ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Καθοδηγούμενη απο Τεκμήρια Επέκταση Ερωτημάτων για Μείωση Παραισθήσεων σε Συστήματα Ερωταπαντήσεων με Μεγάλα Γλωσσικά Μοντέλα

Διπλωματικη Εργασια

του

ΒΑΜΒΑΚΑ ΔΗΜΗΤΡΙΟΥ

Επιβλέποντες: Χρήστος Μακρής
Κώστας Τσίχλας
Σπύρος Σιούτας

Πάτρα, Νοέμβριος 2025

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Καθοδηγούμενη απο Τεκμήρια Επέκταση Ερωτημάτων για Μείωση Παραισθήσεων σε Συστήματα Ερωταπαντήσεων με Μεγάλα Γλωσσικά Μοντέλα

Διπλωματικη Εργασια

του

ΒΑΜΒΑΚΑ ΔΗΜΗΤΡΙΟΥ

Επιβλέποντες: Χρήστος Μακρής
Κώστας Τσίχλας
Σπύρος Σιούτας

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την .

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Χρήστος Μακρής	Κώστας Τσίχλας	Σπύρος Σιούτας
Αναπληρωτής καθηγητής	Αναπληρωτής Καθηγητής	Καθηγητής

Πάτρα, Νοέμβριος 2025

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Υπεύθυνη Δήλωση

Βεβαιώνω ότι είμαι συγγραφέας αυτής της πτυχιακής εργασίας, και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης, βεβαιώνω ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του τμήματος Μηχανικών Ηλεκτρωνικών Υπολογιστών και Πληρογορικής του Πανεπηστημίου Πατρών.

(Υπογραφή)

.....
Δημήτριος Βαμβακάς

Περίληψη

Η ραγδαία ανάπτυξη των Μεγάλων Γλωσσικών Μοντέλων συνοδεύεται από το πρόβλημα των παραισθήσεων, δηλαδή της παραγωγής απαντήσεων που δεν υποστηρίζονται από τα διαθέσιμα δεδομένα. Σε αυτή την διπλωματική εργασία προτείνεται μια πολυεπίπεδη μεθοδολογία για τη δημιουργία συνόλου ερωτήσεων–απαντήσεων με στόχο τη μείωση των παραισθήσεων αλλά και την καλύτερη ποιότητα ερωτήσεων, και την αναζήτηση των επικυρωμένων απαντήσεων τους σε μια συλλογή κειμένων. Η διαδικασία περιλαμβάνει πέντε στάδια: (α) αυτόματη παραγωγή αρχικών ερωτήσεων με βάση τα κείμενα μιας συλλογής κειμένων(κοινής θεματολογίας), (β) αναζήτηση σχετικών, με την κάθε ερώτηση, κειμένων με χρήση BM25, (γ) γείωση/αναδιατύπωση των ερωτήσεων με βάση τα ανακτημένα κείμενα/τεκμήρια και αρχική προσπάθεια επικυρωμένης απάντησης των απαντήσεων, (δ) περικοπή και επικύρωση των απαντήσεων ώστε να παραμένουν σύντομες και αποδεικτικά θεμελιωμένες, και (ε) μια τελική διαδικασία «διάσωσης» όπου επανεξετάζονται τα ερωτήματα που είχαν χαρακτηριστεί αναπάντητα στα προηγούμενα στάδια, για να μειωθεί ο αριθμός τους. Το αποτέλεσμα είναι ένα καθαρότερο σύνολο δεδομένων ερωτήσεων–απαντήσεων, όπου κάθε ερώτηση συνοδεύεται απο σύντομη, επιβεβαιωμένη με βάση κείμενα, απάντηση από ρητή αναφορά τεκμηρίων ή σημειώνεται ως μη απαντήσιμη. Η αξιολόγηση με αυτόματες μετρικές (ROUGE, BERTScore) και η χειροκίνητη αξιολόγηση δείχνουν ότι η προτεινόμενη μέθοδος πετυχαίνει υψηλή ακρίβεια και σημαντική μείωση ψευδών απαντήσεων, επιβεβαιώνοντας ότι η συγκεκριμένη διαδικασία, αποτελεί μια αποτελεσματική στρατηγική περιορισμού παραισθήσεων στα LLMs. Χρησιμοποιώντας τα αποτελέσματα της παραπάνω διαδικασίας, υλοποιήθηκε και μια εφαρμογή, η οποία εμφανίζει στον χρήστη τυχαίες ερωτήσεις απο το σύνολο, και αυτός καλείται να τις απαντήσει. Στο τέλος βαθμολογείται με βάση την ομοιότητα της απάντησης που έδωσε, σε σχέση με την απάντηση του μοντέλου. Αυτό λειτουργεί ως μια πρακτική εφαρμογή του pipeline που σχεδιάσαμε, εστιάζοντας στον τομέα της εκπαίδευσης, ο οποίος μπορεί να εποφεληθεί σε μεγάλο βαθμό απο την ανάπτυξη των τεχνολογιών Ανάκτησης Πληροφορίας και των Μεγάλων Γλωσσικών Μοντέλων.

Λέξεις Κλειδιά

Μεγάλα Γλωσσικά Μοντέλα, RAG, Επέκταση Ερωτημάτων, Ανάκτηση Πληροφορίας, Εκπαίδευση

Abstract

The rapid development of Large Language Models is accompanied by the problem of hallucinations, that is, the generation of answers that are not supported by the available data. In this thesis, a multi-level methodology is proposed for the creation of a set of question–answer pairs, with the aim of reducing hallucinations while also improving question quality, and retrieving validated answers from a text collection. The process includes five stages: (a) automatic generation of initial questions based on the texts of a (thematically related) collection, (b) retrieval of documents relevant to each question using BM25, (c) grounding/reformulation of the questions based on the retrieved texts/documents and an initial attempt to provide validated answers, (d) trimming and validating the answers so that they remain concise and well-supported with evidence, and (e) a final "rescue" process, where previously unanswerable questions are re-examined to reduce their number. The result is a cleaner dataset of question–answer pairs, where each answer is accompanied by a short, text-validated response, explicit reference to evidence, or is marked as unanswerable. Evaluation using automatic metrics (ROUGE, BERTScore) and manual assessment shows that the proposed method achieves high accuracy and a significant reduction of false answers, confirming that this process is an effective strategy for limiting hallucinations in LLMs. Using the results of the above process, an application was also implemented, which presents users with random questions from the dataset, and they are asked to answer them. At the end, they are graded based on the similarity between their answer and the model's answer. This serves as a practical application of the designed pipeline, focusing on the field of education, which can greatly benefit from the development of Information Retrieval technologies and Large Language Models.

Keywords

Large Language Models, RAG, Query Expansion, Information Retrieval, Education

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Χρήστο Μακρή και τον υποψήφιο διδάκτορα κ. Νικήτα-Ρήγα Καλογερόπουλο για την επίβλεψη αυτής της διπλωματικής εργασίας, για την καθοδήγηση και για τις συμβουλές που μου προσέφεραν κατά την εκπόνηση της. Επίσης, θα ήθελα να ευχαριστήσω την οικογένεια μου, και συγκεκριμένα τους γονείς μου, για την στήριξη σε όλα τα χρόνια των σπουδών μου και την προσπάθεια που έκαναν για να με βοηθήσουν να φτάσω εως εδώ. Τέλος, θα ήθελα να ευχαριστήσω τα άτομα που ήταν και είναι κοντά μου σε όλα τα χρόνια των σπουδών μου και τους φίλους μου και συγκεκριμένα την Ζέφη, τον Γιώργο, τον Χρήστο και τον Σωτήρη.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η γρήγορη πρόοδος στα Μεγάλα Γλωσσικά Μοντέλα (LLMs) έχει οδηγήσει σε σημαντικές βελτιώσεις στην επεξεργασία φυσικής γλώσσας, στην παραγωγή κειμένων, στην απάντηση ερωτήσεων και σε εφαρμογές εύρεσης πληροφορίας. Ωστόσο, ένα από τα σημαντικότερα προβλήματα που έχουν εμφανιστεί στον τομέα αυτό είναι το φαινόμενο των παραισθήσεων (hallucinations), δηλαδή η παραγωγή κειμένων ή απαντήσεων από τα LLMs που φαίνονται πειστικές αλλά δεν επιβεβαιώνονται από πραγματικά δεδομένα, γεγονός που θέτει σε κίνδυνο την αξιοπιστία τους. Μια απλή εξήγηση για αυτό το πρόβλημα, είναι το γεγονός πως ένα Μεγάλο Γλωσσικό Μοντέλο, δεν θα «πει» ποτέ απο μόνο του ότι δεν γνωρίζει κάτι. Πάντα προσπαθεί να δώσει απάντηση, με βάση την προτροπή που του έχει δωθεί, ακόμη και αν αυτή δεν ανταποκρίνεται στα πραγματικά δεδομένα.

Το προβλήμα αυτό εντοπίζεται στην καθημερινή χρήση, αλλά γίνεται περισσότερο επικίνδυνο όταν τα LLMs χρησιμοποιούνται σε κρίσιμες εφαρμογές, όπως η υγεία, η νομική ή η εκπαίδευση, όπου η ακρίβεια των πληροφοριών είναι υψίστης σημασίας. Σε αυτές τις περιπτώσεις, μια απάντηση που δεν στηρίζεται σε αποδεδειγμένα τεκμήρια μπορεί να έχει σοβαρές συνέπειες. Ακόμη και στην καθημερινή χρήση, καθώς όλο και περισσότεροι χρήστες αναζητούν πληροφορίες μέσω LLMs, η παραγωγή λανθασμένων απαντήσεων και ψευδών πληροφοριών μπορεί να προκαλέσει σημαντικά προβλήματα.

Για την αντιμετώπιση του ζητήματος των παραισθήσεων στα Μεγάλα Γλωσσικά Μοντέλα, χρησιμοποιούνται τεχνικές όπως η Παραγωγή Ενισχυμένη με Ανάκτηση (Retrieval-Augmented Generation, RAG) για να εντοπιστούν σχετικά κείμενα ή κομμάτια κειμένων, πρακτικές προτροπών (prompting) που ενισχύουν την απόδοση του μοντέλου (π.χ. chain-of-thought), καθώς και προσαρμογή (fine-tuning) μοντέλων ώστε να βελτιωθεί η συμπεριφορά τους απέναντι σε συγκεκριμένες συλλογές κειμένων και οι επιδόσεις τους σε συγκεκριμένες λειτουργίες.

1.1 Σημασία του προβλήματος

Τα Μεγάλα Γλωσσικά Μοντέλα, παρότι έχουν εντυπωσιακές δυνατότητές και χρησιμοποιούνται ευρέως στις μέρες μας, και σε πολλούς τομείς της επιστήμης της Πληροφορικής, συχνά «φαντάζονται» πληροφορίες είτε συμπληρώνοντας κενά είτε παράγοντας ανακρίβειες. Το φαινόμενο αυτό, των «παραισθήσεων», είναι απο τις βασικότερες προκλήσεις στον χώρο των LLMs. Οι παραισθήσεις αυτές μπορούν να λάβουν πολλές μορφές: απλές ανακρίβειες ή παρερμηνείες, είτε παραπληροφόρηση, είτε επινοημένα γεγονότα. Τα hallucinations, λοιπόν, μπορούν να προκληθούν για αρκετούς λόγους. Μερικοί από αυτούς που έχουν συζητηθεί στην βιβλιογραφία [3, 28, 38] είναι οι εξής:

1. Η στατιστική φύση των Μεγάλων Γλωσσικών Μοντέλων: Τα LLMs δεν «γνωρίζουν» πράγματα, με την έννοια που ένας άνθρωπος μπορεί να τα γνωρίζει. Στην ουσία προβλέπουν την επόμενη λέξη, με βάση τα δεδομένα που διαθέτουν. Το γεγονός αυτό, τα κάνει να μην μπορούν να αναγνωρίσουν την άγνοια και ως συνέπεια, μπορεί να επινοήσουν απαντήσεις.

2. Κακή ποιότητα δεδομένων εκπαίδευσης: Τα LLMs βασίζονται σε ένα μεγάλο σύνολο δεδομένων εκπαίδευσης για να παράγουν αποτελέσματα που είναι σχετικά και ακριβή για την κάθε προτρόπη, οποιουδήποτε χρήστη. Ωστόσο, αυτά τα δεδομένα εκπαίδευσης μπορεί να περιέχουν θόρυβο, σφάλματα, προκαταλήψεις ή ασυνέπειες. Κατά συνέπεια, το LLM μπορεί να παράγει λανθασμένα και, σε ορισμένες περιπτώσεις, εντελώς παράλογα αποτελέσματα.

3. Διαδικασία δημιουργίας/γέννησης κειμένου: Ακόμη και αν το σύνολο εκπαίδευσης είναι καλό, χωρίς θόρυβο και ανακρίβειες, μπορεί να δημιουργηθούν παραισθήσεις λόγω προκατάληψης(bias) απο προηγούμενες παραγογές του μοντέλου ή λόγω λάθους αποκωδικοποίησης του μετατροπέα(transformer).

4. Ανθρώπινος παράγοντας: Παρόλο που ο χρήστης δεν ελέγχει τι γίνεται «μέσα» στο μοντέλο, μπορεί να του δημιουργήσει πρόβλημα, αν η προτροπή(prompt) που του δίνει δεν είναι ξεκάθαρη ή είναι αντιφατική. Η εντολή που παίρνει το μοντέλο, και συνεπώς η ποιότητα της, έχει άμεση σχέση και με την ποιότητα της απάντησης που θα δώσει

Είναι εμφανές ότι υπάρχουν αρκετοί λόγοι για τους οποίους ένα LLM, παράγει παραισθήσεις. Έχει αποδειχθεί στην βιβλιογραφία [4] πως για τους παραπάνω λόγους, αλλά και για περιορισμούς στο κομμάτι της ανάκτησης των σχετικών πληροφοριών στο κείμενο με πλήρη ακρίβεια, πως τα Μεγάλα Γλωσσικά Μοντέλα θα αντιμετωπίζουν πάντοτε το πρόβλημα των παραισθήσεων [3]. Έχει, λοιπόν, τεράστια σημασία η προσπάθεια περιορισμού αυτού του φαινομένου. Μερικές γνωστές τεχνικές για τον περιορισμό των παραισθήσεων αναφέρθηκαν και παραπάνω. Όμως, για μια από αυτές, την ακριβή προσαρμογή (fine-tuning), υπάρχουν μερικές αμφιβολίες. Μελετώντας την βιβλιογραφία [22], είναι εμφανές πως το fine-tuning δεν εγγυάται ότι το μοντέλο θα ενσωματώσει σωστά την νέα πληροφορία, ειδικά αν αυτή είναι

πολύ διαφορετική ή άγνωστη από την γνώση που είχε αποκτήσει στην προεκπαίδευση. Καθώς το μοντέλο προσπαθεί να μάθει νέα γνώση, υπάρχει σύγκρουση μεταξύ αυτών των νέων στοιχείων και της ήδη αποθηκευμένης γνώσης. Αυτό μπορεί να οδηγήσει σε λάθη, ή το μοντέλο να αγνοήσει ή να παραποιήσει προηγούμενα γνωστά στοιχεία. Ως αποτέλεσμα, γίνεται εμφανές πως πρέπει να δωθεί προσοχή στις άλλες λύσεις του προβλήματος των παραισθήσεων [6, 4](RAG, Prompt Engineering), οι οποίες δεν επιφυλάσουν κάποιο κίνδυνο για την αξιοπιστία του μοντέλου. Ένας τομέας στην χρήση των Μεγάλων Γλωσσικών Μοντέλων, που κινδυνεύει ιδιαίτερα απο το φαινόμενο των παραισθήσεων, είναι τα συστήματα ερωταπαντήσεων. Αυτό ισχύει και για καθημερινούς χρήστες, που θέλουν να αποκτήσουν προσωπική γνώση, αλλά και για εκπαίδευση μοντέλων η οποία είναι ολοένα και συχνότερη στις μέρες μας, και οι ερωταπαντήσεις χρησιμεύουν αρκετά ως δεδομένα στην εκπαίδευση τους. Είναι σημαντικό λοιπόν να μην υπάρχουν παραισθήσεις στην απάντηση, αλλά ούτε και στην ερώτηση. Σε αυτή την ανάγκη δώθηκε ιδιαίτερη έμφαση, κατά την εκπόνηση της διπλωματικής. Συγκεκριμένα, θεωρήθηκε πως όταν τα Μεγάλα Γλωσσικά Μοντέλα χρησιμοποιούνται για την εκπαίδευση ανθρώπων, οι παραισθήσεις αποτελούν σοβαρό πρόβλημα. Ένα λανθασμένο ή ανυπόστατο στοιχείο που παρουσιάζεται ως σωστή απάντηση μπορεί να παραπλανήσει τον χρήστη και να οδηγήσει σε παγίωση λανθασμένων γνώσεων. Επίσης, υπάρχει και το ζήτημα της ερώτησης που ο χρήστης θα κληθεί να απαντήσει. Αν αυτή έχει παραχθεί από μια απλή προτροπή του LLM, και όχι από κάποιον άνθρωπο ή καθηγητή, που έχει μελετήσει το υλικό, τότε και εκεί υπάρχει το ενδεχόμενο παραισθήσεων. Η συγκεκριμένη ερώτηση μπορεί να μην έχει ξεκάθαρη ή υπαρκτή απάντηση στα κείμενα τα οποία ο χρήστης έχει μελετήσει και συνεπώς να δημιουργηθεί σύγχηση. Η αξιοπιστία του εργαλείου εκμάθησης εξαρτάται άμεσα από την ακρίβεια των παραγόμενων ερωτήσεων και απαντήσεων· Κατα συνέπεια, ο περιορισμός των παραισθήσεων δεν είναι μόνο τεχνικό ζητούμενο αλλά και βασική προϋπόθεση για την αποτελεσματική μάθηση και την εμπιστοσύνη του χρήστη προς το σύστημα αλλά και προς την διαδικασία εκμάθησης στο σύνολο της. Όσο η τεχνολογία εξελίσεται, πρέπει μαζί της να εξελίσεται και η εκπαίδευση. Παρόλο που έχει σημειωθεί τρομερή πρόοδος στον χώρο των LLMs [27], η εφαρμογή τους στον τομέα της εκπαίδευσης δεν έχει εξερευνηθεί στο μέγιστο των δυνατοτήτων που διαθέτουν.

1.2 Στόχοι της Διπλωματικής Εργασίας

Ο κύριος στόχος αυτής της διπλωματικής εργασίας είναι να προταθεί και να υλοποιηθεί ένα πολυεπίπεδο σύστημα παραγωγής ερωτήσεων, πάνω σε μια συλλογή κειμένων, και τεκμηριωμένης απάντησης τους. Πέρα απο την παραγωγή των ερωτήσεων, δίνεται έμφαση και στον κατάλληλο εμπλουτισμό και επιβεβαίωση της κάθε ερώτησης, με βάση τα σχετικά κείμενα. Η διαδικασία αυτή, μπορεί να ονομαστεί γείωση(grounding), σύμφωνα με ερευνιτικό έργο που εξετάστηκε και θα περιγραφεί στην συνέχεια. Γίνεται, λοιπόν, επέκταση των αρχικών ερωτημάτων(query expansion). Δίνεται, επίσης, σημασία στην τεκμηριωμένη απάντηση κάθε ερώτησης, αναζητώντας σχετικά κείμενα όχι μόνο στο κείμενο για το οποίο παράχθηκε η ερώτηση, αλλά και στα υπόλοιπα της συλλογής κειμένων που χρησιμοποείται, έτσι ώστε να υπάρχει όσο πιο ξεκάθαρη, σύντομη και τεκμηριωμένη με βάση τα δεδομένα απάντηση γίνεται. Με αυτό τον τρόπο, θα αποτραπούν οι παραισθήσεις του LLM τόσο στην παραγωγή των ερωτήσεων(μέσω της επέκτασης ερωτημάτων) όσο και στην απάντηση τους. Δίνεται ιδιαίτερη προσοχή στο να εντοπίζεται ξεκάθαρη απάντηση στην κάθε ερώτηση, και αν αυτό δεν επιτευχθεί αποτελεσματικά απο το σύστημά, τότε χαρακτηρίζεται ως μη-απαντήσιμη. Σε πλαίσιο συστημάτων ερωταπαντήσεων, θεωρήθηκε ιδιαίτερα σημαντική η ακρίβεια και για αυτό η διπλωματική αυτή εστιάζει κυρίως στην βελτίωση αυτής. Το τελικό αποτέλεσμα είναι ένα σύνολο ερωτήσεων-απαντήσεων, όπου κάθε ερώτηση συνοδεύεται απο μια σύντομη και πιστή στην συλλογή κειμένων, απάντηση. Για να δώθεί και μια πρακτική λειτουργία στο σύστημα, λήφθηκε η απόφαση για έμφαση στην καθημερινή χρήση αυτού του επικυρωμένου συνόλου ερωτήσεων και απαντήσεων. Συγκεκριμένα, στην εκπαίδευση και εκμάθηση ενός χρήστη. Έφοσον η ροή λειτουργίας (pipeline) που υλοποιήθηκε παράγει, επεκτείνει και επικυρώνει ερωτήσεις οι οποίες υποστηρίζονται απο την συλλογή κειμένων, υλοποιήθηκε μια εφαρμογή που επιλέγει μερικές απο αυτές και ο χρήστης της απαντά. Όταν τις απαντήσει όλες, τότε βαθμολογείται με βάση μετρικές που συγκρίνουν την απάντηση του χρήστη, σε σχέση με την απάντηση που έδωσε. Έτσι λοιπόν, υπάρχει βέβαιοτητα, τόσο για την εγκυρώτητα όσο και για την σχετικότητα της ερώτησης που ο χρήστης θα κλιθεί να απαντήσει, αλλά και για το εαν η απάντηση που έδωσε είναι σύμφωνη με την, τεκμηριωμένη με βάση τα κείμενα, απάντηση του μοντέλου. Τελικός στόχος του έργου, είναι η εργασία να συνεισφέρει τόσο σε μεθοδολογικό επίπεδο, μέσω του προτεινόμενου πολυεπίπεδου συστήματος παραγωγής και γείωσης ερωτήσεων και εντοπισμού τεκμηριωμένων απαντήσεων, όσο και σε εφαρμοστικό, μέσω της ανάπτυξης ενός πρακτικού εργαλείου εκπαίδευσης.

1.3 Συνεισφορά της Διπλωματικής Εργασίας

Μέσω αυτής της διπλωματικής εργασίας, δίνεται έμφαση στην σημασία της ακρίβειας και της τεκμηρίωσης απαντήσεων, στα Μεγάλα Γλωσσικά Μοντέλα. Γίνεται προσπάθεια ενσωματώνοντας τεχνικές Επέκτασης Ερωτημάτων και Retrieval Augmeted Generation, δίνοντας τις κατάλληλες οδηγίες στο μοντέλο και εφαρμόζοντας ελέγχους για το αν η απάντηση είναι σύμφωνη με τα δεδομένα, να περιοριστεί σε μεγάλο βαθμό το φαινόμενο των παραισθήσεων σε ένα LLM. Παρόλο που το φαινόμενο αυτό, δεν μπορεί να εξαφανιστεί τελείως απο την συμπεριφορά των μοντέλων [3], αποδεικνύεται πως με κατάλληλη μεθοδολογία και σωστό έλεγχο, γίνεται να υπάρχει μεγάλη σιγουρία για την ακρίβεια των απαντήσεων που παράγονται, με μόνο κόστος την ποσότητα τους. Θεωρήθηκε καλύτερο, το να απορριφθούν μερικά σετ ερωταπαντήσεων, για τα οποία δεν βρίσκεται κατάλληλο και πιστό τεκμήριο στην συλλογή κειμένων, από το να τα τους επιτραπεί να υπάρχουν. Συνολικά, παρουσιάζεται μια νέα, προτότυπη και πολυεπίπεδη προσέγγιση για το πως μπορεί να γίνει η γέννηση και η απάντηση ερωτήσεων χωρίς τον κίνδυνο των παραισθήσεων, για μια συλλογή κειμένων κοινής θεματολογίας. Κύριος σκοπός της διαδικτυακης εφαρμογής παραγωγής κουίζ είναι να παρουσιαστεί μια ενδεικτική και πρακτική εφαρμογή που θα μπορούσε να έχει το συγκεκριμένο συστήμά. Παράλληλα, σκοπός είναι και να αποδειχθεί πως μπορεί να αυτοματοποιηθεί με χρήση Μεγάλων Γλωσσικών Μοντέλων η διαδικασία εξάσκησης και βαθμολογίας απλών χρηστών ή/και μαθητών χωρίς την ύπαρξη παραισθήσεων του Μοντέλου, που είναι ιδιαίτερα επιβλαβείς στην διαδικασία αυτήν. Με τον τρόπο αυτό οι χρήστες θα μαθαίνουν πιο εύκολα και γρήγορα, διευκολύνοντας παράλληλα και τους φορείς της εκπαίδευσης, αφού δεν θα χρείαζεται να διαβάσουν αναλυτικά την συλλογή κειμένων ή την ύλη, για να παράξουν ερωτήσεις και να εντοπίσουν τις απαντήσεις τους.

