	17
Σχήμα 2.11: Διάγραμμα ροής της μεθόδου RAG με τα βήματα από την ερώτηση του χρήστη έως την τελική απάντηση.	19
Σχήμα 3.1: Διάγραμμα ροής της διαδικασίας RAG.....	23
Σχήμα 3.2: Αρχική οθόνη της εφαρμογής LLM Comparison Chatbot με τα τρία μοντέλα.	28
Σχήμα 3.3: Παράδειγμα εκτέλεσης ερωτήματος.	29
Σχήμα 3.4: Παράδειγμα εκτέλεσης ερωτήματος.	30
Σχήμα 4.1: Κατανομή BERTScore ανά μοντέλο.	43
Σχήμα 4.2: BERTScore ανά ερώτηση.....	43
Σχήμα 4.3: Κατανομή BLEU ανά μοντέλο.	44
Σχήμα 4.4: BLEU ανά ερώτηση.	45
Σχήμα 4.5: Κατανομή ROUGE ανά μοντέλο.	46
Σχήμα 4.6: ROUGE ανά ερώτηση.....	47
Σχήμα 4.7: Συγκριτικός χρόνος απόκρισης ανά μοντέλο.....	48
Σχήμα 4.8: Συγκριτικό μήκος απαντήσεων ανά μοντέλο.....	49

Κατάλογος Πινάκων

Πίνακας 3.1: Ενδεικτική μορφή του ground truth dataset.....	21
Πίνακας 3.2: Σύγκριση των μετρικών.....	41
Πίνακας 3.3: Σύνολο ερωτήσεων-απαντήσεων για πολυμετρική αξιολόγηση.....	41

Συντομογραφίες

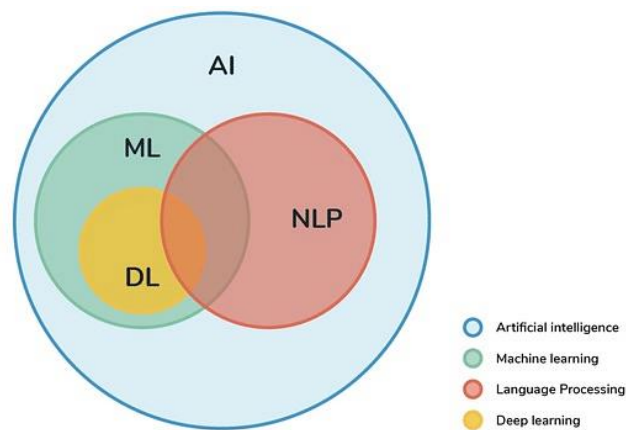
TN	Τεχνητή Νοημοσύνη
ML	Μηχανική Μάθηση (Machine Learning)
DL	Βαθιά Μάθηση (Deep Learning)
NLP	Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing)
LLM	Μεγάλο Γλωσσικό Μοντέλο (Large Language Model)
RAG	Εμπλουτισμένη Δημιουργία με Ανάκτηση (Retrieval-Augmented Generation)
UI	Διεπαφή Χρήστη (User Interface)
API	Διεπαφή Προγραμματισμού Εφαρμογών (Application Programming Interface)
MSE	Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error)
ReLU	Rectified Linear Unit
NER	Αναγνώριση Οντοτήτων (Named Entity Recognition)

Κεφάλαιο 1ο: Εισαγωγή

1.1 Τεχνητή Νοημοσύνη και Μεγάλα Γλωσσικά Μοντέλα

Όλοι έχουμε έρθει σε επαφή με την τεχνητή νοημοσύνη σε κάποιο σημείο της καθημερινότητάς μας. Χρησιμοποιείται σε εφαρμογές που συχνά δεν γνωρίζουμε, όπως οι προτάσεις βίντεο στο YouTube, οι διαφημίσεις στα μέσα κοινωνικής δικτύωσης, οι προσωπικοί βοηθοί όπως η Siri και η Alexa, και σε πιο εξειδικευμένους τομείς όπως η χρηματοοικονομική ανάλυση ή η ιατρική διάγνωση. Η ικανότητα των μηχανών να εκτελούν εργασίες όπως η κατανόηση της γλώσσας, η αναγνώριση εικόνων, η λήψη αποφάσεων και η μάθηση βάσει εμπειρίας, τις οποίες ιστορικά θεωρούσαμε ότι απαιτούσαν ανθρώπινη παρέμβαση, είναι γνωστή ως τεχνητή νοημοσύνη.

Υπάρχουν αρκετά υποπεδία της τεχνητής νοημοσύνης, όπως η μηχανική μάθηση, η βαθιά μάθηση και η επεξεργασία φυσικής γλώσσας. Η σχέση μεταξύ αυτών των σημαντικών περιοχών και των συνόλων στα οποία ανήκει κάθε περιοχή απεικονίζεται στο παρακάτω διάγραμμα. Αν και χωρίζονται σε τομείς, τα σύνολα αλληλεπικαλύπτονται, μοιράζοντας στοιχεία μεταξύ τους. [1]



Σχήμα 1.1: Σχέση μεταξύ Τεχνητής Νοημοσύνης, Μηχανικής Μάθησης, Βαθιάς Μάθησης και Επεξεργασίας Φυσικής Γλώσσας.

Τα μεγάλα γλωσσικά μοντέλα (LLMs) έχουν αναδειχθεί τα τελευταία χρόνια σε αυτόν τον τομέα. Μεγάλες ποσότητες κειμένου —εξ ου και ο όρος "Μεγάλο"— από βιβλία, άρθρα, επιστημονικές δημοσιεύσεις, το διαδίκτυο και άλλες πηγές χρησιμοποιούνται για την εκπαίδευση συστημάτων LLM. Αυτό τους επιτρέπει να "κατανοούν" και να παράγουν φυσική γλώσσα με τρόπο που συχνά μοιάζει ή είναι σχεδόν πανομοιότυπος με τον ανθρώπινο [2]. Ένα βασικό παράδειγμα είναι η ικανότητά τους να απαντούν λογικά και πειστικά σε μια ερώτηση στη φυσική γλώσσα.

Τα μοντέλα GPT (Generative Pre-trained Transformer) από την OpenAI, το Claude από την Anthropic και η σειρά LLaMA από τη Meta είναι τα πιο γνωστά παραδείγματα LLMs. Μπορούν να χρησιμοποιηθούν για ένα ευρύ φάσμα σκοπών, από την παροχή αυτοματοποιημένης εξυπηρέτησης πελατών έως τη διεξαγωγή εις βάθος αναλυτικής έρευνας σε τεράστιους όγκους δεδομένων.

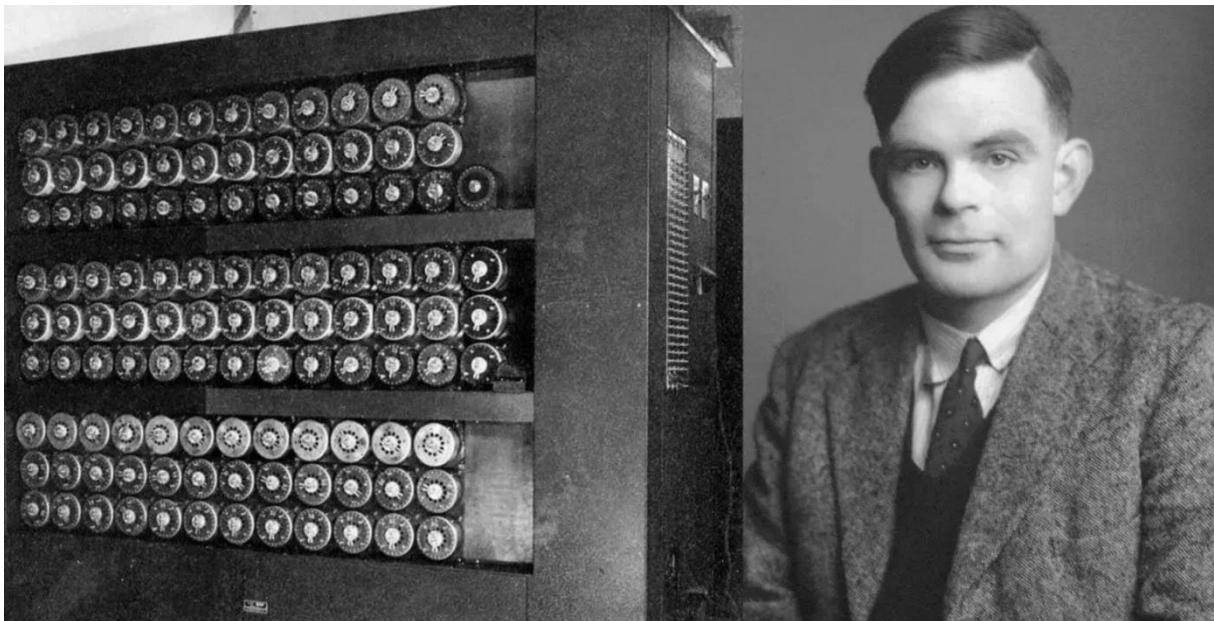
Τα LLM είναι σημαντικά επειδή ανοίγουν νέους δρόμους για την επικοινωνία μεταξύ ανθρώπων και μηχανών. Εκτός από το ότι παρέχουν άμεσες απαντήσεις, είναι σε θέση να γράφουν κώδικα

προγραμματισμού, να συνοψίζουν εκτενή κείμενα και να απλοποιούν δύσκολα θέματα. Είναι επομένως χρήσιμα εργαλεία για την επιστήμη, την εκπαίδευση, την έρευνα και την καθημερινή ζωή.

1.2 Ιστορική Αναδρομή της Τεχνητής Νοημοσύνης

Η εμφάνιση της τεχνητής νοημοσύνης δεν ήταν απότομη. Με ρίζες στην αρχαία φιλοσοφία, η ιδέα ότι οι μηχανές είναι ικανές να σκέφτονται, να μαθαίνουν και να δρουν σαν άνθρωποι επιμένει εδώ και πολλές δεκαετίες. Το ερώτημα αν οι μηχανές θα μπορούσαν τελικά να αναπαράγουν την ανθρώπινη σκέψη υπήρχε από τότε που οι άνθρωποι άρχισαν να τις δημιουργούν.

Στα μέσα του 20ού αιώνα έγινε το πρώτο σημαντικό βήμα. Ένας από τους πιο γνωστούς επιστήμονες της εποχής, ο Άλαν Τούρινγκ, δημοσίευσε το "Computing Machinery and Intelligence" το 1950 [3]. Ρώτησε: «Μπορούν οι μηχανές να σκέφτονται;». Αντί να προσπαθήσει να δώσει έναν θεωρητικό ορισμό της "σκέψης", πρότεινε το διάσημο "Τεστ Τιούρινγκ". Αυτό σήμαινε ότι μια μηχανή είναι έξυπνη αν μπορεί να ξεγελάσει το άτομο που της στέλνει μηνύματα, κάνοντάς το να πιστέψει ότι μιλάει με κάποιον άλλο. Παρά την απλότητά του, αυτό το τεστ αποτέλεσε τη βάση για πολυάριθμες συζητήσεις που συνεχίζονται μέχρι σήμερα.



Σχήμα 1.2: Ο Alan Turing, θεμελιωτής της ιδέας της Τεχνητής Νοημοσύνης, δίπλα στη μηχανή του.

Η τεχνητή νοημοσύνη άρχισε να αναγνωρίζεται ως αυτόνομος ερευνητικός τομέας τη δεκαετία του 1950 και του 1960. Ο Τζον Μακάρθι εισήγαγε τον όρο "Τεχνητή Νοημοσύνη" το 1956 κατά τη διάρκεια ενός συνεδρίου που πραγματοποιήθηκε στο Κολλέγιο Ντάρτμουθ [4]. Οι επιστήμονες εκείνης της εποχής πίστευαν ότι οι υπολογιστές σύντομα θα μπορούσαν να λύσουν ένα ευρύ φάσμα προβλημάτων. Αναπτύχθηκαν προγράμματα ικανά να επιλύουν βασικά μαθηματικά προβλήματα, καθώς και οι πρώτοι αλγόριθμοι αναζήτησης και λογικής.

Ωστόσο, η πρόοδος δεν ήταν τόσο ταχεία όσο είχε αρχικά προβλεφθεί. Η πρώτη περίοδος "Χειμώνα της Τεχνητής Νοημοσύνης" ξεκίνησε τη δεκαετία του 1970, καθώς η χρηματοδότηση και το ενδιαφέρον

μειώθηκαν λόγω των αποτελεσμάτων που υστέρησαν σε σχέση με τις προσδοκίες. Μια δεύτερη περίοδο απογοήτευσης ακολούθησε τη δεκαετία του 1980.

Τη δεκαετία του 1990 και μετά, η μηχανική μάθηση έφερε μια αναγέννηση. Τα συστήματα άρχισαν να αντλούν γνώση από δεδομένα και να βελτιώνονται με την πάροδο του χρόνου. Ο Γκάρι Κασπάροφ, ο παγκόσμιος πρωταθλητής σκακιού, ηττήθηκε από το Deep Blue της IBM [5], το οποίο κυκλοφόρησε για πρώτη φορά το 1997. Αυτό το περιστατικό ήρθε να αντιπροσωπεύσει τη δύναμη των υπολογιστών.

Η πρόοδος ήταν ακόμη πιο εντυπωσιακή τα επόμενα χρόνια. Το AlphaGo παρουσιάστηκε από τη DeepMind το 2016 και νίκησε τον καλύτερο παίκτη στο παιχνίδι Go, το οποίο θεωρείται σημαντικά πιο πολύπλοκο από το σκάκι. Αυτός ο θρίαμβος απέδειξε ότι οι υπολογιστές ήταν ικανοί για στρατηγική σκέψη σε επίπεδα που προηγουμένως θεωρούνταν αδιανόητα.

Σήμερα, η τεχνητή νοημοσύνη είναι παρούσα σε κάθε πτυχή της καθημερινότητας. Από τα κινητά μας τηλέφωνα μέχρι την υγειονομική περίθαλψη, τις μεταφορές, την οικονομία και, φυσικά, την εκπαίδευση, η επιρροή της είναι πανταχού παρούσα. Τα νευρωνικά δίκτυα και νέες αρχιτεκτονικές όπως οι μετασχηματιστές αποτελούν τη βάση της προόδου των τελευταίων ετών, καθιστώντας δυνατή την επεξεργασία φυσικής γλώσσας σε μια κλίμακα άνευ προηγουμένου.

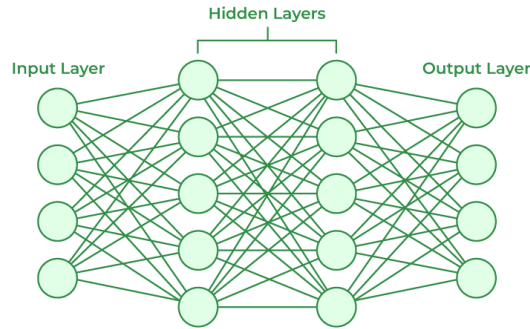
Από τις πρώτες φιλοδοξίες και μέσα από περιόδους απογοήτευσης, η ιστορία της τεχνητής νοημοσύνης χαρακτηρίζεται από σκαμπανεβάσματα. Σήμερα, οι εξελίξεις διαδέχονται η μια την άλλη με τόσο ταχείς ρυθμούς ώστε επηρεάζουν άμεσα την κοινωνία στο σύνολό της.

1.3 Από τη Μηχανική Μάθηση στη Βαθιά Μάθηση

Η μετάβαση από τους συμβατικούς αλγορίθμους στη μηχανική μάθηση (ML) ήταν μια από τις σημαντικότερες εξελίξεις στον τομέα της τεχνητής νοημοσύνης. Τα προγράμματα στις πρώτες μέρες της πληροφορικής ακολουθούσαν αυστηρές οδηγίες που είχαν καθοριστεί από τον προγραμματιστή. Αυτό αποδείχθηκε περιοριστικό, ωστόσο, καθώς δεν ήταν σε θέση να διαχειριστεί τον όγκο και την πολυπλοκότητα των δεδομένων στη σύγχρονη εποχή.

Η ιδέα ότι τα συστήματα μπορούν να "εκπαιδευτούν" με παραδείγματα και να βρουν μοτίβα και συσχετίσεις μόνα τους προτάθηκε για πρώτη φορά από τη μηχανική μάθηση [6]. Οι υπολογιστές έγιναν πιο ευέλικτοι και μπόρεσαν να προσαρμοστούν σε νέα δεδομένα με αυτόν τον τρόπο. Εφαρμογές όπως τα συστήματα συστάσεων, η αυτόματη ταξινόμηση εικόνων και η έξυπνη πλοήγηση, τις οποίες θεωρούμε πλέον δεδομένες, έγιναν δυνατές χάρη σε αυτή την πρόοδο.

Η βαθιά μάθηση (DL), η οποία χρησιμοποιεί πολυεπίπεδα νευρωνικά δίκτυα, ήταν η επόμενη σημαντική εξέλιξη [7]. Επιτρέπει στα συστήματα να εντοπίζουν πολύ πιο περίπλοκα μοτίβα στη φυσική γλώσσα, στους ήχους και στις εικόνες. Πιο εξελιγμένα μοντέλα, όπως τα LLMs της σημερινής εποχής, αναπτύχθηκαν ως αποτέλεσμα αυτής της τεχνολογικής αλλαγής.



Σχήμα 1.3: Απλό διάγραμμα νευρωνικού δικτύου.

Το "άλμα" που αναζωογόνησε την τεχνητή νοημοσύνη ήταν η μετάβαση από τη μηχανική μάθηση στη βαθιά μάθηση. Η βαθιά μάθηση επέτρεψε στους υπολογιστές να κατανοήσουν πιο περίπλοκα μοτίβα και να γίνουν πιο ανθρώπινοι, βασισμένοι στη γνώση που τους είχε διδάξει η μηχανική μάθηση σχετικά με το πώς να μαθαίνουν από τα δεδομένα.

1.4 Τα Μεγάλα Γλωσσικά Μοντέλα και ο Ρόλος τους στην Επεξεργασία Κειμένου

Τα μεγάλα γλωσσικά μοντέλα μπορούν να κάνουν περισσότερα από το να παράγουν απλώς κείμενο ή να παρέχουν απαντήσεις σε ερωτήσεις. Η επεξεργασία και η κατανόηση τεράστιων ποσοτήτων πληροφοριών είναι μία από τις πιο κρίσιμες ευθύνες τους. Φανταστείτε ένα έγγραφο εκατοντάδων σελίδων. Η ανάγνωσή του απαιτεί πολύ χρόνο και ενέργεια για έναν άνθρωπο, αλλά ένα LLM μπορεί να το επεξεργαστεί σε δευτερόλεπτα και να το συνοψίσει ή να επισημάνει τις κύριες ιδέες του [8].

Δύο χαρακτηριστικά αποτελούν τη βάση αυτής της δυνατότητας [8]:

- **Κατανόηση του Πλαισίου:** Για να κατανοήσουν τη σημασία μιας λέξης, τα LLM εξετάζουν ολόκληρο το κείμενο και όχι μόνο τις μεμονωμένες λέξεις. Ως αποτέλεσμα, μπορούν να χειριστούν περίπλοκες λέξεις και προτάσεις που βρίσκονται σε εξειδικευμένα κείμενα.
- **Παραγωγή δομημένων απαντήσεων:** Το μοντέλο μπορεί να διατάξει τις πληροφορίες με πιο άμεσο και κατανοητό τρόπο, όπως μια λίστα ή μια καλογραμμένη παράγραφο, όταν του το ζητήσουμε.

Σε κείμενα όπως οι κανονισμοί των σχολείων, όπου οι πληροφορίες είναι συχνά διάσπαρτες και γραμμένες σε επίσημο ύφος, αυτή η λειτουργία είναι ιδιαίτερα κρίσιμη. Ο αναγνώστης μπορεί να κατανοήσει γρήγορα τα πιο σημαντικά σημεία χρησιμοποιώντας ένα LLM για να προσδιορίσει πού συζητώνται οι συνθήκες, οι διαδικασίες και οι εξαιρέσεις.

Επιπλέον, τα LLM είναι χρήσιμα και για άλλους σκοπούς [9]:

- **Εξυπηρέτηση πελατών:** Μέσω chatbot ερωταπαντήσεων και αυτοματοποιημένων συστημάτων υποστήριξης.
- **Κοινωνικά μέσα:** Ανίχνευση τάσεων, ανάλυση περιεχομένου και συγγραφή αναρτήσεων.
- **Λιανικό εμπόριο και ηλεκτρονικό εμπόριο:** Βελτιωμένη εμπειρία χρήστη και εξατομικευμένες προτάσεις προϊόντων.
- **Χρηματοοικονομικά:** Αυτοματοποίηση διεργασιών, πρόβλεψη κινδύνων και ανάλυση δεδομένων.

- **Μάρκετινγκ και διαφήμιση:** Ανάπτυξη καμπανιών και περιεχομένου ειδικά για το κοινό.
- **Νομική Τεχνολογία:** Βοήθεια σε δικαστικές διαδικασίες, κατανόηση και αναζήτηση νομικών κειμένων.
- **Υγειονομική περίθαλψη:** Βοηθά τους γιατρούς να ερμηνεύουν ιατρικές πληροφορίες και να εκπαιδεύουν τους ασθενείς.



Σχήμα 1.4: Ενδεικτικοί τομείς εφαρμογής των Μεγάλων Γλωσσικών Μοντέλων.

Ως αποτέλεσμα, οι θεωρητικές συνεισφορές των LLM στην επεξεργασία φυσικής γλώσσας είναι άμεσα χρήσιμες στην πράξη. Πρόκειται για εργαλεία που διευκολύνουν την πρόσβαση στη γνώση, μειώνουν τον χρόνο που απαιτείται για την κατανόηση εκτενών και περίπλοκων κειμένων και βοηθούν τους χρήστες να βρουν γρήγορα τις πληροφορίες που χρειάζονται. Με αυτόν τον τρόπο, καθιστούν δυνατή τη χρήση προηγουμένως ακατανόητου περιεχομένου από περισσότερους ανθρώπους.

1.5 Chatbots με LLMs ως Υποστηρικτικά Εργαλεία στην Εκπαίδευση

Τα chatbots, ή αλλιώς τα συστήματα συνομιλίας που μπορούν να απαντούν σε ερωτήσεις χρηστών σε πραγματικό χρόνο, είναι μία από τις πιο δημοφιλείς εφαρμογές των μεγάλων γλωσσικών μοντέλων. Τα σύγχρονα chatbots με τεχνολογία LLM μπορούν να κατανοήσουν καλύτερα το περιεχόμενο μιας ερώτησης και να παρέχουν πιο διεξοδικές και φυσικές απαντήσεις από τα παλαιότερα μοντέλα που βασίζονταν σε προκαθορισμένους κανόνες και περιορισμένες απαντήσεις.

Επίσης, μπορούν να είναι πολύ χρήσιμα στο εκπαιδευτικό περιβάλλον [10]. Αντί να ψάχνει μέσα σε μακροσκελή κείμενα, ένας φοιτητής μπορεί να τα χρησιμοποιήσει για να βρει πληροφορίες σε οδηγούς σπουδών, κανονισμούς σχολών ή ακόμα και σε υλικό μαθημάτων. Μπορούν να κάνουν μια ερώτηση στα chatbots και να λάβουν μια άμεση, σαφή απάντηση, γλιτώνοντάς τους από την ταλαιπωρία να διαβάσουν δεκάδες σελίδες για να καταλάβουν πώς λειτουργεί μια διαδικασία. Τα LLM δίνουν στα chatbots μια φυσική αίσθηση συνομιλίας που μοιάζει περισσότερο με την ανθρώπινη επικοινωνία. Οι αλληλεπιδράσεις που μπορούν πραγματικά να ωφελήσουν τον χρήστη είναι πλέον πιο σημαντικές από τις άψυχες, ρομποτικές απαντήσεις. Επομένως, τα chatbots που βασίζονται σε LLM είναι ένα χρήσιμο εργαλείο που μπορεί να ενσωματωθεί στις καθημερινές ρουτίνες της ακαδημαϊκής κοινότητας, εκτός από το ότι είναι ένα τεχνολογικό θαύμα.

1.6 Σκοπός και Στόχοι

Ο κύριος στόχος αυτής της εργασίας είναι να εξετάσει τρία διακριτά μεγάλα γλωσσικά μοντέλα που εφαρμόζονται σε συγκεκριμένα δεδομένα και συγκεκριμένα έγγραφα σχετικά με τους σχολικούς κανονισμούς. Στόχος αυτής της διαδικασίας είναι να εξαχθούν διορατικά συμπεράσματα για την πρακτική εφαρμογή τους, να αναδειχθούν τα δυνατά και αδύνατα σημεία κάθε μοντέλου, καθώς και οι διαφορές τους.

Η παρουσίαση της χρήσης γλωσσικών μοντέλων για την ανάλυση μεγάλων και πολύπλοκων κειμένων αποτελεί έναν από τους κύριους στόχους αυτής της εργασίας. Ταυτόχρονα, η μελέτη εστιάζει στη σύγκριση τριών διακριτών μοντέλων όσον αφορά την ακρίβεια, την ταχύτητα επεξεργασίας, τη χρηστικότητα και την πληρότητα των παραγόμενων απαντήσεων. Η μελέτη εξετάζει επίσης τα πλεονεκτήματα και τα μειονεκτήματα της εφαρμογής των LLM σε επίσημα και κανονιστικά κείμενα, όπως οι σχολικοί κανονισμοί, και τονίζει πώς αυτές οι πτυχές επηρεάζουν τη χρησιμότητά τους. Προκειμένου να διαπιστωθεί εάν αυτά τα εργαλεία μπορούν να ενσωματωθούν επιτυχώς στις τακτικές διδακτικές πρακτικές, επιδιώκεται να εντοπιστούν οι ευκαιρίες για τη χρήση τους σε ακαδημαϊκά και ερευνητικά περιβάλλοντα.

Ορισμένα ερευνητικά ερωτήματα που καθοδηγούν την εργασία αναπτύσσονται με βάση αυτούς τους στόχους. Το πρώτο ερώτημα είναι πόσο καλά μπορούν τα μεγάλα γλωσσικά μοντέλα να κατανοήσουν και να αναλύσουν ρυθμιστικά κείμενα. Επιπλέον, τίθεται το ερώτημα πώς διαφέρουν τα τρία μοντέλα όσον αφορά το εύρος και την αξιοπιστία των απαντήσεων. Εξίσου κρίσιμο είναι να εξεταστούν οι επιπτώσεις της χρήσης κάθε μοντέλου στον χρόνο εκτέλεσης και στο κόστος, καθώς πρόκειται για πτυχές χρήσιμες στον τελικό χρήστη. Εξετάζονται επίσης οι προκλήσεις και οι περιορισμοί που μπορεί να προκύψουν στην πράξη, καθώς και ο τρόπος με τον οποίο επηρεάζουν την εφαρμογή των LLM σε πρακτικά περιβάλλοντα. Τέλος, εξετάζεται πώς αυτά τα μοντέλα μπορούν να εφαρμοστούν στο μέλλον στην εκπαίδευση, διευκολύνοντας εκπαιδευτικούς και μαθητές στην πρόσβαση και κατανόηση των διαθέσιμων πληροφοριών.

Κεφάλαιο 2ο: Θεωρητικό Πλαίσιο

2.1 Εισαγωγή στο Θεωρητικό Πλαίσιο

Η θεωρητική βάση της παρούσας διατριβής παρουσιάζεται σε αυτό το κεφάλαιο. Είναι αδύνατο να μελετήσουμε και να συγκρίνουμε μεγάλα γλωσσικά μοντέλα χωρίς πρώτα να κατανοήσουμε τις θεμελιώδεις ιδέες της επεξεργασίας φυσικής γλώσσας και της τεχνητής νοημοσύνης. Αυτό το κεφάλαιο, επομένως, καταβάλλει προσπάθεια να παρουσιάσει τα βασικά θεωρητικά συστατικά που σχετίζονται με το θέμα της εργασίας με αναλυτικό τρόπο.

Το πρώτο θέμα που συζητείται είναι η τεχνητή νοημοσύνη, συμπεριλαμβανομένων των διαφόρων μεθόδων της και της εξέλιξής της τις τελευταίες δεκαετίες. Δεδομένου ότι η μηχανική μάθηση και η βαθιά μάθηση είναι οι τομείς που αποτέλεσαν τη βάση για τις πιο πρόσφατες εξελίξεις, διερευνώνται στη συνέχεια τα θεμελιώδη αυτών των πεδίων. Δεδομένου ότι η επεξεργασία φυσικής γλώσσας είναι ο τομέας όπου χρησιμοποιούνται κυρίως τα LLM, δίνεται ιδιαίτερη προσοχή σε αυτόν τον τομέα.

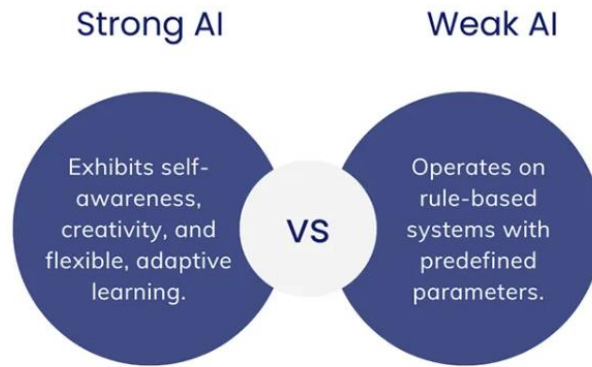
Η αρχιτεκτονική των transformers, η οποία μεταμόρφωσε τον τρόπο με τον οποίο οι μηχανές αναλύουν και παράγουν κείμενο, παρουσιάζεται παρακάτω. Για να δώσουμε μια καλύτερη ιδέα για τις δυνατότητες και τις διακρίσεις τους, παρουσιάζονται στη συνέχεια παραδείγματα μεγάλων γλωσσικών μοντέλων που χρησιμοποιούνται επί του παρόντος – όπως τα GPT, Claude, LLaMA και PaLM.