Συνοψίζοντας, η διπλωματική εργασία συνεισφέρει σε τρία διακριτά επίπεδα:

1. Μεθοδολογικό: Προτείνεται μια νέα, πρωτότυπη και πολυεπίπεδη διαδικασία γέννησης και απάντησης ερωτήσεων, που ενσωματώνει στάδια παραγωγής, επέκτασης(γείωσης), περικοπής και διάσωσης, με στόχο την ελαχιστοποίηση παραισθήσεων.

2. Πρακτικό: Παρουσιάζεται η υλοποίηση μιας εφαρμογής τύπου γεννήτριας κουίζ (quiz generator), η οποία επιλέγει μερικές απο τις επικυρωμένες ερωτήσεις, τις παρουσιάζει σε χρήστες και αξιολογεί αυτόματα τις απαντήσεις τους με βάση μετρικές σύγκρισης (ROUGE, BERTScore). Η εφαρμογή αυτή αποδεικνύει τη χρησιμότητα της μεθοδολογίας σε πραγματικά σενάρια μάθησης.

3. Εκπαιδευτικό: Αναδεικνύεται πώς μπορεί να αυτοματοποιηθεί η διαδικασία δημιουργίας και αξιολόγησης ασκήσεων, υποστηρίζοντας μαθητές και εκπαιδευτικούς. Με αυτό τον τρόπο οι χρήστες έχουν πρόσβαση σε έγκυρες και τεκμηριωμένες ερωτήσεις χωρίς να χρειάζεται να μελετούν εξαντλητικά όλο το υλικό, ενώ μειώνεται ο κίνδυνος διάδοσης λανθασμένων πληροφοριών.

1.4 Διάρθρωση της Διπλωματικής Εργασίας

Η διπλωματική αυτή εργασία έχει οργανωθεί σε επτά κεφάλαια. Στο πρώτο κεφάλαιο, γίνεται αναφορά στο πρόβλημα που προσπαθεί να λύσει το έργο και στον τρόπο με τον οποίο θα πραγματοποιηθεί η προσπάθεια. Παράλληλα αναφέρεται η συνεισφορά του έργου. Στο δεύτερο κεφάλαιο, δίνεται έμφαση στο θεωρητικό υπόβαθρο που πρέπει κάποιος να γνωρίζει και να μελετήσει για να κατανοηθεί η ουσία της εργασίας. Παρουσιάζονται και αναλύονται επαρκώς οι βασικότερες έννοιες που αποτέλεσαν τις βάσεις για να αναπτυχθεί το ολοκληρωμένο, τελικό σύστημα. Στο κεφάλαιο 3, γίνεται αναφορά στην συνεισφορά της Τεχνιτής Νοημοσύνης στον χώρο της εκπαίδευσης. Αυτό το κεφάλαιο παρέχει πληροφορία σχετικά με την αξιοποίηση τέτοιων τεχνολογίων στον εκπαιδευτικό τομέα και την συνεισφορά τους στην ενίσχυση του. Ο τελικός στόχος της εργασίας, εξάλλου, είναι η χρήση του παραγόμενου συνόλου δεδομένων στο πλαίσιο της εκπαίδευσης. Στο κεφάλαιο 4, αρχικά αναλύονται μερικά έργα που μελετήθηκαν και αποτέλεσαν πηγή έμπνευσης της μεθοδολογίας που ακολουθήθηκε για την ανάπτυξη του έργου, αλλά και για τον στόχο του έργου γενικότερα. Στην συνέχεια γίνεται μια γενική περιγραφή για το πως δουλεύει το σύστημα που αναπτύχθηκε, έτσι ώστε να γίνει η λειτουργία του πιο εύκολα κατανοητή. Στο επόμενο κεφάλαιο, το πέμπτο, γίνεται η περιγραφή της υλοποίησης του συστήματος. Αναφέρονται οι βασικές τεχνολογίες που χρησιμοποιήθηκαν κατά την διάρκεια της εκπόνησης της διπλωματικής, καθώς και αναλύεται η ροή του κώδικα που αναπτύθχηκε. Στο κεφάλαιο 6 υποδεικνύονται τα αποτελέσματα της διπλωματικής, τόσο στο κομμάτι των επαληθευμένων και γειωμένων απαντήσεων όσο και στο κομμάτι των έγκυρων απαντήσεων. Επεξηγούνται οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν και ο λόγος για τον οποίο επηλέχθηκαν. Τέλος, στο έβδομο και τελευταίο κεφάλαιο αναφέρονται τα συμπεράσματα που αντλήθηκαν απο την εκπόνηση της διπλωματικής καθώς και μερικοί περιορισμοί που αντιμετωπίστηκαν στην υλοποίηση, μαζί με προτάσεις για μελλοντική έρευνα πάνω στο έργο.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Σε αυτό το κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο που στηρίζει τη διπλωματική εργασία. Αρχικά, γίνεται μια γενική αναφορά στα Μεγάλα Γλωσσικά Μοντέλα, τα βασικά χαρακτηριστικά και τις δυνατότητές τους. Στη συνέχεια εξετάζεται η τεχνική της Παραγωγής Ενισχυμένης με Ανάκτηση (Retrieval Augmented Generation – RAG), η οποία συνδυάζει τις δυνατότητες των μοντέλων με τεχνικές ανάκτησης πληροφορίας. Τέλος, αναλύεται η έννοια της Επέκτασης Ερωτημάτων (query expansion)και γενικά αλλά και ως μέθοδος βελτίωσης της ανάκτησης και της τεκμηρίωσης. Η παρουσίαση αυτών των εννοιών θέτει τις βάσεις για την κατανόηση της μεθοδολογίας που ακολουθεί στα επόμενα κεφάλαια.

2.1 Μεγάλα Γλωσσικά Μοντέλα

Τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs) αποτελούν μια από τις πιο σημαντικές καινοτομίες στον χώρο της Τεχνητής Νοημοσύνης και ειδικότερα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing). Πρόκειται για βαθιά νευρωνικά δίκτυα, τα οποία εκπαιδεύονται σε τεράστιες συλλογές κειμένων και έχουν τη δυνατότητα να κατανοούν και να παράγουν ανθρώπινη γλώσσα με υψηλό βαθμό φυσικότητας. Χάρη στη μεγάλη τους κλίμακα, τόσο σε δεδομένα εκπαίδευσης όσο και σε αριθμό παραμέτρων, τα LLMs έχουν αποδείξει ότι μπορούν να προσεγγίσουν ποικιλία εργασιών (π.χ. μετάφραση, σύνοψη, ερωταποκρίσεις, δημιουργία κειμένου) χωρίς ειδική προσαρμογή για κάθε μία από αυτές. Αυτό έχει οδηγήσει στην δραματική αύξηση στην χρήση των Μεγάλων Γλωσσικών Μοντέλων, τόσο απο απλούς, καθημερινούς χρήστες όσο και απο εταιρίες και επαγγελματίες στον χώρο της Πληροφορικής και της Τεχνητής Νοημοσύνης.
Η λειτουργία των LLMs βασίζεται σε αρχιτεκτονικές τύπου μετασχηματιστή (transformer), οι οποίες εισήγαγαν τον μηχανισμό της αυτοπροσοχής (self-attention) [39]. Η αυτοπροσοχή επιτρέπει στο μοντέλο να λαμβάνει υπόψη του τις συσχετίσεις ανάμεσα σε όλες τις λέξεις μιας πρότασης ή ακόμη και μεγαλύτερων τμημάτων κειμένου, κάτι που τα προηγούμενα μοντέλα δεν μπορούσαν να κάνουν αποτελεσματικά. Συνεπώς, τα LLMs έχουν την ικανότητα να «κατανοούν» τα συμφραζόμενα σε βάθος και να αποδίδουν σημασιολογικά συνεκτικές απαντήσεις. Σύμφωνα με τη βιβλιογραφία, η βασική αρχή που διέπει τη λειτουργία ενός Μεγάλου

Γλωσσικού Μοντέλου είναι η πρόβλεψη της επόμενης γλωσσικής μονάδας (token) με βάση το προηγούμενο κείμενο. Κατά την εκπαίδευση, το μοντέλο προσαρμόζει τα βάρη του ώστε να μεγιστοποιεί την πιθανότητα να προβλέπει σωστά το επόμενο token, δεδομένης μιας ακολουθίας εισόδου. Μέσα από δισεκατομμύρια παραδείγματα τέτοιων προβλέψεων, το μοντέλο εξοικειώνεται με τα στατιστικά μοτίβα της γλώσσας, τα οποία του επιτρέπουν να παράγει συνεκτικό και ρεαλιστικό λόγο [39, 43].
Είναι σημαντικό να αναφερθούμε και στο τι διαφοροποιεί τα LLMs με τα υπόλοιπα γλωσσικά μοντέλα. Το χαρακτηριστικό, λοιπον, που διαφοροποιεί αυτά τα δύο, είναι το τεράστιο μέγεθός των LLM. Η βιβλιογραφία [39] δείχνει ότι η αύξηση του αριθμού παραμέτρων, του όγκου δεδομένων εκπαίδευσης και της υπολογιστικής ισχύος οδηγεί σε εμφάνιση αναφαινόμενων (emergent) ιδιοτήτων. Με απλά λόγια, το μοντέλο αποκτά ικανότητες που δεν ήταν εμφανείς σε μικρότερης κλίμακας εκδοχές. Για παράδειγμα, μεγάλα μοντέλα έχουν δείξει ικανότητα επίλυσης σύνθετων λογικών προβλημάτων ή κατανόησης αφηρημένων εννοιών χωρίς να έχουν εκπαιδευτεί αποκλειστικά γι' αυτό. Σύμφωνα με την ίδια πηγή, ακόμη ένα από τα πιο εντυπωσιακά χαρακτηριστικά τους είναι η ικανότητα μηδενικής εκπαίδευσης και εκπαίδευσης με λίγα παραδείγματα (zero-shot, few-shot learning). Αυτό σημαίνει ότι μπορούν να επιλύσουν εργασίες για τις οποίες δεν έχουν εκπαιδευτεί ειδικά, απλώς ακολουθώντας οδηγίες σε φυσική γλώσσα ή παραδείγματα που παρέχονται στη φάση του prompting. Αυτή η ικανότητα τα διαφοροποιεί ουσιαστικά από τα προηγούμενα συστήματα Επεξεργασίας Φυσικής Γλώσσας, τα οποία απαιτούσαν εξειδικευμένη εκπαίδευση για κάθε ξεχωριστό πρόβλημα.
Παρά τις εξελιγμένες δυνατότητες, τα Μεγάλα Γλωσσικά Μοντέλα αντιμετωπίζουν σημαντικούς περιορισμούς και προκλήσεις που πρέπει να λαμβάνονται υπόψη [43, 29, 30]:

1. Στατικότητα γνώσης / χρονική παλαιότητα: Τα μοντέλα εκπαιδεύονται σε δεδομένα με συγκεκριμένο χρονικό ορίζοντα. Μετά την ολοκλήρωση της εκπαίδευσης, δεν γνωρίζουν γεγονότα που συνέβησαν μετά τη συλλογή των δεδομένων. Αυτό σημαίνει ότι οι απαντήσεις μπορεί να είναι ανακριβείς ή ξεπερασμένες όταν ζητούνται πληροφορίες που αφορούν πρόσφατες εξελίξεις.

2. Ασυνέπειες και αντιφάσεις στα δεδομένα: Τα μεγάλου μεγέθους σύνολα κειμένων συχνά περιέχουν αντιφατικές πληροφορίες, διαφορετικές εκδοχές γεγονότων ή ασυνέπειες ως προς ορολογία και λεπτομέρειες. Το μοντέλο ενδέχεται να χρησιμοποιήσει αυτές τις εκδοχές και να παράγει απαντήσεις που συνδυάζουν στοιχεία από διαφορετικές πηγές με τρόπο που προκαλεί σύγχυση, ανακρίβεια, ή εντελώς λάθος πληροφόρηση.

3. Αναπαραγωγή προκαταλήψεων και μεροληψιών (biases): Εφόσον τα δεδομένα εκπαίδευσης περιέχουν κοινωνικές προκαταλήψεις, στερεότυπα ή ανισότητες, τα μοντέλα συχνά τα αναπαράγουν ή τα ενισχύουν. Αυτό είναι ιδιαίτερα προβληματικό σε περιεχόμενο κοινωνικού/ηθικού ενδιαφέροντος, καθώς μπορεί να οδηγήσει σε άδικες ή λανθασμένες απαντήσεις.

4. Υπολογιστικό κόστος και αποδοτικότητα: Η εκπαίδευση και η εξυπηρέτηση LLMs απαιτούν σημαντικούς υπολογιστικούς πόρους (GPU, μνήμη, ενέργεια). Η χρήση τους σε

πραγματικό χρόνο ή στο τοπικό περιβάλλον των χρηστών συχνά δεν είναι εφικτή χωρίς σημαντική υποδομή ή υποστήριξη. Επιπλέον, η κλιμάκωση του συστήματος (περισσότερα δεδομένα, μεγαλύτερα μοντέλα) συχνά αυξάνει γραμμικά ή εκθετικά τα απαιτούμενα ενεργειακά και υπολογιστικά κόστη.

5. Ερμηνευσιμότητα και διαφάνεια: Τα LLMs λειτουργούν σε «μαύρο κουτί» για πολλούς χρήστες και ερευνητές, δηλαδή δεν υπάρχει εύκολη εξήγηση για το γιατί παράγουν μια συγκεκριμένη απάντηση. Η έλλειψη διαφάνειας καθιστά δύσκολη τη διάγνωση σφαλμάτων ή τη δικαιολόγηση των απαντήσεων σε κρίσιμες εφαρμογές.

6. Απαιτήσεις ρύθμισης και συντήρησης: Τα μοντέλα χρειάζονται τακτική ενημέρωση (fine-tuning, ανανέωση δεδομένων) για να παραμένουν επίκαιρα. Η διαδικασία αυτή είναι χρονοβόρα, απαιτεί επαναδιοργάνωση με νέα δεδομένα και μπορεί να εισαγάγει νέα σφάλματα ή ασυνέπειες στο τελικό αποτέλεσμα

Οι παραπάνω περιορισμοί υπογραμμίζουν ότι τα Μεγάλα Γλωσσικά Μοντέλα, παρά τις επιδόσεις, δεν είναι αλάνθαστα, και χωρίς τα μειονεκτήματα τους. Σε ορισμένες περιπτώσεις, η τυφλή εμπιστοσύνη στα μοντέλα αυτά μπορεί να προκαλέσει περισσότερα πρόβλημματα παρά οφέλη [15]. Καμία μεμονωμένη τεχνική δεν μπορεί να αντιμετωπίσει όλα τα προβλήματα. Η συνδυαστική χρήση τεχνικών όπως ανάκτηση τεκμηρίων, έλεγχοι συνέπειας, φιλτραρίσματα, ενημέρωση δεδομένων και μεθοδολογίες γείωσης είναι κρίσιμη για την αντιμετώπιση αυτών των προκλήσεων. Είναι πολύ σημαντικό, για την πρόοδο της επιστήμης της Πληροφορικής, να επιτευχθεί η μείωση και ο περιορισμός των επικείμενων κινδύνων απο την χρήση και εκπαίδευση των LLMs, και είναι ανάγκη ένα μεγαλύτερο μερίδιο έρευνας να εστιάσει εκεί.

2.2 Παραγωγή Ενισχυμένη Με Ανάκτηση - RAG

2.2.1 Γενική Αναφορά στο RAG

Παραπάνω αναφέρθηκαν τα Μεγάλα Γλωσσικά Μοντέλα και οι εντυπωσιακές τους δυνατότητες, αλλά και οι περιορισμοί οι οποίοι προκύπτουν από την χρήση τους. Ένας καλός τρόπος να βελτιωθεί η απόδοση τους, αλλά να επιτευχθούν και επιπλεόν πράγματα μέσω της χρήσης τους, είναι η Παραγωγή Ενισχυμένη Με Ανάκτηση [23]. Το Retrieval Augmented Generation (RAG) είναι μια αρχιτεκτονική που συνδυάζει τις δυνατότητες Ανάκτησης Πληροφορίας με τη γλωσσική παραγωγή των Μεγάλων Γλωσσικών Μοντέλων, με σκοπό την παραγωγή πιο αξιόπιστων και επικυρωμένων απαντήσεων, σε σχέση με τα δεδομένα που έχει στην διάθεση του το μοντέλο για την διατύπωση και ανάπτυξη τους. Με απλά λόγια, το RAG εισάγει ένα στάδιο ανάκτησης πριν ή κατά τη διάρκεια της παραγωγής κειμένου από το LLM [59, 37]. Συνεπώς, αφού το LLM διαθέτει πλεον τεκμήρια, το τελικό αποτέλεσμα που παράγει μετά απο κάθε query ενισχύεται με βάση αυτά.
Στην ουσία, όταν δίνεται ένα ερώτημα, το σύστημα:

1. Αναζητά στα διαθέσιμα έγγραφα ή συλλογές κειμένων τα πιο σχετικά αποσπάσματα ή τεκμήρια σε σχέση με το ερώτημα

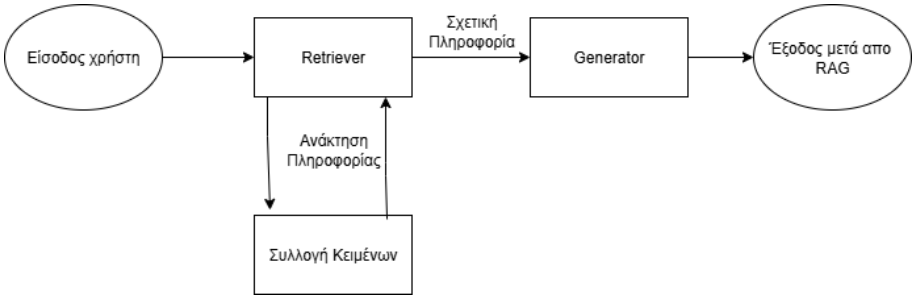
2. Χρησιμοποιεί τα ανακτημένα κομμάτια και τα ενσωματώνει στην διαδικασία απάντησης του ερωτήματος. Στην ουσία, έχουμε ένα ενισχυμένο prompt, με βάση τεκμήρια.

3. Παράγει την τελική απάντηση χρησιμοποιώντας το μοντέλο, το οποίο πλέον έχει πρόσβαση στο ενισχυμένο και τεκμηριομένο περιεχόμενο. Ως αποτέλεσμα, η απάντηση είναι πολύ πιο πιθανό να είναι πιστή στην πραγματικότητα και τεκμηριομένη με βάση τα κείμενα που λειτουργούν ως βάση γνώσης.

Έτσι, το μοντέλο δεν βασίζεται μόνο στην εσωτερική μνήμη των παραμέτρων του, δηλαδή το τι έμαθε κατά την εκπαίδευση, αλλά οδηγείται σε πραγματικά τεκμήρια για να δομήσει απαντήσεις. Αυτή η ιδέα περιγράφεται και στη συλλογική βιβλιογραφία [23, 25] ως τρόπος αντιμετώπισης των προβλημάτων που προκύπτουν από τη στατικότητα των Μεγάλων Γλωσσικών Μοντέλων, τις παραισθήσεις και την έλλειψη ενημερωμένων γνώσεων.

Είναι εμφανές, λοιπόν, πως το Retrieval Augmented Generation επιτρέπει στο μοντέλο να συμβουλεύεται μια βάση γνώσης ως ενδιάμεσο βήμα πριν αποφασίσει τι να απαντήσει για κάθε ερώτημα του χρήστη. Το μοντέλο, εν τέλη, μπορεί να δώσει απαντήσεις βασισμένες όχι μόνο στην εκπαίδευση του, αλλά και σε εξωτερικές, ειδικές ή ενημερωμένες πηγές.

Με την χρήση του RAG, ξεπερνιούνται αρκετοί απο τους περιορισμούς των Μεγάλων Γλωσσικών Μοντέλων. Αρχικά, το RAG επιτρέπει την ενσωμάτωση νέων πληροφοριών που δεν υπήρχαν κατά τη φάση εκπαίδευσης του μοντέλου. Έτσι, μπορεί να απαντήσει σε ερωτήματα για πιο πρόσφατα γεγονότα. Παράλληλα, βοηθάει πολύ στην αύξηση της ακρίβειας και της αξιοπιστίας του μοντέλου και συνεπώς στην μείωση των παραισθήσεων που προκαλούνται.



Σχήμα 2.1: Αρχιτεκτονική του συστήματος Retrieval-Augmented Generation (RAG). Ο χρήστης δίνει είσοδο, ο Retriever αναζητά και κατατάσσει εξωτερική γνώση, η οποία τροφοδοτεί τον Generator, που παράγει την τελική απάντηση.

Όταν το μοντέλο ανακτά πραγματικά τεκμήρια και βασίζει τις απαντήσεις του επάνω τους, μειώνεται ο κίνδυνος να παράξει ανακρίβειες λόγω απουσίας πραγματικών πληροφοριών [6]. Το RAG προσφέρει επίσης μεγάλο βαθμό εξειδίκευσης και προσαρμοστικότητας σε συγκεκριμένους τομείς. Αν τα ανακτόμενα κείμενα, αφορούν κάποιον συγκεκριμένο τομέα, τότε το μοντέλο θα έχει καλύτερη επίδοση για ερωτήματα τέτοιου τύπου [50]. Επιπλέον, το Retrieval Augmented Generation συμβάλλει στην μείωση του κόστος επανεκπαίδευσης του μοντέλου. Αντί να ξαναεκπαιδευτεί ολόκληρο το μοντέλο με νέα δεδομένα, μας παρέχει την δυνατότητα απλώς να ενημερωθεί ή να επεκταθεί η βάση γνώσης που χρησιμοποιεί το κομμάτι της ανάκτησης [25]. Τέλος, κάνει πιο εύκολη την επαλήθευση των απαντήσεων που παράγει το μοντέλο, αφού έχουν προέλθει απο ξεκάθαρη πηγή. Ο χρήστης έχει τη δυνατότητα να ελέγξει τις απαντήσεις, προσδίδοντας μεγαλύτερη εμπιστοσύνη στην έξοδο του μοντέλου. Γίνεται εμφανές, λοιπόν, πως το RAG εξυπηρετεί κυρίως εργασίες όπου η αξιοπιστία και η τεκμηρίωση έχουν μεγάλη σημασία, όπως προβλήματα ερωταπαντήσεων, επικύρωσης πληροφοριών, συστήματα συμβουλευτικής και εκπαιδευτικές εφαρμογές.

Η κύρια αρχιτεκτονική ενός συστήματος Retrieval Augmented Generation, όπως απεικονίζεται στο περίπου και στο σχήμα 2.1, αποτελείται απο τα εξής στάδια:

1. Προεπεξεργασία/Δεικτοδότηση: Τα έγγραφα της βάσης γνώσεων, μετατρέπονται σε λεκτικές μονάδες (tokenized) και αν η συλλογή είναι μεγάλη, μπορούν να διαχωριστούν σε κομμάτια (chunks) για να είναι πιο εύκολα ανακτήσημα. Μπορεί επίσης να σχηματιστεί ένας δείκτης για να διευκολύνει αυτή την διαδικασία. Τα αποσπάσματα, μπορούν επίσης να μετατραπούν σε αναπαραστάσεις(embeddings), και να αποθηκευτούν σε βάση δεδομένων, αν μιλάμε για μεγάλο έργο.

2. Ανάκτηση: Δοθέντος ενός ερωτήματος (prompt, ένας ανακτητής (retriever) αναζητά τα k πιο σχετικά έγγραφα / αποσπάσματα. Η αναζήτηση μπορεί να βασίζεται σε ομοιότητα συνημιτόνων σε αναπαραστάσεις (embeddings) διανυσμάτων, ή σε συνδιαστικές και υβριδικές μεθόδους. Μπορούν ακόμη να εφαρμοστούν τεχνικές αναταξινόμησης(reranking) ή φιλτραρίσματα για να επιλεγούν τα κατάλληλα τεκμήρια για κάθε

ερώτημα.

3. Ενίσχυση (Augmentation): Σε αυτή την φάση, τα ανεκτημένα τεκμήρια ενσωματώνονται στο prompt ή στο context του μοντέλου. Υπάρχουν αρκετοί τρόποι που γίνεται αυτό όπως συνενώσεις και υπολογισμός βαρών (concatenation, weight-based), αλλά έχουν προταθεί και νέοι τρόποι όπως η δυναμική ανάκτηση (dynamic retrieval)

4. Παραγωγή: Το μοντέλο λαμβάνει το ενισχυμένο (prompt) και παράγει την τελική απάντηση. Στην φάση αυτήν μπορούν να γίνουν περαιτέρω βελτιώσεις, όπως επαναληπτικά στάδια ελέγχου, για να γίνει καλύτερη η ποιότητα του τελικού αποτελέσματος.

Στα παραπάνω βήματα, γίνεται να προστεθεί και η ενημέρωση της βάσης γνώσης αφού υπάρχει πάντοτε η δυνατότητα περιοδικής της ενημέρωσης, χωρίς να χρειάζεται κάποια επανεκπαίδευση στο μοντέλο, έτσι ώστε να ποστεθούν σε αυτήν νέα κείμενα, να τροποποιηθούν, ή να αφαιρεθεί κάτι που δεν είναι πλέον επιθυμητό [60].

Με βάση, λοιπόν, την γνώση που υπάρχει στον τομέα του RAG εντοπίζονται μερικά σημαντικά κομμάτια στα οποία μπορεί να συμβάλλει σε μεγάλο βαθμό. Αρχικά, η συμβολή του σε συστήματα ερωταπαντήσεων QA είναι μεγάλη, αφού μπορεί να αντλήσει σημαντικές και έγκυρες πληροφορίες απο την βάση δεδομένων, τόσο για την σύνθεση των ερωτήσεων, όσο και των απαντήσεων [44]. Παράλληλα, βοηθάει σημαντικά στην σύνθεση έγκυρων περιλίψεων κειμένων, αφού με την χρήση του βεβαιώνεται ότι θα συμπλεριληφθούν σημαντικά αποσπάσματα [33]. Επιπλέον, το RAG συμβάλλει στην επαλήθευση γεγονότων και φημών, αφού ενισχύει τον εντοπισμό αξιόπιστων πηγών για να επιβεβαιώσει ή διαψεύσει ισχυρισμούς. Παράλληλα, μπορεί να έχει καλές εφαρμογές σε εταιρικά δεδομένα, αφού εταιρίες μπορούν να κάνουν εύκολα ερωτήσεις πάνω στα δεδομένα της εταιρίας, χωρίς φόβο για ανακρίβειες ή απουσίας σημαντικών πληροφοριών. Κάτι άλλο που θεωρήσαμε πολύ σημαντικό, είναι η εφαρμογή του RAG στον τομέα της εκπαίδευσης. Οι τεκμηριωμένες ερωτήσεις και απαντήσεις που μπορεί να παράγει, έχουν τεράστια σημασία και διευκολύνουν σημαντικά τον χρήστη σε εκπαιδευτικές εφαρμογές όπως κουίζ και γενικά συστήματα μελέτης. Τέλος, η Παραγωγή Ενισχυμένη Με Ανάκτηση, μπορεί να βοηθήσει στην εκμετάλλευση εξειδικευμένων βάσεων γνώσης για την βελτιστοποίηση ενός μοντέλου στον αντίστοιχο τομέα [32] (πχ ιατρική, νομική).

Στο πλαίσιο της διπλωματικής αυτής εργασίας δώθηκε ιδιαίτερη έμφαση στην εφαρμογή του RAG στον χώρο των συστημάτων ερωταπαντήσεων, συνδιάζοντας τον με τον χώρο της εκπαίδευσης για να υπάρχει ένα καλύτερο αποτέλεσμα τόσο στην εγκυρότητα των παραγόμενων ερωτήσεων που βλέπει ο χρήστης, όσο και στον εντοπισμό των κατάλληλων και σωστών απαντήσεων για την κάθε ερώτηση. Χωρίς το Retrieval Augmented Generation, θα ήταν πολύ δύσκολο να επιτευχθεί αυτός ο στόχος, γεγονός που κάνει την συμβολή του στο έργο μας τεράστια. Οι δυνατότητες που παρέχει η εφαρμογή του είναι πολλές και βοηθούν σε πολλούς τομείς, αλλά είναι σημαντικό να γίνουν και βελτιώσεις σε αυτό έτσι ώστε η χρήση του να επιφυλλάσει ακόμα λιγότερες προκλήσεις, και να προσφέρει ακόμα περισσότερα θετικά αποτελέσματα [25].

2.2.2 Σύγχρονα Συστήματα RAG

Τα τελευταία χρόνια έχουν προταθεί πολυάριθμες παραλλαγές και εξελίξεις των συστημάτων Retrieval-Augmented Generation, με σκοπό να ξεπεραστούν οι κυριότεροι περιορισμοί των Μεγάλων Γλωσσικών Μοντέλων και συνεπώς να βελτιωθεί η ακρίβεια, η επεκτασιμότητα και η αξιοπιστία των απαντήσεων που παράγουν. Τα σύγχρονα RAG μοντέλα δεν περιορίζονται πλέον σε μία απλή διαδικασία ανάκτησης και σύνθεσης κειμένου, αλλά εισάγουν μηχανισμούς αξιολόγησης των ανακτημένων τεκμηρίων, επαναληπτική ή διορθωτική ανάκτηση, καθώς και μηχανισμούς αυτοελέγχου. Η εξέλιξη αυτή έχει οδηγήσει στην ανάπτυξη συστημάτων που μπορούν να μειώνουν σε μεγάλο βαθμό τις παραισθήσεις ενός LLM και να προσφέρουν καλύτερη και περισσότερο ολοκληρωμένη ανάκτηση και επαλήθευση πληροφορίας. Παρακάτω θα αναφερθούν μερικά συστήματα που επισημάνθηκαν.

Το πρώτο, άξιο αναφοράς, σύστημα RAG που μελετήθηκε είναι το Self-RAG [2]. Το Self-RAG αποτελεί ένα από τα πιο επιδραστικά μοντέλα RAG νέας γενιάς, καθώς εισάγει την έννοια της ενδοσκόπησης (self-reflection) στα συστήματα ανάκτησης και παραγωγής. Ενώ τα κλασικά, τέτοιου είδους, μοντέλα στηρίζονται σε σταθερή διαδικασία ανάκτησης και σε ένα γλωσσικό μοντέλο το οποίο αξιοποιεί τα ανακτημένα τεκμήρια, το Self-RAG επιτρέπει στο ίδιο το LLM να ελέγχει πότε χρειάζεται ανάκτηση καθώς και να επανα-εξετάζει τις απαντήσεις του και να τις βελτιώνει. Για να το επιτύχει αυτό, εισάγει ειδικά ενδοσκοπικά σύμβολα (reflection tokens), με τα οποία το μοντέλο σηματοδοτεί τρεις ενέργειες: ¨ανέκτησε¨(retrieve), ¨παρήγαγε¨(generate) και ¨κριτίκαρε' (critique). Το Μοντέλο εκπαιδεύεται έτσι ώστε να μαθαίνει όχι μόνο να απαντά, αλλά και να αποφασίζει αν διαθέτει επαρκή γνώση ή αν πρέπει να εκτελέσει νέο κύκλο ανάκτησης τεκμηρίων. Κατά τη φάση της κριτικής, το μοντέλο αξιολογεί τη δική του απάντηση ως προς την ορθότητα και την επάρκεια των πηγών, δημιουργώντας έναν εσωτερικό μηχανισμό αυτοδιόρθωσης. Η αρχιτεκτονική περιλαμβάνει δύο κύριες συνιστώσες: έναν ανακτητή(retriever) που μπορεί να ενημερώνεται συνεχώς και ένα LLM ενισχυμένο με επίπεδα ενδοσκόπησης, τα οποία επιτρέπουν εναλλαγή μεταξύ σκέψης και ανάκτησης. Το Self-RAG εκπαιδεύτηκε πάνω σε σύνολα δεδομένων ανοικτού χώρου και αξιολογήθηκε επαρκώς. Τα αποτελέσματα έδειξαν βελτιωμένη ακρίβεια, λιγότερες παραισθήσεις και υψηλότερη ποιότητα παραπομπών(citation accuracy) σε σχέση με γνωστά μοντέλα. Συμπερασματικά, το Self-RAG αποτελεί σημαντικό βήμα προς συστήματα που δεν εξαρτώνται μόνο από τα τεκμήρια, αλλά μαθαίνουν να αποφασίζουν, να αυτο-αξιολογούν και να αυτο-βελτιώνονται συνεχώς, γεγονός που ενισχύει την ακρίβεια του τελικού αποτελέσματος.

Ένα άλλο έργο που μελετήθηκε, είναι το Corrective Retrieval-Augmented Generation (CRAG) [64]. Στο έργο αυτό, προτείνεται ένας μηχανισμό που επιτρέπει στα RAG συστήματα να ανιχνεύουν και να διορθώνουν λάθη ανάκτησης σε πραγματικό χρόνο. Στην ουσία, το πρόβλημα που επιλύει αυτή η μεθοδολογία είναι ότι σε ένα απλό σύστημα RAG, η διαδικασία ανάκτησης μπορεί να επιστρέψει άσχετα ή μερικώς σχετικά έγγραφα, οδηγώντας το μοντέλο σε παραισθήσεις ή ανακριβείς απαντήσεις. Για να το αντιμετωπίσει αυτό, το CRAG ενσωματώνει έναν Αξιολογητή Ανάκτησης(Retrieval Evaluator), ένα ελαφρύ μοντέλο ταξινόμη-

σης που αξιολογεί τα ανακτημένα τεκμήρια ως σωστά(correct), ασαφή(ambiguous) ή λανθασμένα(incorrect) με βάση την σημασιολογική ομοιότητα με την ερώτηση. Όταν ο αξιολογητής ανιχνεύσει ανεπαρκή τεκμηρίωση, ενεργοποιεί μια διορθωτική φάση (Corrective Retrieval), η οποία αναζητά πρόσθετα αποσπάσματα είτε από τη βάση γνώσης είτε μέσω αναζήτησης στο διαδίκτυο, βελτιώνοντας δυναμικά το σύνολο των πηγών. Στη συνέχεια, εφαρμόζεται ένας μηχανισμός Σύνθεσης-Αποσύνθεσης(Decompose–Recompose), όπου το σύστημα τεμαχίζει τις νέες πηγές σε μικρότερα τμήματα, φιλτράρει τα περιττά ή αντιφατικά δεδομένα και ανασυνθέτει ένα συμπυκνωμένο συμφραζόμενο(context) που τροφοδοτείται ξανά στο LLM για να παραχθεί η τελική απάντηση. Η μέθοδος αυτή δεν απαιτεί κάποια επανεκπαίδευση του μοντέλου και έχει αποδειχθεί πως βελτιώνει την απόδοση μοντέλων σε ορισμένα σύνολα δεδομένων. Αυτό, σε συνδιασμό με όλα τα χαρακτηριστικά που αναφέρθηκαν παραπάνω, την καθιστά ιδιαίτερα χρήσιμη σε μείωση παραισθήσεων σε συστήματα βασισμένα σε LLM, που στοχεύουν σε χρήση πραγματικού χρόνου.

Παράλληλα, στο πλαίσιο των συστημάτων RAG, μελετήθηκε και το έργο RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval [48]. Το RAPTOR αποτελεί μια προσέγγιση στον χώρο των συστημάτων RAG, σχεδιασμένη για να αντιμετωπίσει ένα από τα βασικότερα προβλήματα της παραδοσιακής ανάκτησης, την αδυναμία των μοντέλων να διαχειριστούν εκτενή και πολυεπίπεδα έγγραφα χωρίς να χάνουν σημαντικό κομμάτι πληροφορίας. Σε αντίθεση με τις απλές μεθόδους που ανακτούν λίγα συνεχόμενα αποσπάσματα, το RAPTOR οργανώνει τα δεδομένα σε μια ιεραρχική δενδροειδή δομή που συνδυάζει περιλήψεις σε πολλαπλά επίπεδα αφαίρεσης. Η διαδικασία ξεκινά με τον τεμαχισμό των κειμένων σε μικρά τμήματα περίπου εκατό συμβόλων tokens, τα οποία υφίστανται ενσωμάτωση (embedding) μέσω μοντέλων όπως το SBERT. Συγκεκριμένα, το SBERT (Sentence-BERT) [21] είναι ένα μοντέλο βασισμένο στην αρχιτεκτονική του BERT, το οποίο έχει προσαρμοστεί ώστε να παράγει πυκνές σημασιολογικές αναπαραστάσεις (sentence embeddings) για ολόκληρες προτάσεις ή παραγράφους. Εκπαιδεύεται με τρόπο ώστε οι σημασιολογικά παρόμοιες προτάσεις να βρίσκονται κοντά στο διανυσματικό χώρο. Μετά τον τεμαχισμό αυτόν, τα αποσπάσματα ομαδοποιούνται με χρήση Γκαουσιανών Μοντέλων που κάνουν ομαλή ομαδοποίηση(Gaussian Mixture Models - soft clustering) ώστε να επιτρέπεται η πολλαπλή θεματική συμμετοχή κάθε κομματιού κειμένου chunk και έπειτα συνοψίζονται αφαιρετικά (abstractive summarization) από ένα LLM. Αυτή η διαδικασία επαναλαμβάνεται αναδρομικά και οι νέες περιλήψεις υποβάλλονται ξανά σε ενσωμάτωση, ομαδοποίηση(clustering) και σύνοψη, μέχρι να προκύψει ένα πλήρες δέντρο από λεπτομερείς κόμβους-φύλλα έως έναν συνολικό κόμβο-ρίζα. Κατά τη φάση ανάκτησης, το σύστημα χρησιμοποιεί δύο στρατηγικές: τη δενδρική διάσχιση (tree traversal), όπου αναζητώνται οι πιο σχετικά κόμβοι σε κάθε επίπεδο, και τη συγκεντρωτική προσέγγιση (collapsed tree), όπου αξιολογούνται όλοι οι κόμβοι ως ενιαίο σύνολο. Με αυτόν τον τρόπο, το RAPTOR παρέχει στο LLM συμφραζόμενα (context) και στοιχεία που περιλαμβάνουν τόσο λεπτομερείς πληροφορίες όσο και υψηλού επιπέδου σύνοψη, επιτρέποντας την καλύτερη κατανόηση μεγάλων εγγράφων. Συνεπώς, το RAPTOR αποδεικνύεται ιδιαίτερα αποτελεσματικό σε εφαρμογές όπου απαιτείται συνδυασμός απομακρυσμένων πληροφοριών ή ανάλυση

μεγάλων αναφορών και τεχνικών κειμένων.

Μελετήθηκε, επιπλέον, και το έργο RAFT: Adapting Language Model to Domain Specific RAG [67]. Η εργασία αυτή, εξετάζει πως η βελτιστοποίηση (fine-tuning) μπορεί να ενισχύσει την διαδικασία του RAG, υπο ορισμένες συνθήκες. Συγκεκριμένα, το RAFT (Retrieval-Augmented Fine-Tuning) αποτελεί μια μέθοδο εκπαίδευσης που επιτρέπει στα Μεγάλα Γλωσσικά Μοντέλα να προσαρμόζονται σε εξειδικευμένους τομείς γνώσης, βελτιώνοντας την ικανότητά τους να αξιοποιούν τις πληροφορίες που προέρχονται από εξωτερικά συστήματα ανάκτησης. Αντί να εκπαιδεύει ξανά ολόκληρο το pipeline ενός RAG, το RAFT επικεντρώνεται στην βελτιστοποίηση του ίδιου του LLM πάνω σε δεδομένα τύπου ¨ανοικτού βιβλίου' (open-book), όπου σε κάθε παράδειγμα δίνεται μια ερώτηση, ένα σύνολο εγγράφων που περιλαμβάνει τόσο σχετικά τεκμήρια (golden documents) όσο και παραπλανητικά (distractors), καθώς και η σωστή απάντηση. Κατά την εκπαίδευση, το μοντέλο μαθαίνει να τεκμηριώνει τις απαντήσεις του χρησιμοποιώντας μόνο τις σωστές πηγές και να αγνοεί τα μη σχετικά τεκμήρια, αποκτώντας έτσι ανθεκτικότητα σε περιπτώσεις ατελούς ή θορυβώδους ανάκτησης. Είναι σημαντικό να αναφερθεί, πως το RAFT στηρίζεται σε επιβλεπόμενη βελτιστοποίηση (Supervised Fine-Tuning/SFT) με ρητά επισημασμένα RAG παραδείγματα, που εντόπισαν/συνέθεσαν οι συγγραφείς, ώστε να ενισχύεται η σύνδεση ανάμεσα σε ανάκτηση και παραγωγής απάντησης. Συνολικά, η μέθοδος που παρουσιάζεται στο έργο δίνει μια εικόνα για το πως μπορούν να αξιοποιηθούν οι βελτιστοποιήσεις των Μεγάλων Γλωσσικών Μοντέλων σε συστήματα RAG για ενίσχυση των αποτελεσμάτων σε συγκεκριμένα πεδία.

Ένα άλλο σημαντικό κομμάτι έρευνας που αξίζει να αναφερθεί στο πλαίσιο των RAG συστημάτων, είναι το SELF-ROUTE [34]. Στο έργο αυτό, πραγματοποιείται μια συστηματική σύγκριση μεταξύ των συστημάτων RAG και των Εκτενών Συμφραζομένων Μεγάλων Γλωσσικών Μοντέλων (Long-Context LLMs/LC), όπως τα Gemini 1.5 Pro, GPT-4o . Τα LC LLMs είναι Μεγάλα Γλωσσικά Μοντέλα που μπορούν να επεξεργαστούν εισόδους με πολύ μεγάλο μέγεθος συμφραζομένων (context window), συχνά δεκάδες χιλιάδες έως εκατοντάδες χιλιάδες tokens χωρίς μηχανισμό ανάκτησης, ώστε να χειριστούν εκτενή έγγραφα, πολύπλοκες συνομιλίες ή συλλογιστικά σενάρια που ξεπερνούν τις δυνατότητες των κλασικών μοντέλων [9]. Οι συγγραφείς αξιολογούν τις δύο προσεγγίσεις μετρώντας ακρίβεια (accuracy), κόστος υπολογισμού (efficiency) και αλληλοεπικάλυψη (overlap). Τα αποτελέσματα δείχνουν ότι, όταν υπάρχουν επαρκείς υπολογιστικοί πόροι, τα LC LLMs ξεπερνούν τα RAG σε μέση απόδοση. Ωστόσο, τα RAG παραμένουν πολύ οικονομικότερα, καθώς μειώνουν δραστικά το μήκος εισόδου και συνεπώς το υπολογιστικό κόστος, το οποίο στα σύγχρονα APIs εξαρτάται γραμμικά από τα tokens. Οι ερευνητές διαπιστώνουν, παράλληλα, ότι οι προβλέψεις των δύο συστημάτων είναι ταυτόσημες σε πάνω απο τα μισά ερωτήματα, γεγονός που υποδεικνύει ότι σε πολλές περιπτώσεις η πλήρης επεξεργασία μακρών συμφραζομένων είναι περιττή. Με βάση αυτή την παρατήρηση προτείνεται η μέθοδος SELF-ROUTE, ένας υβριδικός μηχανισμός δρομολόγησης ερωτημάτων που επιτρέπει στο ίδιο το LLM να αποφασίζει (μέσω ενδοσκόπησης) αν μια ερώτηση πρέπει να απαντηθεί μέσω RAG ή μέσω Long-Context προτροπής. Το SELF-

ROUTE επιτυγχάνει αντίστοιχη ακρίβεια με τα LC μοντέλα και με αισθητά μικρότερο κόστος. Το πόρισμα που αντλείται από την μελέτη της συγκεκριμένης εργασίας, είναι πως η εφαρμογή RAG μπορεί να έχει καλύτερα αποτελέσματα απο χρήση LLMs. Επίσης, γίνεται εμφανές πως ο συνδιασμός συστημάτων RAG με ισχυρά LLMs αποδίδει καλύτερα αποτελέσματα απο την μεμονομένη χρήση των μοντέλων, γεγονός που δείχνει την σημασία του RAG στον τομέα της ενίσχυσης τις ακρίβειας και της εξασφάλισης της τεκμηρίωσης, στον χώρο των Μεγάλων Γλωσσικών Μοντέλων.

Η ύπαρξη των αναφερόμενων εργασιών, καθώς και άλλων, δείχνει πόσο σημαντικό πεδίο έρευνας είναι το Retrieval-Augmented Generation. Η συνεχής εξέλιξη των RAG συστημάτων αποδεικνύει ότι η απλή χρήση των μεγάλων γλωσσικών μοντέλων δεν επαρκεί για την παραγωγή αξιόπιστης και τεκμηριωμένης γνώσης. Μέσα από την ενσωμάτωση μηχανισμών ανάκτησης, φιλτραρίσματος και γείωσης της πληροφορίας, τα RAG μοντέλα γεφυρώνουν τη διαφορά ανάμεσα στη γλωσσική ικανότητα των LLMs και στην ανάγκη για ακρίβεια και τεκμηρίωση. Παράλληλα, αποτελούν τη βάση για πλήθος σύγχρονων ερευνητικών κατευθύνσεων που μπορεί να αφορούν υβριδικές προσεγγίσεις ή εντελώς νέες ιδέες και καινοτομίες. Η πρόοδος σε αυτόν τον τομέα μπορεί να αποτελέσει θεμέλιο για περαταίρω εφαρμογές σε κρίσιμα πεδία, όπως η εκπαίδευση, η ιατρική και η νομική. Σε αυτή την ιδέα εστίασε και η παρούσα διπλωματική εργασία, αφού μέσω του συστήματος RAG που αναπτύχθηκε γίνεται προσπάθεια ενίσχυσης του τομέα της εκπαίδευσης, ο οποίος παρουσιάζει ιδιαίτερη ανάγκη όσον αφορά την ακρίβεια και την τεκμηρίωση των δεδομένων.