Τέλος, συζητείται η σχέση μεταξύ των LLM και της εκπαίδευσης, μαζί με τα θεωρητικά ζητήματα που προκύπτουν όταν χρησιμοποιούνται, όπως η προκατάληψη, η ηθική και η αξιοπιστία. Υπό αυτή την έννοια, το θεωρητικό πλαίσιο συνδέεται στενά με το θέμα και τους στόχους της εργασίας, αντί να χρησιμεύει απλώς ως γενική αναφορά στις τεχνολογίες της τεχνητής νοημοσύνης.

2.2 Τεχνητή Νοημοσύνη: Ορισμοί και Βασικές Αρχές

Η τεχνητή νοημοσύνη αποτελεί έναν από τους πιο συναρπαστικούς και ταχύτατα εξελισσόμενους τομείς της επιστήμης των υπολογιστών. Ο όρος περιγράφει την ικανότητα των υπολογιστών και των μηχανών να εκτελούν λειτουργίες που προηγουμένως θεωρούνταν ότι απαιτούσαν ανθρώπινη νοημοσύνη. Αυτές περιλαμβάνουν την κατανόηση της γλώσσας, τη μάθηση βάσει εμπειρίας, την αναγνώριση προτύπων, τη λήψη αποφάσεων και την επίλυση προβλημάτων.

Υπάρχουν πολλές διαφορετικές προσεγγίσεις και υποκατηγορίες στον τομέα της τεχνητής νοημοσύνης. Οι όροι "αδύναμη" ή "στενή" Τεχνητή Νοημοσύνη (weak AI) και "γενική" ή "ισχυρή" Τεχνητή Νοημοσύνη (strong AI) είναι θεμελιωδώς διαφορετικοί [11]. Συστήματα που δημιουργήθηκαν για την εκτέλεση συγκεκριμένων εργασιών, όπως ένα πρόγραμμα σκακιού ή ένα σύστημα αναγνώρισης φωνής, αναφέρονται ως στενή τεχνητή νοημοσύνη. Η ισχυρή Τεχνητή Νοημοσύνη, από την άλλη πλευρά, επιδιώκει να κατασκευάσει μηχανές που να είναι τόσο έξυπνες, προσαρμοστικές και δημιουργικές όσο οι άνθρωποι σε οποιαδήποτε γνωστική εργασία.



Σχήμα 2.1: Σύγκριση ανάμεσα σε StrongAI και WeakAI.

Στις μέρες μας, σχεδόν όλες οι εφαρμογές που συναντάμε εμπίπτουν στην κατηγορία της ασθενούς τεχνητής νοημοσύνης. Οι αλγόριθμοι δημιουργούνται με συγκεκριμένους στόχους κατά νου και στερούνται συνείδησης ή αληθινής κατανόησης, όπως φαίνεται σε όλα, από τους εικονικούς βοηθούς κινητών συσκευών μέχρι τα συστήματα αυτόματης μετάφρασης και τις προτάσεις περιεχομένου στο διαδίκτυο. Αν και η γενική τεχνητή νοημοσύνη ερευνάται και αποτελεί το επίκεντρο έντονης επιστημονικής συζήτησης, παραμένει ακόμα ένας πιο θεωρητικός στόχος.

Οι τεχνικές που χρησιμοποιούνται για τη δημιουργία ευφυών συστημάτων αποτελούν μια ακόμη σημαντική διαφορά. Στις πρώτες μέρες της τεχνητής νοημοσύνης, οι ερευνητές χρησιμοποιούσαν κανόνες και λογικά συστήματα (συστήματα βασισμένα σε κανόνες). Ο προγραμματιστής όριζε με ακρίβεια τα βήματα που έπρεπε να ακολουθήσει το σύστημα σε κάθε περίπτωση. Επειδή όμως οι καταστάσεις του πραγματικού κόσμου είναι πολύ περίπλοκες για να περιγραφούν επαρκώς μόνο από κανόνες, αυτή η προσέγγιση αποδείχθηκε περιορισμένη. Η βασική καινοτομία ήταν η μηχανική μάθηση, η οποία επέτρεψε στα συστήματα να μαθαίνουν και να βελτιώνονται με βάση δεδομένα, χωρίς να απαιτείται ανθρώπινη παρέμβαση σε κάθε στάδιο [12].

Επομένως, η τεχνητή νοημοσύνη δεν είναι μια ενιαία τεχνολογία, αλλά ένα σύνολο στρατηγικών και μεθόδων που επιτρέπουν στους υπολογιστές να μιμούνται ορισμένες πτυχές της ανθρώπινης σκέψης. Για τον λόγο αυτό, η μελέτη πιο εξειδικευμένων εφαρμογών -όπως τα μεγάλα γλωσσικά μοντέλα, που βρίσκονται στην πρώτη γραμμή της τρέχουσας έρευνας- προϋποθέτει κατανόηση αυτών των θεμελιωδών ιδεών.

2.3 Μηχανική Μάθηση και Βαθιά Μάθηση

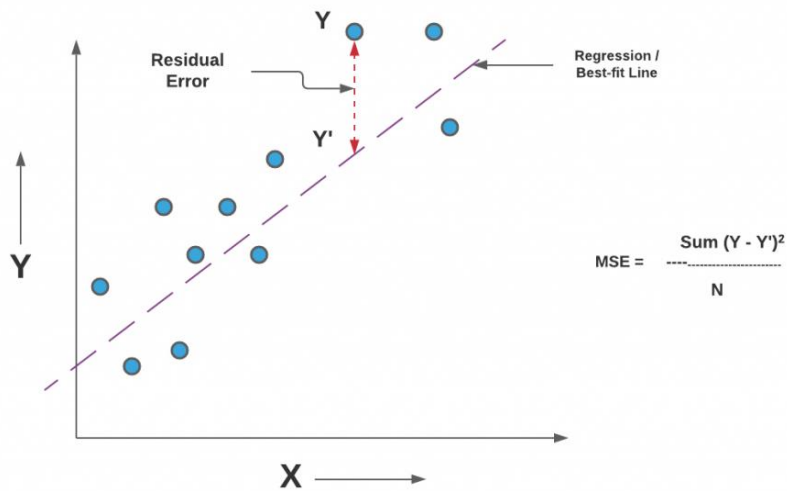
2.3.1 Μηχανική Μάθηση

Ο κλάδος της τεχνητής νοημοσύνης που είναι γνωστός ως μηχανική μάθηση επιτρέπει στα συστήματα να "μαθαίνουν" από δεδομένα και να βελτιώνουν τη λειτουργικότητά τους χωρίς ρητό προγραμματισμό κανόνων. Στόχος της είναι η ανάπτυξη μοντέλων που μπορούν να προβλέπουν ή να αποφασίζουν για νέα, άγνωστα δεδομένα, αφού πρώτα εκπαιδευτούν σε κατάλληλα παραδείγματα.

Μαθηματικά, μπορούμε να πούμε ότι στη μηχανική μάθηση προσπαθούμε να προσεγγίσουμε μια άγνωστη συνάρτηση $f(x)$, έτσι ώστε για κάθε είσοδο x να δίνει την καλύτερη δυνατή εκτίμηση της εξόδου y [13]. Στην πράξη, αυτό συχνά εκφράζεται με την αναζήτηση μιας «καλύτερης ευθείας» (best-fit line) που περνά όσο πιο κοντά γίνεται από τα σημεία των δεδομένων. Η εκπαίδευση γίνεται με πολλά ζεύγη δεδομένων (x_i, y_i) και το μοντέλο βελτιώνεται μειώνοντας το σφάλμα (loss function) [14]. Ένα παράδειγμα συνάρτησης κόστους είναι το μέσο τετραγωνικό σφάλμα (Mean Squared Error, MSE):

$$L = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \quad (2.1)$$

Όσο μικρότερο είναι αυτό το σφάλμα, τόσο καλύτερη είναι η ικανότητα του μοντέλου να μαθαίνει από τα δεδομένα.



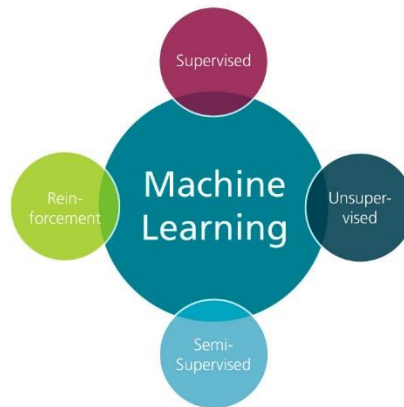
Σχήμα 2.2: Παράδειγμα υπολογισμού του MSE. Η ευθεία προσπαθεί να προσεγγίσει τη σχέση εισόδων-εξόδων, ενώ το MSE μετράει τη διαφορά ανάμεσα στις πραγματικές τιμές (Y) και στις προβλέψεις του μοντέλου (Y').

Η μηχανική μάθηση διακρίνεται σε τέσσερις βασικές κατηγορίες [15]:

- **Εποπτευόμενη μάθηση (Supervised Learning):** Χρησιμοποιούνται δεδομένα που συνοδεύονται από τις σωστές απαντήσεις (labels). Το μοντέλο εκπαιδεύεται πάνω σε αυτά και μαθαίνει τη σχέση εισόδων-εξόδων. Ένα χαρακτηριστικό παράδειγμα είναι η αναγνώριση εικόνων: αν δοθούν σε έναν αλγόριθμο πολλές φωτογραφίες γατών και σκύλων με τις αντίστοιχες ετικέτες, θα μάθει να ξεχωρίζει μόνος του τις δύο κατηγορίες σε νέες φωτογραφίες.
- **Μη εποπτευόμενη μάθηση (Unsupervised Learning):** Τα δεδομένα δεν έχουν έτοιμες απαντήσεις και ο αλγόριθμος προσπαθεί να εντοπίσει μόνος του μοτίβα, συσχετίσεις ή ομάδες. Παράδειγμα αποτελεί η ομαδοποίηση πελατών με βάση τις αγοραστικές τους συνήθειες, όπου ο αλγόριθμος δεν γνωρίζει εκ των προτέρων σε ποια ομάδα ανήκει κάθε πελάτης και προσπαθεί να ανακαλύψει κοινά χαρακτηριστικά.
- **Ενισχυτική μάθηση (Reinforcement Learning):** Το σύστημα «μαθαίνει» μέσα από δοκιμές και λάθη. Ένας πράκτορας (agent) αλληλεπιδρά με ένα περιβάλλον και λαμβάνει ανταμοιβές ή ποινές ανάλογα με τις ενέργειές του, βελτιώνοντας σταδιακά τη στρατηγική του ώστε να

μεγιστοποιήσει τις ανταμοιβές. Τέτοια τεχνική έχει χρησιμοποιείται, για παράδειγμα, σε προγράμματα που μαθαίνουν να παίζουν βιντεοπαιχνίδια ή σε ρομπότ που κινούνται αυτόνομα.

- **Ημι-εποπτευόμενη μάθηση (Semi-Supervised Learning):** Συνδυάζει εποπτευόμενη και μη εποπτευόμενη μάθηση. Σε αυτήν την περίπτωση, το μοντέλο εκπαιδεύεται με λίγα δεδομένα που έχουν ετικέτες και με πολλά περισσότερα χωρίς. Έτσι, αξιοποιεί τις ετικέτες για να κατανοήσει καλύτερα τα μη χαρακτηρισμένα δεδομένα. Ένα χαρακτηριστικό παράδειγμα είναι η ταξινόμηση ιατρικών εικόνων, όπου συνήθως υπάρχει μικρός αριθμός φωτογραφιών που έχουν σχολιαστεί από γιατρούς, αλλά πολύ περισσότερες χωρίς ετικέτες.



Σχήμα 2.3: Βασικές κατηγορίες της Μηχανικής Μάθησης.

Όλοι οι τύποι μηχανικής μάθησης αποτελούν τη βάση για πιο εξελιγμένες μεθόδους όπως η βαθιά μάθηση και τα μεγάλα γλωσσικά μοντέλα (LLMs), τα οποία θα εξετάσουμε στα επόμενα τμήματα.

Για τα LLM, η self-supervised μάθηση είναι μια ιδιαίτερα σημαντική μορφή μάθησης. Σε αυτή την περίπτωση, οι "σωστές απαντήσεις" που πρέπει να μάθει το μοντέλο παράγονται από τα ίδια τα δεδομένα. Το μοντέλο καλείται να μαντέψει τη λέξη "ανατολή", για παράδειγμα, στην πρόταση "Ο ήλιος ανατέλλει από την ____." Αυτό εξαλείφει την ανάγκη για ανθρώπινη επισήμανση, εκπαιδεύοντας το σύστημα να προβλέπει την επόμενη λέξη ή τμήματα κειμένου. Αυτή η στρατηγική είναι σημαντική επειδή καθιστά δυνατή την εκπαίδευση σε μεγάλα σύνολα δεδομένων, κάτι που είναι απαραίτητο για τη δημιουργία LLM και, κατά συνέπεια, για μεθόδους όπως το RAG, το οποίο θα εξετάσουμε αργότερα, που χρησιμοποιούν τη δημιουργία κειμένου και εξωτερικές πηγές γνώσης.

2.3.2 Βαθιά Μάθηση

Τα τεχνητά νευρωνικά δίκτυα αποτελούν τη βάση της βαθιάς μάθησης, ενός από τα σημαντικότερα υποπεδία της μηχανικής μάθησης. Εμπνευσμένα από τη δομή του ανθρώπινου εγκεφάλου, αυτά τα δίκτυα αποτελούνται από "νευρώνες" (κόμβους) που λαμβάνουν, επεξεργάζονται και στέλνουν πληροφορίες σε άλλους νευρώνες [16].

Η βασική ιδέα είναι ότι τα δεδομένα περνούν από πολλαπλά στρώματα, το καθένα από τα οποία εξάγει ένα διαφορετικό χαρακτηριστικό. Για παράδειγμα, το πρώτο επίπεδο μπορεί να εντοπίζει γραμμές και σχήματα σε μια εικόνα, ακολουθούμενα από πιο περίπλοκα χαρακτηριστικά όπως στόματα ή μάτια, και τελικά, ολόκληρα πρόσωπα. Το δίκτυο μπορεί να βρει πιο περίπλοκα μοτίβα όσο περισσότερα στρώματα έχει.

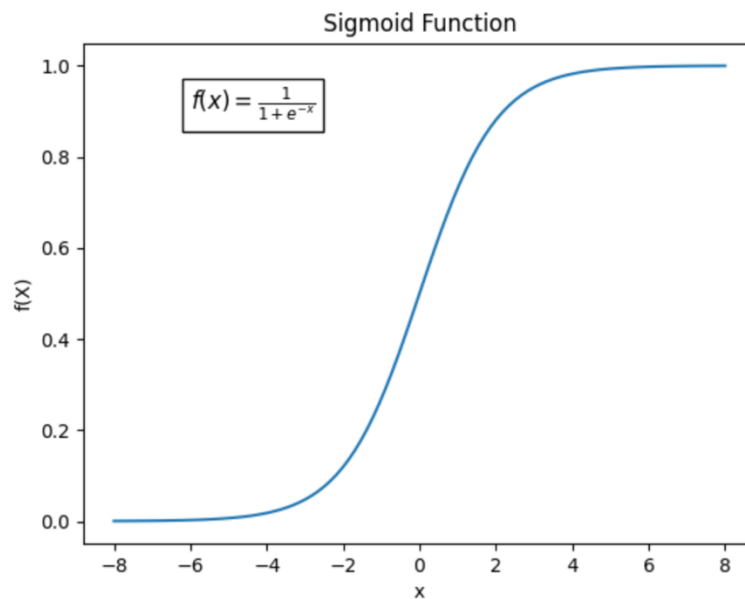
Μαθηματικά, ένας τεχνητός νευρώνας υπολογίζει το εξής:

$$z = \sum_{i=1}^n w_i x_i + b \quad (2.2)$$

Όπου x_i είναι οι εισόδοι, w_i είναι βάρη που καθορίζουν τη σημασία κάθε εισόδου και b είναι μια σταθερά που ονομάζεται bias. Στη συνέχεια, το αποτέλεσμα περνά από μια συνάρτηση ενεργοποίησης (activation function), η οποία αποφασίζει αν και σε ποιο βαθμό θα «ενεργοποιηθεί» ο νευρώνας. Δηλαδή, η συνάρτηση ενεργοποίησης εισάγει μη γραμμικότητα στο δίκτυο, επιτρέποντάς του να μάθει πιο σύνθετες σχέσεις που δεν μπορούν να περιγραφούν μόνο με ευθείες γραμμές.

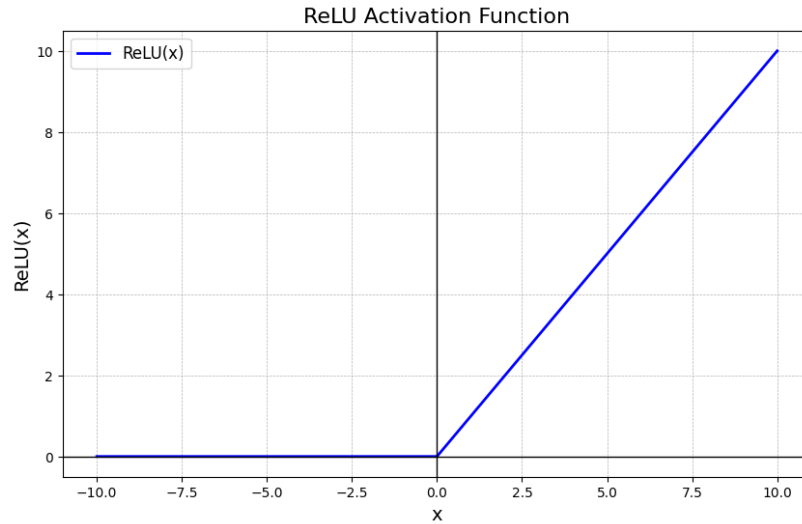
Κάποιες από τις πιο γνωστές συναρτήσεις ενεργοποίησης είναι [17]:

- **Sigmoid:** μετατρέπει τις τιμές σε εύρος από 0 έως 1, χρήσιμη σε προβλήματα δυαδικής ταξινόμησης.



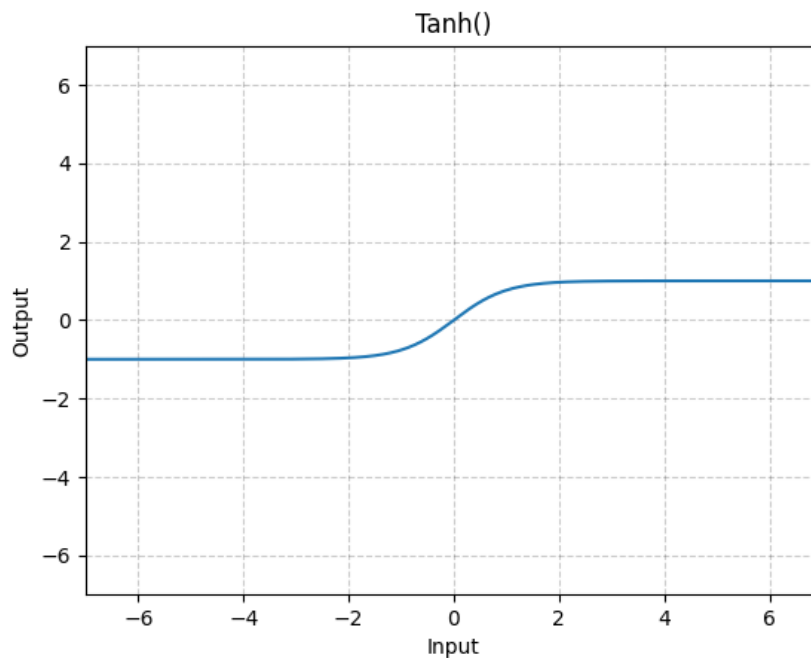
Σχήμα 2.4: Συνάρτηση ενεργοποίησης Sigmoid.

- **ReLU (Rectified Linear Unit):** δίνει τιμή 0 για αρνητικές εισόδους και κρατάει τις θετικές ίδιες. Είναι πολύ διαδεδομένη γιατί είναι απλή και γρήγορη.



Σχήμα 2.5: Συνάρτηση ενεργοποίησης ReLU.

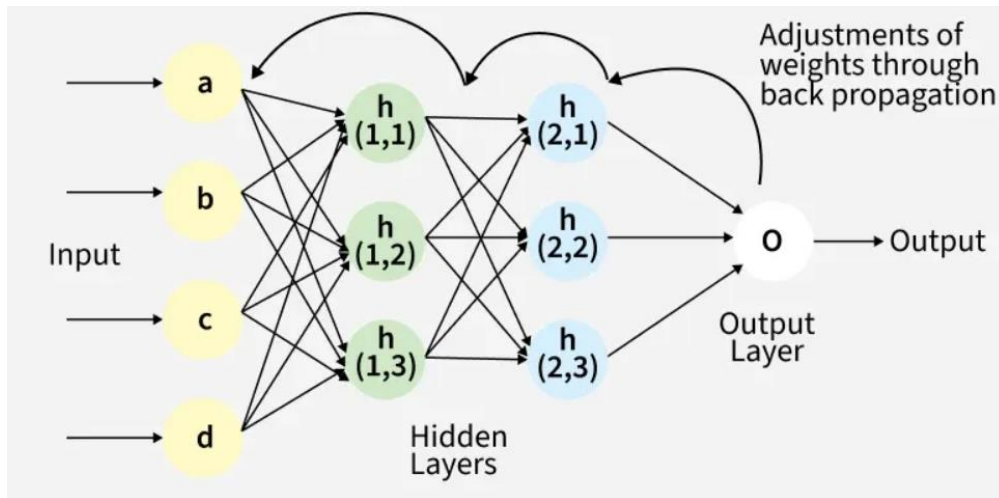
- **Tanh (Hyperbolic Tangent):** μοιάζει με τη sigmoid αλλά δίνει τιμές από -1 έως 1, προσφέροντας καλύτερη ισορροπία σε ορισμένες περιπτώσεις.



Σχήμα 2.6: Συνάρτηση ενεργοποίησης Tanh.

Μέχρι να παραχθεί η τελική έξοδος, αυτή η διαδικασία επαναλαμβάνεται σε κάθε επίπεδο του δικτύου. Η μέθοδος της διάδοσης προς τα πίσω χρησιμοποιείται για την εκπαίδευση, κατά την οποία η διαφορά

μεταξύ των αναμενόμενων και πραγματικών τιμών "επιστρέφει" στο δίκτυο και τροποποιεί τα βάρη, επιτρέποντας στο μοντέλο να βελτιώνεται με την πάροδο του χρόνου.



Σχήμα 2.7: Νευρωνικό δίκτυο με επίπεδο εισόδου, κρυφά επίπεδα και επίπεδο εξόδου. Η εκπαίδευση γίνεται μέσω οπισθοδιάδοσης (backpropagation).

Οι εφαρμογές της βαθιάς μάθησης χρησιμοποιούνται πλέον ευρέως σε αυτόνομα αυτοκίνητα, ανάλυση ιατρικών εικόνων, μετάφραση κειμένου (Google Translate), αναγνώριση φωνής (Siri, Google Assistant) και ακόμη και στα μέσα κοινωνικής δικτύωσης που κάνουν προτάσεις περιεχομένου στους χρήστες.

Το γεγονός ότι τα μεγάλα γλωσσικά μοντέλα (LLMs) αναπτύχθηκαν χρησιμοποιώντας βαθιά μάθηση είναι ιδιαίτερα σημαντικό. Η επεξεργασία φυσικής γλώσσας κατέστη δυνατή σε πρωτοφανή κλίμακα χάρη σε μοναδικές αρχιτεκτονικές νευρωνικών δικτύων, όπως τα Επαναληπτικά Νευρωνικά Δίκτυα (RNNs). Η ανάπτυξη των σημερινών μοντέλων, τα οποία μπορούν να κατανοήσουν και να παράγουν κείμενο με αξιοσημείωτη ακρίβεια, κατέστη δυνατή χάρη σε αυτή την τεχνολογική πρόοδο.

2.4 Επεξεργασία Φυσικής Γλώσσας (NLP)

2.4.1 Ορισμός και Σημασία

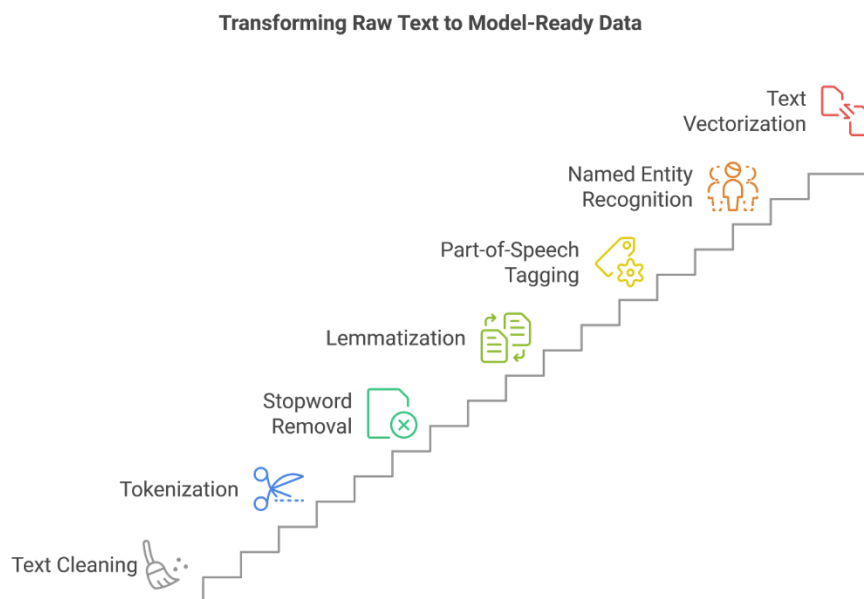
Ο κλάδος της τεχνητής νοημοσύνης που μελετά την επικοινωνία ανθρώπου και μηχανής με βάση τη γλώσσα ονομάζεται Επεξεργασία Φυσικής Γλώσσας, ή NLP. Στόχος του είναι να καταστήσει δυνατή την "κατανόηση" της προφορικής ή γραπτής γλώσσας από τους υπολογιστές και την παραγωγή απαντήσεων σε φυσική γλώσσα. Εφαρμογές όπως τα chatbots, οι προσωπικοί βοηθοί (Alexa, Siri) και οι αυτόματοι μεταφραστές αποτελούν πλέον κοινό μέρος της καθημερινής ζωής χάρη στην επεξεργασία φυσικής γλώσσας.

2.4.2 Βασικά Στάδια Επεξεργασίας

Η διαδικασία επεξεργασίας φυσικής γλώσσας δεν είναι απλή. Ένας υπολογιστής πρέπει να περάσει από διάφορα στάδια για να αποκτήσει "νόημα" [18], ενώ ένας άνθρωπος μπορεί να κατανοήσει ένα κείμενο αμέσως.

Τα πιο κρίσιμα στάδια είναι:

- **Καθαρισμός κειμένου (Text Cleaning):** Αφαίρεση άχρηστων χαρακτήρων, σημείων στίξης ή ειδικών συμβόλων που δεν χρειάζονται.
- **Διαίρεση σε μονάδες (Tokenization):** Το κείμενο χωρίζεται σε μικρότερα κομμάτια (tokens), όπως λέξεις ή προτάσεις.
- **Αφαίρεση άσχετων λέξεων (Stop-word Removal):** Λέξεις όπως «και», «αλλά», «το» δεν προσφέρουν ουσιαστική πληροφορία και συχνά παραλείπονται.
- **Λεμματοποίηση (Lemmatization):** Οι λέξεις μετατρέπονται στη βασική τους μορφή. Για παράδειγμα, «έτρεξα», «τρέχοντας» γίνονται «τρέχω».
- **Γραμματικός χαρακτηρισμός (Part-of-Speech Tagging):** Αναγνώριση του ρόλου κάθε λέξης (ρήμα, ουσιαστικό, επίθετο).
- **Αναγνώριση Οντοτήτων (Named Entity Recognition - NER):** Εντοπισμός συγκεκριμένων στοιχείων μέσα στο κείμενο, όπως ονόματα, τοποθεσίες ή ημερομηνίες.
- **Διανυσματοποίηση κειμένου (Text Vectorization / Embeddings):** Τελικό στάδιο όπου οι λέξεις ή οι προτάσεις μετατρέπονται σε αριθμητικά διανύσματα, ώστε να μπορούν να χρησιμοποιηθούν από νευρωνικά δίκτυα και LLMs.



Σχήμα 2.8: Τα βασικά στάδια προεπεξεργασίας κειμένου στο NLP, από τον καθαρισμό μέχρι τη διανυσματοποίηση.

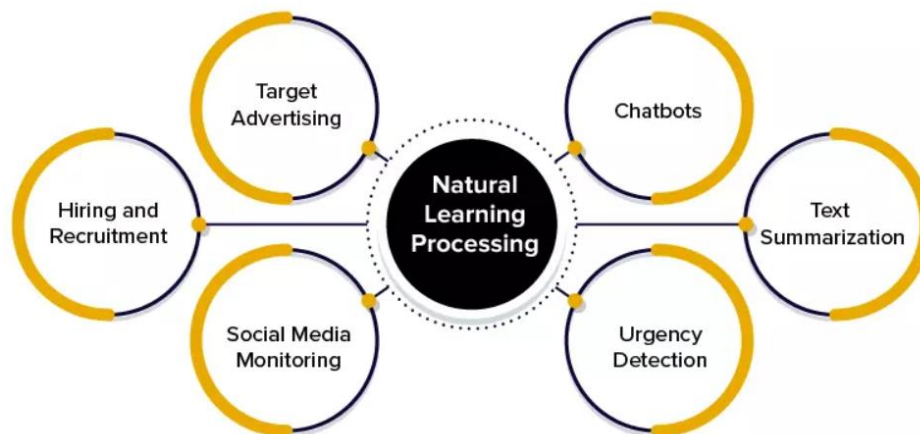
2.4.3 Εφαρμογές του NLP

Με τις καθημερινές της εφαρμογές, η επεξεργασία φυσικής γλώσσας έχει προχωρήσει από τη θεωρία στην πράξη. Ένα από τα καλύτερα παραδείγματα είναι η αυτόματη μετάφραση, όπως το Google Translate, που καθιστά δυνατή τη στιγμιαία μετάφραση κειμένου σε άλλη γλώσσα. Αυτά τα συστήματα επιδιώκουν να αποδώσουν το νόημα, λαμβάνοντας υπόψη το συμφραζόμενο, αντί να μεταφράζουν λέξη προς λέξη, χρησιμοποιώντας εξελιγμένους αλγορίθμους και μεγάλα γλωσσικά μοντέλα [19].