2.3 Επέκταση Ερωτημάτων

Η Επέκταση Ερωτημάτων (Query Expansion) είναι μια κλασική τεχνική στον χώρο της Ανάκτησης Πληροφορίας και της Επεξεργασίας Φυσικής Γλώσσας, που στοχεύει κυρίως στη βελτίωση της κατανόησης ερωτημάτων και απόδοσης των συστημάτων αναζήτησης και ερωταπαντήσεων. Στην απλούστερη μορφή της, περιλαμβάνει τον εμπλουτισμό του αρχικού ερωτήματος με επιπλέον λέξεις ή φράσεις που θεωρούνται σχετικές, ώστε να αυξηθεί η πιθανότητα ταυτοποίησης σχετικών εγγράφων στη βάση δεδομένων αλλά και η συνολική ποιότητα και διαφάνεια του ερωτήματος [58]. Με την έλευση των Μεγάλων Γλωσσικών Μοντέλων, η έννοια αυτή έχει αποκτήσει νέα διάσταση. Τα LLMs μπορούν να δημιουργήσουν συμφραζόμενα, παραφράσεις ή ακόμη και πλήρη υποθετικά έγγραφα (pseudo-documents) που βελτιώνουν την ανάκτηση. Για αυτό, πλέον, είναι σημαντική και συχνή η χρήση της επέκτασης ερωτημάτων σε εφαρμογές βελτιστοποίησης Μεγάλων Γλωσσικών Μοντέλων [66].

Κάποιες γενικές δυνατότητες της Επέκτασης Ερωτημάτων είναι οι εξής:

1. Εύρεση και αναζήτηση συνώνυμων
2. Εύρεση σχετικών λέξεων (πχ αντώνυμα, υπερώνυμα)
3. Εύρεση όλων των διαφόρων μορφολογικών τύπων λέξεων μέσω της αποκοπής stemming κάθε λέξης στο ερώτημα αναζήτησης
4. Διόρθωση ορθογραφικών λαθών και αυτόματη αναζήτηση για τη διορθωμένη μορφή ή πρόταση της στα αποτελέσματα
5. Επαναστάθμιση (re-weighting) των όρων στο αρχικό ερώτημα

Η εφαρμογή της τεχνικής της Επέκτασης Ερωτημάτων έχει εξελιχθεί σε σχέση με τα παλαιότερα χρόνια. Οι αρχικές προσεγγίσεις αφορούσαν κυρίως εισαγωγή λέξεων που σχετίζονται σημασιολογικά με τους όρους του ερωτήματος. Επίσης, γινόταν χρήση των κορυφαίων αποτελεσμάτων μιας αρχικής αναζήτησης για την εξαγωγή επιπλέον όρων, την λεγόμενη τεχνική της ψευδο-ανατροφοδότησης (pseudo-relevance feedback) [55, 5]. Χρησιμοποιούνταν επίσης, κάποια στατιστικά μοντέλα, όπως το μοντέλο διανυσματικού χώρου (vector-space), που αναπροσαρμόζουν τα βάρη των όρων με βάση θετικά/αρνητικά παραδείγματα [24]. Αν και αυτές οι μέθοδοι αύξαναν την ανάκληση, συχνά υπέφεραν από μείωση της ακρίβειας, αφού η προσθήκη ακατάλληλων όρων οδηγούσε σε θόρυβο. Το πρόβλημα αυτό, έρχεται να λύσει η εφαρμογή την Επέκτασης Ερωτημάτων στο πλαίσιο των Μεγάλων Γλωσσικών Μοντέλων. Εφόσον πλέον δεν υπάρχει μονάχα η έννοια της Ανάκτησης Πληροφορίας, αλλά και ερωτήσεις-απαντήσεις μεταξύ χρήστη και μοντέλου, το LLM μπορεί να κάνει query expansion «απο μόνο του», δηλαδή να αναδιατυπώσει την ερώτηση σε διαφορετική μορφή ή/και να προσθέσει λεπτομέρειες σε αυτήν [66].

Γίνεται εμφανές, λοιπόν, πως η τεχνική της επέκτασης ερωτημάτων μπορεί να αποδειχθεί πολύτιμη στην επίλυση προβλημάτων στον χώρο των LLMs, και συγκεκριμένα στο κομμάτι της εφαρμογής του RAG [47]. Αυτό ισχύει, για τους παρακάτω λόγους:

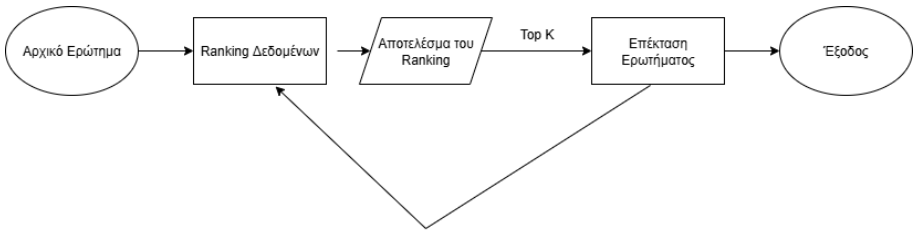
1. Μείωση Παραισθήσεων: Όταν το αρχικό ερώτημα είναι ασαφές ή υπερβολικά γενικό, οι πιθανότητες το LLM να «φανταστεί» πληροφορίες αυξάνονται. Η επέκταση βοηθά στην καλύτερη συμφωνία με σχετικά αποσπάσματα, μειώνοντας έτσι τις παραισθήσεις.

2. Κάλυψη Ποικιλίας Διατυπώσεων: Οι χρήστες μπορούν να εκφράζουν το ίδιο αίτημα με διαφορετικούς όρους. Η επέκταση διευρύνει τον χώρο του ερωτήματος ώστε να εντοπίζονται σχετικές πηγές ακόμη και όταν το αρχικό ερώτημα ήταν περιορισμένο.

3. Βελτίωση Απόδοσης Ανακτητή(Retriever): Σε αραιούς ανακτητές(sparse retrievers) (π.χ. BM25), η προσθήκη συνωνύμων αυξάνει την πιθανότητα λεξικής αντιστοίχισης. Σε πυκνούς ανακτητές(dense retrievers), τα ψευδό-κείμενα(pseudo-documents) οδηγούν σε αναπαραστάσεις(embeddings) πιο κοντά στις σχετικές απαντήσεις.

4. Ευελιξία σε Συγκεκριμένα Πεδία(Domains): Η επέκταση επιτρέπει σε ένα μοντέλο να προσαρμοστεί καλύτερα σε ειδικά πεδία (ιατρική, νομική κ.λπ.), αφού οι όροι του ερωτήματος μπορούν να εξειδικευτούν αυτόματα.

Μια τυπική απεικόνηση για το πως μπορεί να λειτουργεί ένα σύστημα RAG, με Επέκταση ερωτημάτων είναι η εξής:



Σχήμα 2.2: Αρχιτεκτονική ενός συστήματος RAG, με επέκταση ερωτημάτων Το αρχικό ερώτημα μπαίνει ως είσοδος. Μετά, ακολουθεί η κατάταξη των σχετικών με το ερώτημα δεδομένων. Απο το αποτέλεσμα, αντλούνται τα k καλύτερα, και γίνεται η επέκταση ερωτήματος. Υπάρχει η δυνατότητα να γίνει ξανά επέκταση στο ερώτημα για καλύτερο αποτέλεσμα. Τέλος,στην έξοδο υπάρχει το αποτέλεσμα

Είναι σημαντικό να αναφερθεί, πως ο καλύτερος τρόπος να γίνει η αναφερόμενη επέκταση ερωτήματος σε σύστημα RAG, είναι με την χρήση του LLM. Εφόσον το μοντέλο έχει στην διάθεση του τα σχετικά κείμενα για κάθε ερώτημα, η διαδικασία της επέκτασης αυτοματοποιείται και προκύπτει το θεμιτό αποτέλεσμα.

Συνδιάζοντας τα παραπάνω, σε αυτή την εργασία, πραγματοποιήθηκε μια εστιαζόμενη στην ανάκτηση (retireval-based) προσέγγιση για την επέκταση ερωτημάτων. Συγκεκριμένα, η αρχική διατύπωση του ερωτήματος χρησιμοποιείται για να ανακτηθούν σχετικά κείμενα ή αποσπάσματα από τη βάση γνώσης. Στη συνέχεια, τα περιεχόμενα των αποσπασμάτων αξιοποιούνται για να εμπλουτιστεί ή να αναδιατυπωθεί το αρχικό ερώτημα. Η λογική είναι ότι τα σχετικά κείμενα παρέχουν συμφραζόμενα που συχνά απουσιάζουν από την αρχική ερώτηση και καθιστούν δυνατή μια πιο ακριβή και τεκμηριωμένη διατύπωση. Στη βιβλιογραφία, αυτή η διαδικασία σχετίζεται στενά με την έννοια της ψευδό-ανατροφοδότησης (pseudo-relevance feedback) [26], όπου τα αποτελέσματα της πρώτης αναζήτησης χρησιμοποιούνται για να αντληθούν νέοι όροι που τροφοδοτούν μια δεύτερη, βελτιωμένη αναζήτηση. Στην πραγματικότητα, τα ανακτημένα αποσπάσματα δεν παρέχουν μόνο λέξεις-κλειδιά στο LLM αλλά και σημασιολογικό πλαίσιο, με βάση το οποίο το μοντέλο μπορεί να αναδιατυπώσει με ακρίβεια και σαφήνεια το ερώτημα, και στην συνέχεια να το απαντήσει κατάλληλα. Αυτά που αποσκοπεί να πετύχει η συγκεκριμένη εφαρμογή της Επέκτασης Ερωτημάτων είναι:

- να εξαλείφονται στοιχεία που δεν υποστηρίζονται από τα κείμενα και μπορεί να βρίσκονται στο αρχικό ερώτημα, και συνεπώς να μειώνονται οι παραισθήσεις του μοντέλου
- να εμπλουτίζονται με σχετικές πληροφορίες, για να γίνεται το ερώτημα πιο σαφής και εύκολο στην απάντηση
- να δημιουργείται μια πιο αξιόπιστη βάση για την παραγωγή απάντησης απο το μοντέλο αφού ακόμη και σε περιπτώσεις που η ερώτηση δεν επεκταθεί, αυτό σημαίνει ότι είναι ήδη σύμφωνη με τα κείμενα.

Συνοψίζοντας, η εφαρμογή της Επέκτασης Ερωτημάτων με τον καθοδηγούμενο απο ανάκτηση τρόπο μπορεί να έχει σημαντικά οφέλη, αφού η καθοδήγηση απο λέξεις-κλειδία αντικαθιστάται πλεον απο ολοκληρωμένα αποσπάσματα κειμένων τα οποία λειτουργούν ως εξωτερικές πηγές γνώσης για το LLM. Ως αποτέλεσμα, το μοντέλο θα παράξει καλύτερο αποτέλεσμα, αν κριθεί απαραίτητο με βάση τα τεκμήρια απο ανακτήθηκαν. Σε κάθε περίπτωση, αυξάνεται η ακρίβεια και η τεκμηρίωση σύμφωνα με τα δωθέντα κείμενα, γεγονός που αποτελεί τον κυριότερο στόχο της διπλωματικής αυτής.

Κεφάλαιο 3

Χρήση Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης στην Εκπαίδευση

Σε αυτό το κεφάλαιο, θα παρουσιαστούν μερικοί τρόποι που η Τεχνητή Νοημοσύνη και η Μηχανική Μάθηση βοηθούν στο να ενισχυθεί ο τομέας της εκπαίδευσης. Θα γίνει εστίαση, κυρίως σε εφαρμογές που σχετίζονται με αυτόματη δημιουργία ερωτήσεων, εξατομικευμένη μάθηση και υποστήριξη μαθητών-εκπαιδευτικών καθώς το τελικό αποτέλεσμα, μετά απο την μεθοδολογία υλοποίησης της εργασίας αυτής, χρησιμοποιήθηκε για την ανάπτυξη μιας τέτοιας εφαρμογής. Θα αναφερθούν, επίσης, και προκλήσεις που μου μπορεί να εμφανίσει η χρήση της Τεχνητής Νοημοσύνης στον χώρο της εκπαίδευσης, δίνοντας έτσι έμφαση στα υπέρ και στα κατά της. Συνολικά, για τις ανάγκες της συγκεκριμένης εργασίας, αξιοποιήθηκαν και συνδιάστηκαν σωστά οι δυνατότητες που παρέχονται από την εξέλιξη στον τομέα του AI έτσι ώστε να παρουσιάστεί ένα αποτέλεσμα που μπορεί να ενισχύσει τον τομέα της εκπαίδευσης, περιορίζοντας και τους κινδύνους απο την απερίσκεπτη χρήση της. Η τελική εφαρμογή γέννησης κουίζ, εξάλλου, είναι μια εφαρμογή που αξιοποιεί την Τεχνητή Νοημοσύνη για την παραγωγή εκπαιδευτικών δεδομένων, χωρίς να υπάρχουν σημαντικά μειονεκτήματα ή κίνδυνοι απο την χρήση της. Τα δεδομένα που βλέπει ο χρήστης έχουν περάσει απο μια διαδικασία που εξασφαλίζει την εγκυρότητα τους.

3.1 Η Τεχνητή Νοημοσύνη στην Εκπαίδευση

Η Τεχνητή Νοημοσύνη έχει ξεκινήσει να εισέρχεται στον τομέα της εκπαίδευσης, προσφέροντας νέες δυνατότητες σε μαθητές, εκπαιδευτικούς και φορείς. Τα πρώτα Εξυπνα Συστήματα Εκπαίδευσης (Intelligent Tutoring Systems -ITS) στόχευαν στην παροχή εξατομικευμένης υποστήριξης, προσαρμόζοντας την παρουσίαση υλικού στις ανάγκες κάθε μαθητή. Σήμερα, η ραγδαία εξέλιξη της Μηχανικής Μάθησης και των Μεγάλων Γλωσσικών Μοντέλων καθιστά δυνατή την ανάπτυξη πολύ πιο ευέλικτων και ισχυρών εργαλείων, τα οποία υποστηρίζουν τόσο τη μάθηση όσο και τη διδασκαλία σε πραγματικό χρόνο [1]. Ένα σημαντικό παράδειγμα είναι τα συστήματα αυτόματης αξιολόγησης. Μεθοδολογίες μηχανικής μάθησης επιτρέπουν την ανάλυση γραπτών εργασιών, την αναγνώριση μοτίβων λαθών και την παροχή στοχευμένων διορθώσεων. Επίσης, δίνεται η δυνατότητα για άμεση επικύρωση, η οποία βοηθά τον μαθητή να αντιληφθεί τα λάθη του εγκαίρως και να προσαρμόσει τον τρόπο μελέτης του. Παράλληλα, ελαφρύνει τον φόρτο εργασίας των εκπαιδευτικών, οι οποίοι μπορούν να αφιερώσουν περισσότερο χρόνο στην σωστή και ποιοτική παράδοση της ύλης στην διδασκαλία και στην κατάλληλη προετοιμασία τους. Επιπρόσθετα, τα συστήματα συνομιλίας(chatbots) αποτελούν μια άλλη διαδεδομένη εφαρμογή, που μπορεί να συμβάλλει σημαντικά στην εκπαίδευση. Αξιοποιώντας τεχνικές επεξεργασίας φυσικής γλώσσας, μπορούν να λειτουργούν ως βοηθοί που απαντούν σε ερωτήσεις μαθητών, εξηγούν έννοιες και καθοδηγούν στην αναζήτηση υλικού. Τέτοιου είδους εφαρμογές, αυξάνουν την εμπλοκή του μαθητή και ενισχύουν την αυτονομία του, ενώ ταυτόχρονα μειώνουν και τα εμπόδια πρόσβασης στη μάθηση εκτός της αίθουσας, αφού πλεόν ο μαθητής μπορεί να πάρει μια ενδεικτική απάντηση για όλα του τα ερωτήματα. Παράλληλα, υπάρχει και η αυτόματη δημιουργία μαθησιακού υλικού. Μέσω των LLMs, μπορεί να παραχθεί εξατομικευμένο περιεχόμενο όπως ερωτήσεις πολλαπλής επιλογής, ασκήσεις κριτικής σκέψης ή περιλήψεις κειμένων. Δίνεται έτσι η δυνατότητα στον εκπαιδευτικό να προσαρμόσει τη δυσκολία και τη θεματολογία στις ανάγκες διαφορετικών μαθητών ή ομάδων. Συνολικά, τα πλεονεκτήματα της Τεχνητής Νοημοσύνης στην εκπαίδευση είναι αρκετά. Μπορούν να προσφέρουν πολλά στον τομέα της εξατομίκευσης, της προσβασιμότητας και της διευκόλυνσης και απλούστευσης της διαδικασίας μάθησης. Είναι σημαντικό, λοιπόν, να αξιοποιηθεί περαιτέρω και πιο ενεργά στην διαδικασία εκμάθησης.

3.2 Μηχανική Μάθηση και Προσαρμοστική Μάθηση

Η Μηχανική Μάθηση έχει καταστήσει δυνατή την ανάπτυξη προσαρμοστικών μαθησιακών περιβαλλόντων. Τα μοντέλα αναλύουν δεδομένα μαθητών, όπως επίπεδο επίδοσης, ταχύτητα κατανόησης, τύπο και συχνότητα λαθών και προσαρμόζουν τη ροή της διαδικασίας εκμάθησης. Έτσι, κάθε μαθητής μπορεί να λαμβάνει εξατομικευμένη μαθησιακή εμπειρία, αντί για μια γενική και ομοιόμορφη παρουσίαση ύλης. Πλατφόρμες όπως το Coursera, το Khan Academy και το Duolingo ενσωματώνουν ήδη αλγορίθμους προσαρμογής στην κύρια ροή λειτουργίας τους. Συγκεκριμένα, το Duolingo παρακολουθεί συνεχώς την επίδοση κάθε χρήστη και προσαρμόζει το είδος και τη δυσκολία των ασκήσεων, με στόχο τη βέλτιστη πρόοδο. Αυτό υλοποιείται μέσα από έναν βρόγχο ανατροφοδότησης (feedback loop), όπου τα δεδομένα που προκύπτουν από τις απαντήσεις του χρήστη επηρεάζουν την επιλογή του επόμενου περιεχομένου. Συνεπώς, μετά απο κάθε ερώτηση το σύστημα αντιδρά ανάλογα. Ένα ακόμη πλεονέκτημα είναι η ικανότητα των συστημάτων Μηχανικής Μάθησης να αναγνωρίζουν μοτίβα μάθησης. Για παράδειγμα, μπορούν να εντοπίσουν αν ένας μαθητής έχει δυσκολία σε συγκεκριμένη κατηγορία προβλημάτων και να δώσουν έμφαση σε εκείνη την περιοχή [62]. Με αυτόν τον τρόπο, τα μοντέλα αυτά ενισχύουν την αποτελεσματικότητα της εκπαιδευτικής διαδικασίας, βελτιώνουν την εμπλοκή του μαθητή και μειώνουν το ποσοστό εγκατάλειψης, αφού ενισχύεται σημαντικά και η ψυχολογία του βλέποντας ό,τι αναγνωρίζει τα λάθη του, και βελτιώνεται πάνω σε αυτά.

Αν και αυτή η δυνατότητα των συστημάτων Μηχανικής Μάθησης είναι πολύ σημαντική, η παρούσα εργασία επικεντρώνεται στη βελτίωση της ακρίβειας και αξιοπιστίας των παραγόμενων ερωτήσεων και απαντήσεων(δηλαδή του υλικού μάθησης). Όσο σημασία και να έχει η προσαρμοστικότητα, είναι σημαντικό να δωθεί μεγάλη έμφαση και στην ακρίβεια έτσι ώστε να υπάρχει βεβαιότητα για την εγκυρότητα των πληροφοριών που έχει στην διαθεσή του ο μαθητής, αλλά και να εξασφαλιστεί η αξιοπιστία του συστήματος εκπαίδευσης και κατά συνέπεια της ίδιας της διαδικασίας εκπαίδευσης. Σε μελλοντική επέκταση του έργου, θα μπορούσαν να εφαρμοστούν και τεχνικές προσαρμοστικής μάθησης και εξατομίκευσης, για ένα πιο προσωποποιήμενο τελικό αποτέλεσμα.

3.3 Μεγάλα Γλωσσικά Μοντέλα στην Εκπαίδευση

Τα τελευταία χρόνια, τα Μεγάλα Γλωσσικά Μοντέλα, έχουν φέρει και αυτά πρόοδο και αλλαγές στον χώρο της εκπαίδευσης [57]. Η ικανότητά τους να κατανοούν και να παράγουν φυσική γλώσσα σε υψηλό επίπεδο έχει ανοίξει τον δρόμο για εφαρμογές που άλλοτε απαιτούσαν ανθρώπινη εργασία και παρέμβαση, αλλά και πολύ χρόνο προετοιμασίας. Μια από τις πιο χαρακτηριστικές χρήσεις είναι η αυτόματη δημιουργία ερωτήσεων και απαντήσεων. Ένα LLM μπορεί να συνθέσει κουίζ, εξετάσεις ή φύλλα ασκήσεων σε ελάχιστο χρόνο. Σε σύγκριση με παραδοσιακές μεθόδους, το υλικό παράγεται πολύ πιο εύκολα και γρήγορα, και η ποιότητα του μπορεί να είναι ισάξια ή ακόμη και καλύτερη, ενώ πάντα υπάρχει και η δυνατότητα γρήγορης προσαρμογής σε διαφορετικά επίπεδα δυσκολίας ή γνωστικά αντικείμενα. Επιπλέον, τα LLMs μπορούν να προσφέρουν δυνατότητες εξατομίκευσης της μάθησης. Έχουν την δυνατότητα, με τις κατάλληλες οδηγίες (prompts), να προσαρμόσουν τον τρόπο παρουσίασης των πληροφοριών, να αλλάξουν το ύφος της γλώσσας ανάλογα με την ηλικία ή το επίπεδο του μαθητή και να παρέχουν εναλλακτικές διατυπώσεις για την ίδια έννοια. Με αυτόν τον τρόπο διευκολύνεται η κατανόηση, ενώ παράλληλα ενισχύεται η εμπλοκή του μαθητή στην διαδικασία εκμάθησης. Μια ακόμη σημαντική εφαρμογή είναι η υποστήριξη της αναζήτησης και της μελέτης. Χρησιμοποιώντας Μεγάλα Γλωσσικά Μοντέλα οι μαθητές μπορούν να υποβάλλουν ερωτήσεις σε φυσική γλώσσα και να λάβουν κατανοητές απαντήσεις, ακόμα και για πιο σύνθετα θέματα. Αυτό τα καθιστά χρήσιμα όχι μόνο για την παραγωγή νέου περιεχομένου αλλά και ως εργαλεία καθημερινής μελέτης και αυτοκαθοδηγούμενης μάθησης. Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει η διευκόλυνση της προσβασιμότητας . Με την χρήση LLM μπορούν να παραχθούν αυτόματα μεταφράσεις, περιλήψεις ή απλοποιημένα κείμενα, προσφέροντας σε μαθητές με διαφορετικές γλωσσικές ή γνωστικές ανάγκες ένα εργαλείο προσαρμοσμένο στις δυνατότητές τους. Έτσι, η τεχνολογία δεν περιορίζεται στη μαζική παραγωγή υλικού αλλά συμβάλλει ουσιαστικά στη δημιουργία ενός πιο δίκαιου και συμπεριληπτικού εκπαιδευτικού περιβάλλοντος, όπου όλοι έχουν την ίδια πρόσβαση στην γνώση και στην πληροφορία.

Συνολικά, τα Μεγάλα Γλωσσικά Μοντέλα μπορούν να προσφέρουν πάρα πολλά στον χώρο της εκπαίδευσης. Με την χρήση τους, ενισχύεται η διδακτική διαδικασία, διευκολύνεται η πρόσβαση στη γνώση και δίνεται η δυνατότητα για εξατομικευμένη και αποτελεσματικότερη μάθηση. Στην εργασία αυτή, δώθηκε μεγάλη έμφαση στην εξασφάλιση της ποιότητας και εγκυρότητας του παραγόμενου, από το μοντέλο, αποτελέσματος καθώς και στην γρήγορη, εύκολη και πρακτική χρήση των Μεγάλων Γλωσσικών Μοντέλων στο περιβάλλον της εκπαίδευσης.