Η αυτόματη περίληψη κειμένου είναι μια άλλη σημαντική εφαρμογή. Σήμερα, υπάρχουν μοντέλα που μπορούν να διαβάσουν μια εκτενή μελέτη ή άρθρο και να αποστάξουν τις βασικές ιδέες σε μια σύντομη περίληψη. Σε τομείς όπως η έρευνα και η εκπαίδευση, όπου ο χρόνος είναι περιορισμένος και ο όγκος πληροφοριών μεγάλος, αυτό είναι ιδιαίτερα χρήσιμο.

Η ανάλυση συναισθημάτων, που αφορά την ανάλυση της στάσης ή του συναισθήματος που εκφράζεται σε ένα κείμενο, χρησιμοποιεί επίσης την επεξεργασία φυσικής γλώσσας. Για παράδειγμα, μια ανάρτηση στα μέσα κοινωνικής δικτύωσης μπορεί να κατηγοριοποιηθεί ως ουδέτερη, αρνητική ή θετική. Για να βοηθήσει τις επιχειρήσεις να κατανοήσουν καλύτερα τις αντιδράσεις των χρηστών, αυτή η δυνατότητα χρησιμοποιείται κυρίως στο μάρκετινγκ και στην έρευνα κοινού.

Η αναγνώριση ομιλίας είναι μια ακόμα σημαντική εφαρμογή. Σήμερα, οι φωνητικές εντολές μπορούν να γίνουν κατανοητές και να απαντηθούν με φυσικό τρόπο από προσωπικούς βοηθούς όπως ο Βοηθός Google, η Alexa και η Siri. Ταυτόχρονα, τα chatbots έχουν εξελιχθεί σε σημαντικά εργαλεία επικοινωνίας. Τα σύγχρονα chatbots βασίζονται σε εκτεταμένα γλωσσικά μοντέλα και είναι σε θέση να απαντούν με μεγαλύτερη ακρίβεια και φυσικότητα από τα παλαιότερα, που λειτουργούσαν με προκαθορισμένους κανόνες, επιτρέποντας στους μαθητές να αναζητήσουν πληροφορίες σε οδηγούς σπουδών ή σχολικούς κανονισμούς χωρίς να χρειάζεται να διαβάσουν ολόκληρα κείμενα.



Σχήμα 2.9: Ενδεικτικές εφαρμογές της Επεξεργασίας Φυσικής Γλώσσας (NLP), όπως τα chatbots, η σύνοψη κειμένων, η ανάλυση κοινωνικών δικτύων και η στοχευμένη διαφήμιση.

Τέλος, τα συστήματα αναζήτησης βασίζονται επίσης στην επεξεργασία φυσικής γλώσσας. Η Google και παρόμοιες μηχανές αναζήτησης κατανοούν το ερώτημα του χρήστη, αναγνωρίζουν τη σημασία του και ανακτούν τα πιο συναφή δεδομένα. Αν και για τον χρήστη αυτή η διαδικασία αυτή φαίνεται απλή, στηρίζεται σε ιδιαίτερα εξελιγμένες μεθόδους επεξεργασίας φυσικής γλώσσας.

Η ευελιξία του NLP καταδεικνύει ότι πρόκειται για μια τεχνολογία με ιδιαίτερα εκτεταμένες δυνατότητες, η οποία μπορεί να εφαρμοστεί σε εξειδικευμένους τομείς, όπως η ιατρική διάγνωση, η ανάλυση νομικών κειμένων ή η ανάλυση οικονομικών δεδομένων, καθώς και σε πιο κοινούς τομείς όπως η μετάφραση ή η επικοινωνία με chatbots.

2.5 Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs)

2.5.1 Ορισμός και Σημασία

Μία από τις πιο αξιοσημείωτες προόδους στην τεχνητή νοημοσύνη τα τελευταία χρόνια είναι η χρήση μεγάλων γλωσσικών μοντέλων. Για να κατανοήσουν, να επεξεργαστούν και να παράγουν φυσική γλώσσα, αυτά τα συστήματα εκπαιδεύονται σε τεράστιες ποσότητες κειμένου. Διαθέτουν δισεκατομμύρια παραμέτρους, γι' αυτό και χαρακτηρίζονται «ως μεγάλα». Οι παράμετροι αυτές είναι υπεύθυνες για την επεξεργασία και αποθήκευση της γνώσης, γεγονός που επιτρέπει στα LLM να αναπαριστούν περίπλοκες γλωσσικές σχέσεις και να παράγουν απαντήσεις που συχνά φαίνονται φυσικές και ανθρώπινες [20].

Η self-supervised μάθηση αποτελεί την κύρια μέθοδο εκπαίδευσης των LLM. Συγκεκριμένα, το μοντέλο καλείται να συμπληρώσει κενά ή να προβλέψει την επόμενη λέξη σε μια πρόταση, αντλώντας τη δομή και τη σημασιολογία της γλώσσας από εκατομμύρια τέτοια παραδείγματα. Με αυτή την τεχνική εκπαίδευσης, τα μοντέλα αξιοποιούν μεγάλα σύνολα δεδομένων χωρίς να χρειάζονται ανθρώπινη επίσημανση.

Δεδομένου ότι τα LLM μετασχηματίζουν ριζικά τον τρόπο με τον οποίο οι άνθρωποι αλληλεπιδρούν με τις μηχανές, η σημασία τους είναι τεράστια. Οι άνθρωποι επικοινωνούν φυσικά μέσω της γλώσσας, την οποία τα υπολογιστικά συστήματα είναι πλέον σε θέση να "κατανοήσουν". Αυτό συνεπάγεται ότι ο χρήστης χρειάζεται μόνο να υποβάλει ένα ερώτημα στη φυσική γλώσσα και να λάβει απάντηση, χωρίς να απαιτούνται γνώσεις προγραμματισμού. Ως αποτέλεσμα, τα LLM γεφυρώνουν το χάσμα μεταξύ ανθρώπου και υπολογιστή, αυξάνοντας τη χρηστικότητα και την προσβασιμότητα της τεχνητής νοημοσύνης.

Έχουν ένα ευρύ φάσμα εφαρμογών. Μπορούν να γράψουν κώδικα προγραμματισμού, να μεταφράσουν μεταξύ γλωσσών, να λύσουν μαθηματικά προβλήματα, να συνοψίσουν εκτενή κείμενα και να παράγουν περιεχόμενο, όπως αναφορές ή άρθρα. Τα LLM μπορούν επίσης να αξιοποιηθούν στην εκπαίδευση για να υποστηρίξουν τους μαθητές, απαντώντας στις ερωτήσεις που αφορούν οργανωτικά ζητήματα, σχολικούς κανονισμούς ή ακόμα και επιστημονική βιβλιογραφία. Λειτουργούν ως "έξυπνοι βοηθοί" που επιταχύνουν την αναζήτηση πληροφοριών και ενισχύουν την παραγωγικότητα.

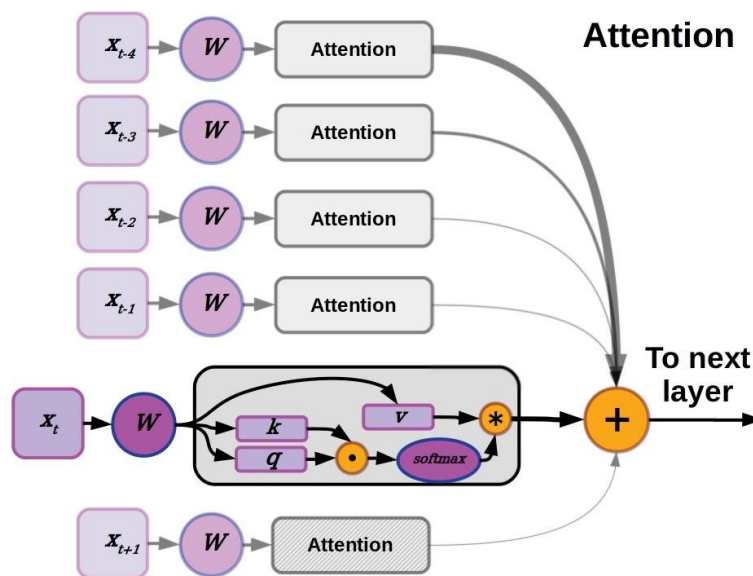
Ωστόσο, η τεράστια ισχύς τους συνοδεύεται από προκλήσεις. Για την εκπαίδευση τους απαιτείται μεγάλη υπολογιστική ισχύς και κατανάλωση ενέργειας, ενώ ενδέχεται να αναπαράγουν προκαταλήψεις που υπάρχουν στα δεδομένα. Επιπλέον, τα μοντέλα μπορούν να παράγουν πληροφορίες που δεν είναι ακριβείς, οπότε παρόλο που οι απαντήσεις τους είναι συχνά εντυπωσιακές, δεν είναι πάντοτε απολύτως αξιόπιστες. Παρ' όλα αυτά, η ανάπτυξή τους αντιπροσωπεύει μια αλλαγή παραδείγματος για το μέλλον της τεχνητής νοημοσύνης, και η αξία τους δεν μπορεί να αμφισβητηθεί [20].

Τα LLM αποτελούν πλατφόρμες που φέρνουν επανάσταση στην αλληλεπίδραση της γνώσης, της εκπαίδευσης και της τεχνολογίας. Είναι κάτι περισσότερο από απλά εργαλεία επεξεργασίας γλώσσας. Η συνεισφορά τους ξεπερνά τις επιμέρους εφαρμογές και ανοίγει τον δρόμο σε μια νέα εποχή, στην οποία οι υπολογιστές θα μπορούν να υποστηρίξουν την ανθρώπινη δημιουργικότητα και την τεχνολογική ανάπτυξη με τρόπους που προηγουμένως θεωρούνταν αδύνατοι.

2.5.2 Αρχιτεκτονική Transformers

Η αρχιτεκτονική Transformer, που παρουσιάστηκε για πρώτη φορά από τους Vaswani et al. στην εργασία τους του 2017 "Attention is All You Need" [21], αποτελεί τη βασική καινοτομία που συνέβαλε στην αποτελεσματικότητα των Μεγάλων Γλωσσικών Μοντέλων (LLMs). Ο μηχανισμός προσοχής (attention) είναι το κεντρικό συστατικό αυτής της αρχιτεκτονικής.

Με απλά λόγια, η προσοχή επιτρέπει στο μοντέλο να "εστιάσει" στις πιο σχετικές λέξεις σε μια πρόταση, προκειμένου να κατανοήσει το νόημά της [21]. Τα Transformers έχουν τη δυνατότητα να επεξεργάζονται όλες τις λέξεις ταυτόχρονα και να αποδίδουν σημασία σε καθεμία ανάλογα με το πλαίσιο, σε αντίθεση με τα προηγούμενα μοντέλα που επεξεργάζονταν τις λέξεις διαδοχικά (μία μετά την άλλη).



Σχήμα 2.10: Ο μηχανισμός Προσοχής (Attention) σε ένα Transformer.

Κάθε λέξη (είσοδος x) μετατρέπεται σε τρία διακριτά διανύσματα, Ερώτημα (Q), Κλειδί (K) και Τιμή (V), όπως φαίνεται στο παραπάνω διάγραμμα. Οι βαθμοί προσοχής (attention weights), οι οποίοι δείχνουν πόσο σημαντική είναι μια λέξη σε σχέση με τις άλλες, υπολογίζονται συγκρίνοντας αυτά τα διανύσματα. Όλες οι λέξεις που προχωρούν στο επόμενο επίπεδο συνδυάζονται και σταθμίζονται με βάση τη σημασία τους.

Για παράδειγμα, στην πρόταση "Ο μαθητής διάβασε το βιβλίο επειδή είχε εξετάσεις", το μοντέλο πρέπει να κατανοήσει ότι η φράση "είχε εξετάσεις" αναφέρεται στον "μαθητή" και όχι στο "βιβλίο". Η προσοχή το επιτυγχάνει δίνοντας περισσότερο βάρος στη λέξη "μαθητής" και λιγότερο στη λέξη "βιβλίο".

Επειδή η αρχιτεκτονική Transformer υποστηρίζει την παράλληλη επεξεργασία, έχει επίσης το πλεονέκτημα ότι μπορεί να εκπαιδευτεί σε μεγάλα κείμενα πολύ πιο γρήγορα. Τα σημερινά LLM βασίζονται σε μοντέλα με δισεκατομμύρια παραμέτρους, κάτι που έγινε δυνατό χάρη σε αυτή την αρχιτεκτονική.

2.6 Retrieval-Augmented Generation (RAG)

2.6.1 Ορισμός και Σημασία

Μια σχετικά πρόσφατη μέθοδος αξιολόγησης των Μεγάλων Γλωσσικών Μοντέλων (LLMs) είναι το Retrieval-Augmented Generation (RAG). Η μέθοδος αυτή στοχεύει κυρίως στην αντιμετώπιση ενός από τα μεγαλύτερα μειονεκτήματα των LLM: την αδυναμία τους να "γνωρίζουν" πιο πρόσφατες πληροφορίες και την εξάρτησή τους από όσα έχουν μάθει κατά τη διάρκεια της εκπαίδευσης. Με απλά λόγια, το RAG δημιουργεί μια τελική έξοδο που είναι πιο ακριβής, επίκαιρη και τεκμηριωμένη, συνδυάζοντας τις γενετικές δυνατότητες των LLM με τη δύναμη της ανάκτησης πληροφοριών.

Το RAG είναι σημαντικό επειδή επιτρέπει στα LLM να χρησιμοποιούν εξωτερικές πηγές γνώσης, αντί να βασίζονται αποκλειστικά στη "μνήμη" τους. Έτσι, προσφέρει ένα πιο "ανοιχτό" σύστημα που αντλεί και χρησιμοποιεί πληροφορίες σε πραγματικό χρόνο, σε αντίθεση με ένα "κλειστό" μοντέλο που μπορεί να ξεχάσει ή να δώσει εσφαλμένες απαντήσεις.

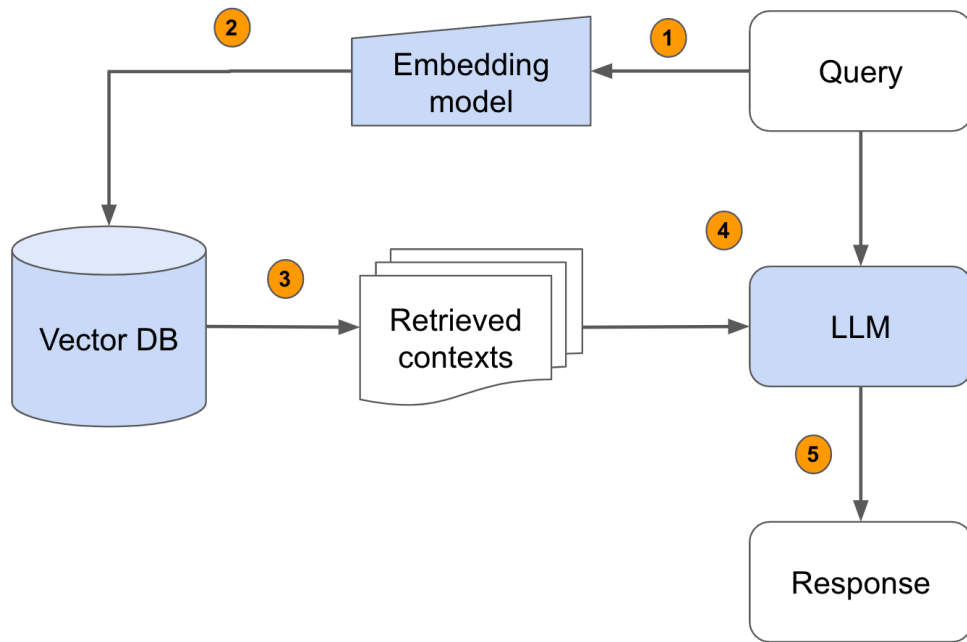
2.6.2 Πως Λειτουργεί το RAG

Η ανάκτηση πληροφοριών και η δημιουργία απαντήσεων είναι δύο διακριτές διεργασίες που συνδυάζονται για να αποτελέσουν τη βάση της μεθόδου RAG [22]. Με την ενσωμάτωση νέων, τρεχόντων και εξειδικευμένων κειμένων, το σύστημα παρακάμπτει το μειονέκτημα των μεγάλων γλωσσικών μοντέλων που βασίζονται μόνο σε δεδομένα εκπαίδευσης.

Όταν ένας χρήστης υποβάλλει μια ερώτηση, μετατρέπεται αρχικά σε ένα embedding, το οποίο είναι μια αριθμητική αναπαράσταση. Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, τα embeddings είναι διανύσματα που αντιπροσωπεύουν τη σημασία μιας πρότασης σε έναν πολυδιάστατο χώρο, οργανώνοντας τις σχετικές έννοιες με ομαδοποιημένο τρόπο. Για παράδειγμα, η πρόταση "Πόσες φορές μπορώ να κάνω μάθημα;" θα μετατραπεί σε ένα διάνυσμα που αποτυπώνει το πνεύμα της ιδέας.

Στη συνέχεια, το σύστημα αντιπαραβάλλει αυτό το embedding με embeddings αποσπασμάτων ή εγγράφων που διατηρούνται σε μια βάση γνώσεων. Η ομοιότητα μετράται προκειμένου να γίνει η σύγκριση, τυπικά χρησιμοποιώντας μετρικές όπως η ομοιότητα συνημίτονου. Όσο πιο κοντά είναι τα διανύσματα στο αρχικό ερώτημα, τόσο πιο σχετικό θεωρείται το έγγραφο. Τα πιο σχετικά κείμενα — συνήθως τα καλύτερα αποτελέσματα — επιλέγονται από όλα όσα είναι διαθέσιμα.

Ένα εμπλουτισμένο ερώτημα παράγεται συνδυάζοντας το αρχικό ερώτημα του χρήστη με τα σχετικά αποσπάσματα που έχουν ανακτηθεί. Στο τέλος, αυτό το νέο κείμενο προστίθεται στο γλωσσικό μοντέλο, ώστε η παραγόμενη απάντηση να βασίζεται τόσο σε πραγματικά, συγκεκριμένα δεδομένα όσο και στη γενική γνώση του μοντέλου. Ως αποτέλεσμα, η τελική πρόβλεψη είναι μια απάντηση που περιλαμβάνει πληροφορίες από αξιόπιστες πηγές, αντί για απλώς μια πιθανή συνέχεια της πρότασης.



Σχήμα 2.11: Διάγραμμα ροής της μεθόδου RAG με τα βήματα από την ερώτηση του χρήστη έως την τελική απάντηση.

Μαθηματικά, το RAG μπορεί να θεωρηθεί ως η εκτίμηση μιας πιθανότητας απάντησης $P(y|q, d_1, \dots, d_k)$, όπου η απάντηση y εξαρτάται τόσο από την ερώτηση q όσο και από τα αποσπάσματα d_1, \dots, d_k που ανακτήθηκαν. Η διαφορά σε σχέση με τα κλασικά LLMs είναι ότι η πιθανότητα αυτή υπολογίζεται όχι μόνο με βάση τη μνήμη του μοντέλου, αλλά και με βάση την εξωτερική γνώση.

Αυτή η λειτουργία αποτελεί παράδειγμα του πώς το RAG γεφυρώνει δύο κόσμους: την ακρίβεια που παρέχεται από εξωτερικές πηγές γνώσης και την ικανότητα των LLM να παράγουν φυσική και συνεκτική γλώσσα. Ως αποτέλεσμα, οι παραγόμενες απαντήσεις είναι κείμενα που υποστηρίζονται από πραγματικά έγγραφα και όχι απλώς πιθανές προβλέψεις βασισμένες στην "μνήμη" του μοντέλου. Γι' αυτό, το RAG είναι ιδιαίτερα χρήσιμο σε περιπτώσεις όπου η ακρίβεια των πληροφοριών είναι ζωτικής σημασίας, όπως στην νομική ανάλυση, την ακαδημαϊκή έρευνα ή τη δημιουργία διοικητικών κανονισμών.

2.7 Αξιολόγηση Μεγάλων Γλωσσικών Μοντέλων

Επειδή μας επιτρέπει να αξιολογήσουμε την ακρίβεια, την αποτελεσματικότητα και τη χρησιμότητα των απαντήσεων που παράγουν τα μεγάλα γλωσσικά μοντέλα, η αξιολόγηση αποτελεί έναν κρίσιμο τομέα έρευνας. Καθώς αυτά τα μοντέλα πρέπει να λειτουργούν σε ποικίλα περιβάλλοντα και εφαρμογές, απαιτούν μετρικές που λαμβάνουν υπόψη την ακρίβεια, την ταχύτητα και την πληρότητα της εξόδου [23].

Η ακρίβεια, δηλαδή η αναλογία των απαντήσεων που ταιριάζουν με τις σωστές σύμφωνα με κάποια αναφορά αλήθειας, αποτελεί κρίσιμη μέτρηση, καθώς δείχνει πόσο καλά το μοντέλο μπορεί να παράγει το επιθυμητό αποτέλεσμα.

Επιπλέον, είναι κοινή πρακτική να εξετάζεται η ομοιότητα δύο απαντήσεων. Το ίδιο νόημα μπορεί να μεταφερθεί ακόμα και με διαφορετική διατύπωση. Η μέτρηση της ομοιότητας προσφέρει μια πιο ευέλικτη άποψη της ποιότητας του αποτελέσματος και μπορεί να υπολογιστεί χρησιμοποιώντας μια ποικιλία τεχνικών, συμπεριλαμβανομένης της ομοιότητας συνημιτόνου μεταξύ των αναπαραστάσεων των διανυσμάτων κειμένου.

Ένας άλλος σημαντικός παράγοντας είναι ο χρόνος απόκρισης (καθυστέρηση), δηλαδή πόσο χρόνο χρειάζεται στο μοντέλο για να δημιουργήσει μια απάντηση. Ο χρόνος απόκρισης είναι μια κρίσιμη μέτρηση για την αξιολόγηση της λειτουργικότητας σε εφαρμογές που απαιτούν αλληλεπίδραση σε πραγματικό χρόνο, όπως τα chatbots.

Το μήκος της απάντησης χρησιμεύει ως πρόσθετη μέτρηση. Οι απαντήσεις μπορεί να είναι υπερβολικά εκτενείς και να περιλαμβάνουν άσχετες λεπτομέρειες, ή πολύ σύντομες, παραλείποντας σημαντικές πληροφορίες. Η παρακολούθηση του μήκους διευκολύνει τον προσδιορισμό του κατά πόσο το μοντέλο παράγει αποτελέσματα που είναι τόσο χρήσιμα όσο και ισορροπημένα.

Μετρικές που δημιουργήθηκαν ειδικά για εργασίες επεξεργασίας κειμένου χρησιμοποιούνται επίσης στη διεθνή βιβλιογραφία. Για παράδειγμα, το εργαλείο μηχανικής μετάφρασης BLEU (Bilingual Evaluation Understudy) συγκρίνει την έξοδο του μοντέλου με μία ή περισσότερες μεταφράσεις αναφοράς. Η αξιολόγηση της περίληψης κειμένου χρησιμοποιεί συχνά το ROUGE (Recall-Oriented Understudy for Gisting Evaluation), το οποίο ποσοτικοποιεί την επικάλυψη λέξεων και φράσεων μεταξύ της περίληψης του μοντέλου και των ανθρώπινων περιλήψεων. Μια πιο ισορροπημένη εικόνα της απόδοσης προσφέρουν άλλες μετρικές, όπως το METEOR ή το F1-score, οι οποίες συνδυάζουν την ακρίβεια και την ανάκληση, ιδιαίτερα σε εργασίες ταξινόμησης ή εξαγωγής πληροφοριών [23].

Απαιτείται μια ποικιλία μετρικών που καλύπτουν διάφορες πτυχές της απόδοσης, όπως η ακρίβεια των απαντήσεων, η σημασιολογική ομοιότητα, η ταχύτητα παραγωγής και η έκταση των πληροφοριών, για την αξιολόγηση των LLM. Αυτή η ποικιλομορφία διασφαλίζει ότι τα αποτελέσματα της αξιολόγησης δεν βασίζονται σε μια μόνο μετρική, αλλά αντιπροσωπεύουν με ακρίβεια την ικανότητα του μοντέλου να ικανοποιεί τους σκοπούς για τους οποίους χρησιμοποιείται.

Κεφάλαιο 3ο: Μεθοδολογία

3.1 Εισαγωγή

Αυτό το κεφάλαιο περιγράφει λεπτομερώς τη μεθοδολογία που ακολουθήθηκε, από τη συλλογή και προετοιμασία των δεδομένων έως την αξιολόγηση των αποτελεσμάτων. Στόχος είναι να δοθεί μια σαφής και συνεκτική περιγραφή της διαδικασίας που χρησιμοποιήθηκε για τη σύγκριση των τριών μεγάλων γλωσσικών μοντέλων.

3.2 Δεδομένα

Οι πληροφορίες που χρησιμοποιήθηκαν σε αυτή την εργασία συγκεντρώθηκαν από 50 αρχεία PDF, τα οποία περιλαμβάνουν διοικητικά κείμενα, κανονισμοί και οδηγούς σπουδών από διάφορα πανεπιστημιακά τμήματα. Τα έγγραφα αυτά ξεχωρίζουν για το ευρύ φάσμα θεματικών εννοιών που καλύπτουν, το μέγεθός τους και την επίσημη γλώσσα τους, χαρακτηριστικά που τα καθιστούν κατάλληλα για πρακτική αξιολόγηση της ικανότητας των μεγάλων γλωσσικών μοντέλων να εντοπίζουν, να αναγνωρίζουν και να οργανώνουν δεδομένα από πολυάριθμες και περίπλοκες πηγές.

Αναπτύχθηκε ένα σύνολο δεδομένων αναφοράς (ground truth) για την αξιόπιστη αξιολόγηση των μοντέλων. Οι ερωτήσεις και οι αντίστοιχες απαντήσεις προήλθαν απευθείας από τα έγγραφα και καταγράφηκαν σε μορφή CSV. Για κάθε καταχώριση καταγράφονται οι εξής πληροφορίες: η ερώτηση και η απάντηση, το όνομα του αρχείου PDF από το οποίο ελήφθησαν οι πληροφορίες και ο αριθμός σελίδας όπου βρισκόταν η πληροφορία.

Η δυνατότητα επαλήθευσης κάθε ερωτήματος στο αρχικό κείμενο εξασφαλίζει διαφάνεια και τεκμηρίωση. Επιπλέον, η δομημένη μορφή του συνόλου δεδομένων, καθιστά δυνατή την αυτοματοποιημένη σύγκριση των απαντήσεων που παρέχουν τα LLM με τις καταγεγραμμένες σωστές απαντήσεις.

Παρακάτω παρουσιάζεται ενδεικτικά η μορφή του ground truth dataset, όπως αποθηκεύτηκε σε πίνακα CSV:

ID	Ερώτηση	Απάντηση	Πηγαίο Αρχείο	Σελίδα
..
..

Πίνακας 3.1: Ενδεικτική μορφή του ground truth dataset.

3.3 Επεξεργασία και Οργάνωση Δεδομένων

Ένα κρίσιμο βήμα ήταν η προεπεξεργασία των δεδομένων, η οποία κατέστη δυνατή τη μετατροπή των εγγράφων PDF σε μορφή που τα μεγάλα γλωσσικά μοντέλα θα μπορούσαν να αναζητήσουν και να επεξεργαστούν. Για να διασφαλιστεί η αξιοπιστία και η αποτελεσματικότητα του συστήματος, τα 50 αρχεία PDF μετατράπηκαν σε επεξεργάσιμο κείμενο και υποβλήθηκαν σε μια σειρά δοκιμών.

Τμηματοποίηση Κειμένου

Τα εκτενή κανονιστικά κείμενα χωρίστηκαν σε διαχειρίσιμα τμήματα. Για τον σκοπό αυτό, χρησιμοποιήθηκε επικάλυψη χαρακτήρων (CHUNK_OVERLAP) = 150 και μέγεθος τμημάτων (CHUNK_SIZE) = 1200. Με αυτόν τον τρόπο, η επικάλυψη διασφαλίζει ότι σημαντικές πληροφορίες στα όρια των τμημάτων δεν χάνονται, ενώ κάθε τμήμα είναι αρκετά μεγάλο ώστε να διατηρήσει τη σημασιολογική συνοχή.

Η βιβλιογραφία για τα συστήματα RAG υπογραμμίζει τη σημασία αυτής της πρακτικής. Όπως αναφέρεται, στο [22], η επιλογή του μεγέθους των τμημάτων επηρεάζει σημαντικά την απόδοση των LLM στην ανάκτηση πληροφοριών, με τα τμήματα μεσαίου μεγέθους (500–1000 tokens) να παρέχουν μια ισορροπία μεταξύ πληρότητας και ακρίβειας.

Μετατροπή σε Embeddings

Μετά τη διαίρεση του κειμένου σε τμήματα, χρησιμοποιήθηκε το μοντέλο Gemini Embeddings της Google – μια διανυσματική αναπαράσταση που ενσωματώνει το νόημα του κειμένου σε έναν πολυδιάστατο χώρο –για να μετατρέψει κάθε τμήμα σε ένα embedding. Τα embeddings επιτρέπουν στο σύστημα να προσδιορίσει τη σημασιολογική ομοιότητα μεταξύ ερωτήσεων και αποσπασμάτων, καθώς και την ομοιότητα λέξεων.

Τα embeddings αποτελούν θεμελιώδες συστατικό της ανάπτυξης συστημάτων ανάκτησης στην διεθνή έρευνα. Για παράδειγμα, χρησιμοποιώντας ενσωματώσεις Sentence-BERT, οι Reimers & Gurevych (2019) απέδειξαν σημαντική βελτίωση στη σημασιολογική ομοιότητα σε σύγκριση με τις παραδοσιακές μεθόδους που βασίζονται σε TF-IDF [24]. Η ίδια αρχή εφαρμόζεται σε ακόμη μεγαλύτερη κλίμακα δεδομένων με μοντέλα όπως τα Gemini Embeddings, που χρησιμοποιούνται στη παρούσα εργασία.

Διατήρηση στη Βάση Δεδομένων (VectorDB)

Η βάση δεδομένων ChromaDB, η οποία έχει σχεδιαστεί για αναζήτηση βάσει ομοιότητας, ήταν εκεί που αποθηκεύτηκαν τα embeddings. Ακόμα και σε μεγάλα σύνολα δεδομένων, τα πιο σχετικά αποσπάσματα μπορούν να ανακτηθούν άμεσα χάρη σε αυτό το είδος βάσης. Οι βάσεις δεδομένων φορέων διευκολύνουν την αποτελεσματική αποθήκευση και την ανάκτηση των embeddings από εκτεταμένες συλλογές κειμένου, καθιστώντας τες ουσιαστικό μέρος της αρχιτεκτονικής RAG [22].

Ανάκτηση Πληροφοριών

Χρησιμοποιώντας το ίδιο μοντέλο που χρησιμοποιήθηκε για τη δημιουργία των embeddings των εγγράφων, το ερώτημα του χρήστη μετατρέπεται σε embedding κατά τη διαδικασία αναζήτησης. Στη συνέχεια, αυτή η ενσωμάτωση συγκρίνεται με κάθε απόσπασμα που έχει αποθηκευτεί στη βάση δεδομένων διανυσμάτων, χρησιμοποιώντας μετρικές ομοιότητας, όπως η ομοιότητα συνημιτόνου, η οποία υπολογίζει τη γωνία μεταξύ δύο διανυσμάτων και δείχνει πόσο σημασιολογικά παρόμοια είναι. Όσο πιο κοντά η τιμή στο 1, τόσο πιο σχετικό θεωρείται ότι είναι το απόσπασμα σε σχέση με το ερώτημα.

Με αυτόν τον τρόπο, το σύστημα επιλέγει τα k πιο κατάλληλα αποσπάσματα που θεωρούνται αντιπροσωπευτικά του περιεχομένου των εγγράφων. Ο αριθμός k επιλέγεται για να επιτευχθεί ισορροπία μεταξύ της αποφυγής της υπερφόρτωσης του γλωσσικού μοντέλου με πληροφορίες και της

παροχής αρκετών αποσπασμάτων για την απάντηση στην ερώτηση του χρήστη. Η επολογή αυτή επηρεάζει άμεσα την ποιότητα των απαντήσεων, διότι μια υπερβολικά μεγάλη τιμή μπορεί να προκαλέσει θόρυβο και καθυστερήσεις στην επεξεργασία, ενώ μια πολύ μικρή τιμή μπορεί να οδηγήσει σε ελλείψεις πληροφοριών.

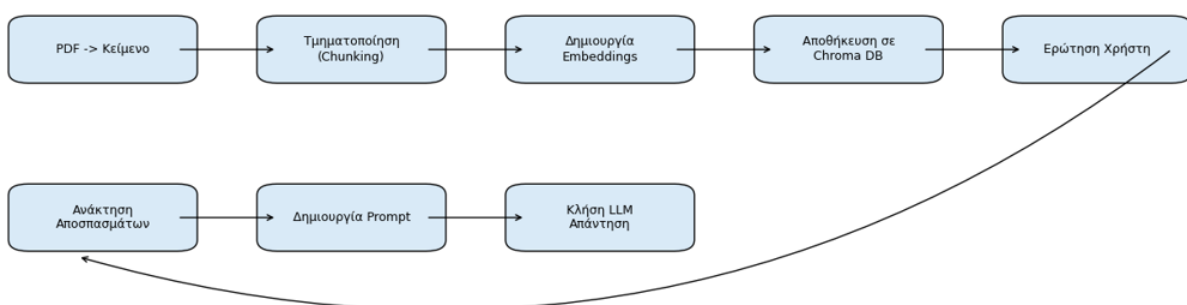
Ένα εμπλουτισμένο ερώτημα δημιουργείται συνδυάζοντας την αρχική ερώτηση με τα σχετικά αποσπάσματα που έχουν ανακτηθεί. Αυτή η προτροπή είναι το κείμενο που τροφοδοτείται στο μεγάλο γλωσσικό μοντέλο, το οποίο πλέον καθοδηγείται από πραγματικά δεδομένα προερχόμενα από τους κανονισμούς, συμπληρώνοντας τις γενικές του γνώσεις. Επειδή βασίζεται σε συγκεκριμένες πηγές, η παραγόμενη απάντηση είναι πιο ακριβής, αξιόπιστη και καλά τεκμηριωμένη.

Η διαδικασία αυτή, που συνδυάζει την ικανότητα ανάκτησης πραγματικών εγγράφων με τις γλωσσικές δυνατότητες του μοντέλου για την παραγωγή φυσικών και συνεκτικών απαντήσεων, αποτελεί τον πυρήνα της λογικής του RAG. Ως αποτέλεσμα, ο μηχανισμός ανάκτησης λειτουργεί ως γέφυρα μεταξύ της βάσης γνώσεων και του LLM, διασφαλίζοντας ότι ο χρήστης λαμβάνει τις πιο ακριβείς και σχετικές πληροφορίες..

Συνολική Ροή

Η συνολική ροή της διαδικασίας προεπεξεργασίας και ανάκτησης μπορεί να περιγραφεί ως εξής:

- **Μετατροπή PDF σε κείμενο:** Κάθε έγγραφο μετατρέπεται σε μορφή που μπορεί να υποστεί επεξεργασία.
- **Τμηματοποίηση (Chunking):** Το κείμενο χωρίζεται σε αποσπάσματα συγκεκριμένου μεγέθους με επικάλυψη, ώστε να διατηρείται η νοηματική συνοχή.
- **Δημιουργία Embeddings:** Τα chunks μετατρέπονται σε διανυσματικές αναπαραστάσεις που αποτυπώνουν το σημασιολογικό περιεχόμενο.
- **Αποθήκευση σε Vector Database:** Τα embeddings αποθηκεύονται σε ChromaDB, επιτρέποντας την ταχεία αναζήτηση με βάση τη σημασιολογική ομοιότητα.
- **Υποβολή Ερώτησης:** Ο χρήστης διατυπώνει μια ερώτηση, η οποία επίσης μετατρέπεται σε embedding.
- **Ανάκτηση Αποσπασμάτων:** Το embedding της ερώτησης συγκρίνεται με τα αποθηκευμένα embeddings και ανακτώνται τα πιο σχετικά αποσπάσματα.
- **Δημιουργία Prompt:** Τα αποσπάσματα ενσωματώνονται στην ερώτηση ώστε να σχηματιστεί ένα πλήρες prompt.
- **Κλήση LLM:** Το prompt δίνεται στο μοντέλο και παράγεται η τελική απάντηση.



Σχήμα 3.1: Διάγραμμα ροής της διαδικασίας RAG.

3.4 Αρχιτεκτονική Συστήματος

Η λογική RAG αποτελεί τη βάση για την αρχιτεκτονική που αναπτύχθηκε σε αυτή τη μελέτη, η οποία συνδυάζει πτυχές γλωσσικής παραγωγής και ανάκτησης πληροφοριών. Η θεμελιώδης ιδέα ήταν η δημιουργία ενός συστήματος ικανού να απαντά σε ερωτήσεις χρηστών αξιοποιώντας τόσο τα πραγματικά δεδομένα από τους κανονισμούς της μελέτης όσο και τις γενικές γνώσεις του μοντέλου.

Για τη διευκόλυνση της γρήγορης αναζήτησης, τα δεδομένα που συλλέχθηκαν μορφοποιήθηκαν κατάλληλα και αποθηκεύτηκαν σε μια βάση δεδομένων. Στη συνέχεια δημιουργήθηκε ένας μηχανισμός ανάκτησης, ο οποίος επιλέγει τα πιο κατάλληλα αποσπάσματα συγκρίνοντας το ερώτημα του χρήστη με όλα τα αποθηκευμένα. Αυτό το βήμα είναι κρίσιμο, καθώς καθορίζει ποιο τμήμα του περιεχομένου θα εισαχθεί τελικά στο γλωσσικό μοντέλο, με άμεσο αντίκτυπο στην ποιότητα της παραγόμενης απάντησης.

Το επόμενο βήμα είναι η δημιουργία ενός εμπλουτισμένου μηνύματος(prompt), το οποίο ενσωματώνει το ερώτημα του χρήστη μαζί με τα σχετικά αποσπάσματα που ανακτήθηκαν. Το σχετικό LLM λαμβάνει αυτό το μήνυμα, αφομοιώνει τις πληροφορίες και παράγει την τελική απάντηση. Με αυτόν τον τρόπο, το μοντέλο χτίζει πάνω σε συγκεκριμένες πηγές αντί να βασίζεται αποκλειστικά στην εκπαίδευσή του, βελτιώνοντας την ακρίβεια και την αξιοπιστία των αποτελεσμάτων.

Η συνολική λειτουργία μπορεί να περιγραφεί μέσα από τέσσερα επίπεδα:

- **Επίπεδο Δεδομένων:** περιλαμβάνει τα αρχεία PDF, τα τμήματα (chunks) που προκύπτουν από την τμηματοποίηση και τις διανυσματικές τους αναπαραστάσεις (embeddings) στη βάση δεδομένων.
- **Επίπεδο Ανάκτησης:** υλοποιεί τον μηχανισμό που αντιστοιχίζει το ερώτημα στα πιο σχετικά αποσπάσματα, αξιοποιώντας μετρικές ομοιότητας.
- **Επίπεδο Επεξεργασίας:** το γλωσσικό μοντέλο λαμβάνει το prompt και συνθέτει την απάντηση, αξιοποιώντας τόσο τη δική του γνώση όσο και τα δεδομένα που ανακτήθηκαν.
- **Επίπεδο Διεπαφής Χρήστη:** παρέχει το γραφικό περιβάλλον μέσω του οποίου ο χρήστης μπορεί να υποβάλει ερωτήσεις και να λάβει απαντήσεις σε κατανοητή μορφή.

Προκειμένου το σύστημα να είναι επεκτάσιμο και εύκολα προσαρμόσιμο σε διάφορα LLMs, επιλέχθηκε αυτή η αρχιτεκτονική. Για τη συγκριτική αξιολόγηση, δοκιμάστηκαν τρία διακριτά μοντέλα στο ίδιο πλαίσιο, καθένα από τα οποία έλαβε τις ίδιες προτροπές και δημιούργησε απαντήσεις.

Η διαφάνεια ολόκληρης της διαδικασίας διασφαλίζεται από αυτόν τον σχεδιασμό, ξεκινώντας με την υποβολή της ερώτησης, προχωρώντας στην αναζήτηση στα αρχεία, καταλήγοντας στην δημιουργία της προτροπής και τελειώνοντας με την απάντηση. Ταυτόχρονα, η αρχιτεκτονική αυτή επιτρέπει την ανίχνευση πιθανών τομέων βελτίωσης, είτε στη διαδικασία ανάκτησης είτε στον τρόπο με τον οποίο τα μοντέλα χρησιμοποιούν το διαθέσιμο περιεχόμενο.

3.4.1 Επίπεδο Δεδομένων

Το πρώτο επίπεδο της αρχιτεκτονικής καλύπτει τα δεδομένα που χρησιμοποιήθηκαν και τον τρόπο προετοιμασίας τους για το σύστημα. Συγκεκριμένα, συγκεντρώθηκαν 50 αρχεία PDF, τα οποία

περιλάμβαναν κανονισμούς σπουδών, εγχειρίδια και σχετικά διοικητικά κείμενα από πανεπιστημιακές σχολές. Τα έγγραφα αυτά αποτελούν κατάλληλο πεδίο δοκιμής για την πρακτική ικανότητα των γλωσσικών μοντέλων, καθώς συνδυάζουν μεγάλο μήκος, ποικιλία θεματικών ενοτήτων και επίσημο ύφος.

Για τη διευκόλυνση της επεξεργασίας των εγγράφων χρησιμοποιήθηκε η μέθοδος chunking. Το κείμενο χωρίστηκε σε τμήματα των 1000 χαρακτήρων με επικάλυψη 150 χαρακτήρων. Οι συγκεκριμένες τιμές επιλέχθηκαν προκειμένου να επιτευχθεί ισορροπία μεταξύ της διατήρησης κρίσιμων πληροφοριών και της σημασιολογικής πληρότητας κάθε τμήματος. Η επικάλυψη εξασφαλίζει ότι σημαντικά σημεία στα όρια δύο ενοτήτων περιλαμβάνονται και στα δύο τμήματα, αποτρέποντας την απώλεια πληροφοριών.

Κάθε κομμάτι μετατράπηκε στη συνέχεια σε embedding, δηλαδή σε διανυσματική αναπαράσταση. Η σημασία του κειμένου αποτυπώνεται σε έναν πολυδιάστατο χώρο, τοποθετώντας κοντά μεταξύ τους τα αποσπάσματα με παρόμοιο περιεχόμενο. Αυτό είναι απαραίτητο, διότι επιτρέπει στο σύστημα να συγκρίνει μια ερώτηση με βάση το βαθύτερο σημασιολογικό της περιεχόμενο και όχι μόνο τις λέξεις που περιέχει. Όλα τα embeddings, αποθηκεύτηκαν σε μια εξειδικευμένη βάση δεδομένων, σχεδιασμένη για ταχεία και ακριβή αναζήτηση ακόμα και σε μεγάλα σύνολα δεδομένων.

Υπό αυτή την έννοια, το Επίπεδο Δεδομένων κάνει περισσότερα από το να συλλέγει και να αποθηκεύει αρχεία, καθώς θέτει το απαραίτητο πλαίσιο πάνω στο οποίο χτίζεται η υπόλοιπη αρχιτεκτονική. Στην ουσία, λειτουργεί ως η "βάση γνώσεων" του συστήματος, συλλέγοντας και οργανώνοντας τα ακατέργαστα δεδομένα που θα χρησιμοποιηθούν για την παραγωγή απαντήσεων από το γλωσσικό μοντέλο.

3.4.2 Επίπεδο Ανάκτησης

Το Επίπεδο Ανάκτησης αφορά τον μηχανισμό με τον οποίο το σύστημα καθορίζει ποια αποσπάσματα κειμένου από τη βάση δεδομένων είναι τα πιο σχετικά με το ερώτημα του χρήστη και αποτελεί το δεύτερο επίπεδο της αρχιτεκτονικής. Το βήμα αυτό είναι ιδιαίτερα κρίσιμο, καθώς η ποιότητα των απαντήσεων που παράγει το γλωσσικό μοντέλο εξαρτάται άμεσα από την καταλληλότητα των αποσπασμάτων που ανακτώνται.

Όταν ένας χρήστης υποβάλλει μια ερώτηση σε φυσική γλώσσα, η διαδικασία ξεκινά. Το ίδιο μοντέλο που χρησιμοποιήθηκε για τη δημιουργία των embeddings των αποσπασμάτων κειμένου εφαρμόζεται και στην ερώτηση, μετατρέποντας την σε ενσωμάτωση. Με αυτόν τον τρόπο καθίσταται δυνατή η σημασιολογική σύγκριση μεταξύ του ερωτήματος και κάθε τμήματος κειμένου που έχει αποθηκευτεί στη βάση δεδομένων.

Η ομοιότητα συνημιτόνου αποτελεί τη βασική μετρική που χρησιμοποιείται για τη σύγκριση. Μέσω του υπολογισμού της γωνίας μεταξύ των διανυσμάτων, η μέθοδος αυτή αποδίδει έναν δείκτη που υποδεικνύει τον βαθμό συνάφειας: όσο πιο κοντά βρίσκεται η τιμή στο 1, τόσο μεγαλύτερη είναι η σημασιολογική εγγύτητα μεταξύ ερωτήματος και αποσπάσματος. Ο αριθμός k επιλέγεται για να διασφαλιστεί ισορροπία ανάμεσα στην παροχή επαρκούς πληροφορίας και στην αποφυγή υπερφόρτωσης του γλωσσικού μοντέλου. Στην πράξη, το σύστημα επιλέγει τα k πιο συναφή αποσπάσματα.

Οι πληροφορίες που θα συμπεριληφθούν στο επόμενο βήμα επιλέγονται χρησιμοποιώντας την διαδικασία αυτή ως "φίλτρο". Η ανακριβής ανάκτηση ενέχει τον κίνδυνο το LLM να παράγει απαντήσεις βασισμένες σε ανακριβή ή ελλιπή δεδομένα. Αντιθέτως, μια επιτυχημένη ανάκτηση

διασφαλίζει ότι η τελική απάντηση θα υποστηρίζεται από αυθεντικές και αξιόπιστες πηγές, ενισχύοντας έτσι τη νομιμότητα και την τεκμηρίωση του αποτελέσματος.

Το Επίπεδο Ανάκτησης λειτουργεί έτσι ως ο κρίσιμος σύνδεσμος μεταξύ του γλωσσικού μοντέλου και της βάσης γνώσεων. Ουσιαστικά, αποτελεί το στάδιο που "γεφυρώνει" το ερώτημα του χρήστη με τις πλέον σχετικές πληροφορίες που περιέχονται στα έγγραφα, θέτοντας το πλαίσιο για μια απάντηση εύστοχη και επιστημονικά τεκμηριωμένη.

3.4.3 Επίπεδο Επεξεργασίας

Η επεξεργασία αποτελεί το τρίτο επίπεδο της αρχιτεκτονικής, στο οποίο παράγονται οι τελικές απαντήσεις μέσω των LLMs. Σε αυτό το στάδιο, το γλωσσικό μοντέλο καθοδηγείται από το εμπλουτισμένο prompt που αναπτύχθηκε στο προηγούμενο επίπεδο, αντί να λειτουργεί αυτόνομα ή αποκλειστικά βάσει των γνώσεων που απέκτησε κατά τη διάρκεια της εκπαίδευσής του.

Πιο συγκεκριμένα, το σύστημα δημιουργεί ένα μήνυμα που ενσωματώνει το ερώτημα του χρήστη όσο και τα σχετικά αποσπάσματα που αντλήθηκαν από τη βάση δεδομένων. Αυτός ο συνδυασμός είναι απαραίτητος, καθώς τα αποσπάσματα παρέχουν στο LLM τις βασικές πληροφορίες που χρειάζεται για την παραγωγή απαντήσεων ακριβών, συνεκτικών και προσαρμοσμένων στο εκάστοτε ερώτημα. Χωρίς αυτή την ενίσχυση, το μοντέλο θα μπορούσε να χρησιμοποιήσει μόνο τις γενικές του γνώσεις, με κίνδυνο να παράγει ψευδείς ή ακόμα και φανταστικές πληροφορίες (παραισθήσεις).

Τρία διακριτά γλωσσικά μοντέλα —Gemini, LLaMA και Gemma— δοκιμάστηκαν παράλληλα χρησιμοποιώντας την ίδια διαδικασία στο πλαίσιο της παρούσας εργασίας. Δεδομένου ότι όλα τα μοντέλα κλήθηκαν να απαντήσουν στις ίδιες ερωτήσεις χρησιμοποιώντας το ίδιο πλαίσιο πληροφοριών, κατέστη δυνατή μια αξιόπιστη συγκριτική αξιολόγηση. Αυτή η μέθοδος εγγυάται ότι τα αποτελέσματα είναι συγκρίσιμα και ότι οι παρατηρούμενες διαφορές οφείλονται κυρίως στα χαρακτηριστικά κάθε μοντέλου.

Κατά την αλληλεπίδραση με τα LLMs, σημειώθηκαν επίσης ορισμένες τεχνικές προκλήσεις. Ένα βασικό ζήτημα αφορά το όριο στον αριθμό των tokens που μπορεί να επεξεργαστεί κάθε μοντέλο, γεγονός που καθιστά αναγκαία την προσεκτική επιλογή του μεγέθους των αποσπασμάτων που εντάσσονται στο prompt. Επιπλέον, ο χρόνος απόκρισης διαφοροποιείται από μοντέλο σε μοντέλο και έχει μεγάλο αντίκτυπο στο πόσο χρήσιμο είναι το σύστημα, ιδίως σε εφαρμογές που απαιτούν άμεση αλληλεπίδραση.

Συνεπώς, το Επίπεδο Επεξεργασίας λειτουργεί ως ο "κεντρικός μηχανισμός" του συστήματος, συνδυάζοντας την ακρίβεια των ανακτημένων δεδομένων με τη γλωσσική ικανότητα των LLM, προκειμένου να παραχθούν απαντήσεις που είναι τόσο συνεκτικές όσο και καλά τεκμηριωμένες.

3.4.4 Επίπεδο Διεπαφής Χρήστη (UI)

Η Διεπαφή Χρήστη, δηλαδή το τμήμα του συστήματος με το οποίο αλληλεπιδρά ο τελικός χρήστης, αποτελεί το τέταρτο επίπεδο της αρχιτεκτονικής. Για την υλοποίηση της διεπαφής χρησιμοποιήθηκε η βιβλιοθήκη Gradio, η οποία επιτρέπει τη γρήγορη δημιουργία απλών αλλά χρήσιμων γραφικών περιβαλλόντων σε εφαρμογές μηχανικής μάθησης.

Ο βασικός στόχος του σχεδιασμού της διεπαφής ήταν η απλότητα και η ευκολία στη χρήση, ώστε οι χρήστες να μπορούν να υποβάλλουν ερωτήσεις χωρίς να χρειάζεται να είναι εξοικειωμένοι με τεχνικές λεπτομέρειες. Μπορούν να διατυπώσουν την ερώτησή τους σε φυσική γλώσσα χρησιμοποιώντας ένα απλό πλαίσιο κειμένου, ενώ τα αποτελέσματα εμφανίζονται διατεταγμένα ακριβώς από κάτω. Με αυτόν τον τρόπο αποφεύγονται περίπλοκες διαδικασίες και καλλιεργείται η εντύπωση μιας "συνομιλίας" με το σύστημα.

Η διεπαφή καθιστά επίσης δυνατή την προβολή απαντήσεων από πολλά LLM ταυτόχρονα. Με τον τρόπο αυτόν, ο χρήστης μπορεί να συγκρίνει άμεσα την ποιότητα, την πληρότητα και τη σαφήνεια της εξόδου κάθε μοντέλου, στοιχείο που είναι απαραίτητο για τους στόχους της παρούσας εργασίας. Σε αντίθεση με τα θεωρητικά τους χαρακτηριστικά, αυτή η παρουσίαση παρέχει μια πιο πρακτική οπτική, βοηθώντας στον εντοπισμό των πλεονεκτημάτων και μειονεκτημάτων των μοντέλων σε πραγματικά σενάρια.

Η δυνατότητα του Gradio για απλή ενσωμάτωση σε ένα περιβάλλον ιστού ήταν ένας άλλος παράγοντας στην απόφαση. Για ευρύτερη χρήση, το σύστημα μπορεί να εγκατασταθεί σε διακομιστή εκτός από τη λειτουργία του τοπικά. Η προσβασιμότητα αυτή ενισχύει τη χρηστικότητα του chatbot, καθώς μπορούν να το αξιοποιήσουν και μη τεχνικοί χρήστες, όπως φοιτητές ή καθηγητές που χρειάζονται άμεση πρόσβαση σε πληροφορίες σχετικά με τους κανονισμούς σπουδών.

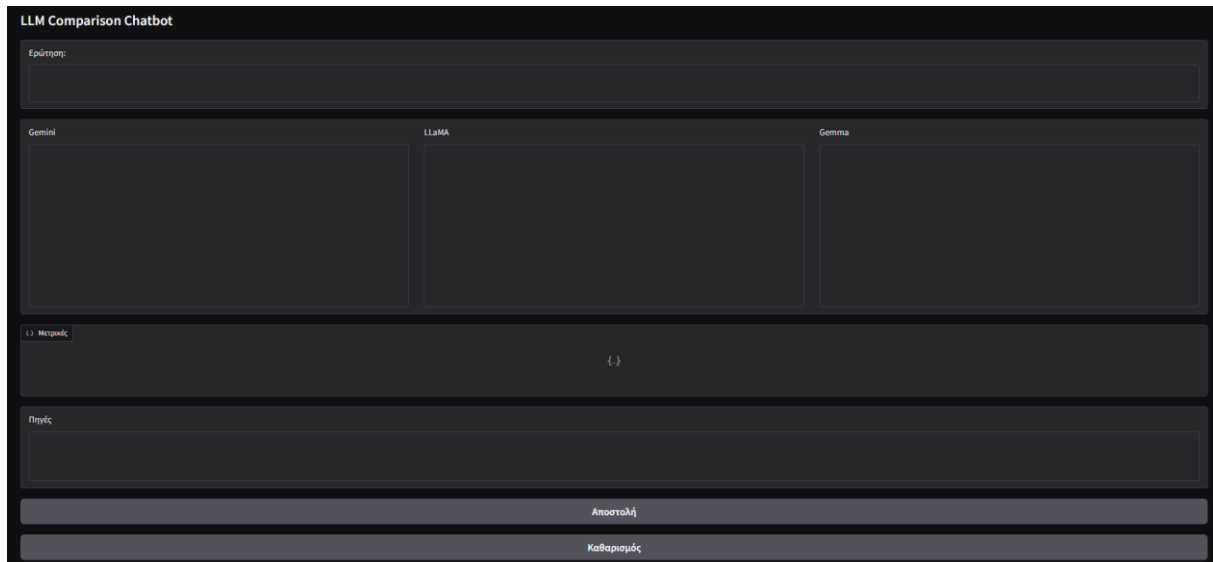
Λαμβάνοντας υπόψη όλα τα παραπάνω, το επίπεδο της διεπαφής χρήστη είναι απαραίτητο για την ολοκλήρωση της αρχιτεκτονικής, καθώς συνδέει την τεχνική πολυπλοκότητα του συστήματος με την καθημερινή ζωή του χρήστη. Ως αποτέλεσμα, το chatbot παύει να είναι ένα θεωρητικό εργαλείο και αντ' αυτού γίνεται ένας χρήσιμος και διαισθητικός τρόπος αναζήτησης και οργάνωσης πληροφοριών.

3.4.5 Παράδειγμα Χρήσης της Εφαρμογής

Η αρχική οθόνη του συστήματος παρουσιάζεται παρακάτω. Το περιβάλλον είναι απλό, καθαρό και σχεδιασμένο ώστε να διευκολύνει τη σύγκριση των απαντήσεων μεταξύ των διαφορετικών μοντέλων. Στο επάνω μέρος υπάρχει το πλαίσιο εισαγωγής της ερώτησης του χρήστη, ενώ ακριβώς από κάτω εμφανίζονται τρεις διακριτές περιοχές εξόδου, καθεμία αφιερωμένη σε ένα μοντέλο: Gemini, LLaMA και Gemma. Οι απαντήσεις τοποθετούνται οριζόντια, ώστε να είναι άμεσα συγκρίσιμες ως προς την πληρότητα, την ακρίβεια και το ύφος της διατύπωσης.

Κάτω από τα πεδία των απαντήσεων εμφανίζονται δύο επιπλέον ενότητες: η πρώτη αφορά τις μετρικές, όπου παρουσιάζονται οι χρόνοι απόκρισης κάθε μοντέλου, ενώ η δεύτερη περιλαμβάνει τις πηγές, δηλαδή τα αποσπάσματα των εγγράφων από τα οποία αντλήθηκαν οι πληροφορίες. Στο κάτω μέρος της διεπαφής βρίσκονται δύο κουμπιά λειτουργίας: «Αποστολή», για την αποστολή του ερωτήματος και την εμφάνιση των αποτελεσμάτων, και «Καθαρισμός» για την επαναφορά των πεδίων και την εισαγωγή νέου ερωτήματος.

Η διάταξη αυτή καθιστά τη σύγκριση των μοντέλων άμεση και οπτικά ευδιάκριτη, προσφέροντας στον χρήστη μια ολοκληρωμένη εικόνα τόσο για την ταχύτητα όσο και για την ποιότητα των απαντήσεων που παράγει κάθε LLM:



Σχήμα 3.2: Αρχική οθόνη της εφαρμογής LLM Comparison Chatbot με τα τρία μοντέλα.

Κατά τη λειτουργία της εφαρμογής, ο χρήστης εισάγει την ερώτησή του στο πλαίσιο κειμένου στο επάνω μέρος της οθόνης. Με την επιλογή του κουμπιού «Αποστολή», η ερώτηση αποστέλλεται ταυτόχρονα στα τρία μεγάλα γλωσσικά μοντέλα μέσω των αντίστοιχων API. Κάθε μοντέλο λαμβάνει το ίδιο prompt, εμπλουτισμένο με τα σχετικά αποσπάσματα από τη βάση δεδομένων κανονισμών (μέσω του μηχανισμού RAG), ώστε η απάντηση να βασίζεται σε πραγματικά δεδομένα και όχι μόνο στη γνώση εκπαίδευσης του μοντέλου.