3.4 Προκλήσεις και Περιορισμοί

Παρά τα σημαντικά οφέλη που προσφέρει η εισαγωγή της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης στην εκπαίδευση, η χρήση τους συνοδεύεται από μια σειρά προκλήσεων και περιορισμών που δεν μπορούν να αγνοηθούν. Η κατανόηση αυτών των προκλήσεων είναι απαραίτητη, ώστε να διασφαλιστεί ότι η τεχνολογία αξιοποιείται με τρόπο που πραγματικά ενισχύει τη μαθησιακή διαδικασία και δεν την υπονομεύει. Μερικές απο τις σημαντικές προκλήσεις που αναγνωρίζονται, είναι οι εξής [65]:

- Ανακρίβειες και παραισθήσεις: Ένα από τα πιο κρίσιμα προβλήματα στον χώρο των Μεγάλων Γλωσσικών Μοντέλων είναι το φαινόμενο των παραισθήσεων, δηλαδή η παραγωγή λανθασμένων ή ανυπόστατων πληροφοριών από τα μοντέλα. Στο πλαίσιο της εκπαίδευσης, αυτό μπορεί να οδηγήσει σε μετάδωση και επικύρωση λανθασμένων γνώσεων ή παραπληροφόρηση. Ειδικά όταν οι μαθητές δεν διαθέτουν ακόμη τις κατάλληλες δεξιότητες κριτικής αξιολόγησης, υπάρχει ο κίνδυνος να δεχτούν χωρίς να αμφισβητήσουν ό,τι τους παρουσιάζει το σύστημα.

- Κοινωνικο-ψυχολογικές επιπτώσεις: Η συνεχής χρήση εργαλείων Τεχνητής Νοημοσύνης ενδέχεται να οδηγήσει σε υπερβολική εξάρτηση από αυτά. Οι μαθητές μπορεί να βασιστούν υπερβολικά σε αυτόματες απαντήσεις και να παραμελήσουν την ανάπτυξη κρίσιμων δεξιοτήτων, όπως η κριτική σκέψη, η δημιουργική επίλυση προβλημάτων και η ανεξάρτητη μελέτη. Επιπλέον, η διαρκής αλληλεπίδραση με ψηφιακούς βοηθούς ενδέχεται να περιορίσει την επαφή μαθητή-καθηγητή, η οποία αποτελεί βασικό στοιχείο της εκπαιδευτικής εμπειρίας.

- Ανισότητες στην πρόσβαση: Αν και η Τεχνητή Νοημοσύνη μπορεί να διευρύνει την προσβασιμότητα στη γνώση, υπάρχει ταυτόχρονα ο κίνδυνος να εντείνει τις κοινωνικές ανισότητες. Τα συστήματα που βασίζονται σε προηγμένα μοντέλα απαιτούν ισχυρή υπολογιστική ισχύ και αξιόπιστη σύνδεση στο διαδίκτυο, κάτι που δεν είναι δεδομένο σε όλες τις σχολικές μονάδες ή για όλους τους μαθητές. Έτσι, δημιουργείται ένα νέο χάσμα ανάμεσα σε όσους έχουν πρόσβαση σε αυτά τα εργαλεία και σε όσους όχι.

- Κόστος και βιωσιμότητα: Η υλοποίηση και συντήρηση προηγμένων συστημάτων απαιτεί σημαντικούς πόρους, τόσο οικονομικούς όσο και ενεργειακούς. Η εκπαίδευση, ειδικά σε δημόσιο επίπεδο, συχνά περιορίζεται από προϋπολογισμούς γεγονός που καθιστά δύσκολη την συνολική εφαρμογή τέτοιων λύσεων σε μεγάλη έκταση, στον χώρο του εκπαιδευτικού ιδρύματος.

- Ζητήματα δεοντολογίας και διαφάνειας: Τα περισσότερα μοντέλα λειτουργούν ως «μαύρα κουτιά» (black-box), καθιστώντας αδύνατο να εξηγηθεί με σαφήνεια πώς κατέληξαν σε μια απάντηση ή σε μια πρόταση περιεχομένου. Η έλλειψη ερμηνευσιμότητας μειώνει την εμπιστοσύνη των χρηστών και δημιουργεί ερωτήματα για την αντικειμενικότητα και την αξιοπιστία των συστημάτων. Επιπλέον, ενισχύεται ο κίνδυνος μεταφοράς προκαταλήψεων που μπορεί υπάρχουν στα δεδομένα εκπαίδευσης, με αποτέλεσμα αυτές να αναπαράγονται μέσα στη μαθησιακή διαδικασία.

Είναι εμφανές, λοιπόν, πως η χρήση της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης στον μαθησιακό χώρο συνοδεύεται απο αρκετούς κινδύνους, οι οποίοι όμως μπορούν να αντιμετωπιστούν. Στο πλαίσιο αυτής της εργασίας, επιτυγχάνεται μέσω της σωστής χρήσης του LLM και της καλής εφαρμογής Retrieval Augmented Generation η βεβαίωση σε μεγάλο βαθμό, πως πολλές απο τις προαναφερόμενες προκλήσεις δεν ισχύουν στην εκπαιδευτική εφαρμογή που απαπτύχθηκε. Μελετώντας διάφορα συστήματα RAG, εξήχθη το συμπέρασμα πως σε γενικό βαθμό η τεχνική του RAG μπορεί να συμβάλλει σημαντικά στον περιορισμό, ακόμη και στην εξάλλειψη, των περιορισμών που εισάγει η χρήση συστημάτων Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης, στον χώρο της εκπαίδευσης. Συνεπώς, είναι πολύ σημαντική η σωστή εφαρμογή του στο συγκεκριμένο πλαίσιο για να επιτευχθέι η ενίσχυση της διαδικασίας εκπαίδευσης αξιοποιώντας τις δυνατότες που προσφέρουν οι νέες τεχνολογίες.

Κεφάλαιο 4

Περιγραφή Θέματος και Σχετικές Μελέτες

Σε αυτό το κεφάλαιο θα γίνει μια γενική περιγραφή του θέματος της διπλωματικής, δηλαδή του συστήματος που αναπτύχθηκε. Αρχικά, θα γίνει αναφορά σε σχετικές δημοσιεύσεις(θα δωθεί ιδιαίτερη έμφαση σε μια) που προσέλκησαν ενδιαφέρον και οδήγησαν στο να χτιστούν οι βάσεις για την μεθοδολογία που τελικά ακολουθήθηκε. Στην συνέχεια θα περιγραφεί, σε γενικό πλαίσιο, πως λειτουργεί το σύστημα και τέλος θα αναλυθεί και το που αποσκοπεί η χρήση του.

4.1 Εργασία που Προκάλεσε Ενδιαφέρον

Μια από τις εργασίες που αποτέλεσαν άμεση πηγή έμπνευσης για την προσέγγιση που ακολουθήθηκε στην παρούσα διπλωματική εργασία είναι η δημοσίευση "Generate-then-Ground in Retrieval-Augmented Generation for Multi-Hop Question Answering" [49] . Η δημοσίευση αυτή, εστιάζει σε ένα από τα κεντρικά προβλήματα των Μεγάλων Γλωσσικών Μοντέλων , την παραγωγή ανακριβών ή ατεκμηρίωτων απαντήσεων, γνωστών και ως παραισθήσεις. Παρότι τα LLMs εμφανίζουν εντυπωσιακές ικανότητες στην κατανόηση και παραγωγή φυσικής γλώσσας, συχνά αποτυγχάνουν να βασίσουν τις απαντήσεις τους σε αξιόπιστες πηγές, ιδίως όταν καλούνται να απαντήσουν ερωτήσεις γνώσης, που η απάντηση βρίσκεται μέσα σε κάποια πηγή. Οι συγγραφείς του άρθρου προτείνουν μια διαδικασία δύο βημάτων, την οποία ονομάζουν Generate-then-Ground. Σε αντίθεση με κλασικές μεθόδους Retrieval-Augmented Generation, όπου πρώτα γίνεται η ανάκτηση αποσπασμάτων και στη συνέχεια η παραγωγή απάντησης, εδώ η διαδικασία αντιστρέφεται. Συγκεκριμένα, το μοντέλο πρώτα παράγει μια αρχική εκδοχή απάντησης και στη συνέχεια αυτή η απάντηση ελέγχεται και «γειώνεται» με χρήση τεκμηρίων απο εξωτερικές πηγές μέσω της διαδικασίας ανάκτησης. Η κύρια λογική πίσω από αυτή την ιδέα είναι ότι τα Μεγάλα Γλωσσικά Μοντέλα έχουν συχνά τη δυνατότητα να συνθέτουν υποθετικές ή μερικώς σωστές απαντήσεις, ακόμη και αν δεν διαθέτουν πλήρως την απαραίτητη πληροφορία. Χρησιμοποιώντας την αρχική παραγόμενη απάντηση ως οδηγό, το σύστημα αναζητά αποσπάσματα που επιβεβαιώνουν, διορθώνουν ή απορρίπτουν το αρχικό

αποτέλεσμα. Με αυτόν τον τρόπο, η τελική απάντηση είναι πιο πιστή στις διαθέσιμες πηγές και η πιθανότητα παραισθήσεων μειώνεται σημαντικά.
Τα βασικά στοιχεία της μεθοδολογίας που προτείνουν οι συγγραφείς της μελέτης αυτής, είναι τα εξής:

1. Αρχική παραγωγή (Generate): Το μοντέλο δημιουργεί μια υποψήφια απάντηση σε ερώτηση που δέχεται. Αυτή η απάντηση μπορεί να περιέχει τόσο σωστά όσο και ατεκμηρίωτα στοιχεία. Σε αυτή την φάση, δεν έχει σημασία αν η απάντηση είναι σωστή, λάθος ή μερικά ανακριβής.

2. Ανάκτηση αποσπασμάτων (Retrieval): Με βάση την παραγόμενη απάντηση, το σύστημα ανακτά σχετικά τεκμήρια από μια βάση γνώσης ή συλλογή κειμένων. Η ανάκτηση καθοδηγείται από το περιεχόμενο της απάντησης και όχι μόνο από την αρχική ερώτηση. Επιτυγχάνεται έτσι η ανάκτηση καλύτερων τεκμηρίων, αφού πλέον είναι διαθέσιμη μια ενδεικτική απάντηση που μπορεί να βοηθήσει στον εντοπισμό, σχετικών με αυτήν, αποσπασμάτων στην συλλογή κειμένων.

3. Γείωση (Ground): Η αρχική απάντηση συγκρίνεται και συνδέεται με τα ανακτημένα τεκμήρια. Το LLM καλείται να παραγάγει μια νέα εκδοχή της απάντησης, αυτή τη φορά προσαρμοσμένη και «γειωμένη» στις πηγές. Εάν η αρχική απάντηση ήταν ελλιπής ή λανθασμένη, η διαδικασία αυτή εξασφαλίζει διόρθωση και τεκμηρίωση με βάση τα κείμενα που ανακτήθηκαν.

4. Τελική έξοδος: Το αποτέλεσμα είναι μια απάντηση που διατηρεί τη γλωσσική ποιότητα και την δυνατότητα του Μεγάλου Γλωσσικού Μοντέλου να παράγει μερικώς σωστές απαντήσεις εισάγωντας, ταυτόχρονα, την έννοια της τεκμηρίωσης και της επαλήθευσης τους. Με τον τρόπο αυτό, επιτυγχάνεται ισορροπία ανάμεσα στην φυσική ροή της λειτουργίας του LLM, δηλαδή της παραγωγής κειμένου, και την ακρίβεια των δεδομένων εξόδου μέσω της διαδικάσίας της ανάκτησης.

Η συμβολή της συγκεκριμένης μελέτης, και συνεπώς της προτεινόμενης μεθολογίας, στον χώρο της απάντησης ερωτήσεων με χρήση Μεγάλων Γλωσσικών Μοντέλων είναι τεράστια. Στην ουσία, αποδεικνύει ότι με την εφαρμογή της αντίστροφης, απο την κλασσική, σειράς βημάτων μπορεί να βελτιωθεί σημαντικά η αξιοπιστία των συστημάτων ερώτησης–απάντησης. Επιπλέον, ανοίγει τον δρόμο για μεθόδους όπου η παραγωγή και η ανάκτηση δεν είναι απομονωμένα στάδια, αλλά λειτουργούν συνδυαστικά, συμπληρώνοντας και ενισχύοντας η μία την άλλη. Είναι σημαντικό να αξιοποιηθούν τα πλεονεκτήματα και των δύο αυτών λειτουργικών φάσεων, και η μεθοδολογία αυτή το πετυχαίνει σε αξιέπαινο βαθμό.

4.2 Συσχέτιση με την Παρούσα Εργασία

Στην αρχική της μορφή, η μεθοδολογία Generate-then-Ground εφαρμόστηκε στο πεδίο της απάντησης ερωτήσεων πολλαπλών αλμάτων (multi-hop question answering), δηλαδή σε ερωτήματα που απαιτούν συνδυασμό πληροφοριών από πολλαπλά αποσπάσματα ή πηγές για να δοθεί μια ολοκληρωμένη απάντηση. Σε τέτοιες περιπτώσεις τα παραδοσιακά συστήματα RAG συχνά δυσκολεύονται, επειδή η ανάκτηση με βάση το αρχικό ερώτημα μπορεί να φέρει αποσπασματικά ή άσχετα κείμενα, χωρίς να καλύπτουν όλα τα απαραίτητα βήματα συλλογισμού για την απάντηση του εκάστοτε ερωτήματος. Όπως αναφέρθηκε και παραπάνω, οι συγγραφείς προτείνουν πρώτα τη δημιουργία μιας υποψήφιας απάντησης από το μοντέλο η οποία ακόμη κι αν δεν είναι πλήρως ακριβής, εμπεριέχει ενδείξεις για τα ενδιάμεσα στοιχεία που χρειάζονται. Αυτή η παραγόμενη απάντηση χρησιμοποιείται στη συνέχεια ως οδηγός για την ανάκτηση σχετικών αποσπασμάτων, τα οποία επιτρέπουν στο μοντέλο να «γειώσει» την τελική του απάντηση σε πραγματικά δεδομένα και να καλύψει πολλαπλά βήματα συλλογισμού. Με αυτό τον τρόπο, η μέθοδος στοχεύει να βελτιώσει την αξιοπιστία και την πληρότητα των απαντήσεων σε σύνθετα ερωτήματα.

Η παρούσα εργασία άντλησε έμπνευση από την κεντρική αυτή ιδέα και υλοποιήθηκε μια παραλλαγή της η οποία, όμως, δεν αφορά την απάντηση ερωτήσεων πολλαπλών βημάτων. Συγκεκριμένα, εξετάζεται η παραγωγή και επεξεργασία ερωτήσεων αντί για απαντήσεις. Θεωρήθηκε πως είναι σημαντικό για οποιοδήποτε πλαίσιο χρήσης, είτε για ένα απλό σύστημα ερωτήσεων-απαντήσεων, είτε κάποια συλλογή για εκπαίδευση μοντέλου, είτε οτιδήποτε άλλο, οι ερωτήσεις να είναι σωστές και βασισμένες πάντοτε σε πραγματικά δεδομένα κειμένου. Το φαινόμενο των παραισθήσεων στα Μεγάλα Γλωσσικά Μοντέλα, είναι ακόμη πιο επιβλαβές όταν αφορά την παραγόμενη ερώτηση, αφού μπορεί να παραπλανήσει και τον απλό χρήστη που αναζητά την απάντηση, αλλά και κάποιο άλλο μοντέλο ή ακόμη και το ίδιο. Έτσι, επιλέχθηκε η εστίαση στην διαδικασία παραγωγής ερωτήσεων που να είναι γειωμένες σε πραγματικά δεδομένα, ακριβείς και αξιοποιήσιμες σε οποιοδήποτε περιβάλλον. Με αυτό τον τρόπο, η φιλοσοφία του Generate-then-Ground μεταφέρεται δημιουργικά σε ένα διαφορετικό αλλά συναφές πεδίο εφαρμογής, και συμβάλλει στην επίλυση του καίριου προβλήματος των παραισθήσεων στα Μεγάλα Γλωσσικά Μοντέλα.

Συγκεκριμένα, για την ανάπτυξη της προσέγγισης που ακολουθεί η παρούσα διπλωματική εργασία πατώντας στην ιδέα του Generate-Then-Ground, υιοθετήθηκε η εξής λογική:

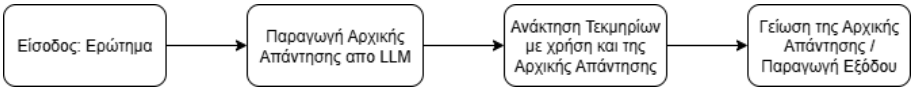
1. Παραγωγή πρόχειρων ερωτήσεων: Αρχικά δημιουργούνται αυτόματα, με LLM prompt, ερωτήσεις με βάση μια συλλογή κειμένων. Αυτές οι ερωτήσεις μπορεί να είναι αρκετά ποιοτικές(στην περίπτωση που το μοντέλο λειτούργησε σωστά), αλλά μπορεί να είναι και πρόχειρες, να περιέχουν ασάφειες ή να μην είναι πλήρως σύμφωνες με το περιεγχόμενο των εκάστοτε κειμένων.

2. Ανάκτηση σχετικών αποσπασμάτων: Για κάθε παραγόμενη ερώτηση, αναζητούνται τα πιο σχετικά αποσπάσματα από την συλλογή κειμένων.

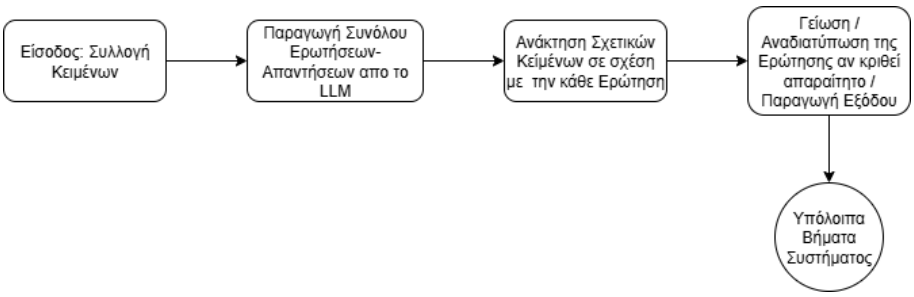
3. Γείωση/Αναδιατύπωση (Επέκταση Ερωτήματος): Οι αρχικές ερωτήσεις αναδιατυπώνονται από το μοντέλο με βάση τα ανακτημένα αποσπάσματα. Στο σημείο αυτό γίνεται, ουσιαστικά, η επέκταση ερωτήματος. Η ερώτηση είτε παραμένει ίδια, αν το μοντέλο κρίνει πως είναι σύμφωνη με τα ανακτόμενα τεκμήρια, είτε είναι καλύτερα διατυπωμένη ή περισσότερο ακριβής εφόσον πλέον είναι βασισμένη σε τεκμήρια και λιγότερο επιρρεπής σε παραισθήσεις.

Η βασική διαφοροποίηση, λοιπόν, είναι ότι ενώ το αρχικό άρθρο εστιάζει στην μεθοδολογία του Generate-Then-Ground για απαντήσεις, η παρούσα εργασία υιοθετεί την ίδια φιλοσοφία για τις ερωτήσεις. Το σύστημα που αναπτύθηκε, αντί να προσπαθεί εξαρχής να διατυπώσει τέλεια ερωτήματα, δημιουργεί πρώτα πρόχειρες εκδοχές και στη συνέχεια τις προσαρμόζει(αν κριθεί απαραίτητο) σε πραγματικά αποσπάσματα κειμένων, επιτυγχάνοντας πιο ακριβείς και χρήσιμες ερωτήσεις για οποιαδήποτε χρήση. Είναι σημαντικό να αναφερθεί, πως με αυτόν τον τρόπο η μεθοδολογία συνδέεται άμεσα με την εκπαίδευση και τη δημιουργία κουίζ. Οι ερωτήσεις που προκύπτουν είναι αξιόπιστες, στηρίζονται σε πραγματικό υλικό και αποφεύγουν τον κίνδυνο των παραισθήσεων, που μπορεί να προκύψει από την απλή χρήση Μεγάλων Γλωσσικών Μοντέλων. Η μελέτη που αναλύθηκε παραπάνω, επομένως, αποτέλεσε το θεωρητικό πλαίσιο πάνω στο οποίο σχεδιάστηκε και προσαρμόστηκε η παρούσα εργασία.

Παρακάτω απεικονείζεται σχηματικά η ροή των δύο μεθολογιών:



Σχήμα 4.1: Ροή λειτουργίας του συστήματος που προτείνουν οι συγγραφείς του αντίστοιχου έργου.



Σχήμα 4.2: Ροή λειτουργίας της προτεινόμενης παραλλαγής.

Γίνονται εμφανείς, λοιπόν, οι ομοιότητες αλλά και οι διαφορές στο πλαίσιο εφαρμογής ανάμεσα στην υπάρχουσα προσέγγιση και στην παραλλαγή που αναπτύχθηκε. Τα αποτελέσματα που σημείωσαν οι συγγραφείς απο την χρήση της μεθοδολογίας Generate-Then-Ground, ήταν αρκετά ενθαρρυντικά και παρουσιάζουν ιδιαίτερο ενδιαφέρον. Για αυτό και αντλήθηκε έμπνευση από το έργο τους, για την ανάπτυξη του συστήματος που πραγματεύεται η παρούσα εργασία. Όμως, η παραγωγή και η ενίσχυση/επέκταση ερωτήσεων δεν ήταν το μόνο κομμάτι στο οποίο δώθηκε έμφαση. Η ποιότητα και η εγγυρότητα των απαντήσεων των ενισχυμένων ερωτήσεων, είναι ένα ακόμη κομμάτι που εξετάστηκε σημαντικά.

4.3 Εργασίες Σχετικά με τον Περιορισμό Παραισθήσεων σε Συστημάτα Ερωτήσεων-Απαντήσεων

Ένα σημαντικό κομμάτι στο οποίο δώθηκε ιδιαίτερη έμφαση στην συγκεκριμένη διπλωματική, είναι ο περιορισμός των παραισθήσεων ενός Μεγάλου Γλωσσικού Μοντέλου, συγκεκριμένα σε συστήματα ερωτήσεων-απαντήσεων. Στην επίλυση αυτού του προβλήματος έχει γίνει σημαντική έρευνα και συνεχίζεται να γίνεται και σήμερα, σε μεγαλύτερο βαθμό απο ποτέ.

Μια εργασία για τον περιορισμό των παραισθήσεων σε Μεγάλα Γλωσσικά Μοντέλα είναι η δημοσίευση "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models" [36]. Η μέθοδος που προτείνουν οι συγγραφείς, αποτελεί μια μέθοδο εντοπισμού παραισθήσεων σε μοντέλα ¨μαύρα κουτιά' black-box. Στην ουσία, δειγματοληπτούνται πολλαπλές απαντήσεις του μοντέλου και μετρούνται ασυνέπειες και αντιφάσεις μεταξύ τους. Όταν αυτές οι ασυνέπεις είναι υψηλές, τότε υπάρχει σημαντική πιθανότητα παραίσθησης του μοντέλου. Η λειτουργία του μοντέλου αυτού, θα μπορούσε να προσαρμοστεί πάνω σε αξιολόγηση συστημάτων ερωτήσεων-απαντήσεων και να αντλήσει κάποια συμπεράσματα για το τι είναι παραίσθηση και τι όχι.

Η μεθοδολογία παραπάνω, αποτελεί μια ενδιαφέρουσα προσέγγιση στην επίλυση των παραισθήσεων των LLM. Για τις ανάγκες τις συγκεκριμένης εργασίας, όμως, επιλέχθηκε η χρήση του RAG για την βελτίωση του συγκεκριμένου ζητήματος αφού, σύμφωνα με την βιβλιογραφία [53], η χρήση του είναι μια απο τις καλύτερες λύσεις για την επίλυση.