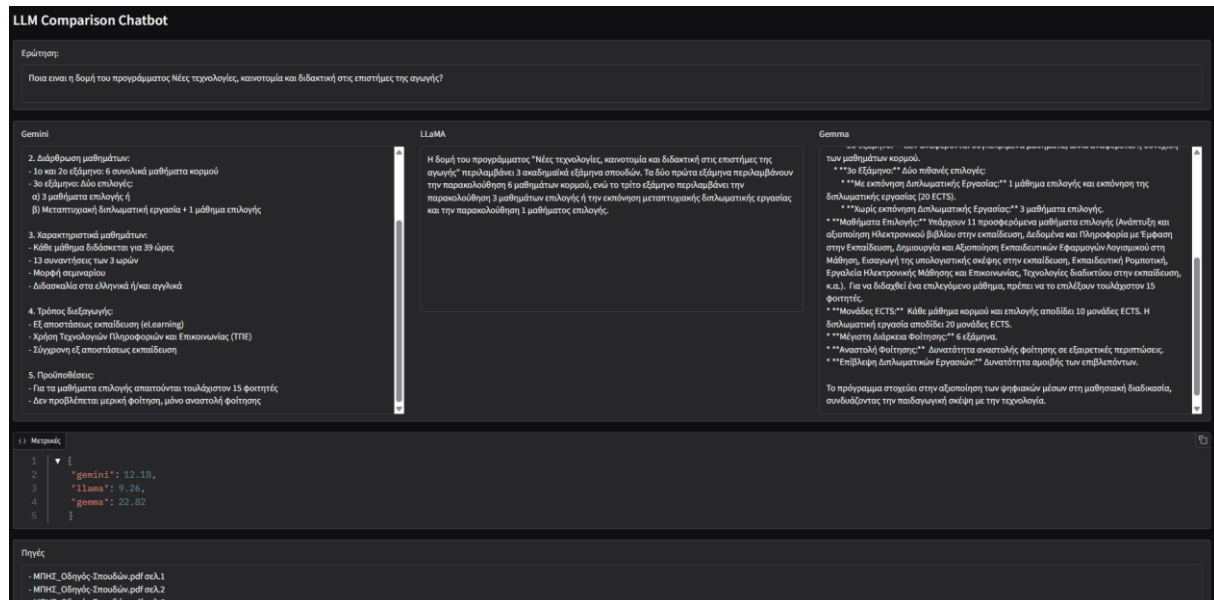
Οι απαντήσεις εμφανίζονται παράλληλα στα τρία διακριτά πλαίσια της διεπαφής, επιτρέποντας στον χρήστη να συγκρίνει άμεσα το περιεχόμενο, τη σαφήνεια και τη δομή τους. Κάτω από τις απαντήσεις εμφανίζονται οι μετρικές χρόνου απόκρισης, που δείχνουν πόσο γρήγορα απάντησε κάθε μοντέλο, καθώς και οι πηγές των αποσπασμάτων, ώστε να τεκμηριώνεται η προέλευση της πληροφορίας.

Η συνολική διαδικασία είναι πλήρως αυτοματοποιημένη. Ο χρήστης δεν χρειάζεται να γνωρίζει τεχνικές λεπτομέρειες για embeddings ή APIs, καθώς όλη η εσωτερική λειτουργία (ανάκτηση αποσπασμάτων, δημιουργία prompt και αποστολή ερωτήματος) εκτελείται στο παρασκήνιο. Η διεπαφή έχει σχεδιαστεί έτσι ώστε να προσομοιώνει μια φυσική “συνομιλία” με το σύστημα, προσφέροντας μια απλή και φιλική εμπειρία χρήσης.

Στην πράξη, μετά την αποστολή ενός ερωτήματος, η εφαρμογή εμφανίζει τα αποτελέσματα των τριών μοντέλων σε παράλληλη διάταξη, όπως φαίνεται στην παρακάτω εικόνα. Κάθε πλαίσιο περιλαμβάνει την απάντηση που παρήγαγε το αντίστοιχο μοντέλο, επιτρέποντας στον χρήστη να συγκρίνει άμεσα τη σαφήνεια, την πληρότητα και την πιστότητα του περιεχομένου σε σχέση με τα κανονιστικά έγγραφα.

Στο παράδειγμα που παρουσιάζεται, το ερώτημα αφορά τη δομή ενός προγράμματος μεταπτυχιακών σπουδών. Το Gemini παρήγαγε μια εκτενή, αναλυτική απάντηση που περιλαμβάνει πλήρη διάρθρωση των μαθημάτων, τρόπο διδασκαλίας και προϋποθέσεις φοίτησης. Το LLaMA παρείχε πιο συνοπτική αλλά δομημένη περιγραφή, εστιάζοντας κυρίως στον αριθμό των εξαμήνων και την κατανομή των μαθημάτων. Αντίθετα, το Gemma έδωσε πιο αφηγηματική και περιγραφική απάντηση, η οποία, παρότι εκφραστικά αναλυτική, περιλάμβανε λιγότερες ακριβείς αναφορές στα στοιχεία του προγράμματος.

Ακριβώς κάτω από τις απαντήσεις, το πεδίο «Μετρικές» εμφανίζει τον χρόνο απόκρισης κάθε μοντέλου σε δευτερόλεπτα, επιτρέποντας τη σύγκριση της ταχύτητας επεξεργασίας. Στο συγκεκριμένο παράδειγμα, το LLaMA απάντησε ταχύτερα, ενώ το Gemma χρειάστηκε περισσότερο χρόνο λόγω του μεγαλύτερου όγκου κειμένου που παρήγαγε. Το πεδίο «Πηγές» παρουσιάζει τις σελίδες των αρχείων PDF από όπου αντλήθηκαν τα αποσπάσματα, προσδίδοντας διαφάνεια και αξιοπιστία στα παραγόμενα αποτελέσματα.



Σχήμα 3.3: Παράδειγμα εκτέλεσης ερωτήματος.

Ως δεύτερο παράδειγμα, παρουσιάζεται η εκτέλεση του ερωτήματος «Ποια είναι τα μαθήματα του προγράμματος Νέες τεχνολογίες, καινοτομία και διδακτική στις επιστήμες της αγωγής?».

Η ερώτηση αυτή στοχεύει στην ανάκτηση συγκεκριμένων πληροφοριών από τον οδηγό σπουδών, γεγονός που καθιστά σαφή τη σημασία της χρήσης των embeddings και της βάσης διανυσμάτων για την ακριβή εντόπιση των σχετικών αποσπασμάτων.

Το Gemini παρήγαγε την πιο πλήρη και οργανωμένη απάντηση, παρουσιάζοντας τόσο τα μαθήματα κορμού όσο και τα μαθήματα επιλογής, συνοδευμένα από πληροφορίες σχετικά με τις πιστωτικές μονάδες (ECTS). Η απάντησή του περιείχε επίσης σύντομη περιγραφή της δομής και των κατηγοριών μαθημάτων, προσφέροντας ένα πλήρες και εύληπτο αποτέλεσμα.

Το LLaMA παρείχε πιο συνοπτική αλλά συνεκτική απάντηση, διατηρώντας μεν τη διάκριση μεταξύ μαθημάτων κορμού και επιλογής, αλλά χωρίς πρόσθετες πληροφορίες για τις πιστωτικές μονάδες ή τη διάρθρωση του προγράμματος. Η απάντηση ήταν πιο προσανατολισμένη στην ακριβή παράθεση τίτλων μαθημάτων, με μικρότερη επεξήγηση.

Το Gemma, αντίθετα, παρήγαγε μια πιο εκτενή και περιγραφική απάντηση, χρησιμοποιώντας διακριτή μορφοποίηση και αναλυτική παρουσίαση των εννοιών. Παρότι κάλυψε πλήρως τα μαθήματα, το ύψος του κειμένου ήταν πιο αφηγηματικό, προσθέτοντας ερμηνευτικά σχόλια που δεν υπήρχαν στο αρχικό έγγραφο.

Κεφάλαιο 3

Στο κάτω μέρος της διεπαφής, το πεδίο «Μετρικές» κατέγραψε τους χρόνους απόκρισης των μοντέλων, δείχνοντας παρόμοια ταχύτητα εκτέλεσης, με ελαφρώς μεγαλύτερη καθυστέρηση για το Gemma λόγω της εκτενέστερης εξόδου. Οι πηγές που εμφανίζονται στο τέλος αντιστοιχούν σε σελίδες του οδηγού σπουδών όπου εντοπίστηκαν τα σχετικά αποσπάσματα, γεγονός που επιβεβαιώνει τη σωστή λειτουργία του μηχανισμού RAG.

LLM Comparison Chatbot

Ερώτηση:
Ποια είναι τα μαθήματα του προγράμματος "Νέες τεχνολογίες, καινοτομία και διδακτική στις επιστήμες της αγωγής";

Gemini	LLaMA	Gemma
<p>Μαθήματα Κορμού:</p> <ol style="list-style-type: none">Ψηφιακό περιεχόμενο και επικοινωνία στην εκπαίδευσηΨηφιακοί πόροι ήχου και εικόνας στην εκπαιδευτική διαδικασίαΨηφιακή ΠαίδευσηΔημιουργία διαδραστικού εκπαιδευτικού περιεχομένουΔιδακτική ΜεθοδολογίαΜεθοδολογία εκπαιδευτικής έρευνας <p>Μαθήματα Επιλογής:</p> <ol style="list-style-type: none">Ανάπτυξη και αξιοποίηση Ηλεκτρονικού Βιβλίου στην εκπαίδευσηΔεδομένα και Πληροφορία με Έμφαση στην ΕκπαίδευσηΔημιουργία και Αξιοποίηση Εκπαιδευτικών Εφαρμογών Λογισμικού στη ΜάθησηΕισαγωγή της υπολογιστικής σκέψης στην εκπαίδευσηΕκπαιδευτική ΡομποτικήΕργαλεία Ηλεκτρονικής Μάθησης και ΕπικοινωνίαςΤεχνολογίες διαδίκτυο στην εκπαίδευση <p>Κάθε μάθημα αντιστοιχεί σε 10 πιστωτικές μονάδες (ECTS).</p>	<p>Τα μαθήματα του προγράμματος διακρίνονται σε μαθήματα κορμού και επιλογής. Μαθήματα Κορμού:</p> <ol style="list-style-type: none">Ψηφιακό περιεχόμενο και επικοινωνία στην εκπαίδευσηΨηφιακοί πόροι ήχου και εικόνας στην εκπαιδευτική διαδικασίαΨηφιακή ΠαίδευσηΔημιουργία διαδραστικού εκπαιδευτικού περιεχομένουΔιδακτική ΜεθοδολογίαΜεθοδολογία εκπαιδευτικής έρευνας <p>Μαθήματα Επιλογής:</p> <ol style="list-style-type: none">Ανάπτυξη και αξιοποίηση Ηλεκτρονικού Βιβλίου στην εκπαίδευσηΔεδομένα και Πληροφορία με Έμφαση στην ΕκπαίδευσηΔημιουργία και Αξιοποίηση Εκπαιδευτικών Εφαρμογών Λογισμικού στη ΜάθησηΕισαγωγή της υπολογιστικής σκέψης στην εκπαίδευσηΕκπαιδευτική ΡομποτικήΕργαλεία Ηλεκτρονικής Μάθησης και ΕπικοινωνίαςΤεχνολογίες διαδίκτυο στην εκπαίδευση	<p>Το πρόγραμμα "Νέες τεχνολογίες, καινοτομία και διδακτική στις επιστήμες της αγωγής" περιλαμβάνει τα ακόλουθα μαθήματα:</p> <p>**Μαθήματα Κορμού**</p> <ol style="list-style-type: none">Ψηφιακό περιεχόμενο και επικοινωνία στην εκπαίδευσηΨηφιακοί πόροι ήχου και εικόνας στην εκπαιδευτική διαδικασίαΨηφιακή ΠαίδευσηΔημιουργία διαδραστικού εκπαιδευτικού περιεχομένουΔιδακτική ΜεθοδολογίαΜεθοδολογία εκπαιδευτικής έρευνας <p>**Μαθήματα Επιλογής (οι φοιτητές επιλέγουν 3 από αυτά ή εκπονούν διπλωματική εργασία και επιλέγουν 1)**</p> <ol style="list-style-type: none">Ανάπτυξη και αξιοποίηση Ηλεκτρονικού Βιβλίου στην εκπαίδευσηΔεδομένα και Πληροφορία με Έμφαση στην ΕκπαίδευσηΔημιουργία και Αξιοποίηση Εκπαιδευτικών Εφαρμογών Λογισμικού στη ΜάθησηΕισαγωγή της υπολογιστικής σκέψης στην εκπαίδευση

Μετρικές:

```
1 {
2   "gemini": 11.7,
3   "llama": 10.05,
4   "gemma": 11.22
5 }
```

Πηγές:

- ΜΠΗΓ_Οδηγός_Σπουδών.pdf σελ.1
- ΜΠΗΓ_Οδηγός_Σπουδών.pdf σελ.4
- ΜΠΗΓ_Οδηγός_Σπουδών.pdf σελ.6

Σχήμα 3.4: Παράδειγμα εκτέλεσης ερωτήματος.

Η σύγκριση των απαντήσεων δείχνει ότι, ενώ και τα τρία μοντέλα εντόπισαν τις σωστές πληροφορίες, διαφέρουν ως προς το επίπεδο ανάλυσης και τη μορφή παρουσίασης. Το Gemini απέδωσε την πληρέστερη και πιο δομημένη απάντηση, το LLaMA παρουσίασε μια σύντομη αλλά ακριβή περίληψη, ενώ το Gemma εστίασε περισσότερο στην περιγραφική διάσταση του περιεχομένου. Το παράδειγμα αυτό αναδεικνύει την αποτελεσματικότητα της προσέγγισης RAG και την πρακτική αξία της διεπαφής για τη συγκριτική αξιολόγηση μεγάλων γλωσσικών μοντέλων.

Η δυνατότητα οπτικής σύγκρισης των απαντήσεων και των χρόνων απόκρισης καθιστά τη διεπαφή ένα ιδιαίτερα χρήσιμο εργαλείο αξιολόγησης. Ο χρήστης μπορεί να διαπιστώσει με άμεσο και κατανοητό τρόπο τις διαφορές ανάμεσα στα μοντέλα τόσο ως προς την ακρίβεια της πληροφορίας όσο και ως προς τη συνοχή και την ταχύτητα παραγωγής της απάντησης.

3.5 Εργαλεία και Τεχνολογίες

3.5.1 Γλώσσα Προγραμματισμού (Python)

Το σύστημα υλοποιήθηκε με χρήση Python, μια γλώσσα προγραμματισμού που χρησιμοποιείται συχνά σε εφαρμογές που σχετίζονται με την επεξεργασία φυσικής γλώσσας και την τεχνητή νοημοσύνη. Η Python είναι ιδανική για την υλοποίηση έργων που απαιτούν πειραματισμό και προσαρμοστικότητα, καθώς χαρακτηρίζεται από σαφή σύνταξη, πλούσιο οικοσύστημα βιβλιοθηκών και ευκολία στην κατανόηση του κώδικα.

Η γλώσσα χρησιμοποιήθηκε καθ' όλη τη διάρκεια της μεθοδολογίας της παρούσας εργασίας, συμπεριλαμβανομένης της φόρτωσης και ανάγνωσης αρχείων PDF, της διαίρεσης του περιεχομένου σε διαχειρίσιμα τμήματα, της μετατροπής τους σε αναπαραστάσεις διανυσμάτων και της αποθήκευσή τους στη βάση δεδομένων διανυσμάτων. Επιπλέον, χρησιμοποιήθηκε για την οργάνωση της επικοινωνίας με τα γλωσσικά μοντέλα, επιτρέποντας την ανταλλαγή ερωτημάτων και απαντήσεων σε μορφή συγκρίσιμη με το σύνολο δεδομένων αναφοράς.

Η γλώσσα υπήρξε επίσης κρίσιμη για τη διαχείριση ευαίσθητων δεδομένων, όπως τα κλειδιά API, τα οποία διατηρήθηκαν χωριστά χρησιμοποιώντας το αρχείο .env, γεγονός που βελτίωσε την ασφάλεια και απλοποίησε τη μεταφορά της εφαρμογής μεταξύ διαφορετικών περιβαλλόντων. Τέλος, τα αποτελέσματα οργάνωθηκαν και αποθηκεύτηκαν χρησιμοποιώντας βιβλιοθήκες όπως η json και η csv, γεγονός που επέτρεψε την αυτόματη σύγκριση με το σύνολο δεδομένων αναφοράς.

3.5.2 Βιβλιοθήκες

Για την ανάπτυξη του συστήματος χρησιμοποιήθηκαν πολυάριθμες βιβλιοθήκες της Python, καθεμία με μια διακριτή λειτουργία στη διαδικασία ανάπτυξης. Λειτουργώντας ως αγωγός μεταξύ των θεωρητικών βημάτων της μεθοδολογίας και της εφαρμογής τους στον πραγματικό κόσμο, αυτές οι βιβλιοθήκες ενίσχυσαν την αυτοματοποίηση και την αξιοπιστία του συστήματος.

Το περιεχόμενο κειμένου των εγγράφων διαβάστηκε και εξάχθηκε χρησιμοποιώντας τη βιβλιοθήκη pdfplumber. Με αυτόν τον τρόπο, οδηγοί μελέτης και κανονισμοί που περιείχονταν σε αρχεία PDF μετατράπηκαν σε επεξεργάσιμο κείμενο που θα μπορούσε να χρησιμοποιηθεί για περαιτέρω ανάλυση.

Το κείμενο χωρίστηκε σε μικρότερες ενότητες χρησιμοποιώντας τη βιβλιοθήκη LangChain, η οποία προσφέρει προκατασκευασμένες συναρτήσεις για αναδρομικό τεμαχισμό και για την αναπαράσταση των εννοιών ως αντικείμενα Document. Επιπλέον, η βιβλιοθήκη αυτή λειτούργησε ως γέφυρα προς τη βάση δεδομένων διανυσμάτων ChromaDB, επιτρέποντας την αποθήκευση και ανάκτηση ενσωματώσεων σύμφωνα με τη σημασιολογική τους ομοιότητα.

Για τη δημιουργία των αναπαραστάσεων διανυσμάτων (embeddings) χρησιμοποιήθηκε η βιβλιοθήκη langchain_google_genai, η οποία παρέχει πρόσβαση στο μοντέλο GeminiEmbeddings της Google.

Τα Rouge_scorer, sacrebleu και bert_score ενσωματώθηκαν για την αξιολόγηση της ποιότητας των απαντήσεων χρησιμοποιώντας αναγνωρισμένους δείκτες φυσικής γλώσσας, ενώ τα scikit-learn και άλλες βιβλιοθήκες χρησιμοποιήθηκαν για τον υπολογισμό μετρικών ομοιότητας (συνημιτονική ομοιότητα).

Κατά τη φάση ανάπτυξης της διεπαφής χρήστη χρησιμοποιήθηκε το Gradio, ένα πλαίσιο που επιτρέπει τη γρήγορη δημιουργία απλών και χρήσιμων γραφικών περιβαλλόντων, όπως αναφέρθηκε και προηγουμένως. Το Gradio επέτρεψε στον χρήστη να κάνει ερωτήσεις, να συγκρίνει τις απαντήσεις διαφόρων γλωσσικών μοντέλων και να αλληλεπιδρά με το σύστημα σε πραγματικό χρόνο.

Τέλος, το κύριο εργαλείο για την ανάπτυξη και τον πειραματισμό ήταν το περιβάλλον PyCharm CE. Η δυνατότητα του να εκτελεί κώδικα σε μικρά βήματα και να οπτικοποιεί τα αποτελέσματα σε πραγματικό χρόνο συνέλαβαν ουσιαστικά τόσο στην τεκμηρίωση όσο και στην επαλήθευση της μεθοδολογίας.

Από την ανάγνωση και τον διαχωρισμό των εγγράφων μέχρι τη δημιουργία embeddings, την αποθήκευσή τους σε μια βάση δεδομένων, την επεξεργασία τους με γλωσσικά μοντέλα και, τέλος, την

παρουσίαση των αποτελεσμάτων στον τελικό χρήστη, οι βιβλιοθήκες που χρησιμοποιήθηκαν κάλυψαν όλο το φάσμα της μεθοδολογίας.

Η ανάπτυξη του pipeline βασίστηκε, εν μέρει, σε βέλτιστες πρακτικές και τεχνικές που αναφέρονται στη διεθνή βιβλιογραφία. Ενδεικτικά, οι Topsakal & Akinçi (2023) [25] περιγράφουν την ταχεία ανάπτυξη εφαρμογών βασισμένων σε μεγάλα γλωσσικά μοντέλα με αξιοποίηση της βιβλιοθήκης LangChain, προσφέροντας ένα πλαίσιο που επηρέασε και την εργασία.

3.5.3 Embeddings και Βάση Διανυσμάτων

Η μετατροπή των αποσπασμάτων κειμένου σε αναπαραστάσεις διανυσμάτων (embeddings) αποτέλεσε μια από τις πιο κρίσιμες φάσεις στην ανάπτυξη του συστήματος. Τοποθετώντας αποσπάσματα με παρόμοιο περιεχόμενο πιο κοντά το ένα στο άλλο, τα embeddings αποτυπώνουν το νόημα ενός κειμένου σε έναν πολυδιάστατο μαθηματικό χώρο. Η εύρεση αποσπασμάτων κειμένου που σχετίζονται εννοιολογικά με το ερώτημα του χρήστη –ακόμη κι αν δεν χρησιμοποιούν τις ακριβώς ίδιες λέξεις– καθίσταται δυνατή από τη σημασιολογική αναζήτηση.

Η συγκεκριμένη εργασία χρησιμοποίησε τη βιβλιοθήκη langchain_google_genai για τη δημιουργία embeddings μέσω του μοντέλου GeminiEmbeddings της Google. Η βιβλιοθήκη αυτή εφαρμόζει ένα εσωτερικό σύστημα προεπεξεργασίας και καθαρισμού κειμένου πριν τη μετατροπή των αποσπασμάτων σε διανυσματικές αναπαραστάσεις. Η διαδικασία αυτή είναι απαραίτητη, καθώς διασφαλίζει ότι το εισερχόμενο κείμενο είναι ομοιογενές, καθαρό και απαλλαγμένο από μορφοποιητικό θόρυβο που προκύπτει κατά την εξαγωγή από τα αρχεία PDF.

Αρχικά, το κείμενο υφίσταται κανονικοποίηση χαρακτήρων και μετατροπή όλων των λέξεων σε πεζά, ώστε να αποφεύγονται διαφορές μεταξύ γραφών με κεφαλαία ή τονισμένα γράμματα. Στη συνέχεια, αφαιρούνται μη αναγνωρίσιμοι χαρακτήρες, αλλαγές γραμμής, κενά και σημεία στίξης που δεν προσφέρουν σημασιολογική πληροφορία.

Η βιβλιοθήκη χρησιμοποιεί κανόνες κανονικοποίησης Unicode και regex pattern matching, ώστε το κείμενο να αποκτά καθαρή και σταθερή μορφή.

Μετά τον αρχικό καθαρισμό, εφαρμόζεται tokenization μέσω του SentencePiece tokenizer της Google, ο οποίος έχει εκπαιδευτεί σε πολυγλωσσικά σύνολα δεδομένων και μπορεί να αναγνωρίσει σωστά ελληνικές λέξεις και σύνθετες εκφράσεις. Κατά το στάδιο αυτό, αγνοούνται οι stopwords (π.χ. «και», «το», «είναι»), ώστε το τελικό embedding να βασίζεται μόνο στις λέξεις που έχουν ουσιαστικό νόημα. Το καθαρισμένο και κανονικοποιημένο κείμενο χωρίζεται στη συνέχεια σε chunks συγκεκριμένου μήκους (περίπου 1000 χαρακτήρες), με επικάλυψη 150 χαρακτήρων, ώστε να διατηρείται η θεματική συνοχή και να αποφεύγεται απώλεια πληροφορίας στις μεταβάσεις.

Τέλος, κάθε chunk εισάγεται στο Gemini Embeddings API, το οποίο εφαρμόζει έναν προεκπαιδευμένο sentence encoder για τη δημιουργία dense vectors που αποτυπώνουν το σημασιολογικό περιεχόμενο του κειμένου. Η διαδικασία αυτή εκτελείται αυτόματα και έχει ως αποτέλεσμα τη δημιουργία καθαρών, σταθερών και συνεπών αναπαραστάσεων, έτοιμων να αποθηκευτούν στη βάση διανυσμάτων ChromaDB.

Η ύπαρξη αυτού του εσωτερικού μηχανισμού καθαρισμού αποτελεί σημαντικό πλεονέκτημα της βιβλιοθήκης langchain_google_genai, καθώς μειώνει την ανάγκη εξωτερικού preprocessing και εξασφαλίζει υψηλή ακρίβεια στην αντιστοίχιση νοημάτων κατά την αναζήτηση στη βάση δεδομένων.

Τα embeddings διατηρήθηκαν σε μια βάση δεδομένων τύπου ChromaDB, η οποία είναι ένα μικρό και αποτελεσματικό πρόγραμμα open source που δημιουργήθηκε ειδικά για αναζητήσεις βάσει ομοιότητας. Ακόμη και με χιλιάδες αποσπάσματα, η ChromaDB επέτρεψε γρήγορη αναζήτηση, διασφαλίζοντας την ανάκτηση αποτελεσμάτων σε πραγματικό χρόνο.

Ο βαθμός ομοιότητας μεταξύ του embedding της ερώτησης και των embeddings των αποσπασμάτων χρησιμοποιήθηκε για την πραγματοποίηση της αναζήτησης στη βάση δεδομένων. Η ομοιότητα συνημιτόνου, η οποία μετρά τη γωνία μεταξύ δύο διανυσμάτων, αποτέλεσε τη βασική μετρική σύγκρισης. Όσο μικρότερη είναι η γωνία μεταξύ τους, τόσο μεγαλύτερη είναι η σημασιολογική ομοιότητα των περιεχομένων. Μέσω αυτής της διαδικασίας, το σύστημα μπορούσε να εντοπίζει τα πιο συναφή αποσπάσματα και να τα αποστέλλει στο μεγάλο γλωσσικό μοντέλο, προκειμένου να παραχθεί η τελική απάντηση.

Μετά τη δημιουργία των embeddings και τον υπολογισμό των ομοιοτήτων, τα διανύσματα αποθηκεύτηκαν σε μια βάση διανυσμάτων τύπου ChromaDB, η οποία αποτελεί ένα ελαφρύ και ιδιαίτερα αποδοτικό σύστημα ανοιχτού κώδικα, σχεδιασμένο ειδικά για αναζητήσεις βάσει σημασιολογικής συνάφειας. Ακόμη και με χιλιάδες αποσπάσματα, η ChromaDB εξασφάλισε ταχύτατη αναζήτηση και ανάκτηση αποτελεσμάτων σε πραγματικό χρόνο, επιτρέποντας στο σύστημα να παραμένει λειτουργικό και αποδοτικό ακόμη και σε μεγάλα σύνολα εγγράφων.

Η προσέγγιση Retrieval-Augmented Generation (RAG) που εφαρμόστηκε στη συγκεκριμένη εργασία βασίστηκε ακριβώς σε αυτόν τον συνδυασμό: embeddings υψηλής ποιότητας σε συνδυασμό με τη βάση διανυσμάτων ChromaDB. Μέσω αυτής της αρχιτεκτονικής, το σύστημα ήταν σε θέση να εξάγει πληροφορίες απευθείας από τα κανονιστικά έγγραφα, παρακάμπτοντας τους περιορισμούς των μεγάλων γλωσσικών μοντέλων, τα οποία στηρίζονται αποκλειστικά στα δεδομένα εκπαίδευσής τους. Η ενσωμάτωση πραγματικών εγγράφων στη διαδικασία παραγωγής απαντήσεων βελτίωσε σημαντικά την ακρίβεια, μείωσε τις πιθανότητες εμφάνισης λανθασμένων ή μη τεκμηριωμένων πληροφοριών και ενίσχυσε την αξιοπιστία του συστήματος. Χωρίς αυτό το βήμα, το chatbot δεν θα μπορούσε να αξιοποιήσει αποτελεσματικά τα κανονιστικά δεδομένα ούτε να παρέχει απαντήσεις βασισμένες σε αξιόπιστες και επαληθεύσιμες πηγές.

Η ChromaDB λειτουργεί ως εξειδικευμένο σύστημα αποθήκευσης διανυσμάτων και υποστηρίζει αναζητήσεις βάσει σημασιολογικής ομοιότητας.

Σε αντίθεση με μια απλή βάση δεδομένων όπου η αναζήτηση γίνεται με ταυτόσημες λέξεις ή φράσεις, η ChromaDB εκτελεί υπολογισμούς σε πολυδιάστατο χώρο: κάθε απόσπασμα κειμένου αντιστοιχίζεται σε ένα σημείο αυτού του χώρου, και οι αποστάσεις μεταξύ σημείων καθορίζουν τη νοηματική τους εγγύτητα.

Εσωτερική λειτουργία και αποθήκευση δεδομένων

Η ChromaDB αποθηκεύει τα embeddings ως πυκνά διανύσματα (dense vectors) μεγέθους d , όπου d είναι η διάσταση του εκάστοτε embedding. Όλα τα διανύσματα τοποθετούνται σε έναν πίνακα τύπου $n \times d$, όπου n είναι το πλήθος των αποσπασμάτων (chunks) που προέκυψαν από τα αρχεία PDF. Κάθε διάνυσμα συνοδεύεται από metadata, όπως το όνομα του αρχείου, η σελίδα, η θέση και το περιεχόμενο του αποσπάσματος. Έτσι, όταν το σύστημα ανακτήσει ένα embedding, μπορεί να γνωρίζει από ποιο αρχείο και ποια σελίδα προέρχεται η πληροφορία.

Η ChromaDB χρησιμοποιεί εσωτερικά δομές ευρετηρίασης (indexing structures) ώστε να επιτρέπει γρήγορη ανάκτηση χωρίς να ελέγχει ολόκληρη τη βάση κάθε φορά.

Η πιο σημαντική από αυτές είναι η HNSW (Hierarchical Navigable Small World Graph), μια μέθοδος που οργανώνει τα διανύσματα σε έναν γράφο πολλαπλών επιπέδων. Κάθε κόμβος του γράφου συνδέεται με κοντινά διανύσματα (γειτονικούς κόμβους), σχηματίζοντας ένα δίκτυο που επιτρέπει αναζήτηση με πολυπλοκότητα σχεδόν $O(\log n)$.

Αυτό σημαίνει ότι ακόμη και με δεκάδες χιλιάδες αποσπάσματα, ο χρόνος απάντησης παραμένει κάτω από ένα δευτερόλεπτο.

Μηχανισμός Αναζήτησης

Κατά την υποβολή ενός ερωτήματος από τον χρήστη:

- Η ερώτηση μετατρέπεται σε embedding μέσω του ίδιου μοντέλου.
- Υπολογίζεται η ομοιότητα συνημίτονου μεταξύ του embedding της ερώτησης και όλων των αποθηκευμένων διανυσμάτων:

$$\text{similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (3.1)$$

- Η βάση επιστρέφει τα k πιο σχετικά αποσπάσματα.
- Αυτά τα αποσπάσματα ενώνονται στο prompt που δίνεται στο LLM, έτσι ώστε η απάντηση να βασίζεται σε πραγματικά δεδομένα.

Ο μηχανισμός αυτός είναι σημασιολογικός, όχι απλά συντακτικός: δύο προτάσεις που έχουν διαφορετικές λέξεις αλλά ίδιο νόημα θα εντοπιστούν ως παρόμοιες, γιατί τα διανύσματά τους βρίσκονται πολύ κοντά στον χώρο των embeddings.

Πλεονεκτήματα της ChromaDB

Η χρήση της ChromaDB παρουσίασε πολλαπλά πλεονεκτήματα για το έργο:

- **Ταχύτητα:** η αναζήτηση βασίζεται σε ευρετηρίαση με HNSW, επιτυγχάνοντας πολύ χαμηλό χρόνο απόκρισης ακόμη και σε μεγάλα σύνολα δεδομένων.
- **Ακρίβεια:** οι απαντήσεις στηρίζονται σε πραγματικά αποσπάσματα με τη μεγαλύτερη σημασιολογική συνάφεια.
- **Ευκολία ενσωμάτωσης:** η ChromaDB ενσωματώνεται άμεσα με τη βιβλιοθήκη LangChain και δεν απαιτεί εξωτερικό server.

- **Αποθήκευση και επαναχρησιμοποίηση:** η λειτουργία persistence επιτρέπει την τοπική αποθήκευση των embeddings, ώστε να μη χρειάζεται επανυπολογισμός κάθε φορά που ανοίγει η εφαρμογή.

Σημασία για τη μεθοδολογία RAG

Η ενσωμάτωση της βάσης διανυσμάτων αποτέλεσε καθοριστικό στοιχείο για την επιτυχία του συστήματος, καθώς συνέβαλε ουσιαστικά στην πρακτική εφαρμογή της μεθοδολογίας RAG. Μέσω της ChromaDB, το LLM δεν περιορίζεται πλέον αποκλειστικά στη γενική γνώση που απέκτησε κατά τη φάση της εκπαίδευσής του, αλλά αποκτά τη δυνατότητα να αντλεί άμεσα ακριβείς και τεκμηριωμένες πληροφορίες από τα πραγματικά έγγραφα PDF των κανονισμών. Με τον τρόπο αυτό, κάθε απάντηση βασίζεται σε πραγματικό περιεχόμενο και όχι σε υποθέσεις ή γενικεύσεις που μπορεί να παράγει το μοντέλο.

Η προσέγγιση αυτή βελτίωσε σημαντικά την ποιότητα των απαντήσεων, μειώνοντας τα φαινόμενα hallucinations, δηλαδή περιπτώσεων όπου το LLM κατασκευάζει μη υπαρκτές πληροφορίες ή λανθασμένα δεδομένα. Παράλληλα, ενίσχυσε την αξιοπιστία και τη συνέπεια του συστήματος, καθώς οι απαντήσεις συνοδεύονται από αναφορές στις πραγματικές πηγές από τις οποίες αντλήθηκαν. Το αποτέλεσμα είναι ένα σύστημα που μπορεί να λειτουργήσει όχι απλώς ως συνομιλιακό εργαλείο, αλλά ως ένα τεκμηριωμένο βοήθημα ακαδημαϊκής υποστήριξης, παρέχοντας στους φοιτητές σαφείς, επικαιροποιημένες και επαληθεύσιμες πληροφορίες. Με αυτόν τον τρόπο, η RAG μετατρέπεται από θεωρητική μεθοδολογία σε μια ουσιαστικά εφαρμόσιμη λύση για το πανεπιστημιακό περιβάλλον, αποδεικνύοντας τη δυνατότητα των μεγάλων γλωσσικών μοντέλων να συνδυάζουν τη φυσική γλώσσα με έγκυρα δεδομένα και να παρέχουν απαντήσεις υψηλής ακρίβειας και επιστημονικής εγκυρότητας.

3.5.4 Διαχείριση API Tokens

Η πρόσβαση στα γλωσσικά μοντέλα που χρησιμοποιήθηκαν στην παρούσα εργασία πραγματοποιήθηκε μέσω API tokens, τα οποία αποτελούν μοναδικούς κωδικούς ταυτοποίησης που εκδίδονται από τους αντίστοιχους παρόχους. Τα tokens αυτά επιτρέπουν στον χρήστη να επικοινωνεί με τα μοντέλα μέσω διαδικτυακών υπηρεσιών (APIs), να αποστέλλει ερωτήματα και να λαμβάνει απαντήσεις σε πραγματικό χρόνο. Με τον τρόπο αυτό, η υλοποίηση κατέστη δυνατή χωρίς να απαιτείται η τοπική εκτέλεση των μοντέλων, η οποία θα απαιτούσε ιδιαίτερα μεγάλη υπολογιστική ισχύ.

Στην εργασία αξιοποιήθηκαν τρία διαφορετικά μοντέλα, το καθένα με τη δική του διαδικασία πρόσβασης:

- Για το Gemini, τα API tokens εκδόθηκαν μέσω της πλατφόρμας Google AI Studio, που παρέχει πρόσβαση στα μοντέλα της Google.
- Για τα LLaMA και Gemma, τα tokens αποκτήθηκαν μέσω της υπηρεσίας OpenRouter, μιας πλατφόρμας που συγκεντρώνει διαφορετικά μεγάλα γλωσσικά μοντέλα και επιτρέπει την πρόσβαση σε αυτά μέσα από ένα ενιαίο API.

Η χρήση του OpenRouter προσέφερε σημαντικά πλεονεκτήματα, καθώς έδωσε τη δυνατότητα να ενσωματωθούν πολλαπλά LLMs στην ίδια εφαρμογή με ελάχιστες αλλαγές στον κώδικα, μειώνοντας τον χρόνο ανάπτυξης και απλοποιώντας τη διαδικασία ενοποίησης.

Για λόγους ασφάλειας, τα API tokens δεν ενσωματώθηκαν απευθείας στον πηγαίο κώδικα, αλλά αποθηκεύτηκαν σε ξεχωριστό αρχείο περιβάλλοντος (.env). Μέσω αυτού, τα κλειδιά φορτώνονταν δυναμικά κατά την εκτέλεση του προγράμματος, χωρίς να εμφανίζονται στο script. Η πρακτική αυτή εξασφάλισε τόσο την προστασία των ευαίσθητων πληροφοριών όσο και την ευκολία στη μεταφορά και εγκατάσταση της εφαρμογής σε διαφορετικά περιβάλλοντα.

Κατά τη χρήση των API, λήφθηκαν υπόψη οι περιορισμοί πρόσβασης που επιβάλλουν οι πάροχοι των μοντέλων, καθώς κάθε υπηρεσία εφαρμόζει συγκεκριμένα όρια για να εξασφαλίσει τη σταθερότητα και τη συνεχή διαθεσιμότητα των πόρων της. Οι περιορισμοί αυτοί χωρίζονται σε τρεις κύριες κατηγορίες:

- **Περιορισμοί χωρητικότητας (token limits):** αφορούν το μέγιστο μέγεθος του κειμένου που μπορεί να σταλεί ή να παραχθεί σε κάθε αίτημα. Περιλαμβάνουν τόσο το ερώτημα (prompt) του χρήστη όσο και την παραγόμενη απάντηση του μοντέλου. Αν το σύνολο υπερβεί το επιτρεπόμενο, το API απορρίπτει το αίτημα με μήνυμα σφάλματος.
- **Περιορισμοί ταχύτητας (rate limits):** καθορίζουν πόσα αιτήματα μπορούν να αποστέλλονται σε συγκεκριμένο χρονικό διάστημα (π.χ. ανά δευτερόλεπτο ή ανά λεπτό). Όταν ο ρυθμός υποβολής αιτημάτων ξεπερνά το όριο, η υπηρεσία απαντά με σφάλμα προσωρινής υπέρβασης (Too Many Requests), ζητώντας να γίνει εκ νέου προσπάθεια μετά από μικρό χρονικό διάστημα.
- **Πολιτικές κατανάλωσης (usage quotas):** καθορίζουν τον συνολικό αριθμό αιτημάτων ή tokens που μπορούν να χρησιμοποιηθούν μέσα σε μια ημέρα ή έναν μήνα. Αυτές οι πολιτικές εφαρμόζονται για να αποτραπεί η υπερβολική χρήση του API από μεμονωμένους χρήστες ή εφαρμογές.

Στην πράξη, κάθε πλατφόρμα καθορίζει τα δικά της όρια, τα οποία εξαρτώνται από το είδος του λογαριασμού και το επίπεδο πρόσβασης. Το Gemini API της Google εφαρμόζει όρια τόσο στο πλήθος των αιτημάτων που μπορούν να σταλούν ανά χρονική περίοδο, όσο και στο μέγεθος του περιεχομένου (tokens) που επεξεργάζεται σε κάθε κλήση. Αντίστοιχα, η πλατφόρμα OpenRouter, μέσω της οποίας πραγματοποιήθηκε η πρόσβαση στα μοντέλα LLaMA και Gemma, επιβάλλει περιορισμούς στη συχνότητα των αιτημάτων και στην ημερήσια κατανάλωση, διαφοροποιώντας τα όρια μεταξύ ελεύθερης και επαγγελματικής χρήσης. Οι περιορισμοί αυτοί είναι δυναμικοί και ενδέχεται να αναθεωρούνται από τους παρόχους ανάλογα με τη ζήτηση ή τις τεχνικές αναβαθμίσεις των υπηρεσιών τους.

Επιπλέον, σε ορισμένα APIs εφαρμόζεται και περιορισμός ταυτόχρονων αιτημάτων (concurrent requests), δηλαδή του πλήθους αιτημάτων που μπορούν να εκτελούνται παράλληλα από την ίδια εφαρμογή. Αυτός ο τύπος περιορισμού συμβάλλει στην αποφυγή υπερφόρτωσης του διακομιστή και στη σταθερότητα της ανταπόκρισης.

Κατά τη φάση των δοκιμών και της αξιολόγησης των μοντέλων, οι παραπάνω περιορισμοί ελήφθησαν υπόψη τόσο στον σχεδιασμό των πειραμάτων όσο και στη συχνότητα των ερωτημάτων. Με αυτόν τον τρόπο αποφεύχθηκαν σφάλματα υπέρβασης των διαθέσιμων πόρων και διατηρήθηκε η αξιοπιστία των αποτελεσμάτων.

3.5.5 Υλοποίηση με Πολλαπλά LLMs

Η ιδέα πίσω από την ανάπτυξη του συστήματος ήταν ότι, προκειμένου να διασφαλιστεί η αντικειμενική σύγκριση των αποτελεσμάτων, το ίδιο pipeline θα πρέπει να μπορεί να συνδεθεί με διάφορα μεγάλα

γλωσσικά μοντέλα. Εκτός από την παραγωγή αποτελεσμάτων από τα μοντέλα, ο στόχος ήταν να δειχθεί μεθοδικά πώς ποικίλλει η απόδοση των μοντέλων όταν χρησιμοποιούνται στο ίδιο σύνολο δεδομένων.

Τα μοντέλα επιλέχθηκαν με βάση την προσβασιμότητά τους μέσω API, τη διαθεσιμότητά τους και τη θέση τους στην παγκόσμια βιβλιογραφία. Το Gemini επιλέχθηκε ως κύριο παράδειγμα της επιχειρηματικής στρατηγικής της Google. Το LLaMA είναι ένα παράδειγμα μοντέλου ανοιχτού κώδικα, το οποίο χρησιμοποιείται ευρέως για πειραματισμό και έχει σημαντική ερευνητική αξία. Τέλος, η Gemma, ένα πιο πρόσφατο μοντέλο της Google προσβάσιμο μέσω της πλατφόρμας OpenRouter, θεωρήθηκε επωφελές να συμπεριληφθεί.

Η μέθοδος που χρησιμοποιήθηκε για την ενσωμάτωση του συστήματος εγγυάται την ενιαία διαχείριση. Στην πραγματικότητα, αναπτύχθηκαν συναρτήσεις "wrapper", οι οποίες ανατέθηκε ο ρόλος της διαμόρφωσης του ερωτήματος του χρήστη και των αποσπασμάτων σε ένα μήνυμα πριν από την ενεργοποίηση του κατάλληλου API. Επομένως, η εμπειρία του τελικού χρήστη ήταν συνεπής, παρόλο που οι διαδικασίες επικοινωνίας με τα API (LLaMA–Gemma μέσω OpenRouter και Gemini μέσω Google AI Studio) διέφεραν τεχνικά: την ίδια ερώτηση έστελναν και στα τρία μοντέλα κάθε φορά, με το ίδιο πλαίσιο πληροφοριών.

Το περιβάλλον διεπαφής ενσωμάτωνε επίσης συγκριτική λογική. Για να επιτευχθεί άμεση σύγκριση, οι απαντήσεις από καθένα από τα τρία μοντέλα εμφανίστηκαν για κάθε ερώτηση. Επειδή ο χρήστης μπορούσε να δει τα αποτελέσματα δίπλα-δίπλα, αυτό απλοποίησε τη διαδικασία αξιολόγησης και μείωσε την πιθανότητα μεροληψίας.

Πρόεκυψαν πρακτικά προβλήματα κατά την υλοποίηση, όπως οι διαφορετικοί χρόνοι απόκρισης μεταξύ των API ή τα διάφορα όρια token που υποστηρίζονται από κάθε μοντέλο. Με την τροποποίηση των παραμέτρων και τη διαχείριση των δεδομένων εισόδου, αυτά τα προβλήματα επιλύθηκαν. Για να γίνει η αξιολόγηση όσο το δυνατόν πιο αντικειμενική, καταγράφηκαν ταυτόχρονα πρόσθετες μετρήσεις για κάθε ερώτηση, συμπεριλαμβανομένου του χρόνου απόκρισης (latency), του μήκους της απάντησης σε λέξεις και των πηγών που χρησιμοποιήθηκαν. Για να αποφευχθεί η υπερφόρτωση του κύριου κειμένου, οι παράμετροι και ο σχετικός κώδικας αναλύονται λεπτομερώς στο Παράρτημα Α.

Το πλαίσιο για τη συγκριτική τους αξιολόγηση καθορίστηκε με την ολοκλήρωση της υλοποίησης χρησιμοποιώντας τα τρία διακριτά μοντέλα. Μια διεξοδική εξέταση των αποτελεσμάτων της υλοποίησης του συστήματος παρέχεται στο επόμενο κεφάλαιο. Για να καταδειχθούν οι ομοιότητες και οι διαφορές μεταξύ των LLM, αξιολογούνται η ακρίβεια, η πληρότητα, ο χρόνος απόκρισης και το στυλ των απαντήσεων χρησιμοποιώντας συγκεκριμένα παραδείγματα και μετρικές. Το κύριο κριτήριο για τον προσδιορισμό των δυνατοτήτων και των αδυναμιών κάθε μοντέλου είναι αυτή η ανάλυση.

3.6 Μετρικές Αξιολόγησης και Μαθηματικοί Ορισμοί

Η αξιολόγηση των αποτελεσμάτων που παράγουν τα LLMs αποτελεί κρίσιμο στάδιο σε κάθε πειραματική μελέτη που αφορά επεξεργασία φυσικής γλώσσας. Οι μετρικές αξιολόγησης προσφέρουν έναν αντικειμενικό τρόπο μέτρησης της ποιότητας, της ακρίβειας και της σημασιολογικής πληρότητας των απαντήσεων που παράγει ένα μοντέλο. Με τον τρόπο αυτό, οι ερευνητές μπορούν να συγκρίνουν διαφορετικά συστήματα και να εντοπίζουν ποιο μοντέλο είναι καταλληλότερο για κάθε τύπο εφαρμογής.

Στην παρούσα εργασία χρησιμοποιήθηκαν τρεις ευρέως καθιερωμένες μετρικές αξιολόγησης: BLEU, ROUGE και BERTScore. Οι δείκτες αυτοί εξετάζουν διαφορετικές πτυχές της απόδοσης: ο BLEU επικεντρώνεται στη λεκτική ακρίβεια, ο ROUGE στην πληρότητα του περιεχομένου και ο BERTScore στη σημασιολογική συνάφεια. Ο συνδυασμός τους επιτρέπει μια πολύπλευρη και ισορροπημένη αξιολόγηση των LLMs.

3.6.1 Θεωρητικό Υπόβαθρο των Μετρικών

Οι μετρικές που χρησιμοποιούνται για την αξιολόγηση συστημάτων φυσικής γλώσσας μπορούν να ταξινομηθούν σε δύο βασικές κατηγορίες:

- **Λεκτικές ή στατιστικές μετρικές**, οι οποίες υπολογίζουν την επικάλυψη λέξεων ή ακολουθιών λέξεων (n-grams) μεταξύ των προτάσεων που παρήγαγε το μοντέλο και των προτάσεων-αναφοράς. Τυπικά παραδείγματα αυτής της κατηγορίας είναι οι BLEU και ROUGE.
- **Σημασιολογικές μετρικές**, οι οποίες χρησιμοποιούν διανυσματικές αναπαραστάσεις για να υπολογίσουν τη νοηματική εγγύτητα ανάμεσα σε δύο προτάσεις, ανεξάρτητα από τη συντακτική τους ομοιότητα. Σε αυτή την κατηγορία ανήκει ο BERTScore.

Η ταυτόχρονη χρήση και των δύο τύπων μετρικών κρίνεται απαραίτητη, διότι οι λεκτικές μετρικές μετρούν την πιστότητα στη διατύπωση, ενώ οι σημασιολογικές την πιστότητα στο νόημα. Με αυτόν τον τρόπο επιτυγχάνεται μια πιο ρεαλιστική εκτίμηση της πραγματικής απόδοσης ενός LLM.

3.6.2 BLEU

Η μετρική BLEU υπήρξε η πρώτη συστηματική μέθοδος αυτόματης αξιολόγησης μεταφράσεων και παραγόμενου κειμένου. Υπολογίζει πόσο μεγάλο ποσοστό των λέξεων ή των ακολουθιών λέξεων της απάντησης του μοντέλου υπάρχουν και στο κείμενο-αναφορά, μετρώντας ουσιαστικά την ακρίβεια (precision).

Ο μαθηματικός ορισμός του δείκτη δίνεται από τη σχέση:

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (3.2)$$

όπου

- $p_n = \frac{\text{αριθμός κοινών } n\text{-grams}}{\text{σύνολο } n\text{-grams της πρόβλεψης}}$
- w_n : τα βάρη κάθε n-gram

- *BP*: brevity penalty, συντελεστής που μειώνει την τιμή του δείκτη όταν η παραγόμενη απάντηση είναι υπερβολικά σύντομη σε σχέση με την αναφορά:

$$BP = \begin{cases} 1, & \text{αν } c > r \\ e^{(1-\frac{r}{c})}, & \text{αν } c \leq r \end{cases}$$

όπου c είναι το μήκος της πρόβλεψης και r το μήκος της αναφοράς.

3.6.3 ROUGE

Η μετρική ROUGE αναπτύχθηκε αρχικά για την αξιολόγηση συστημάτων αυτόματης περίληψης. Επικεντρώνεται στην ανάκληση, δηλαδή στο ποιο ποσοστό των λέξεων ή φράσεων της αναφοράς περιλαμβάνεται στο παραγόμενο κείμενο.

Η πιο συνηθισμένη εκδοχή, ROUGE-N, ορίζεται ως:

$$ROUGE - N = \frac{\sum_{gram_n \in Ref} Count_{match}(gram_n)}{\sum_{gram_n \in Ref} Count(gram_n)} \quad (3.3)$$

όπου

- $Count_{match}(gram_n)$: πλήθος κοινών n-grams μεταξύ αναφοράς και πρόβλεψης,
- $Count(gram_n)$: πλήθος συνολικών n-grams στην αναφορά.

Οι εκδόσεις ROUGE-1 και ROUGE-2 μετρούν αντίστοιχα την επικάλυψη μονόλεξων και δίλεξων, ενώ η ROUGE-L βασίζεται στη μεγαλύτερη κοινή ακολουθία λέξεων (Longest Common Subsequence).

Πλεονεκτήματα:

- Καταγράφει το ποσοστό πληρότητας της πληροφορίας.
- Εξαιρετικά χρήσιμη για περιλήψεις και απαντήσεις πολλαπλών προτάσεων.

Μειονεκτήματα:

- Δεν κατανοεί τη σημασία των λέξεων.
- Δύο νοηματικά ισοδύναμες απαντήσεις με διαφορετικές λέξεις μπορεί να λάβουν χαμηλό σκορ.

3.6.4 BERTScore

Η μετρική BERTScore αξιοποιεί τις δυνατότητες των νευρωνικών μοντέλων τύπου Transformer για τη σημασιολογική σύγκριση προτάσεων. Αντί να ελέγχει αν δύο προτάσεις έχουν κοινές λέξεις, υπολογίζει την ομοιότητα των embeddings που παράγονται από το μοντέλο BERT για κάθε λέξη ή φράση.

Για κάθε λέξη x_i της πρόβλεψης και κάθε λέξη y_i της αναφοράς, υπολογίζεται η συνημιτονική ομοιότητα:

$$\text{sim}(x_i, y_i) = \frac{E(x_i) \cdot E(y_i)}{\|E(x_i)\| \|E(y_i)\|} \quad (3.4)$$

Από αυτήν υπολογίζονται:

$$P = \frac{1}{m} \sum_i \max \text{sim}(x_i, y_i), \quad R = \frac{1}{n} \sum_j \max \text{sim}(x_i, y_i)$$

και το τελικό σκορ δίνεται από το αρμονικό μέσο των δύο (F1):

$$F1 = 2 \times \frac{P \times R}{P + R}$$

όπου P (Precision) μετρά τη σημασιολογική ακρίβεια και R (Recall) τη σημασιολογική πληρότητα.

Πλεονεκτήματα:

- Ανιχνεύει συνώνυμα και παραφράσεις.
- Ανθεκτικός σε μορφολογικές διαφοροποιήσεις.
- Αντικατοπτρίζει καλύτερα την ανθρώπινη κρίση.

Μειονεκτήματα:

- Απαιτεί αυξημένους υπολογιστικούς πόρους.
- Εξαρτάται από το ποιο προεκπαιδευμένο μοντέλο BERT χρησιμοποιείται.

3.6.5 Συγκριτική Ερμηνεία των Μετρικών

Μετρική	Τύπος	Εστίαση	Εύρος Τιμών	Πλεονεκτήματα	Περιορισμοί
BLEU	Lexical	Ακρίβεια	0-1	Ταχύτητα, σταθερότητα	Ευαίσθητη σε διατύπωση
ROUGE	Lexical	Ανάκληση	0-1	Ανιχνεύει ελλείψεις	Όχι σημασιολογική ανάλυση
BERTScore	Semantic	Νοηματική Συνάφεια	0-1	Αναγνωρίζει συνώνυμα	Υπολογιστικά βαριά

Πίνακας 3.2: Σύγκριση των μετρικών.

Η ταυτόχρονη χρήση και των τριών εξασφαλίζει πληρότητα:

- Η BLEU δείχνει πόσο κοντά στη διατύπωση του ground truth είναι το κείμενο.
- Η ROUGE δείχνει αν καλύπτεται πλήρως η πληροφορία.
- Η BERTScore δείχνει αν το νόημα είναι σωστό, ακόμη κι αν οι λέξεις διαφέρουν.

3.6.6 Εφαρμογή των Μετρικών στην Παρούσα Εργασία

Στο πλαίσιο της παρούσας εργασίας, οι τρεις μετρικές εφαρμόστηκαν στο ίδιο σύνολο δεδομένων ερωτήσεων–απαντήσεων, προκειμένου να αξιολογηθεί συγκριτικά η απόδοση των τριών Μεγάλων Γλωσσικών Μοντέλων. Για κάθε ερώτηση, οι απαντήσεις που παρήχθησαν από τα μοντέλα συγκρίθηκαν με το αντίστοιχο κείμενο/αναφορά, ώστε να υπολογιστούν οι δείκτες λεκτικής ακρίβειας, πληρότητας πληροφορίας και σημασιολογικής συνάφειας.

ID	Question	Answer
1	Πόσα εξάμηνα διαρκεί το προπτυχιακό πρόγραμμα σπουδών του Τμήματος Μηχανολόγων Μηχανικών ΔΙΠΑΕ;	Δέκα (10) ακαδημαϊκά εξάμηνα, με δυνατότητα παράτασης έως έξι (6) επιπλέον.
2	Πόσα ECTS απαιτούνται για τη λήψη του διπλώματος Μηχανολόγου Μηχανικού στο ΔΙΠΑΕ;	300 ECTS (30 ECTS ανά εξάμηνο).
3	Πόσες εβδομάδες διδασκαλίας περιλαμβάνει κάθε εξάμηνο;	Κατ' ελάχιστο, 13 εβδομάδες διδασκαλίας και 1 κενή εβδομάδα πριν την εξεταστική περίοδο.
4	Ποιο είναι το ανώτατο όριο πιστωτικών μονάδων (ECTS) που μπορεί να δηλώσει ένας φοιτητής ανά εξάμηνο;	Στο 1ο έτος 30 ECTS, στο 2ο έτος 48 ECTS, και από το 3ο έτος και μετά μέχρι 60 ECTS.
5	Ποια είναι η ελάχιστη βαθμολογία για να θεωρηθεί ότι πέρασε ένας φοιτητής σε ένα μάθημα;	Βαθμός πέντε (5) σε κλίμακα 0–10.
6	Τι συμβαίνει αν ένας φοιτητής αποτύχει τρεις φορές στο ίδιο μάθημα;	Έχει δικαίωμα να εξεταστεί από τριμελή επιτροπή διδασκόντων.
7	Ποιες είναι οι ποινές για αντιγραφή κατά τη διάρκεια των εξετάσεων;	Ποινή αποκλεισμού από όλα τα μαθήματα της επόμενης εξεταστικής.
8	Πότε πραγματοποιούνται οι ορκυμοσίες των αποφοίτων;	Το αργότερο εντός τριών (3) μηνών μετά το πέρας των εξεταστικών περιόδων.
9	Τι προϋποθέσεις υπάρχουν για τη διενέργεια προαιρετικής πρακτικής άσκησης;	Απαιτείται η ολοκλήρωση όλων των μαθημάτων κορμού και των μαθημάτων κατεύθυνσης (240 ECTS).
10	Ποιος είναι ο Πρόεδρος του Τμήματος Μηχανολόγων Μηχανικών ΔΙΠΑΕ σύμφωνα με τον κανονισμό;	Κώστας Κλειδής, Καθηγητής Α' Βαθμίδας.

Πίνακας 3.3: Σύνολο ερωτήσεων-απαντήσεων για πολυμετρική αξιολόγηση.

Η ανάλυση των αποτελεσμάτων κατέδειξε ότι οι τρεις μετρικές προσφέρουν συμπληρωματικές οπτικές γωνίες αξιολόγησης. Ο δείκτης BLEU επικεντρώνεται στη λεκτική ομοιότητα, αποτυπώνοντας το βαθμό στον οποίο η παραγόμενη απάντηση ακολουθεί τη φρασεολογία του κειμένου/αναφοράς. Ο δείκτης ROUGE αξιολογεί την πληρότητα της απάντησης, υπολογίζοντας το ποσοστό των πληροφοριών του αρχικού κειμένου που αποδίδονται από το μοντέλο. Τέλος, ο δείκτης BERTScore

εκτιμά τη σημασιολογική εγγύτητα μεταξύ των δύο κειμένων, βασιζόμενος στα embeddings που ενσωματώνουν τη νοηματική συσχέτιση των λέξεων.

Η πολυμετρική προσέγγιση επιτρέπει την εξαγωγή πιο αντικειμενικών και ρεαλιστικών συμπερασμάτων, αποφεύγοντας τη μονοδιάστατη αξιολόγηση που θα προέκυπτε από τη χρήση μόνο ενός δείκτη. Ο συνδυασμός των BLEU, ROUGE και BERTScore προσφέρει μια ολιστική εικόνα της απόδοσης των μοντέλων, καλύπτοντας τόσο τη λεκτική όσο και τη σημασιολογική διάσταση των απαντήσεων. Με αυτόν τον τρόπο, ενισχύεται η αξιοπιστία της πειραματικής διαδικασίας και τεκμηριώνεται με σαφήνεια η συνολική λειτουργικότητα του συστήματος.

Κεφάλαιο 4ο: Αποτελέσματα

4.1 Εισαγωγή

Η αξιολόγηση αποτελεί ένα από τα σημαντικότερα στάδια της παρούσας εργασίας, καθώς επιτρέπει να εκτιμηθεί κατά πόσο το σύστημα που αναπτύχθηκε ανταποκρίνεται στον στόχο του, δηλαδή την παροχή ακριβών και τεκμηριωμένων απαντήσεων με βάση τους κανονισμούς σπουδών. Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή του pipeline στα 50 αρχεία PDF, καθώς και η συγκριτική αξιολόγηση των τριών μεγάλων γλωσσικών μοντέλων (Gemini, LLaMA και Gemma). Αντίστοιχη προσέγγιση ακολουθείται και στη μελέτη HELM του Stanford [26], η οποία συγκρίνει πολλαπλά LLMs υπό κοινό πλαίσιο αξιολόγησης, χρησιμοποιώντας μετρικές όπως η ακρίβεια, ο χρόνος απόκρισης και το υπολογιστικό κόστος.