Όσον αφορά την έρευνα πάνω στον εντοπισμό και περιορισμό των παραισθήσεων με χρήση RAG, έχει σημειωθεί αρκετή πρόοδος. Αρχικά είναι σημαντικό να αναφερθεί το έργο ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability [51]. Οι συγγραφείς του συγκεκριμένου έργου, υποστήριξαν πως οι παραισθήσεις σε συστήματα RAG είναι συχνό να εμφανιστούν όταν τα μοντέλα δίνουν υπερβολικό βάρος στην παραμετρική γνώση παρά στο ανακτόμενο περιεργχόμενο. Η μέθοδος που προτείνουν, αντιμετωπίζει το πρόβλημα από το εσωτερικό των LLMs, αποσυνδέοντας τη χρήση της παραμετρικής γνώσης από την εξωτερική ανάκτηση για να εντοπίσει παραισθητικές αποκρίσεις με μεγαλύτερη αποτελεσματικότητα. Επιπλέον, σε περιπτώσεις που το ζητούμενο είναι να παραχθούν δομημένες απαντήσεις, όπως ένα σύστημα ερωτήσεων-απαντήσεων, μια μελέτη που μας αφορά είναι η "Reducing hallucination in structured outputs via Retrieval-Augmented Generation" [6]. Οι συγγραφείς του συγκεκριμένου έργου δείχνουν πως αν εφαρμόστει RAG σε JSON εξόδους και η ανάκτηση γίνει σωστά, παρουσιάζεται σημαντική βελτίωση στην αποφυγή μη υπαρκτών πεδίων. Παράλληλα, στην βιβλιογραφία "Leveraging the Domain Adaptation of Retrieval Augmented Generation Models for Question Answering and Reducing Hallucination" [44], δίνεται έμφαση στο γεγονός ότι η προσαρμογή ενός συστήματος RAG σε ειδικό πλαίσιο (domain) όχι μόνο βελτιώνει την απόδοση του συστήματος ερωταπαντήσεων συνολικά, αλλά και μειώνει τις παραισθήσεις που μπορεί να υπάρχουν.

Μια άλλη φιλοσοφία στην προσέγγιση του προβλήματος των παραισθήσεων σε συστήματα RAG, αποτελεί το έργο "DelucionQA: Detecting Hallucinations in Domain-specific Question Answering" [46]. Οι συγγραφείς του, δημιουργούν ένα σύνολο δεδομένων που περιλαμβάνει ερωτήσεις και απαντήσεις σε εξειδικευμένα γνωστικά πεδία, όπου τα LLMs και τα συστήματα RAG συχνά παράγουν απαντήσεις που φαίνονται πειστικές αλλά είναι ανακριβείς ή ελλιπείς. Το σύνολο δεδομένων αυτό, στρέφει το βλέμα στο ότι πρέπει να γίνεται έλεγχος στην αξιοπιστία όλου του συστήματος ερωτήσεων-απαντήσεων. Παράλληλα, η εργασία "Hallucination Detection in Large Language Models with Metamorphic Relations" [63] προτείνει την μεθοδολογία MetaQA, μια μέθοδο ανίχνευσης παραισθήσεων χωρίς εξωτερικούς πόρους. Η βασική ιδέα εδώ, είναι η χρήση μεταμορφικών σχέσεων. Το ίδιο ερώτημα παραφράζεται ή μετασχηματίζεται σε διαφορετικές μορφές, και στη συνέχεια ελέγχεται αν οι απαντήσεις του μοντέλου παραμένουν συνεπείς μεταξύ τους. Αν εντοπιστούν ασυνέπειες, αυτό αποτελεί ένδειξη ότι η αρχική απάντηση εμπεριέχει παραισθήσεις.

Εκτός απο τον εντοπισμό των παραισθήσεων, είναι σημαντικό δωθεί ιδιαίτερη έμφαση και στην αποφυγή τους, όσο αυτή είναι δυνατή, αφού είναι αδύνατον να υπάρχει πάντοτε απόλυτη βεβαιότητα οτι ένα Μεγάλο Γλωσσικό Μοντέλο δεν θα παράξει παραισθήσεις. Συγκεκριμένα, στο πλαίσιο των συστημάτων QA έχει μεγάλη σημασία το αν οι ερωτήσεις, είτε αυτές παράγονται απο το μοντέλο είτε όχι, είναι απαντήσιμες. Πρόσφατες μελέτες, όπως η εργασία "Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of Large Language Models" [35], διερευνούν το κατά πόσο τα Μεγάλα Γλωσσικά Μοντέλα μπορούν να απέχουν από την παραγωγή απάντησης όταν δεν διαθέτουν επαρκή στοιχεία ή όταν η πληροφορία δεν υπάρχει στο πεδίο γνώσης τους. Οι συγγραφείς δείχνουν ότι αν και τα LLMs μπορούν σε ορισμένες περιπτώσεις να αναγνωρίσουν την αβεβαιότητά τους, τείνουν συχνά να απαντούν με παραισθησιακό περιεχόμενο αντί να δηλώσουν άγνοια. Αυτό καθιστά απαραίτητη την ανάπτυξη μηχανισμών ελέγχου που ενισχύουν την ικανότητα αποχής των μοντέλων. Αντίστοιχα, η εργασία "Unanswerability Evaluation for Retrieval-Augmented Generation" [41] εξετάζει την αξιολόγηση απαντησιμότητας σε συστήματα τύπου RAG. Οι ερευνητές προτείνουν μετρικές και διαδικασίες που επιτρέπουν να εντοπιστεί αν ένα ερώτημα διαθέτει επαρκή τεκμηρίωση στα ανακτημένα αποσπάσματα, πριν παραχθεί η τελική απάντηση. Έτσι, το μοντέλο μπορεί να απορρίψει ερωτήσεις χωρίς επαρκή υποστήριξη, μειώνοντας σε μεγάλο βαθμό τις παραισθήσεις.

Συμπερασματικά, μελετώντας την σχετική βιβλιογραφία, οι κυριότερες προσεγγίσεις για τον περιορισμό των παραισθήσεων σε συστήματα RAG με εφαρμογές στο πλαίσιο των συστημάτων ερωταπαντήσεων, μπορούν να συνοψιστούν σε δύο βασικούς πόλους:

- την εναλλασσόμενη διαδικασία γέννησης/επαλήθευσης, ώστε η παραγόμενη πληροφορία να στηρίζεται σε τεκμήρια και να αποφεύγεται η δημιουργία περιεχομένου εκτός των πηγών.
- την ικανότητα αποχής του συστήματος, όταν δεν υπάρχει επαρκής τεκμηρίωση στη βάση γνώσης.

Για την ανάπτυξη του συστήματος που αναλύεται στην παρούσα διπλωματική, ακολουθήθηκε αυτή η προσέγγιση δίνοντας ιδιαίτερη έμφαση στην εγκυρότητα των παραγόμενων ερωτήσεων μετά απο την φάση της γείωσης, αλλά και στην απάντηση τους. Επιδιώχθηκε οι απαντήσεις να είναι ξεκάθαρα αποσπάσματα του, σχετικότερου με την ερώτηση, κειμένου και αν δεν βρεθεί κατάλληλο απόσπασμα που να απαντά ξεκάθαρα την ερώτηση, αυτή να σημειώνεται ως μη απαντήσιμη. Η επαλήθευση γίνεται κάνοντας εκ νέου την διαδικασία RAG, αλλά με τοπική εφαρμογή για κάθε σετ ερώτησης-απάντησης-κειμένου, έτσι ώστε να υπάρχει βεβαιότητα πως δεν σημειώνονται ερωτήσεις με απάντηση ως μη απαντήσιμες. Με αυτόν τον τρόπο, η μεθοδολογία που αναπτύχθηκε συνθέτει στοιχεία από σύγχρονες προσεγγίσεις στην υπάρχουσα βιβλιογραφία, παρέχοντας ένα ολοκληρωμένο, αξιόπιστο και τεκμηριωμένο σύστημα ερωταπαντήσεων με ελαχιστοποιημένο κίνδυνο παραισθήσεων και με πρακτική εφαρμογή στον εκπαιδευτικό χώρο.

4.4 Γενική Περιγραφή Συστήματος

Το σύστημα που αναπτύχθηκε στο πλαίσιο της παρούσας διπλωματικής εργασίας αποτελεί ένα πολυεπίπεδο σύστημα παραγωγής ερωτήσεων και απαντήσεων, το οποίο βασίζεται στην χρήση Μεγάλων Γλωσσικών Μοντέλων με μηχανισμούς Ανάκτησης Πληροφορίας, κυρίως Retrieval-Augmented Generation, χρησιμοποιώντας και την έννοια της Επέκτασης Ερωτημάτων για να πετύχει καλύτερα αποτελέσματα. Στόχος του είναι η παραγωγή ποιοτικών και σχετικών με την συλλογή κειμένων ερωτήσεων και η επικυρωμένη απο τεχμήρια επιβεβαίωση απαντήσεων πάνω σε μια συλλογή κειμένων, με τρόπο που να ελαχιστοποιεί το φαινόμενο των παραισθήσεων, το οποίο εντοπίζεται συχνά στα Μεγάλα Γλωσσικά Μοντέλα, και να εξασφαλίζει την ακρίβεια και την αξιοπιστία του τελικού αποτελέσματος, δηλαδή της παραγόμενης ερώτησης και απάντησης. Το σύστημα λειτουργεί σε διαδοχικά στάδια, όπου κάθε βήμα έχει σαφή ρόλο στην επεξεργασία, εμπλουτισμό και επαλήθευση των παραγόμενων δεδομένων και γενικά σε ολόκληρη την ορθή λειτουργία του συστήματος. Η συνολική ροή ακολουθεί μια λογική παρόμοια με αυτή που περιγράφεται στη βιβλιογραφία, αλλά με προσαρμογές που την καθιστούν πιο κατάλληλη για την παραγωγή και αξιολόγηση συνόλων ερωτήσεων-απαντήσεων με τρόπο που εστιάζει στην ακρίβεια της απάντησης και στην εγκυρότητα της ερώτησης.

Παρακάτω θα περιγραφούν συνοπτικά όλες τις φάσεις του συστήματος, για να γίνει κατανοητή η λειτουργία του:

4.4.1 Συλλογή δεδομένων εισόδου

Αφετηρία του συστήματος αποτελεί μια συλλογή κειμένων που περιέχει πληροφορίες γύρω από ένα ενιαίο θεματικό πεδίο. Επιλέχθηκε μια συλλογή απο κείμενα που αναφέρονται στον τομέα της ιατρικής, και συγκεκριμένα αφορούν την νόσο της Κυστικής Ίνωσης. Αυτή η επιλογή έγινε, κυρίως, για δύο λόγους. Όπως αναφέρθηκε και παραπάνω, σύμφωνα με την βιβλιογραφία, η χρήση της τεχνικής του Retrieval-Augmented Generation στα Μεγάλα Γλωσσικά Μοντέλα έχει καλύτερη επίδοση σε εργασίες συγκεκριμένου πεδίου (domain-specific tasks). Συνεπώς, επιλέχθηκε μια τέτοια συλλογή κειμένων για να επιτευχθεί καλύτερη επίδοση του μοντέλου. Παράλληλα, αφού το τελικό αποτέλεσμα της εργασίας είναι μια εκπαιδευτική εφαρμογή, θεωρήθηκε ορθό να χρησιμοποιηθεί μια συλλογή κειμένων που θα μπορούσε κάποιος μαθητής ή γενικά χρήστης να μελετήσει και να εξασκηθεί ή να εξεταστεί πάνω στις γνώσεις του, για το αντικείμενο που πραγματεύεται. Η χρήση μιας τέτοιας συλλογής στο σύστημα, δίνει ακριβώς αυτή την δυνατότητα στον χρήστη.

Στην εργασία αυτή, τα δεδομένα οργανώνονται σε μορφή αρχείου CSV, όπου κάθε εγγραφή αντιστοιχεί σε ένα απόσπασμα της συλλογής κειμένων.

4.4.2 Παραγωγή αρχικών ερωτήσεων και απαντήσεων

Στο επόμενο στάδιο, το σύστημα αξιοποιεί το Μεγάλο Γλωσσικό Μοντέλο μέσω στοχευμένων προτροπών (prompting), με σκοπό την παραγωγή αρχικών ερωτήσεων που σχετίζονται άμεσα με το περιεχόμενο κάθε αποσπάσματος. Για κάθε εγγραφή της συλλογής κειμένων(σε αυτή την περίπτωση του αρχείου CSV), δημιουργείται ένα σύνολο ερωτήσεων που αντανακλούν τα κύρια σημεία του κειμένου, με στόχο να καλυφθεί η ουσία του αποσπάσματος. Συνεπώς, επιλέχθηκαν δύο ερωτήσεις για κάθε εγγραφή για να πετύχουμε μεγαλύτερη θεματική κάλυψη.

Παράλληλα, κατα την διάρκεια του prompting, το μοντέλο οδηγείται στο να παράξει και μια αρχική απάντηση για την κάθε ερώτηση που δημιουργεί καθώς και μια βασική αιτιολόγηση αλλά και έναν βαθμό δυσκολίας για την κάθε ερώτηση. Η οδηγία για παροχή αιτιολόγησης βοηθάει το μοντέλο να παράξει καλύτερο αποτέλεσμα, από το να ζητήσουμε απλά την απάντηση. Η προσθήκη του χαρακτηριστικού της δυσκολίας της ερώτησης, παρότι δεν χρησιμοποιείται στα επόμενα στάδια του συστήματος, παρέχει πρόσθετη πληροφορία για μελλοντικές χρήσεις όπως η αυτόματη ταξινόμηση των ερωτήσεων ανά επίπεδο δυσκολίας σε εκπαιδευτικές εφαρμογές, και όχι μόνο.

Οι ερωτήσεις που παράγονται σε αυτή τη φάση ονομάζονται πρόχειρες ερωτήσεις (draft questions) και αποτελούν την πρώτη μορφή του συνόλου δεδομένων που χρησιμοποιείται στα επόμενα βήματα του συστήματος. Σε αυτό το σημείο, δεν έχει ακόμη εφαρμοστεί Ανάκτηση Πληροφορίας ή έλεγχος τεκμηρίωσης, επομένως οι ερωτήσεις και οι απαντήσεις μπορεί να περιλαμβάνουν ασαφείς ή μη επιβεβαιωμένες λεπτομέρειες. Για τον λόγο αυτό λοιπον, τα επόμενα στάδια του συστήματος επικεντρώνονται στη γείωση, τον εμπλουτισμό και την επαλήθευση των ερωτήσεων και των απαντήσεων τους, ώστε το τελικό αποτέλεσμα να είναι όσο περισσότερο αξιόπιστο γίνεται.

4.4.3 Ανάκτηση σχετικών τεκμηρίων

Μετά τη δημιουργία των αρχικών ερωτήσεων, το σύστημα χρησιμοποιεί τεχνικές Ανάκτησης Πληροφορίας, και συγκεκριμένα τον αλγόριθμο BM25(Best Matching 25), για να εντοπίσει τα αποσπάσματα της συλλογής κειμένων που είναι περισσότερο σχετικά με κάθε παραγόμενη ερώτηση. Η ανάκτηση αυτή προσφέρει στο μοντέλο εξωτερική γνώση πέραν του αποσπάσματος από το οποίο προήλθε η ερώτηση, βελτιώνοντας έτσι την τεκμηρίωσή της.

Με απλά λόγια, πλέον η κάθε ερώτηση επικυρώνεται με χρήση ολόκληρης της συλλογής κειμένων, σε περίπτωση που κάποιο άλλο κομμάτι της απαντάει καλύτερα την ερώτηση και γενικά είναι πιο σχετικό με αυτήν. Στην πραγματικότητα, δεν υπάρχει πλεον περιορισμός

σχετικά με το σε πιο απόσπασμα της συλλογής κειμένων αντιστοιχεί η κάθε ερώτηση, αφού τώρα εξετάζεται ολόκληρη η συλλογή για να εντοπιστούν τα πιο σχετικά κείμενα. Με βάση αυτή την εξωτερική γνώση, βελτιώνεται η τεκμηρίωση της κάθε ερώτησης. Η γνώση του μοντέλου παύει να είναι τοπική, περιορισμένη αυστηρά σε συγκεκριμένο απόσπασμα και γίνεται καθολική, αντλώντας πληροφορίες απο ολόκληρη την συλλογή κειμένων.

Το αποτέλεσμα αυτού του σταδίου είναι ένα σύνολο ερωτήσεων εμπλουτισμένων με τα αντίστοιχα συμφραζόμενα (contexts) που περιλαμβάνουν τα πιο σχετικά αποσπάσματα. Αυτή η φάση λειτουργεί ως συνδετικός κρίκος μεταξύ της γλωσσικής παραγωγής μιάς απλής προτροπής του Μεγάλου Γλωσσικού Μοντέλου και της πραγματικής γνώσης που προσφέρει η καθολική ανάκτηση πάνω στην συλλογή κειμένων. Χρησιμοποιώντας τα εμπλουτισμένα, με σχετικά αποσπάσματα, δεδομένα το σύστημα οδηγείται στην φάση της επέκτασης των ερωτημάτων και αρχικής απάντησης τους, έτσι ώστε να επιτευχθεί η θεμιτή η εγκυρότητα και αξιοπιστία.

4.4.4 Γείωση / Επέκταση / Αναδιατύπωση ερωτήσεων και αρχική απάντηση

Η φάση της γείωσης των ερωτήσεων αποτελεί ένα απο τα κύρια σημεία συστήματος που υλοποιήθηκε. Στην φάση αυτήν, κάθε ερώτηση εμπλουτίζεται ή αναδιατυπώνεται με βάση τα αποσπάσματα που ανακτήθηκαν, έτσι ώστε να είναι πιστή στα δεδομένα και απαλλαγμένη από ασαφή ή φανταστικά στοιχεία. Παρέχονται στο μοντέλο τα εμπλουτισμένα, με συμφραζόμενα, δεδομένα μαζί με την αρχική ερώτηση και αυτό παράγει μια νέα εκδοχή της. Αυτή η νέα εκδοχή της ερώτησης μπορεί και να είναι ίδια με την προηγούμενη ή λίγο αλλαγμένη, αν κριθεί απο το μοντέλο ότι είναι σύμφωνη με τα τεκμήρια που δώθηκαν. Όμως, το τελικό αποτέλεσμα είναι πάντα σύμφωνο με τα ανεκτημένα δεδομένα, γεγονός που προσδίδει αξιοπιστία σε αυτή την διαδικασία γείωσης. Είναι σημαντικό να αναφερθεί πως το στάδιο αυτό λειτουργεί ουσιαστικά ως εφαρμογή επέκτασης ερωτημάτων καθοδηγούμενη από ανάκτηση, καθώς αξιοποιούνται τα περιεχόμενα των ανακτημένων τεκμηρίων για να βελτιωθεί η διατύπωση και η πληρότητα του ερωτήματος. Με αυτόν τον τρόπο, κάθε ερώτηση καθίσταται περισσότερο αξιόπιστη και απόλυτα συνδεδεμένη με την πραγματική πληροφορία της βάσης γνώσης.

Παράλληλα, στην ίδια υλοποιητική φάση, γίνεται και η παραγωγή μιας αρχικής απάντησης στην κάθε γειωμένη ερώτηση. Το μοντέλο επιχειρεί να εντοπίσει το σχετικότερο απόσπασμα από τη συλλογή κειμένων και να παράξει μια πρώτη περιγραφική απάντηση, συνοδευόμενη από το τεκμήριο και τα αναγνωριστικά των εγγράφων που υποστηρίζουν την παραγόμενη πληροφορία. Σε αυτή την φάση, δεν δίνεται ιδιαίτερη σημασία στο μέγεθος, στην ποιότητα και στην εγκυρότητα της απάντησης αφού η εστίαση γίνεται κυρίως στην επέκταση της ερώτησης.

Οι επόμενες φάσεις του συστήματος επικεντρώνονται στην επικύρωση, περικοπή και βελτίωση των απαντήσεων, αλλά και στην επιβεβαίωση ότι υπάρχει ξεκάθαρη και σύντομη απάντηση για την κάθε ερώτηση.

4.4.5 Περικοπή και αρχική Επικύρωση απαντήσεων

Μετά τη γείωση και την αρχική παραγωγή απάντησης, ακολουθεί ένα κρίσιμο στάδιο περικοπής και αρχικής επικύρωσης. Αυτή η ανάγκη προκύπτει απο την συμπεριφορά του LLM. Συγκεκριμένα, όταν του δίνονται τα συμφραζόμενα και του ζητήται να απαντήσει με βάση αυτά, έχει την τάση να παραθέτει ολόκληρη πρόταση ή παράγραφο που περιέχει την απάντηση, για να "δικαιολογήσει" τη λογική του. Σε περίπτωση που δεν βρεθεί αμέσως μια κατάλληλη και σύντομη πρόταση, μπορεί να παραθέσει μια τεράστια ή ακόμη και ολόκληρο το κείμενο, όπως παρατηρήθηκε ότι συνέβη σε μερικές περιπτώσεις. Στόχος εδώ είναι να μετατραπεί η απάντηση του μοντέλου σε μια σύντομη, ακριβή και ρητά τεκμηριωμένη φράση, η οποία να μπορεί να σταθεί μόνη της ως αντιστοίχιση στην ερώτηση. Με αυτόν τον τρόπο, το σύστημα αποκτά δύο σημαντικά πλεονεκτήματα:

- αυξάνεται η σαφήνεια/ακρίβεια των απαντήσεων, μειώνοντας περιττές πληροφορίες
- καθιστάται το σύνολο ερωτήσεων-απαντήσεων φιλικό σε αξιολόγηση με μετρικές και χρήσιμο για εκπαιδευτική χρήση, αφού οι απαντήσεις πλεον είναι σύντομες και περιεκτικές.

Στην πράξη, το βήμα αυτό λαμβάνει ως είσοδο την αρχική απάντηση και το συνοδευτικό τεκμήριο (evidence) ,δηλαδη το απόσπασμα που δηλώνει το ίδιο το μοντέλο ως πηγή και επιχειρεί να εντοπίσει τη μικρότερη φράση που απαντά ευθέως στο ερώτημα. Η λογική είναι η εξής: Αντί να κρατηθεί ολόκληρη η παραγόμενη απάντηση, απομονώνεται το ελάχιστο απαραίτητο τμήμα(σύμφωνα με ενα λογικό όριο χαρακτήρων που τέθηκε), όπως ένα κύριο όρο, μια ονοματική φράση ή μια σύντομη πρόταση. Αυτό μειώνει δραστικά το ρίσκο περιττών πληροφοριών ή παραισθήσεων μέσα στην απάντηση της εκάστοτε ερώτησης και βελτιώνει την αντιστοίχιση σημαντικών όρων στην αξιολόγηση. Κεντρική σχεδιαστική επιλογή του σταδίου αυτού, είναι ότι η απάντηση στην ερώτηση πρέπει να υπάρχει αυτούσια μέσα στο κείμενο. Αν η αρχική απάντηση του μοντέλου δεν εντοπίζεται με σιγουριά σε μία σύντομη φράση του κειμένου (π.χ. αποτυγχάνει το ταίριασμα, λείπουν κρίσιμα αριθμητικά, υπάρχουν αντιφάσεις ή η μόνη απάντηση που μπορεί να δώσει το μοντέλο είναι ολόκληρο το κείμενο), τότε το παράδειγμα χαρακτηρίζεται ως μη απαντήσιμο. Πρόκειται για συνειδητή επιλογή υπέρ της αξιοπιστίας αφού προτιμήθηκε να απορριφθεί ένα ζεύγος ερώτησης-απάντησης που δεν στηρίζεται με σαφή και μικρή φράση στο τεκμήριο, παρά να κρατηθεί μια ασαφή ή επισφαλή απάντηση που θα ενίσχυε τον κίνδυνο παραισθήσεων ή θα προκαλούσε πρόβλημα ο εντοπισμός της απάντησης της απο έναν απλό χρήστη.