Η αξιολόγηση πραγματοποιήθηκε με δύο βασικές μεθόδους:

- **Ποιοτική αξιολόγηση:** Εξετάστηκαν δείγματα ερωτήσεων και συγκρίθηκαν οι απαντήσεις που παρήγαγαν τα μοντέλα με τις αναμενόμενες απαντήσεις (ground truth). Μέσα από αυτή τη διαδικασία εντοπίστηκαν φαινόμενα πληρότητας, ακρίβειας αλλά και πιθανά σφάλματα, όπως γενικεύσεις ή δημιουργία μη τεκμηριωμένων πληροφοριών (hallucinations).
- **Ποσοτική αξιολόγηση:** Υπολογίστηκαν καθιερωμένοι δείκτες της φυσικής γλώσσας, όπως BERTScore, BLEU και ROUGE, οι οποίοι μετρούν διαφορετικές πτυχές της ποιότητας των απαντήσεων. Παράλληλα, καταγράφηκαν μετρικές συστήματος, όπως ο χρόνος απόκρισης (latency) και το μήκος της παραγόμενης απάντησης.

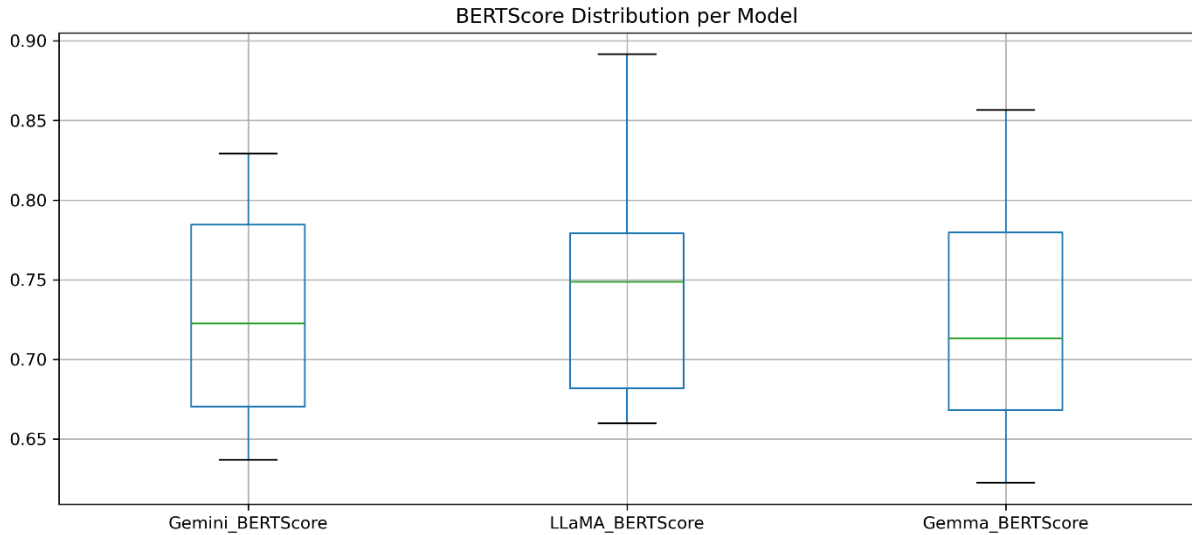
Η συνδυαστική ανάλυση ποιοτικών και ποσοτικών αποτελεσμάτων προσφέρει μια πιο ολοκληρωμένη εικόνα της απόδοσης των μοντέλων. Έτσι, δεν αξιολογείται μόνο η ικανότητα τους να παράγουν «ορθολογικά» κείμενα, αλλά και το κατά πόσο οι απαντήσεις τους είναι συμβατές με τις πραγματικές πληροφορίες που περιέχονται στα έγγραφα.

Η παρουσίαση των αποτελεσμάτων οργανώνεται ως εξής: αρχικά αναλύονται οι μετρικές BERTScore, BLEU και ROUGE, στη συνέχεια συζητούνται τα πλεονεκτήματα και οι αδυναμίες κάθε μοντέλου, και τέλος παρατίθεται μία συνολική σύγκριση.

4.2 BERTScore

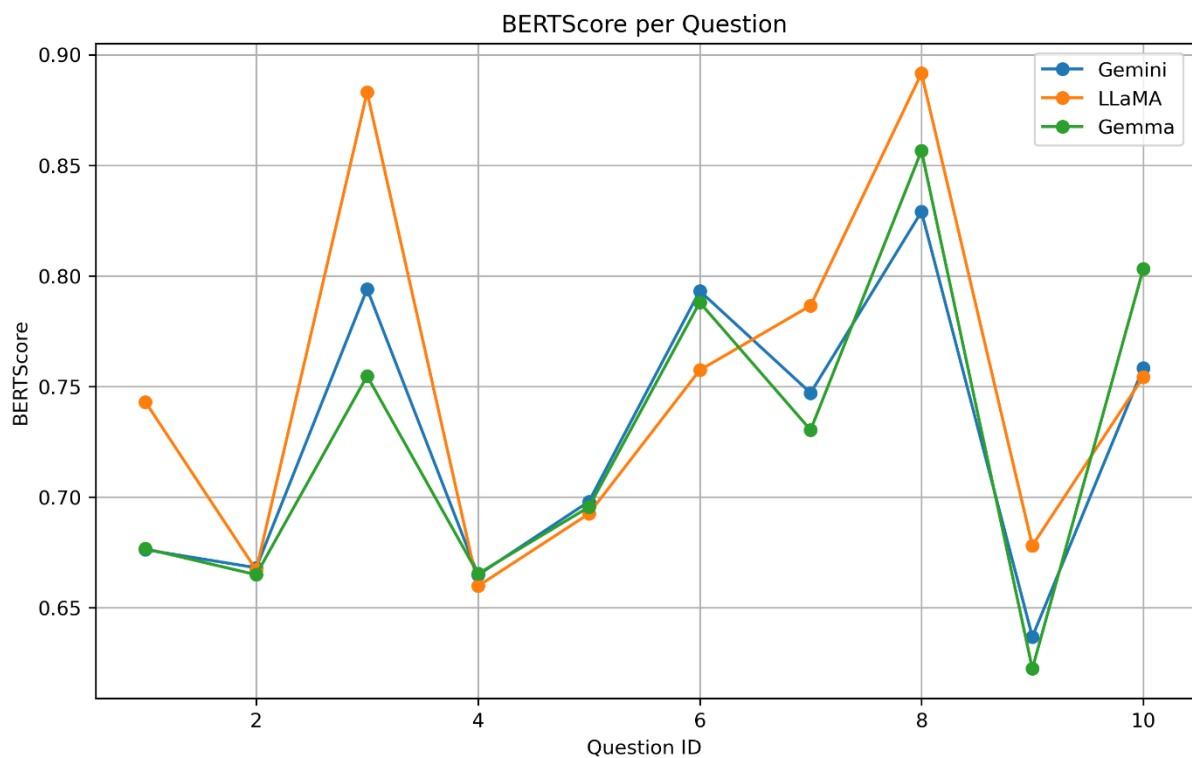
Ο δείκτης BERTScore χρησιμοποιείται ευρέως για την αξιολόγηση παραγόμενου κειμένου, καθώς βασίζεται στη σημασιολογική εγγύτητα λέξεων και φράσεων αντί στην απλή αντιστοίχιση χαρακτήρων. Στην παρούσα εργασία εφαρμόστηκε για να μετρηθεί η ομοιότητα ανάμεσα στις απαντήσεις των μοντέλων και στις σωστές απαντήσεις (ground truth) που περιέχονταν στο dataset αξιολόγησης.

Όπως φαίνεται στο Διάγραμμα 4.1 (boxplot), οι τιμές του BERTScore για όλα τα μοντέλα κινούνται σε παρόμοιο εύρος, από 0.63 έως 0.85 περίπου, με λίγες εξαιρέσεις που φτάνουν κοντά στο 0.90. Ο μέσος όρος του LLaMA βρίσκεται γύρω στο 0.75, ελαφρώς υψηλότερος από τον μέσο όρο του Gemini (~0.72) και του Gemma (~0.71). Το εύρημα αυτό δείχνει ότι το LLaMA, σε γενικές γραμμές, παράγαγε απαντήσεις με μεγαλύτερη σημασιολογική εγγύτητα σε σχέση με το ground truth.



Σχήμα 4.1: Κατανομή BERTScore ανά μοντέλο.

Στο Διάγραμμα 4.2 (BERTScore ανά ερώτηση), παρατηρείται ότι τα μοντέλα εμφανίζουν παρόμοια συμπεριφορά στις περισσότερες ερωτήσεις, αλλά με ορισμένες αξιοσημείωτες αποκλίσεις.



Σχήμα 4.2: BERTScore ανά ερώτηση.

Στην Ερώτηση 3, το LLaMA ξεχώρισε με πολύ υψηλή επίδοση (σχεδόν 0.89), σε αντίθεση με το Gemma που κινήθηκε χαμηλότερα (~0.75).

Στην Ερώτηση 8, και τα τρία μοντέλα είχαν υψηλές επιδόσεις, με το Gemma να φτάνει πάνω από 0.85.

Στην Ερώτηση 9, και τα τρία μοντέλα εμφάνισαν πτώση, με χαμηλές τιμές γύρω στο 0.63–0.67, γεγονός που δείχνει δυσκολία στην αναπαραγωγή της σωστής απάντησης.

Συνολικά, τα αποτελέσματα δείχνουν ότι:

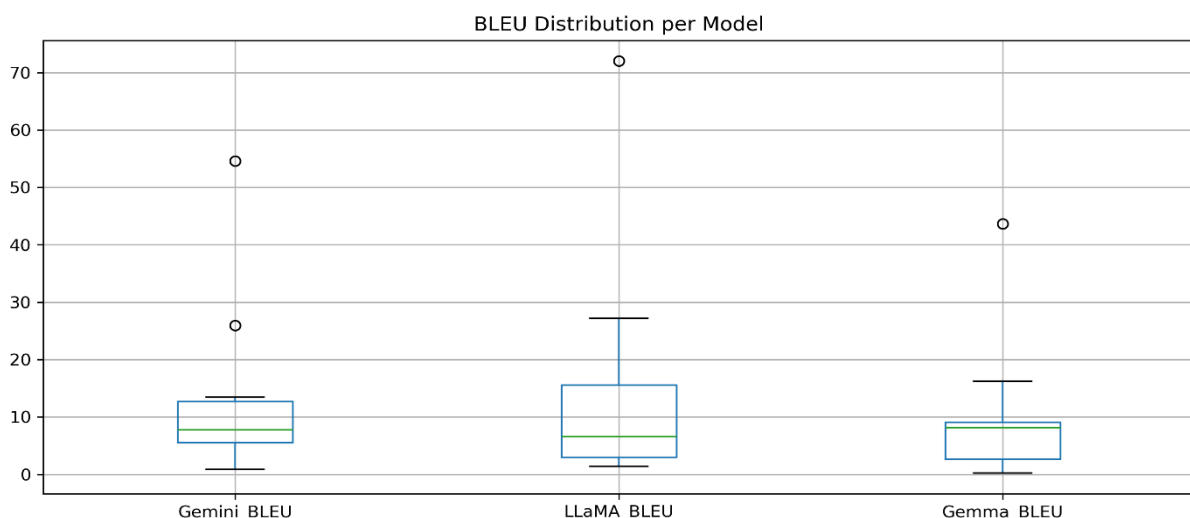
- Το LLaMA υπερέχει ελαφρώς σε μέσο όρο, παρουσιάζοντας πιο συνεπείς απαντήσεις με καλύτερη σημασιολογική συνέπεια.
- Το Gemini κινείται σε σταθερά επίπεδα, χωρίς ακραίες τιμές, γεγονός που δείχνει συνεκτικότητα αλλά και λιγότερες κορυφαίες επιδόσεις.
- Το Gemma εμφανίζει μεγαλύτερη μεταβλητότητα: σε κάποιες ερωτήσεις έχει πολύ καλή απόδοση (π.χ. Ερώτηση 8), ενώ σε άλλες χαμηλότερη από τα άλλα δύο μοντέλα.

Η ανάλυση του BERTScore δείχνει ότι όλα τα μοντέλα ήταν σε θέση να δώσουν απαντήσεις κοντά στο ground truth, με το LLaMA να ξεχωρίζει οριακά, αλλά χωρίς σημαντικές αποκλίσεις από τα άλλα δύο. Αυτό υποδηλώνει ότι, από πλευράς σημασιολογικής εγγύτητας, το σύστημα μπορεί να θεωρηθεί επαρκώς αξιόπιστο ανεξαρτήτως μοντέλου.

4.3 BLEU

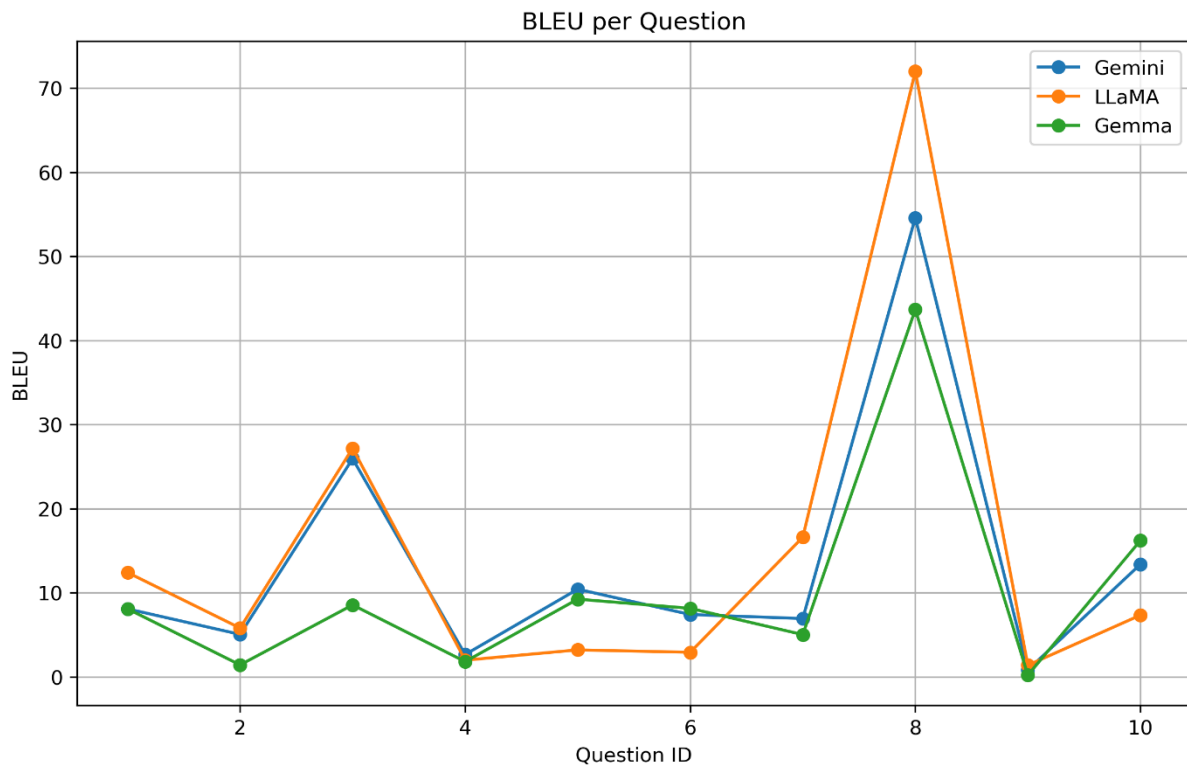
Ο δείκτης BLEU μετρά την ομοιότητα ανάμεσα σε παραγόμενα και αναμενόμενα κείμενα με βάση την ακριβή επικάλυψη λέξεων και φράσεων (n-grams). Σε αντίθεση με το BERTScore, ο BLEU είναι πιο «αστηρός», καθώς αξιολογεί εάν το μοντέλο χρησιμοποίησε τις ίδιες λέξεις με την απάντηση-στόχο.

Όπως φαίνεται στο Διάγραμμα 4.3 (boxplot), οι τιμές BLEU παρουσιάζουν μεγαλύτερη διασπορά σε σχέση με το BERTScore. Το LLaMA ξεχωρίζει επειδή εμφανίζει το μεγαλύτερο εύρος τιμών: σε ορισμένες περιπτώσεις καταγράφει πολύ υψηλές επιδόσεις (έως και 70+), ενώ σε άλλες πέφτει σε χαμηλότερα επίπεδα από Gemini και Gemma. Το Gemini κινείται σε πιο σταθερό εύρος (5–15 κατά μέσο όρο), ενώ το Gemma παρουσιάζει επίσης ενδιάμεσες τιμές με μικρότερες κορυφώσεις.



Σχήμα 4.3: Κατανομή BLEU ανά μοντέλο.

Στο Διάγραμμα 4.4 (BLEU ανά ερώτηση), παρατηρούνται ορισμένα ενδιαφέροντα μοτίβα.



Σχήμα 4.4: BLEU ανά ερώτηση.

Στο Διάγραμμα 4.4 (BLEU ανά ερώτηση), παρατηρούνται έντονες διαφοροποιήσεις:

- Στην Ερώτηση 3 και στην Ερώτηση 8, το LLaMA πέτυχε εξαιρετικά υψηλά σκορ, δείχνοντας ικανότητα να αναπαράγει σχεδόν αυτούσια την απάντηση του ground truth.
- Σε άλλες ερωτήσεις όμως (π.χ. Ερώτηση 4 και Ερώτηση 9), το BLEU του LLaMA έπεσε χαμηλότερα από Gemini και Gemma, υποδηλώνοντας δυσκολία στην ακριβή επιλογή λέξεων.
- Το Gemini έδωσε πιο συνεπή αποτελέσματα, χωρίς ακραίες αποκλίσεις, αλλά και χωρίς πολύ υψηλές κορυφές.
- Το Gemma κινήθηκε ενδιάμεσα, με καλές επιδόσεις σε ορισμένες ερωτήσεις αλλά χωρίς τη σταθερότητα του Gemini ή τις κορυφές του LLaMA.

Συνολικά, τα αποτελέσματα δείχνουν ότι:

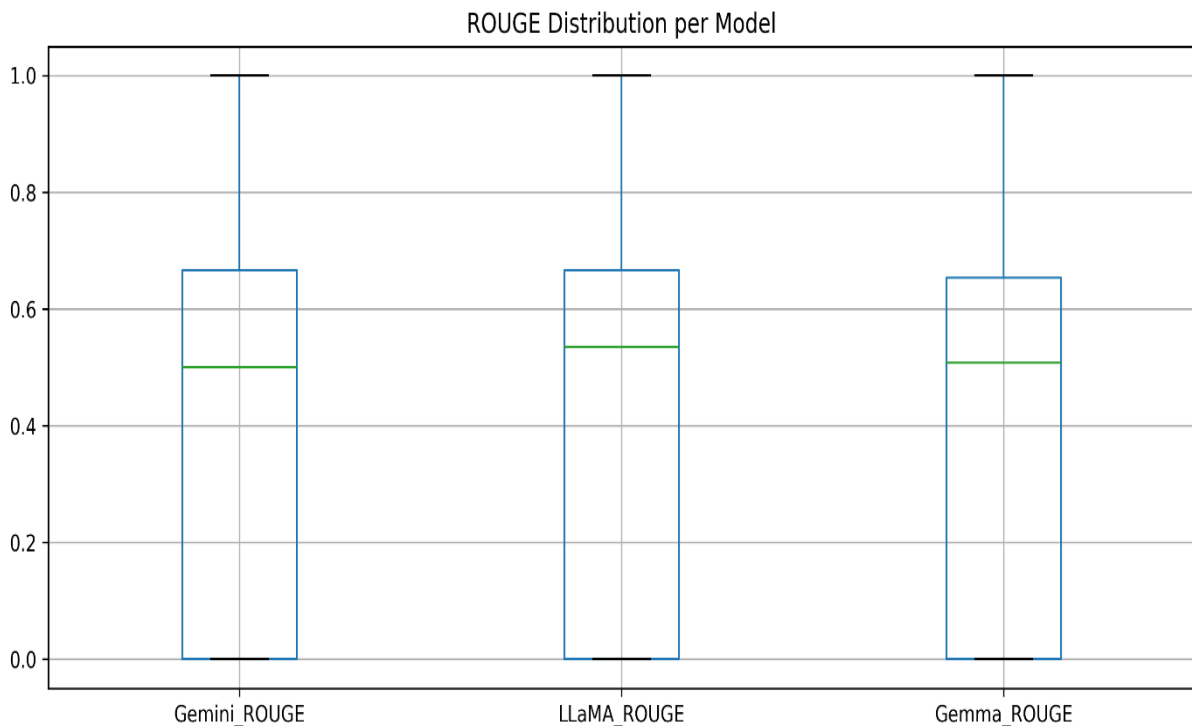
- Το LLaMA μπορεί να θεωρηθεί το πιο «ισχυρό» μοντέλο σε επίπεδο BLEU, καθώς σε ορισμένες περιπτώσεις αποδίδει με εξαιρετική ακρίβεια.
- Ωστόσο, η αστάθειά του μειώνει τη συνολική του αξιοπιστία.
- Το Gemini υπερτερεί σε συνεκτικότητα, με πιο σταθερές απαντήσεις, αν και με χαμηλότερη ακρίβεια.
- Το Gemma λειτουργεί ως ενδιάμεση περίπτωση, προσφέροντας ισορροπημένες αλλά λιγότερο εντυπωσιακές επιδόσεις.

Η εικόνα που προκύπτει είναι ότι το LLaMA ξεχωρίζει για τις κορυφαίες επιδόσεις του, αλλά η αστάθειά του το καθιστά λιγότερο προβλέσιμο σε σχέση με τα άλλα δύο μοντέλα, τα οποία προτιμώνται όταν απαιτείται μεγαλύτερη συνέπεια.

4.4 ROUGE

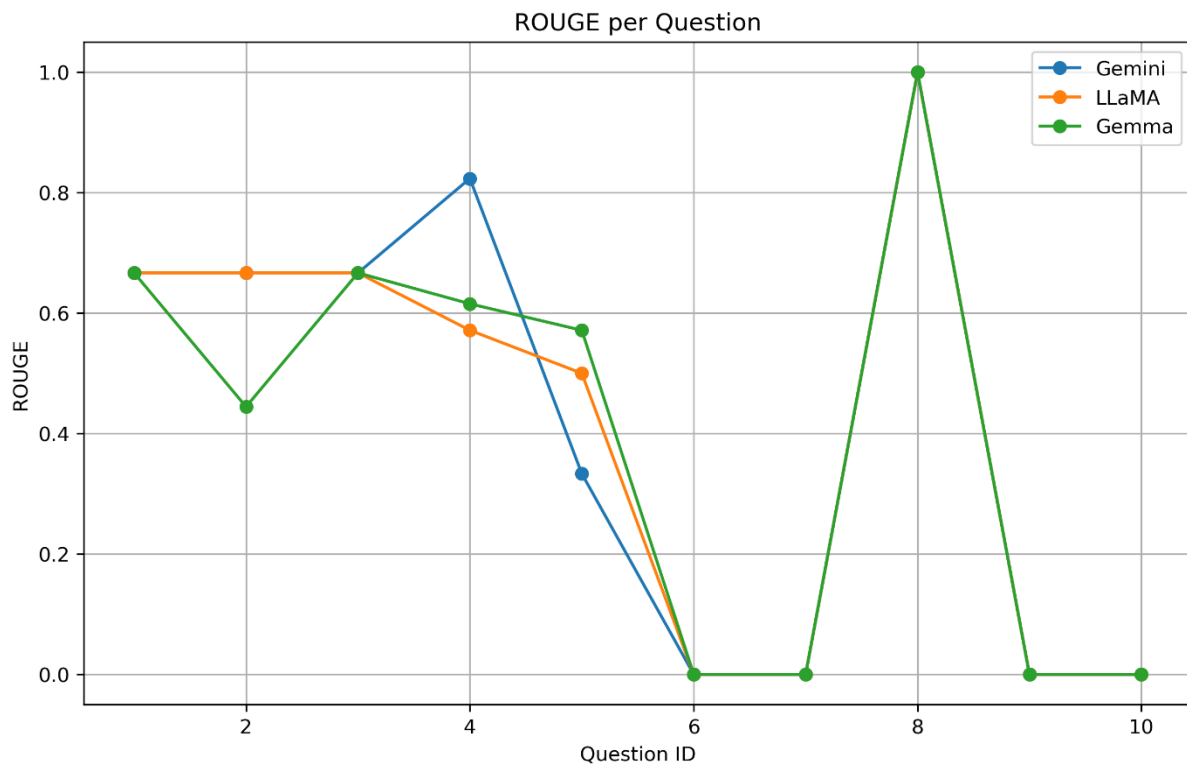
Ο δείκτης ROUGE (Recall-Oriented Understudy for Gisting Evaluation) χρησιμοποιείται για την αξιολόγηση της ποιότητας παραγόμενου κειμένου με βάση την επικάλυψη λέξεων, φράσεων και n-grams. Σε αντίθεση με τον BLEU, που είναι πιο «αυστηρός» και εστιάζει στην ακρίβεια, ο ROUGE δίνει έμφαση στην ανάκληση (recall), δηλαδή στο κατά πόσο η απάντηση του μοντέλου περιέχει τα κύρια στοιχεία της αναμενόμενης απάντησης, ακόμη κι αν δεν υπάρχει πλήρης ταύτιση λέξεων.

Όπως φαίνεται στο Διάγραμμα 4.5 (boxplot), οι τιμές ROUGE για όλα τα μοντέλα κινούνται σε παρόμοιο εύρος (0.45–0.65). Το Gemma παρουσιάζει μεγαλύτερη διασπορά, με υψηλές τιμές σε ορισμένες ερωτήσεις αλλά και πολύ χαμηλές σε άλλες, κάτι που δείχνει αστάθεια. Το Gemini και το LLaMA κινούνται πιο σταθερά γύρω από τη μέση τιμή (~0.55), χωρίς ακραίες αποκλίσεις.



Σχήμα 4.5: Κατανομή ROUGE ανά μοντέλο.

Στο Διάγραμμα 4.6 (ROUGE ανά ερώτηση) παρατηρούνται σημαντικές διαφοροποιήσεις.



Σχήμα 4.6: ROUGE ανά ερώτηση.

Όπως φαίνεται στο Διάγραμμα 4.6, τα τρία μοντέλα παρουσιάζουν διαφοροποιήσεις μέχρι την Ερώτηση 5, ενώ από την Ερώτηση 6 και μετά συγκλίνουν πλήρως.

- Στην Ερώτηση 8, και τα τρία μοντέλα κατέγραψαν την καλύτερη δυνατή επίδοση (ROUGE = 1.0), δείχνοντας ότι η απάντησή τους κάλυψε πλήρως το ground truth.
- Στις Ερωτήσεις 6, 7, 9 και 10, και τα τρία μοντέλα κατέγραψαν μηδενικές τιμές, γεγονός που δείχνει ότι δυσκολεύτηκαν να αναπαράγουν τα απαραίτητα στοιχεία της απάντησης.
- Στις πρώτες ερωτήσεις (1–5), παρατηρείται κάποια διαφοροποίηση: το Gemini και το LLaMA κινήθηκαν πιο σταθερά γύρω από τις μέσες τιμές (0.5–0.7), ενώ το Gemma εμφάνισε μεγαλύτερες διακυμάνσεις.

Συνολικά, τα αποτελέσματα του δείκτη ROUGE δείχνουν ότι:

- Τα τρία μοντέλα έχουν παρόμοια συνολική συμπεριφορά, ειδικά από την Ερώτηση 6 και μετά όπου οι τιμές τους ταυτίζονται.
- Το Gemini και το LLaMA παρουσιάζουν μεγαλύτερη συνέπεια στις πρώτες ερωτήσεις, ενώ το Gemma καταγράφει έντονες διακυμάνσεις.
- Το ROUGE = 1.0 στην Ερώτηση 8 για όλα τα μοντέλα δείχνει ότι το σύστημα μπορεί, υπό συγκεκριμένες συνθήκες, να αναπαράγει πλήρως τη σωστή απάντηση.

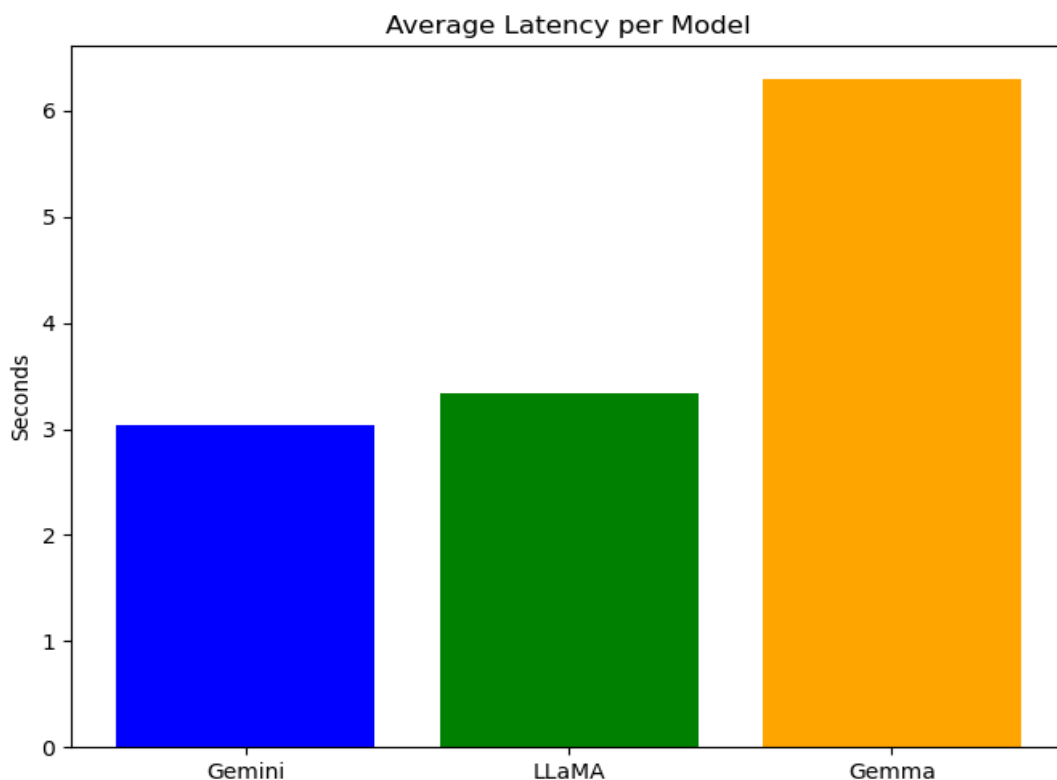
Η γενική εικόνα είναι ότι, όσον αφορά τον δείκτη ROUGE, τα τρία μοντέλα δεν εμφανίζουν μεγάλες διαφορές στη συνολική τους απόδοση. Το Gemini και το LLaMA τείνουν να είναι πιο σταθερά, ενώ το Gemma εμφανίζει μεγαλύτερη μεταβλητότητα στις πρώτες ερωτήσεις, αλλά συγκλίνει πλήρως με τα άλλα δύο στο δεύτερο μισό του dataset.