Το τίμημα αυτής της αυστηρότητας είναι ότι μερικά έγκυρα ζεύγη μπορεί να απορρίπτονται επειδή δεν εντοπίζεται αρκετά σύντομο και ξεκάθαρο κομμάτι κειμένου. Ωστόσο, αυτός ο συμβιβασμός είναι σκόπιμος αφού η διπλωματική εστιάζει στην ακρίβεια και τεκμηρίωση και όχι στη μέγιστη κάλυψη. Εξάλλου, η σχεδιαστική επιλογή συμπληρώνεται από το επόμενο στάδιο, όπου επιχειρείται στοχευμένη επανά-ανάκτηση για να διασωθεί ένα μέρος των μη-απαντήσιμων περιπτώσεων που μπορεί, λόγω ελλειπής ανάκτησης στο πρώτο στάδιο, να μην εντοπίστηκε μια σύντομη φράση που τις απαντάει κατάλληλα.

4.4.6 Επανεξέταση μη απαντήσιμων περιπτώσεων

Το στάδιο αυτό εφαρμόζεται στις περιπτώσεις όπου, μετά την φάση της περικοπής, ένα ζεύγος ερώτησης–απάντησης έχει χαρακτηριστεί ως μη απαντήσιμο. Στόχος είναι να διαπιστωθεί αν πράγματι δεν υπάρχει απάντηση μέσα στα τεκμήρια, ή αν η αποτυχία προήλθε από ατελή ανάκτηση ή μη εντοπισμό κατάλληλης σύντομης φράσης στο αρχικό απόσπασμα. Όπως αναφέρθηκε και στην περιγραφή της προηγούμενης φάσης (την περικοπή και σήμανση ερωτήσεων ως μη απαντήσημων) κάποια σετ ερωτήσεων-απαντήσεων απορρίφθηκαν. Μετά και απο χειροκίνητη αξιολόγηση μερικών απο τα σετ, παρατηρήθηκε ότι κάποια είχαν λανθασμένα σημανθεί ως μη απαντήσημα. Ένας λόγος για αυτό, ήταν το αρχικό στάδιο της ανάκτησης. Μπορεί κάποια σημαντικά και σχετικά αποσπάσματα, που βρίσκονταν στο αρχικό κείμενο για το οποίο γεννήθηκε η ερώτηση, να παραλείφθηκαν κατα την διαδικασία της αρχικής ανάκτησης τεκμηρίων. Υπήρχαν περιπτώσεις όπου το σχετικότερο κείμενο σε σχέση με την ερώτηση ήταν διαφορετικό απο το κείμενο για το οποίο το μοντέλο παρήγαγε την ερώτηση. Το γεγονός αυτό, προφανώς και είναι θεμιτό διότι έτσι εξασφαλίζεται η εγκυρότητα και η έλλειψη παραισθήσεων στο σύστημα, αλλά μπορεί και να εγκυμονεί κινδύνους όπως η απόρριψη κάποιων πραγματικά απαντήσιμων σετ ερωτήσεων-απαντήσεων.

Για αυτό, λοιπόν, ενσωματώθηκε στην ροή λειτουργίας το συγκεκριμένο στάδιο, στο οποίο ελέγχονται οι σημασμένες ως μη απαντήσιμες περιπτώσεις με τοπικά περιορισμένη ανάκτηση. Για κάθε ερώτηση που έχει σημανθεί ως μη απαντήσιμη:

1. Το σύστημα εντοπίζει το αρχικό κείμενο απο το οποίο προήλθε η ερώτηση, μέσω του αναγνωριστικού source id.

2. Πραγματοποιείται τοπική ανάκτηση εντός του εντοπισμένου κειμένου για την εκ νέου προσπάθεια απάντησης της ερώτησης. Σε αυτό το βήμα, εφαρμόζεται ένας συνδυασμός BM25, για εκτίμηση της σχετικότητας των επιμέρους προτάσεων ή παραθύρων του κειμένου και επικάλυψης λεκτικών μονάδων(token overlap), ώστε να εντοπιστούν τα τμήματα που περιέχουν λέξεις ή όρους κοινούς με την ερώτηση. Μετά και απο δοκιμές, παρατηρήθηκε ότι ο συνδυασμός των δύο αυτών μεθόδων έχει καλύτερα αποτελέσματα από την απλή εφαρμογή ενός εξ' αυτών.

3. Το LLM λαμβάνει ως είσοδο την ερώτηση και τα αποσπάσματα που προέκυψαν από την τοπική ανάκτηση και καλείται να παράξει ξανά μια σύντομη απάντηση, βασισμένη στα νέα τεκμήρια. Γίνεται, λοιπόν, μια δέυτερη φάση RAG για παραγωγή νέων απαντήσεων.

4. Η νέα απάντηση υποβάλλεται εκ νέου στην ίδια διαδικασία περικοπής με πριν.Η απάντηση πρέπει να αποτελεί σύντομη φράση, αυτούσια από το κείμενο, που απαντάει ικανοποιητικά την ερώτηση. Αν εντοπιστεί τέτοια φράση, η ερώτηση χαρακτηρίζεται ως 'διασωθείσα' (salvaged) και μετατρέπεται σε απαντήσιμη, στο τελικό σύνολο δεδομένων . Διαφορετικά, παραμένει μη απαντήσιμη.

Είναι σημαντικό να αναφερθεί πως για να αυξηθούν οι πιθανότητες εύρεσης έγκυρης φράσης ως απάντηση χωρίς να επεκταθεί η αναζήτηση σε όλη τη συλλογή, η διαδικασία τοπικής ανάκτησης και εξαγωγής απάντησης εκτελείται σε δύο περάσματα. Ένα με μικρότερο μέγεθος παραθύρου λεκτικών μονάδων, και ένα με μεγαλύτερο. Έτσι, εξετάζουμε σε ικανοποιητικό βαθμό το κάθε απόσπασμα για φράση-απάντηση.

Συνολικά, η διαδικασία αυτή ανακτά μερικές απαντήσεις που μπορεί να χάθηκαν λόγω του πρώτου στάδιου της ανάκτησης. Παράλληλα, διατηρεί την εγκυρότητα, καθώς δεν χρησιμοποιεί νέα ή άσχετα κείμενα ως βάση γνώσης για την ανάκτηση, αλλά περιορίζεται στην πηγή για την οποία γεννήθηκε η ερώτηση. Αυτό βοηθάει στις περιπτώσεις που υπάρχει σχετική απάντηση στο συγκεκριμένο κείμενο, παρόλου που μπορεί να μην βρέθηκε ως το πιο σχετικό σε σχέση με την ερώτηση. Έτσι, καλύπτεται και η περίπτωση που η καλύτερη απάντηση βρίσκεται σε άλλο κείμενο(στην πρώτη φάση της ανάκτησης) αλλά και η περίπτωση που η απάντηση μπορεί να βρίσκεται στο κείμενο για το οποίο δημιουργήθηκε αρχικά, απο το μοντέλο, η ερώτηση. Ως μη απαντήσιμες παραμένουν οι ερωτήσεις που δεν ανήκουν σε μια απο τις δύο περιπτώσεις, δηλαδή οι παραισθήσεις ή οι ερωτήσεις που δεν έχουν ξεκάθαρη απάντηση μέσα στο κείμενο. Τελικά, τα σετ ερωτήσεων-απαντήσεων που απομένουν αποτελούν ένα ολοκληρωμένο, έγκυρο, αξιόπιστο και επικυρωμένο, μέσα απο μια πολυφασική διαδικασία, σύνολο δεδομένων έτσι ώστε η χρήση του να είναι εύκολη και πρακτική για εκπαιδευτικές εφαρμογές, και όχι μόνο.

Είναι σημαντικό να αναφερθεί, επίσης, και ο μόνος θεωρητικός περιορισμός του σταδίου αυτού. Όπως επεξηγήθηκε παραπάνω, σε αυτή την φάση δεν εξετάζονται αποσπάσματα εκτός του αρχικού κειμένου-πηγής της κάθε ερώτησης για την διαδικασία RAG. Αν η σωστή απάντηση βρίσκεται αλλού, δεν θα εντοπιστεί. Οι περιπτώσεις αυτές, όμως, παραμένουν ελάχιστες ή και ανύπαρκτες για το συγκεκριμένο σύνολο δεδομένων που εξετάστηκε αφού αν υπήρχε ξεκάθαρη απάντηση στα θεμιτά πλαίσια, θα είχε εντοπιστεί και παραχθεί απο το μοντέλο σε κάποιο στάδιο της ροής του συστήματος. Σε γενικότερο πλαίσιο, αυτή η επιλογή εξασφαλίζει

ότι όλες οι τελικές απαντήσεις παραμένουν πλήρως τεκμηριωμένες και ότι η ακρίβεια υπερισχύει της κάλυψης, χαρακτηριστικά που είναι πολύ σημαντικά στον τομέα των εκπαιδευτικών εφαρμογών.

4.4.7 Εφαρμογή για Εκπαίδευση

Για να δωθεί έμφαση στην πρακτικότητα του συστήματος που αναπτύχθηκε, δημιουργήθηκε μια εφαρμογή αυτόματης παραγωγής ερωτήσεων (quiz), η οποία αξιοποιεί το τελικό επικυρωμένο σύνολο ερωτήσεων–απαντήσεων. Η εφαρμογή αυτή έχει ως στόχο να προσφέρει ένα εργαλείο εκπαίδευσης και αυτοαξιολόγησης των χρηστών, βασισμένο σε πραγματικά και τεκμηριωμένα δεδομένα που προέκυψαν απο την χρήση της ροής λειτουργίας (pipeline) που παρουσιάζεται. Παρακάτω, θα περιγραφεί σε γενικό πλαίσιο η λειτουργία της.

Η εφαρμογή φορτώνει το αρχείο με το τελικό σύνολο δεδομένων, το οποίο περιέχει τα ζεύγη γειωμένης ερώτησης και απάντησης που έχουν περάσει όλους τους ελέγχους επικύρωσης και περικοπής. Από το σύνολο αυτό, επιλέγεται τυχαία ένα υποσύνολο (στην παρούσα φάση δέκα ερωτήσεων), το οποίο παρουσιάζεται στον χρήστη. Για κάθε ερώτηση, ο χρήστης καλείται να δώσει τη δική του απάντηση σε πεδίο ελεύθερου κειμένου. Μετά την υποβολή όλων των απαντήσεων, η εφαρμογή υπολογίζει αυτόματα τη βαθμολογία χρησιμοποιώντας μετρικές ακρίβειας, έτσι ώστε να διαπιστωθεί το πόσο όμοια είναι με την απάντηση που έδωσε στις ίδιες ερωτήσεις το σύστημα. Το τελικό αποτέλεσμα μετατρέπεται σε βαθμό κλίμακας 1–10 και παρουσιάζεται στον χρήστη ως συνολική αξιολόγηση, όταν πατήσει το αντίστοιχο κουμπί και έχει υποβάλλει απαντήσεις σε όλες τις ερωτήσεις. Η εφαρμογή ενσωματώνει, επίσης, μηχανισμό αποθήκευσης των απαντήσεων σε εξωτερικό αρχείο, ώστε τα αποτελέσματα να μπορούν να αξιοποιηθούν για εκπαιδευτική αξιολόγηση ή ανάλυση επίδοσης σε μετέπειτα χρόνο απο τον ίδιο τον χρήστη ή κάποιον φορέα εκπαίδευσης.

Συνολικά, η εφαρμογή αποτελεί ένα πρακτικό παράδειγμα αξιοποίησης του RAG-based QA συστήματος που αναπτύχθηκε σε πραγματικό εκπαιδευτικό πλαίσιο. Μέσα απο το περιβάλλον αυτό επιτυγχάνεται η επαναχρησιμοποίηση του παραγόμενου συνόλου ερωτήσεων–απαντήσεων για εξάσκηση του χρήστη, ο διαδραστικός τρόπος μάθησης και η αυτόματη αξιολόγηση γνώσεων βασισμένη σε επαληθευμένες απαντήσεις. Παράλληλα, η παρουσίαση της εφαρμογής αυτής δεν αποσκοπεί μόνο στην επίδειξη της λειτουργικότητας του συστήματος, αλλά και στην ανάδειξη της σημασίας της τεκμηρίωσης και της ακρίβειας στις απαντήσεις που προέρχονται από Μεγάλα Γλωσσικά Μοντέλα. Η εφαρμογή βασίζεται αποκλειστικά σε ερωτήσεις και απαντήσεις που έχουν επαληθευτεί μέσω μηχανισμών Ανάκτησης Πληροφορίας και RAG. Δεν είναι απλές απαντήσεις που έχει δώσει ένα Μεγάλο Γλωσσικό Μοντέλο, μετά απο μια απλή προτροπή ενός χρήστη. Έτσι, περιορίζεται δραστικά ο κίνδυνος αλληλεπίδρασης του χρήστη

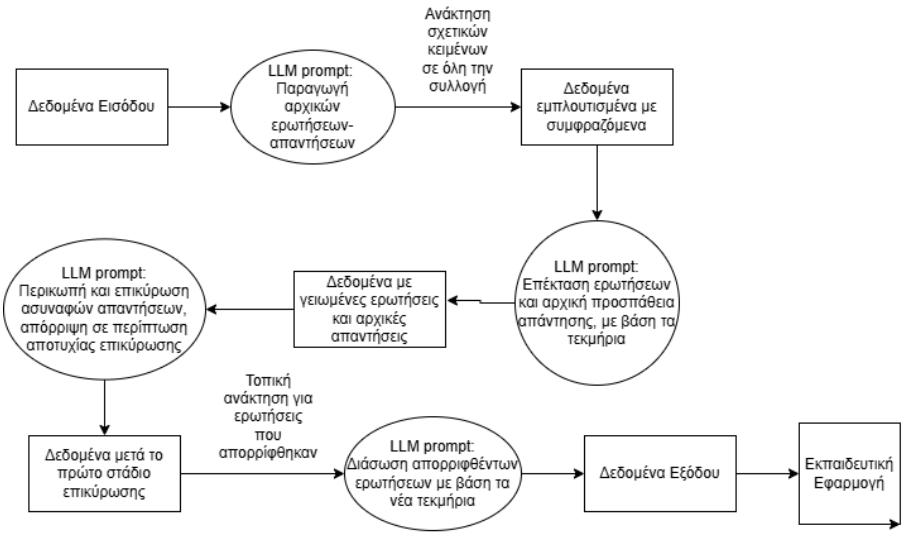
με περιεχόμενο που μπορεί να αποτελεί παραισθήσεις του μοντέλου. Τελικά, η εφαρμογή ενσωματώνει και αποδεικνύει στην πράξη τον κεντρικό στόχο της διπλωματικής εργασίας, δηλαδή τον περιορισμό των παραισθήσεων σε συστήματα ερωτήσεων-απαντήσεων και πετυχαίνει την δημιουργία ενός αξιόπιστου συνόλου που μπορεί να χρησιμοποιηθεί με ασφάλεια.

Κεφάλαιο 5

Υλοποίηση

Στο κεφάλαιο αυτό θα αναλύθει η διαδικασία υλοποίησης του συστήματος που περιγράφηκε. Αρχικά, θα δωθεί μια περιγραφή των τεχνολογιών που χρησιμοποιήθηκαν για την υλοποίηση. Στην συνέχεια, θα περιγραφεί ο κώδικας της παρούσας διπλωματικής, έτσι ώστε να γίνει κατανοητός ο τρόπος λειτουργίας του συστήματος καθώς και τα επιμέρους κομμάτια του.

Πρωτού αναφερθούν τα παραπάνω, ακολουθεί ένα διάγραμμα ροής για το σύστημα, έτσι ώστε να οπτικοποιηθεί η λειτουργία του:



Σχήμα 5.1: Ροή λειτουργίας του pipeline που υλοποιήθηκε.

5.1 Τεχνολογίες και εργαλεία που χρησιμοποιήθηκαν

Παρακάτω θα αναφερθούν οι τεχνολογίες και τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση του κώδικα της παρούσας διπλωματικης.

5.1.1 Γλώσσα Προγραμματισμού Python

Η Python είναι μια ισχυρή, γενικής χρήσης, υψηλού επιπέδου γλώσσα προγραμματισμού, η οποία σχεδιάστηκε με έμφαση στην απλότητα, την αναγνωσιμότητα και την παραγωγικότητα. Ορισμένα στοιχεία της αποτελούν:

- Είναι ερμηνευόμενη (interpreted), δηλαδή ο κώδικας εκτελείται γραμμή-γραμμή από έναν διερμηνέα, χωρίς να απαιτείται προαπαιτούμενη μετάφραση σε κώδικα μηχανής
- Είναι υψηλού επιπέδου (high level), χαρακτηριστικό που σημαίνει ότι ο προγραμματιστής δεν χρειάζεται να ασχολείται με λεπτομέρειες χαμηλού επιπέδου, όπως διαχείριση μνήμης
- Υποστηρίζει πολλαπλά παραδείγματα προγραμματισμού. Αντικειμενοστραφή, διαδικασιακή, ακόμη και λειτουργική προσέγγιση μπορούν να συγχωνευθούν.
- Λόγω της απλής και καθαρής σύνταξής της, ο κώδικας είναι ευανάγνωστος και πιο συντηρήσιμος, χαρακτηριστικό που αποτελεί μεγάλο πλεονέκτημα ειδικά σε έργα μεγάλης κλίμακας.

Παράλληλα, η γλώσσα Python είναι απο τις πιο δημοφιλείς γλώσσες προγραμματισμού, λόγω της μεγάλης της ευελιξίας. Μπορεί να χρησιμοποιηθεί αποτελεσματικά σε πολλούς διαφορετικούς τομείς όπως ανάλυση δεδομένων, τεχνητή νοημοσύνη, αυτοματισμούς και υπηρεσίες web είτε front είτε back end. Για τους παραπάνω λόγους, επιλέχθηκε η χρήση της, για την εκπόνηση της συγκεκριμένης διπλωματικής.

Η Python από μόνη της προσφέρει ένα ισχυρό υποσύνολο εργαλείων μέσω της standard library. Ωστόσο, ένα από τα μεγαλύτερα πλεονεκτήματα της είναι ότι υποστηρίζεται από ένα ολοκληρωμένο οικοσύστημα βιβλιοθηκών που επεκτείνουν τη γλώσσα για πολύ πιο εξειδικευμένα έργα, και κάνουν την χρήση της ακόμη πιο απλή και ευέλικτη. Παρακάτω, θα παρουσιαστούν οι βιβλιοθήκες που χρησιμοποιήθηκαν για την υλοποίηση.

5.1.2 Βιβλιοθήκες της Python που χρησιμοποιήθηκαν

Για την υλοποίηση του συστήματος χρησιμοποιήθηκαν διάφορες βιβλιοθήκες της Python, τόσο από την τυπική βιβλιοθήκη της γλώσσας όσο και εξωτερικές, οι οποίες διευκολύνουν την επεξεργασία δεδομένων, την ανάκτηση πληροφορίας και την επικοινωνία με γλωσσικά μοντέλα. Παρακάτω παρουσιάζονται συνοπτικά οι βασικές βιβλιοθήκες που αξιοποιήθηκαν:

- **argparse** : Η βιβλιοθήκη αυτή χρησιμοποιείται για τη διαχείριση παραμέτρων και ορισμάτων γραμμής εντολών. Επιτρέπει την εκτέλεση των επιμέρους αρχείων του συστήματος με διαφορετικές επιλογές εισόδου ή λειτουργίας, καθιστώντας τη διαδικασία πιο ευέλικτη και επαναχρησιμοποιήσιμη.

- **json** :Η βιβλιοθήκη αυτή χρησιμοποιείται για την ανάγνωση και εγγραφή δεδομένων σε μορφή JSON, ένα από τα πιο συχνά χρησιμοποιούμενα formats για αποθήκευση δομημένων δεδομένων. Στο παρόν σύστημα χρησιμοποιείται κυρίως για την αποθήκευση των παραγόμενων ερωτήσεων, απαντήσεων και συμφραζομένων.

- **re** : Η βιβλιοθήκη κανονικών εκφράσεων αξιοποιείται για τον καθαρισμό και τον προκαταρκτικό έλεγχο των κειμένων. Μέσω αυτής, αφαιρούνται περιττοί χαρακτήρες, σύμβολα ή ειδικές σημάνσεις από τα δεδομένα εισόδου και τις παραγόμενες απαντήσεις.

- **os , sys** : Αυτές οι βιβλιοθήκες παρέχουν πρόσβαση σε λειτουργίες του λειτουργικού συστήματος και του περιβάλλοντος εκτέλεσης. Η os χρησιμοποιήθηκε για τη διαχείριση αρχείων και φακέλων, ενώ η sys για τη ρύθμιση παραμέτρων εκτέλεσης και την ομαλή διαχείριση εξαιρέσεων.

- **csv** : Η βιβλιοθήκη αυτή χρησιμοποιείται κυρίως για την ανάγνωση και εγγραφή δεδομένων σε αρχεία μορφής comma separated values. Στην εργασία, χρησιμοποιήθηκε σε κάποια σημεία αυτή η δομή αρχείου καθώς είναι εύκολα αναγνώσιμη και επεξεργάσιμη.

- **math** : Η βιβλιοθήκη αυτή χρησιμοποιήθηκε για βασικούς μαθηματικούς υπολογισμούς, όταν αυτοί χρειάστηκαν

- **dataclasses** : Η βιβλιοθήκη αυτή επιτρέπει τον καθαρό ορισμό αντικειμένων που αποθηκεύουν πληροφορίες (όπως ένα ζεύγος ερώτησης–απάντησης)

- **typing** : Η βιβλιοθήκη αυτή παρέχει υποστήριξη για στατικούς τύπους και προσθέτει type hints, βελτιώνοντας την αναγνωσιμότητα και τη συντηρησιμότητα του κώδικα.

- **datetime** : Η βιβλιοθήκη αυτή, παρέχει ημερομηνίες, ώρες και χρονικές σημάνσεις(timestamps). Διευκολύνει την παρακολούθηση και επαναληψιμότητα των εκτελέσεων του κώδικα.

- **random** : Η βιβλιοθήκη αυτή, χρησιμοποιείται για δειγματοληψία ή τυχαία επιλογή στοιχείων. Εδώ, η χαρακτηριστική χρήση της ήταν ή επιλογή ερωτήσεων για αξιολόγηση και η τυχαία παρουσίασή τους στην τελική εκπαιδευτική εφαρμογή.

- **pathlib** : Η βιβλιοθήκη αυτή παρέχει αντικειμενοστραφή διαχείριση αρχείων και διαδρομών. Χρησιμοποιείται για ασφαλέστερη αναφορά σε paths ανεξαρτήτως λειτουργικού συστήματος.

Οι παραπάνω βιβλιοθήκες είναι ενσωματωμένες στην βασική έκδοση της Python. Οι επόμενες απαιτούν εγκατάσταση μέσω pip:

- **pandas** : Μια από τις σημαντικότερες βιβλιοθήκες του έργου, η pandas χρησιμοποιήθηκε για την οργάνωση, φιλτράρισμα και επεξεργασία των δεδομένων σε μορφή πινάκων (DataFrames). Διευκολύνει τη διαχείριση πολύπλοκων δομών δεδομένων, την αντιστοίχιση ερωτήσεων–απαντήσεων και την επεξεργασία των αποτελεσμάτων του μοντέλου. Τα δεδομένα εξάγονται και αποθηκεύονται σε αρχεία μορφής JSON Lines (.jsonl), τα οποία επιτρέπουν αποδοτική ανάγνωση και εγγραφή μεγάλου όγκου πληροφοριών κατά τη διάρκεια του pipeline.

- **requests** : Η βιβλιοθήκη αυτή χρησιμοποιείται για την επικοινωνία με εξωτερικά APIs μέσω αιτημάτων HTTP. Στο πλαίσιο της παρούσας εργασίας, αξιοποιείται για την αποστολή των prompts και τη λήψη απαντήσεων από το τοπικό LLM endpoint, που λειτουργεί ως μηχανισμός παραγωγής ερωτήσεων και απαντήσεων. Με αυτόν τον τρόπο, επιτυγχάνεται η απρόσκοπτη σύνδεση του συστήματος με το γλωσσικό μοντέλο χωρίς την ανάγκη εξωτερικών υπηρεσιών.