4.5 Συνολική Σύγκριση

Τα τρία μοντέλα (Gemini, LLaMA και Gemma) έχουν διαφορετικά δυνατά και αδύνατα σημεία, αλλά παρόμοια συνολική απόδοση, σύμφωνα με την ανάλυση των μετρικών BERTScore, BLEU και ROUGE. Σε ορισμένα ερωτήματα, το LLaMA έλαβε τις υψηλότερες βαθμολογίες (ειδικά στη μέτρηση BLEU), αποδεικνύοντας ότι μπορούσε να αναπαράγει συγκεκριμένες απαντήσεις με ακρίβεια. Ωστόσο, απέδωσε ασυνεπώς, καθώς έλαβε χαμηλότερες βαθμολογίες σε άλλες ερωτήσεις. Το Gemma επέδειξε μεγαλύτερη μεταβλητότητα, παρουσιάζοντας περιστασιακά εξαιρετική απόδοση (όπως ROUGE=1.0 στην Ερώτηση 8) και περιπτώσεις ολικής αποτυχίας, ενώ το Gemini ξεχώρισε για τη σταθερότητά του, προσφέροντας συνεπείς απαντήσεις χωρίς αξιοσημείωτες διακυμάνσεις.

Μετρικές συστήματος, οι οποίες είναι απαραίτητες για την εφαρμογή ενός LLM στον πραγματικό κόσμο, προστέθηκαν στην αξιολόγηση εκτός από τις γλωσσικές μετρικές.

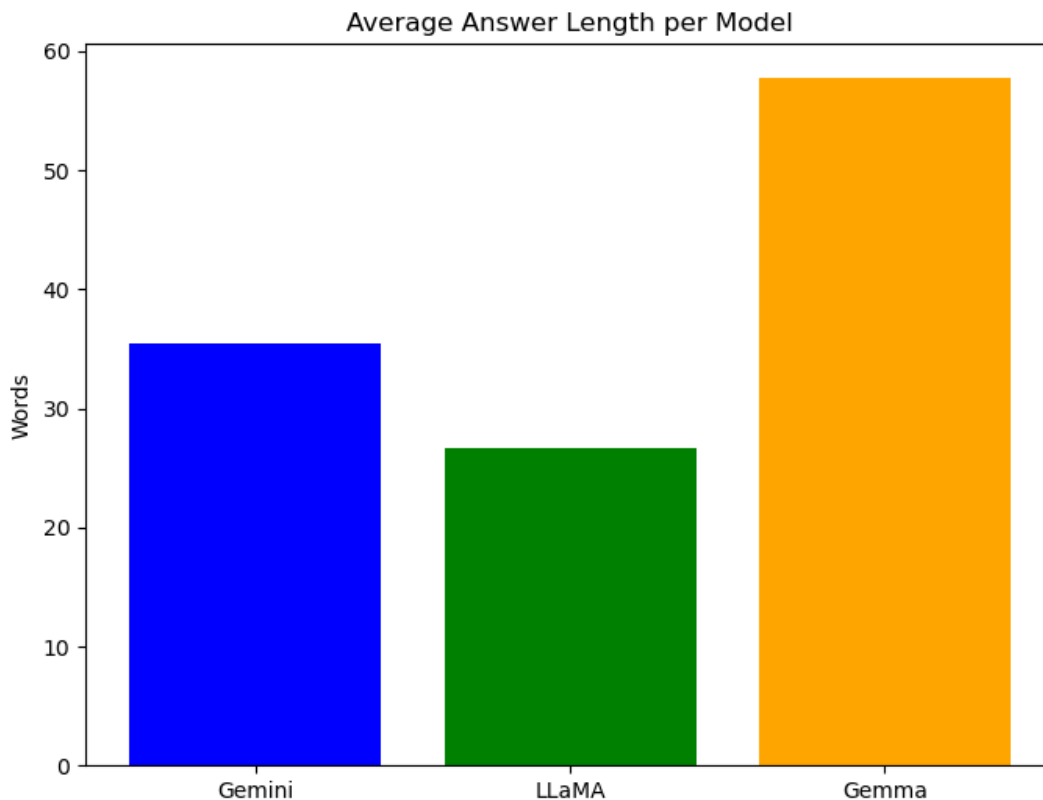
Με μέσο χρόνο απόκρισης περίπου 3 δευτερόλεπτα, το Gemini έδωσε τις ταχύτερες απαντήσεις, ακολουθούμενο από το LLaMA με ελαφρώς υψηλότερο latency (~3,3 δευτερόλεπτα) και το Gemma με σημαντικά υψηλότερο latency (πάνω από 6 δευτερόλεπτα κατά μέσο όρο), όπως φαίνεται στο Διάγραμμα 4.7. Σύμφωνα με αυτή την έρευνα, το Gemini αποδίδει καλύτερα σε πραγματικό χρόνο (όπως στην αλληλεπίδραση με τον χρήστη), ενώ το Gemma μπορεί να προκαλέσει καθυστερήσεις που υποβαθμίζουν την εμπειρία του χρήστη. Στη μέση, το LLaMA έχει αξιοπρεπείς χρόνους αντίδρασης, αλλά του λείπει το πλεονέκτημα ταχύτητας του Gemini.



Σχήμα 4.7: Συγκριτικός χρόνος απόκρισης ανά μοντέλο.

Αντίστοιχα, στο Διάγραμμα 4.8 παρουσιάζεται το μέσο μήκος απαντήσεων ανά μοντέλο. Το Gemma ξεχώρισε ως το πιο «αναλυτικό», με απαντήσεις που έφτασαν κατά μέσο όρο τις 58 λέξεις. Το Gemini κινήθηκε σε ενδιάμεσο επίπεδο (~36 λέξεις), παρέχοντας πιο ισορροπημένες απαντήσεις, ενώ το LLaMA παράγαγε τις πιο σύντομες, με μέσο μήκος περίπου 27 λέξεις. Αυτή η διαφοροποίηση δείχνει τη στρατηγική κάθε μοντέλου:

- Το Gemini στοχεύει σε περιεκτικότητα και ισορροπία,
- Το LLaMA δίνει πιο συνοπτικές και «κοφτές» απαντήσεις,
- Το Gemma τείνει να επεκτείνει το κείμενο, προσφέροντας περισσότερες λεπτομέρειες αλλά με το κόστος της ταχύτητας.



Σχήμα 4.8: Συγκριτικό μήκος απαντήσεων ανά μοντέλο.

Η διαφορά ανάμεσα σε latency και μήκος απαντήσεων αναδεικνύει έναν σημαντικό συμβιβασμό (trade-off). Τα πιο σύντομα μοντέλα (όπως το Gemini και σε μικρότερο βαθμό το LLaMA) είναι ταχύτερα και συνεπέστερα, ενώ τα πιο αναλυτικά (Gemma) είναι πιο αργά αλλά προσφέρουν λεπτομερέστερες πληροφορίες. Η επιλογή του «καταλληλότερου» μοντέλου εξαρτάται επομένως από τις προτεραιότητες κάθε εφαρμογής: αν ζητείται ταχύτητα και συνέπεια, προτιμάται το Gemini· αν ζητείται μέγιστη λεπτομέρεια, τότε το Gemma μπορεί να είναι πιο χρήσιμο, παρότι πιο αργό· ενώ το LLaMA προσφέρει έναν ενδιάμεσο συνδυασμό.

Τέλος, πρέπει να σημειωθεί ότι οι μετρικές BLEU και ROUGE έχουν περιορισμούς, καθώς βασίζονται σε ακριβή αντιστοίχιση λέξεων. Παρατηρήθηκε ότι απαντήσεις με παρόμοιο νόημα, αλλά διαφορετική διατύπωση, λάμβαναν χαμηλές ή μηδενικές τιμές. Αντίθετα, το BERTScore, που μετρά σημασιολογική ομοιότητα, κατέγραψε πιο ρεαλιστικά τη συνάφεια αυτών των απαντήσεων με το ground truth. Για τον

Κεφάλαιο 4

λόγο αυτό, η αξιολόγηση πρέπει να γίνεται συνδυαστικά, με διαφορετικές μετρικές και όχι μονοδιάστατα.

Συνοψίζοντας:

- Το LLaMA είναι το πιο «ασχυρό» σε επίπεδο ακρίβειας (BLEU), αλλά με σημαντική αστάθεια.
- Το Gemini είναι το πιο σταθερό και ταχύτερο, με καλή ισορροπία ποιότητας–ταχύτητας.
- Το Gemma είναι το πιο αναλυτικό αλλά και το πιο αργό, με μεγάλη μεταβλητότητα.

Κεφάλαιο 5ο: Συμπεράσματα και Μελλοντική Εργασία

5.1 Συμπεράσματα

Η παρούσα εργασία είχε ως στόχο την ανάπτυξη και αξιολόγηση ενός συστήματος ερωταπαντήσεων (RAG -Retrieval-Augmented Generation) βασισμένου σε τρία διαφορετικά μεγάλα γλωσσικά μοντέλα: το Gemini, το LLaMA και το Gemma. Το σύστημα σχεδιάστηκε ώστε να αξιοποιεί πραγματικά δεδομένα από κανονισμούς σπουδών πανεπιστημιακών σχολών, παρέχοντας απαντήσεις που δεν προέρχονται αποκλειστικά από τη γενική γνώση του εκάστοτε μοντέλου, αλλά ενισχύονται από συγκεκριμένα αποσπάσματα κειμένων.

Από την ανάλυση των αποτελεσμάτων προέκυψε ότι τα τρία μοντέλα παρουσίασαν συγκρίσιμη συνολική απόδοση, με διαφοροποιήσεις σε επιμέρους πτυχές. Σε όρους σημασιολογικής ποιότητας, όπως μετράται με το BERTScore, όλα τα μοντέλα κατέγραψαν παρόμοιο βαθμό συνάφειας με το ground truth, με το LLaMA να σημειώνει ελαφρώς υψηλότερο μέσο όρο. Στον δείκτη BLEU, το LLaMA εμφάνισε ιδιαίτερα υψηλές τιμές σε ορισμένες ερωτήσεις, γεγονός που αναδεικνύει την ικανότητά του να αναπαράγει με ακρίβεια την αναμενόμενη απάντηση, αλλά η επίδοσή του χαρακτηρίστηκε από αστάθεια. Ο δείκτης ROUGE ανέδειξε το Gemma ως ικανό να επιτυγχάνει κορυφαία αποτελέσματα (π.χ. στην Ερώτηση 8 με πλήρη κάλυψη), αν και η συνολική του απόδοση παρουσίασε έντονη μεταβλητότητα. Το Gemini και το LLaMA έδειξαν πιο σταθερή συμπεριφορά στον ίδιο δείκτη, χωρίς ακραίες διακυμάνσεις.

Όσον αφορά τις μετρικές συστήματος, το Gemini αναδείχθηκε ως το ταχύτερο μοντέλο, με μέσο χρόνο απόκρισης περίπου τρία δευτερόλεπτα. Το LLaMA ακολούθησε με 3,3 δευτερόλεπτα, ενώ το Gemma παρουσίασε μεγαλύτερη καθυστέρηση, που ξεπέρασε τα έξι δευτερόλεπτα. Σε επίπεδο μήκους απαντήσεων, το Gemma παρήγαγε τα πιο εκτενή κείμενα (περίπου 58 λέξεις κατά μέσο όρο), το Gemini κινήθηκε σε ενδιάμεσο επίπεδο (36 λέξεις), ενώ το LLaMA παρήγαγε τις πιο σύντομες απαντήσεις (27 λέξεις). Τα ευρήματα αυτά αναδεικνύουν διαφορετικές στρατηγικές παραγωγής: το LLaMA προτιμά συνοπτικές και άμεσες διατυπώσεις, το Gemini προσφέρει ισορροπία ανάμεσα στη σαφήνεια και την ανάλυση, ενώ το Gemma δίνει έμφαση στη λεπτομέρεια και την εκτενή περιγραφή.

Συνολικά, δεν μπορεί να υποστηριχθεί ότι κάποιο μοντέλο υπερτερεί απόλυτα σε όλες τις κατηγορίες. Το LLaMA υπερέχει σε ακρίβεια αλλά με αστάθεια, το Gemini διακρίνεται για τη σταθερότητα και την ταχύτητά του, ενώ το Gemma ξεχωρίζει για την αναλυτικότητα με σημαντικό κόστος στον χρόνο απόκρισης. Η επιλογή του καταλληλότερου μοντέλου εξαρτάται από τις απαιτήσεις της εκάστοτε εφαρμογής: εάν ζητείται ταχύτητα και συνέπεια, το Gemini αποτελεί την καλύτερη λύση· εάν ζητείται ακρίβεια σε συγκεκριμένα ερωτήματα, το LLaMA έχει πλεονέκτημα ενώ όταν προτεραιότητα είναι η λεπτομερής απάντηση, το Gemma μπορεί να θεωρηθεί πιο χρήσιμο.

Τέλος, τα αποτελέσματα ανέδειξαν ουσιαστικές διαφορές μεταξύ των μοντέλων που μπορούν να αποδοθούν σε σχεδιαστικές και αρχιτεκτονικές επιλογές. Το Gemini υπερέχει σε ταχύτητα, πιθανώς λόγω βελτιστοποίησης για real-time εφαρμογές. Το LLaMA παρήγαγε εκτενέστερες και πιο περιγραφικές απαντήσεις, στοιχείο που συνδέεται με την εκπαίδευσή του σε μεγάλα και ποικίλα σύνολα δεδομένων. Το Gemma, αν και αργότερο, παρείχε απαντήσεις πιο ισορροπημένες σε περιεχόμενο και ακρίβεια. Παράλληλα, ο περιορισμένος όγκος δεδομένων και η χρήση ελεύθερων εκδόσεων API

αποτελούν παράγοντες που περιορίζουν τη γενίκευση των ευρημάτων, στοιχείο που καθιστά αναγκαία τη διεύρυνση της μελέτης σε μελλοντική εργασία.

5.2 Μελλοντική Εργασία

Η παρούσα εργασία ανέδειξε τόσο τις δυνατότητες όσο και τους περιορισμούς της προσέγγισης RAG με LLMs, ανοίγοντας τον δρόμο για περαιτέρω μελέτη. Σε μελλοντικό στάδιο, η εργασία μπορεί να επεκταθεί με την ενσωμάτωση μεγαλύτερου όγκου και ποικιλίας δεδομένων, όπως πρόσθετους κανονισμούς, πρακτικά συνεδριάσεων ή διοικητικές εγκυκλίους, ώστε να διευρυνθεί το πεδίο εφαρμογής και να ελεγχθεί η ανθεκτικότητα των μοντέλων σε πιο σύνθετα κείμενα. Παράλληλα, η διαδικασία ανάκτησης θα μπορούσε να βελτιωθεί μέσω πιο εξελιγμένων τεχνικών, όπως η μέθοδος Maximal Marginal Relevance ή υβριδικές προσεγγίσεις που συνδυάζουν dense και sparse retrieval, με στόχο τη βελτίωση της ακρίβειας και της συνάφειας.

Η συγκριτική αξιολόγηση δύναται να εμπλουτιστεί με την ένταξη νέων μοντέλων (π.χ. GPT, Claude), ώστε να εξεταστεί η συμπεριφορά τους στο ίδιο πλαίσιο, καθώς και με τη χρήση επιπλέον μετρικών αξιολόγησης, όπως ο δείκτης METEOR ή πιο σύγχρονες σημασιολογικές μετρικές που λαμβάνουν υπόψη παραφράσεις. Αυτό θα επέτρεπε την εξαγωγή πιο σφαιρικών και αξιόπιστων συμπερασμάτων.

Μία ακόμη σημαντική κατεύθυνση αφορά τη βελτίωση της διεπαφής χρήστη. Η ανάπτυξη πιο ολοκληρωμένων περιβαλλόντων αλληλεπίδρασης, που θα υποστηρίζουν ιστορικό ερωτήσεων, δυνατότητα φιλτραρίσματος πηγών και στοχευμένη αναζήτηση σε συγκεκριμένα έγγραφα, θα ενίσχυε τη χρηστικότητα του συστήματος.

Εξίσου κρίσιμη είναι η διερεύνηση της αξιοποίησης του συστήματος σε εκπαιδευτικά πλαίσια. Η δυνατότητα των LLMs να απαντούν σε ερωτήσεις πάνω σε κανονιστικά κείμενα μπορεί να προσφέρει στους φοιτητές άμεση και στοχευμένη πρόσβαση σε κανονισμούς σπουδών, μειώνοντας τον χρόνο αναζήτησης πληροφοριών. Παράλληλα, οι διδάσκοντες θα μπορούσαν να χρησιμοποιήσουν τέτοια εργαλεία για να καθοδηγούν τους φοιτητές σε πολύπλοκα διοικητικά ή θεσμικά ζητήματα, αλλά και ως βάση για την ανάπτυξη σεναρίων μάθησης που αξιοποιούν πραγματικά έγγραφα. Η πιλοτική εφαρμογή σε πανεπιστημιακές τάξεις θα έδινε πολύτιμα ποιοτικά δεδομένα σχετικά με τη χρησιμότητα, τη φιλικότητα και την αποδοχή του συστήματος από τους τελικούς χρήστες.

Μια τέτοια μελέτη θα μπορούσε να αποδώσει πολύτιμα ποιοτικά στοιχεία, τα οποία, σε συνδυασμό με τις αυτόματες μετρικές, θα οδηγούσαν σε πιο ολοκληρωμένη κατανόηση της αξίας, των προοπτικών αλλά και των περιορισμών του συστήματος, τόσο σε ερευνητικό όσο και σε εκπαιδευτικό επίπεδο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Journal Articles

- [1] Mishra, Akanksha. (2024). A Comprehensive Review of Artificial Intelligence and Machine Learning: Concepts, Trends, and Applications. *International Journal of Scientific Research in Science and Technology*. 11. 126-142. 10.32628/IJSRST2411587.
- [2] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M.A., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey. *ArXiv*, abs/2402.06196.
- [4] McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1956). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.
- [5] Korf RE. Does Deep Blue use Artificial Intelligence?1. *ICGA Journal*. 1997;20(4):243-245. doi:10.3233/ICG-1997-20404.
- [6] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160. Springer.
- [7] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117.
- [8] Godbole, Aditi. (2025). Leveraging Long-Context Large Language Models for Multi-Document Understanding and Summarization in Enterprise Applications. 10.48550/arXiv.2409.18454.
- [9] Chen, Zhiyu & Ma, Jing & Zhang, Xinlu & Hao, Nan & Yan, An & Nourbakhsh, Armineh & Yang, Xianjun & McAuley, Julian & Petzold, Linda & Wang, William. (2024). A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law.
- [10] Hsain, Achraf & El Housni, Hamza. (2024). LARGE LANGUAGE MODEL-POWERED CHATBOTS FOR INTERNATIONALIZING STUDENT SUPPORT IN HIGHER EDUCATION. 10.48550/arXiv.2403.14702.
- [11] Liu, Bin. (2021). "Weak AI" is Likely to Never Become "Strong AI", So What is its Greatest Value for us?. 10.48550/arXiv.2103.15294.
- [12] Olabiyi, Winner & Akinyele, Docas & Joel, Emmanuel. (2025). The Evolution of AI: From Rule-Based Systems to Data-Driven Intelligence.
- [13] Antoska Knights, Vesna & Prchkovska, Marija. (2024). From Equations to Predictions: Understanding the Mathematics and Machine Learning of Multiple Linear Regression. *Journal of Mathematical & Computer Applications*. 3. 1-8. 10.47363/JMCA/2024(3)137.
- [14] Jadon, Aryan & Patil, Avinash & Jadon, Shruti. (2022). A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting. 10.48550/arXiv.2211.02989.
- [15] Rathod, Yogesh. (2024). A Survey of Machine Learning Techniques for Artificial Intelligence. *International Journal of Computer Techniques*. 11.
- [16] Schmidhuber, Juergen. (2014). Deep Learning in Neural Networks: An Overview. *Neural Networks*. 61. 10.1016/j.neunet.2014.09.003.

- [17] Dubey, Shiv Ram & Singh, Satish Kumar & Chaudhuri, Bidyut. (2022). Activation Functions in Deep Learning: A comprehensive Survey and Benchmark. *Neurocomputing*. 503. 10.1016/j.neucom.2022.06.111.
- [18] Lourdasamy, Ravi & Abraham, Stanislaus. (2018). A Survey on Text Pre-processing Techniques and Tools. *International Journal of Computer Sciences and Engineering*. 06. 148-157. 10.26438/ijcse/v6si3.148157.
- [19] Torfi, Amirsina & Shirvani, Rouzbeh & Keneshloo, Yaser & Tavvaf, Nader & Fox, Edward. (2020). Natural Language Processing Advancements by Deep Learning: A Survey. 10.48550/arXiv.2003.01200.
- [20] Hadi, Muhammad Usman & Al-Tashi, Qasem & Qureshi, Rizwan & Shah, Abbas & Muneer, Amgad & Irfan, Muhammad & Zafar, Anas & Shaikh, Muhammad & Akhtar, Naveed & Wu, Jia & Mirjalili, Seyedali. (2023). Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. 10.36227/techrxiv.23589741.v1.
- [21] Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need. 10.48550/arXiv.1706.03762.
- [22] Lewis, Patrick & Perez, Ethan & Piktus, Aleksandara & Petroni, Fabio & Karpukhin, Vladimir & Goyal, Naman & Küttler, Heinrich & Lewis, Mike & Yih, Wen-tau & Rocktäschel, Tim & Riedel, Sebastian & Kiela, Douwe. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 10.48550/arXiv.2005.11401.
- [23] Bommasani, Rishi & Liang, Percy & Lee, Tony. (2023). Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*. 1525. 10.1111/nyas.15007.
- [24] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.
- [25] Topsakal, Oguzhan & Akinci, T. Cetin. (2023). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. *International Conference on Applied Engineering and Natural Sciences*. 1. 1050-1056. 10.59287/icaens.1127.
- [26] Liang, Percy & Bommasani, Rishi & Lee, Tony & Tsipras, Dimitris & Soylu, Dilara & Yasunaga, Michihiro & Zhang, Yian & Narayanan, Deepak & Wu, Yuhuai & Kumar, Ananya & Newman, Benjamin & Yuan, Binhang & Yan, Bobby & Zhang, Ce & Cosgrove, Christian & Manning, Christopher & Ré, Christopher & Acosta-Navas, Diana & Hudson, Drew & Koreeda, Yuta. (2022). Holistic Evaluation of Language Models. 10.48550/arXiv.2211.09110.

Βιβλία

- [3] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 49, 433–460.

ΠΑΡΑΡΤΗΜΑ Α : Βασικός Κώδικας Python

Στο παρόν παράρτημα παρατίθεται ο κώδικας Python που αναπτύχθηκε για την υλοποίηση του συστήματος. Ο κώδικας αξιοποιεί τις βιβλιοθήκες LangChain και ChromaDB, καθώς και τα API των μεγάλων γλωσσικών μοντέλων LLMs (Gemini, LLaMA, Gemma), προκειμένου να πραγματοποιείται η ανάγνωση εγγράφων PDF, η δημιουργία βάσης διανυσμάτων (vector database), η ανάκτηση σχετικών αποσπασμάτων και η παρουσίαση συγκριτικών απαντήσεων μέσω περιβάλλοντος Gradio.

```
import os, time, requests, pdfplumber, shutil, gc
import pandas as pd
import chromadb
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain_community.vectorstores import Chroma
from langchain.docstore.document import Document
from langchain_google_genai import GoogleGenerativeAIEmbeddings
import gradio as gr

#Φόρτωση .env αρχείου
load_dotenv()
# Ρυθμίσεις φακέλων και API keys
PDF_FOLDER = "./documents"
VECTOR_STORE_DIR = "chroma_db"
GOOGLE_API_KEY = os.getenv("GOOGLE_API_KEY")
OPENROUTER_API_KEY = os.getenv("OPENROUTER_API_KEY ")

# Ορισμός embedding μοντέλου και LLMs
EMBEDDING_MODEL = GoogleGenerativeAIEmbeddings(model="models/gemini-embedding-
001")
MODEL_GEMINI = "google/gemini-2.0-flash-exp:free"
MODEL_LLAMA = "meta-llama/llama-3.3-70b-instruct"
MODEL_GEMMA = "google/gemma-3-27b-it:free"

# Παράμετροι chunking κειμένου
CHUNK_SIZE = 1000
CHUNK_OVERLAP = 150

# Συνάρτηση ανάγνωσης PDF και εξαγωγής κειμένου
def read_pdf_pages(file_path):
    pages = []
    with pdfplumber.open(file_path) as pdf:
        for i, page in enumerate(pdf.pages, start=1):
            t = page.extract_text() or ""
            if t.strip():
                pages.append({"page": i, "text": t})
    return pages

# Συνάρτηση διαχωρισμού κειμένου σε κομμάτια (chunks)
def chunk_text_recursive(text, chunk_size=CHUNK_SIZE, chunk_overlap=CHUNK_OVERLAP):
    splitter = RecursiveCharacterTextSplitter(
        chunk_size=chunk_size, chunk_overlap=chunk_overlap,
        separators=["\n\n", "\n", ".", " ", ""]
    )
    return splitter.split_text(text)
```

```

# Δημιουργία ή φόρτωση βάσης δεδομένων vector embeddings
def build_vector_db():
    if os.path.exists(VECTOR_STORE_DIR):
        return Chroma(persist_directory=VECTOR_STORE_DIR,
embedding_function=EMBEDDING_MODEL)

    docs = []
    pdf_files = [f for f in os.listdir(PDF_FOLDER) if f.lower().endswith(".pdf")]
    for pdf_file in pdf_files:
        pdf_path = os.path.join(PDF_FOLDER, pdf_file)
        pages = read_pdf_pages(pdf_path)
        for p in pages:
            chunks = chunk_text_recursive(p["text"])
            for idx, c in enumerate(chunks):
                docs.append(Document(page_content=c, metadata={"source": pdf_file, "page": p["page"], "chunk":
idx}))
    vectordb = Chroma.from_documents(docs, embedding=EMBEDDING_MODEL,
persist_directory=VECTOR_STORE_DIR)
    vectordb.persist()
    return vectordb

# Συναρτήσεις για ερωτήματα στα LLMs μέσω OpenRouter
def query_model(prompt, model):
    url = "https://openrouter.ai/api/v1/chat/completions"
    headers = {"Authorization": f"Bearer {OPENROUTER_API_KEY}", "Content-Type":
"application/json"}
    data = {"model": model, "messages": [{"role": "user", "content": prompt}], "temperature": 0.7}
    start = time.time()
    resp = requests.post(url, headers=headers, json=data)
    end = time.time()
    try: answer = resp.json()["choices"][0]["message"]["content"]
except: answer = "Σφάλμα API"
    return answer, round(end - start, 2)

# Ανάκτηση context και εκτέλεση queries σε όλα τα μοντέλα
def query_all_models(vectordb, question, k=10):
    results = vectordb.similarity_search(question, k=k)
    context = "\n\n---\n\n".join([r.page_content for r in results])
    sources = [f"- {r.metadata['source']} σελ. {r.metadata['page']}" for r in results]
    prompt = f"Ερώτηση: {question}\n\nΑποσπάσματα: \n{context}\n\nΑπάντησε (στα ελληνικά):"

    g_ans, g_t = query_model(prompt, MODEL_GEMINI)
    l_ans, l_t = query_model(prompt, MODEL_LLAMA)
    gm_ans, gm_t = query_model(prompt, MODEL_GEMMA)

    return {
        "answers": {"gemini": g_ans, "llama": l_ans, "gemma": gm_ans},
        "latencies": {"gemini": g_t, "llama": l_t, "gemma": gm_t},
        "sources": "\n".join(sources)
    }

# GradioUI για συγκριτική παρουσίαση απαντήσεων
def launch_gradio():
    with gr.Blocks() as demo:

```

```

gr.Markdown("## LLM Comparison Chatbot")
query_input = gr.Textbox(label="Ερώτηση:", lines=2)
output_gemini = gr.Textbox(label="Gemini", lines=12)
output_llama = gr.Textbox(label="LLaMA", lines=12)
output_gemma = gr.Textbox(label="Gemma", lines=12)
metrics_box = gr.JSON(label="Μετρικές")
sources_box = gr.Textbox(label="Πηγές", lines=3)
submit_btn = gr.Button("Αποστολή")
clear_btn = gr.Button("Καθαρισμός")

def run_all_models(question):
    result = query_all_models(vectordb, question)
    return result["answers"]["gemini"], result["answers"]["llama"], result["answers"]["gemma"],
    result["latencies"], result["sources"]

submit_btn.click(fn=run_all_models, inputs=[query_input],
                outputs=[output_gemini, output_llama, output_gemma, metrics_box, sources_box])
clear_btn.click(fn=lambda: ("", "", "", {}, ""), inputs=[],
               outputs=[output_gemini, output_llama, output_gemma, metrics_box, sources_box])

demo.launch(inbrowser=True)

# Εκκίνηση συστήματος
vectordb = build_vector_db()
launch_gradio()

```