- **rank_bm25** : Η βιβλιοθήκη αυτή υλοποιεί τον αλγόριθμο BM25, έναν από τους πιο διαδεδομένους αλγορίθμους ανάκτησης πληροφορίας. Θα γίνει περεταίρω αναφορά σε αυτόν, στην συνέχεια

- **streamlit** : Η βιβλιοθήκη streamlit αποτελεί ένα σύγχρονο περιβάλλον (framework) για τη δημιουργία διαδραστικών web. Θα γίνει περεταίρω αναφορά στο περιβάλλον αυτό, στην συνέχεια.

Όλες οι βιβλιοθήκες που αναφέρθηκαν, βοήθησαν σημαντικά στην συγγραφή του κώδικα της διπλωματικής. Η δυνατότητα αξιοποίησης αυτών των βιβλιοθηκών που προσφέρει η γλώσσα Python, ήταν και ένας απο τους βασικότερους λόγους που επιλέχθηκε ως γλώσσα υλοποίησης του συστήματος.

5.1.3 Η πλατφόρμα Ollama

Το Ollama είναι μια πλατφόρμα ανοιχτού λογισμικού που επιτρέπει την τοπική εκτέλεση Μεγάλων Γλωσσικών Μοντέλων σε προσωπικούς υπολογιστές, χωρίς την ανάγκη αποστολής των δεδομένων σε εξωτερικές υπηρεσίες. Διαθέτει βιβλιοθήκη προεκπαιδευμένων μοντέλων τα οποία μπορούν να εξαχθούν και να χρησιμοποιηθούν, με εύκολο τρόπο. Σε αυτή την εργασία, επιλέχθηκε το μοντέλο llama 3.1. Παράλληλα, δίνεται και η δυνατότητα στον χρήστη να ρυθμίσει παραμέτρους μοντέλου, για πιο συγκεκριμένα αποτελέσματα. Επιπλεόν, ένας βασικός λόγος για την επιλογή της πλατφόρμας ήταν οι ενσωματωμένες δυνατότητες της για χρήση μέσω Python βιβλιοθήκες. Τέλος, η χρήση του συμβάλει στην αποφυγή κόστων ανά αίτημα που έχουν κάποια cloud API, και στην δυνατότητα αποτελεσματικής τοπικής εκτέλεσης ενός προεκπαιδευμένου LLM. Για αυτούς τους λόγους, το Ollama αποτελεί αξιόπιστη και κατάλληλη επιλογή για υλοποίηση του συγκεκριμένου pipeline.

5.1.4 Ο αλγόριθμος Ανάκτησης Πληροφορίας BM25

Ο αλγόριθμος BM25 (Best Matching 25) αποτελεί εξέλιξη των μοντέλων ανάκτησης πληροφορίας που βασίζονται στην λογική του TF-IDF. Στην ουσία, αξιολογεί πόσο καλά ταιριάζει ένα έγγραφο με ένα ερώτημα λαμβάνοντας υπόψη όχι μόνο τη συχνότητα ενός όρου αλλά και το μήκος του εγγράφου και τον κορεσμό όρων. Για να αποτυπωθεί η λειτουργία του, θα αναφερθούμε στις έννοιες TF-IDF

Η Συχνότητα Όρου (Term Frequency-TF) μετρά το πόσο συχνά εμφανίζεται ένας όρος του ερωτήματος μέσα σε ένα έγγραφο. Ωστόσο, ο αλγόριθμος BM25 εισάγει ένα στοιχείο κορεσμού (saturation effect), δηλαδή πέρα από ένα ορισμένο σημείο, επιπλέον εμφανίσεις ενός όρου συνεισφέρουν όλο και λιγότερο στη βαθμολογία του εγγράφου. Με αυτό τον τρόπο αποφεύγεται το φαινόμενο όπου τα πολύ μεγάλα έγγραφα ευνοούνται άδικα.

Μαθηματικά, το συνιστώμενο μέρος της TF κανονικοποιείται σύμφωνα με τον τύπο:

$$TF(t,d) = \frac{freq(t,d) \cdot (k_1 + 1)}{freq(t,d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)}$$

όπου:

- t: όρος του ερωτήματος
- d: έγγραφο
- $freq(t,d)$: αριθμός εμφανίσεων του όρου t στο έγγραφο d
- $|d|$: μήκος του εγγράφου d (σε αριθμό όρων)
- avgdl: μέσο μήκος εγγράφων στη συλλογή

- k_1: παράμετρος που ελέγχει το φαινόμενο κορεσμού (συνήθως 1.2–2.0)
- b: παράμετρος που ελέγχει την επίδραση του μήκους εγγράφου (περίπου 0.75)

Η Ανάστροφη Συχνότητα Εγγράφου (Inverse Document Frequency-IDF) μετρά τη σπουδαιότητα ενός όρου σε ολόκληρο το σύνολο εγγράφων. Οι σπάνιοι όροι θεωρούνται πιο ενημερωτικοί από τους συχνούς.

Ο συντελεστής IDF υπολογίζεται ως εξής:

$$IDF(t) = \log\left(\frac{N - n_t + 0.5}{n_t + 0.5} + 1\right)$$

όπου:

- N: συνολικός αριθμός εγγράφων στο σύνολο
- n_t: αριθμός εγγράφων που περιέχουν τον όρο t

Η τελική βαθμολογία BM25 ενός εγγράφου d ως προς ένα ερώτημα q υπολογίζεται αθροίζοντας τις συνεισφορές όλων των όρων του ερωτήματος. Κάθε όρος βαθμολογείται με βάση τη συχνότητα εμφάνισής του στο έγγραφο (TF) και τη σπανιότητά του στο σύνολο των εγγράφων (IDF). Ο συνολικός τύπος δίνεται από τη σχέση:

$$Score(q, d) = \sum_{t \in q} IDF(t) \cdot TF(t, d)$$

Αυτό το άθροισμα εκφράζει τη συνολική συνάφεια του εγγράφου d σε σχέση με το ερώτημα q. Τα έγγραφα ταξινομούνται με βάση αυτή τη βαθμολογία, και τα αποτελέσματα με τη μεγαλύτερη τιμή θεωρούνται τα πιο σχετικά με το ερώτημα.

Για τις ανάγκες της συγκεκριμένης εργασίας, επιλέχθηκε ο αλγόριθμος BM25 λόγω της απλότητας του και της ευελιξίας του. Έχει θεωρηθεί ως ένας απο τους κατάλληλοτερους αλγορίθμους ανάκτησης για εργασίες ερωταπαντήσεων [31]. Για αυτό τον λόγο, στο πλαίσιο της εργασίας αυτής, θεωρήθηκε πως βοηθάει στην διαδικασία επέκτασης της αρχικής ερώτησης αλλά και εντοπισμού της απάντησης της. Τέλος, ένα απο τα κυριότερα μειονεκτήματα του, δηλαδή η έλλειψη σημασιολογικής κατανοήσης, δεν επηρεάζει το συγκεκριμένο έργο, αφού η εστίαση γίνεται στην εγκυρότητα και στον εντοπισμό τεκμηρίων σε συλλογή κειμένων για περιορισμό παραισθήσεων του μοντέλου, τόσο στην παραγωγή της ερώτησης όσο και στην απάντηση μας. Δεν απασχολούν συνώνυμα και σημασιολογικές έννοιες, αφού έμφαση δίνεται στην ακρίβεια του συνόλου ερωτήσεων-απαντήσεων.

5.1.5 Streamlit

Το Streamlit είναι ένα περιβάλλον ανοικτού λογισμικού (open source framework) για την ανάπτυξη διαδραστικών διαδικτυακών εφαρμογών απευθείας μέσω Python. Σε αντίθεση με τις παραδοσιακές μεθόδους ανάπτυξης ιστοσελίδων, η χρήση του Streamlit δεν απαιτεί γνώσεις ή χρήση τεχνολογιών όπως HTML, CSS ή JavaScript. Αντί αυτών, ο προγραμματιστής μπορεί να δημιουργήσει πλήρως λειτουργικές διεπαφές χρήστη χρησιμοποιώντας αποκλειστικά Python scripts, γεγονός που μειώνει σημαντικά την πολυπλοκότητα και επιταχύνει την ανάπτυξη πρωτοτύπων.Η βασική φιλοσοφία της βιβλιοθήκης είναι να επιτρέπει στον ερευνητή ή στον προγραμματιστή να επικεντρώνεται στη λογική της εφαρμογής και στα δεδομένα, χωρίς να χρειάζεται να ασχοληθεί με το σχεδιασμό του γραφικού περιβάλλοντος. Με ελάχιστο κώδικα, μπορούν να προστεθούν κουμπιά, πίνακες, διαγράμματα και γενικά ό,τι στοιχείο μπορεί να θέλει κάποιος. Επιπλέον, το Streamlit υποστηρίζει άμεση εκτέλεση και ενημέρωση των αποτελεσμάτων σε πραγματικό χρόνο, καθώς και εύκολη ενσωμάτωση βιβλιοθηκών της Python.

Στο πλαίσιο της παρούσας διπλωματικής χρησιμοποιήθηκε το Streamlit για να παρουσιαστεί η εκπαιδευτική εφαρμογή, που είναι και το τελικό προϊόν της εργασίας. Μέσω αυτού, αποφεύχθηκε η χρήση HTML, CSS και JavaScript, και ενσωματώθηκε το GUI απευθείας μέσω Python. Επιτεύχθηκε έτσι, η ανάπτυξη μιας λειτουργικής διεπαφής χρήστη, χωρίς να εμποδιστεί η ροή εργασίας του έργου.

5.2 Περιγραφή Κώδικα

Παρακάτω θα εξηγηθεί η λειτουργία των scripts που αποτελούν την ροή λειτουργίας του συστήματος. Θα αναφερθεί το τι κάνει και που στοχεύει το καθένα, το τι παράγει στην έξοδο του και το πως ο συνδιασμός όλων οδηγεί στο τελικό σύνολο ερωταπαντήσεων.

5.2.1 Φάση παραγωγής - gen_questions_ollama.py

Το script αυτό, παίρνει ως είσοδο την συλλογή κειμένων σε μορφή csv (εδώ, είναι το αρχείο cf_passages.csv) και, για κάθε απόσπασμα, παράγει ερωτήσεις που μπορούν να απαντηθούν απο το κάθε απόσπασμα, μαζί με αρχική "προτεινόμενη απάντηση", δυσκολία και ενδεικτική αιτιολόγηση. Η έξοδος είναι JSONL αρχείο (cf_drafts.jsonl) όπου κάθε γραμμή αντιστοιχεί στο αποτέλεσμα για ένα κείμενο (passage).
Τα βασικότερα στοιχεία του script, έιναι τα εξής:

- GenConfig : συγκεντρώνει όλες τις επιλογές που καθορίζουν την συμπεριφορά του LLM-endpoint, όπως temperature και num_ctx (context window.
- SYSTEM_PROMPT: καθορίζει το ρόλο που θα έχει το LLM (εδώ "expert assessment designer") και απαιτεί αυστηρό JSON με κλειδιά question, answer, difficulty (easy—medium—hard), rationale. Παράλληλα, ζητά ερωτήσεις που απαντώνται μόνο από το κάθε κείμενο και αποφεύγουν εξωτερικές γνώσεις.
- USER_TEMPLATE: θέτει το εκάστοτε passage σε ένα σαφές πλαίσιο και ζητά την παραγωγή ερωτήσεων, και την έξοδο. Αυτό αποσκοπεί στην σταθερή μορφή και στην πιστότητα στο κείμενο.
- `call_ollama(cfg, prompt) -> str`: χτίζει payload για ποστ στο Ollama api, έτσι ώστε να λειτουργίσει το επιλεγμένο LLM. Επιστρέφει σκέτο κείμενο όχι JSON αντικείμενο αφού η μορφοποίηση επιβάλλεται από το prompt.
- `_read_df(path, input_col, id_col) -> pd.DataFrame`: διαβάζει .json ή .csv σε dataframe και εξασφαλίζει την κατάλληλη μορφοποίηση του. Κανονικοποιεί, λοιπόν, την είσοδο.
- `parse_json_list(text) -> List[Dict]]`: Προσπαθεί αρχικά να φορτώσει json μέσω json.loads(text). Αν αποτύχει, απομονώνει το υποσύνολο από τον πρώτο [έως το τελευταίο] και επιχειρεί εκ νέου json.loads. Αυτό μπορεί να φανεί σημαντικό σε περιπτώσεις που το μοντέλο αποτύχει να παράξει αυστηρό json.

Ακολουθεί η βασική ροή της main:

- Parse των ορισμάτων, και κατασκευή GenConfig.
- Ανάγνωση εισόδου μέσω _read_df.
- Άνοιγμα εξόδου, και για κάθε γραμμή, κλήση μοντέλου μέσω call_ollama. Ακολουθεί έλεγχος μέσω parse_json_list και αν πετύχει, αποθηκεύεται στο items[], όπου κάθε item έχει question, answer, difficulty, rationale.
- Δομείται ένα record ανα passage.

Τα records που προκύπτουν είναι της μορφής:

```
{
  "source_id": <id>,
  "generated_at": "<timestamp>",
  "model": "<tag>",
  "num_requested": <num>,
  "items": [...],
  "passage": "<raw text>"
}
```

Στην ουσία, το script αυτό, παίρνει ως είσοδο ένα csv αρχείο, με στήλες πχ id, passage και παράγει ως έξοδο ένα jsonl (cf_drafts.jsonl), με records όπως το παραπάνω

Είναι σημαντικό να αναφερθεί πως ο καθορισμός ρόλου στο LLM και η προσπάθεια για περιορισμό ερωτήσεων που απαντόνται απο το κείμενο και δεν απαιτούν εξωτερικές γνώσεις, αποσκοπούν στο να ενισχύσουν το αρχικό prompt και να παραχθούν καλύτερα αρχικά αποτελέσματα. Η διαδικασία θα μπορούσε να γίνει και με πιο αφελές prompting, αλλά επιλέχθηκε αυτή την προσέγγιση. Παράλληλα, δίνουμε έμφαση και στην δομή των δεδομένων, ζητόντας δομημένο (structured)json απο το μοντέλο, έτσι ώστε η χρήση του αρχείου εξόδου, να είναι πιο εύκολη και σταθερή κατά την συνέχεια της υλοποίησης.

5.2.2 Φάση πρώτης Ανάκτησης - bm25_retireval.py

Το script αυτό, παίρνει ως είσοδο την συλλογή κειμένων, και την συλλογή ερωταπαντήσεων που παράχθηκε στο προηγούμενο στάδιο. Για κάθε αρχική (draft) ερώτηση ανακτά τα k πιο σχετικά αποσπάσματα από ολόκληρη τη συλλογή, χρησιμοποιώντας BM25. Το αποτέλεσμα γράφεται σε jsonl αρχείο με μια εγγραφή (record) ανά ερώτηση, η οποία περιέχει τα στοιχεία της καθώς και τα σχετικά με αυτήν κείμενα της συλλογής κειμένων. Σκοπός είναι η συσχέτιση συλλογής κειμένων και ερώτησης, πέρα απο το prompting του LLM. Παρόλο που του ζητάται οι ερωτήσεις να έχουν απάντηση μέσα απο το κείμενο, δεν γίνεται να αποκλειστεί ο κίνδυνος των παραισθήσεων, στο στάδιο του prompting. Η ανάκτηση σχετικών κειμένων, αποτελεί σημαντικό βήμα προς την επικύρωση, και στην συνέχεια βελτίωση, των παραγόμενων ερωτήσεων. Ακολουθούν τα βασικά σημεία του script:

Tokenization και BM25 index:

Το tokenization γίνεται με χρήση κανονικής έκφρασης regular expression - regex και πραγματοποιείται καθαρά, λεξιλογική ανάκτηση

```
WORD_RE = re.compile(r"[A-Za-z0-9']+")

def tokenize(text: str):
    return [t.lower() for t in WORD_RE.findall(text or "")]
```

Δημιουργείται BM25 index πάνω σε ολόκληρο το corpus. Επιστρέφει και τα tokenized passages για αποδοτικότητα, αποφεύγοντας έτσι το re-tokenization.

```
def build_bm25(passages):
    tokenized = [tokenize(p) for p in passages]
    return BM25Okapi(tokenized), tokenized
```

Ανάκτηση:

Γίνεται tokenize της ερώτησης και υπολογισμός BM25 scores όλων των εγγράφων της συλλογής. Γίνεται top-k επιλογή με ταξινόμηση κατά φθίνουσα βαθμολογία. Επιστρέφονται ζεύγη (doc_index, score)

```
def retrieve(bm25, tokenized_corpus, question: str, k: int = 5):
    q = tokenize(question)
    scores = bm25.get_scores(q)
    idxs = sorted(range(len(scores)), key=lambda i: scores[i], reverse=True)[:k]
    return [(i, float(scores[i])) for i in idxs]
```

Κύρια ροή της main:

- Διαβάζει την συλλογή κειμένων, και φτιάχνει το BM25 index με όλα τα passages
- Για κάθε γραμμή του cf_drafts.json (ένα record με items = λίστα ερωτήσεων), έχει μια έξοδο ανα ερώτηση, παίρνει τα top-k hits και κατασκευάζει τα contexts ως λίστα αντικειμένων
- Τελικά, γράφει ένα record με τα, source_id (το passage απο το οποίο προήλθε η ερώτηση), question_ix (θέση της ερώτησης μέσα στα items), question, draft_answer, draft_difficulty και contexts (η λίστα top-k συμφραζομένων). Όλα σε μια γραμμή JSON ανα ερώτηση.

Η λίστα αντικειμένων contexts έχει την μορφή:

```
{
  "rank": 1..k,
  "doc_index": <index in passages>,
  "doc_id": <df.id[doc_index]>,
  "score": <bm25 score>,
  "passage": "<raw text>"
}
```

Ως έξοδο, για κάθε draft ερώτηση, έχουμε ένα JSON object της μορφής:

```
{
  "source_id": <id>,
  "question_ix": <0|1>,
  "question": "<draft question>",
  "draft_answer": "<first answer from the llm>",
  "draft_difficulty": "easy|medium|hard",
  "contexts": [
    {
      "rank": 1,
      "doc_index": <int>,
      "doc_id": <int>,
      "score": <float>,
      "passage": "<passage from corpys>"
    },
    ...
  ]
}
```

Η έξοδος αυτού του script, είναι το **cf_drafts_with_ctx.jsonl** και θα χρησιμοποιηθεί στο επόμενο στάδιο για την επέκταση των αρχικών ερωτήσεων. Είναι σημαντικό να αναφερθεί πως επιλέχθηκε ο αλγόριθμος BM25, και γενικά η λεξιλογική ανάλυση, για λόγους απλότητας, ταχύτητας και ευχρηστίας. Επίσης το κύριο μειονέκτημα του BM25, δηλαδή η μη-σημασιολογική κάλυψη, δεν παίζει ρόλο στον σκοπό της συγκεκριμένης εργασίας αφού δεν απασχολούν τα συνώνυμα και παρόμοιες έννοιες, αλλά η αυστηρή τεκμηρίωση σύμφωνα με τα κείμενα, για περιορισμό των παραισθήσεων σε Μεγάλο Γλωσσικό Μοντέλο.

Παράλληλα, επιλέχθηκε default τιμή στο k το 5, που θεωρήθηκε ως μια καλή τιμή για να ύπαρξη πλούσιας ποικιλία τεκμηρίων, αλλά όχι υπερβολική.

Συνολικά, το **bm25_retrieval.py** συνδέει την γλωσσική παραγωγή του LLM, με τα πραγματικά δεδομένα των κειμένων. Μετατρέπει κάθε draft ερώτηση, που μπορεί να μην ανταποκρίνεται στο πραγματικό περιεγχόμενο του αντίστοιχου κειμένου, σε ένα πακέτο ερώτησης και συμφραζομένων το οποίο θα χρησιμοποιηθεί στην συνέχεια για την επικύρωση και επέκταση της κάθε ερώτησης και της αντίστοιχης απάντησης της.

5.2.3 Φάση επέκτασης/γείωσης ερωτήσεων και αρχικής απάντησης - ground_questions.py

Αυτό το script, παίρνει ως είσοδο κάθε εγγραφή του **cf_drafts_with_ctx.jsonl** και ζητά απο το μοντέλο να:

1. γειώσει/επεκτείνει την ερώτηση ώστε να είναι αποκλειστικά απαντήσιμη από την συλλογή κειμένων, και σχετικότερη με αυτά
2. να κάνει την αρχική προσπάθεια για να εντοπίσει αν μια ερώτηση είναι μη απαντήσιμη
3. να κάνει την αρχική προσπάθεια για να δώσει σύντομη απάντηση και evidence(φράση απο το απόσπασμα που την απαντά καλύτερα)

Στην συνέχεια εφαρμόζει το δικό του επικυρωτή (validator), προσπαθώντας να επαληθεύσει αν υπάρχει σαφές τεκμήριο στα συμφραζόμενα και με την σειρά του να εξάγει σύντομη απάντηση από τα τεκμήρια (evidence). Αν λείπουν κρίσιμα πεδία, μπορεί να ανατρέψει την ένδειξη απαντησιμότητας. Γίνεται προσπάθεια, όμως, να μην αντιστραφούν σε αυτό το στάδιο πολλές ερωτήσεις (ο έλεγχος είναι χαλαρός), αφού το θέμα της αντιστρεψιμότητας εξετάζεται αναλυτικότερα σε επόμενη φάση. Ακολουθούν τα βασικά σημεία του κώδικα:

Prompting:

- SYSTEM_PROMPT: Δίνονται οι ρόλοι προσεκτικού βαθμολογιτή ("careful grader") και επανα-συντάκτη ερωτήσεων ("question rewriter") στο μοντέλο. Αυτό πρέπει να παραγάγει (με βάση την πρόχειρη ερώτηση (draft question) και τα ανεκτημένα τεκμήρια) μια γειωμένη ερώτηση (grounded question) που είναι απαντήσιμη αποκλειστικά απο τα κείμενα. Δεν πρέπει να γίνει καμία εικασία. Αν το draft ζητά περισσότερη ακρίβεια απ' ό,τι υπάρχει η ερώτηση, είτε χαλαρώνεται (coarser granularity) είτε δηλώνεται ως μη απαντήσιμη. Αν είναι απαντήσιμη, δίνεται απάντηση και τεκμήριο (evidence), αποκλειστικά σύμφωνα με το αντίστοιχο κείμενο. Επιστρέφεται αυστηρό JSON.
- USER_TEMPLATE: Δίνει στο μοντέλο την αρχική ερώτηση, τα ανεκτημένα τεκμήρια και οδηγίες σε περίπτωση που η ερώτηση είναι ήδη σύμφωνη με τα κείμενα(να την κρατήσει), ή δεν υποστηρίζεται(ή να περιοριστεί σε αυτά που λέει το κείμενο ή να θέσει μη-απαντήσιμη την ερώτηση)

Σημαντικό να αναφερθεί πως το script χρησιμοποιεί χαμηλή θερμοκρασία (temperature) ίση με 0.2 για ελαχιστοποίηση τυχαιότητας, έτσι ώστε το grounding να είναι σταθερό. Επιπλέον, περιορίζει πόσα συμφραζόμενα βλέπει το μοντέλο (maxctx με default τιμή 5, αφού παράγουμε 5 contexts και στο BM25) και περιορίζει κάθε απόσπασμα σε ctx-max-chars χαρακτήρες (default 1200 αφού τα κείμενα εδώ είναι μικρά και δεν υπάρχει κίνδυνος να μείνει σημαντική πληροφορία εκτός), ώστε το prompt να μένει εντός παραθύρου συμφραζομένων (context window) και να μην υπάρχει πρόβλημα στο LLM.

Κύριες Βοηθητικές Συναρτήσεις:

- _norm, contains_norm: κανονικοποιούν κείμενο και ελέγχουν αν ένα evidence εμπεριέχεται στα contexts.
- _sentences(text): σπάει ένα κείμενο σε προτάσεις
- _key_tokens(text) + STOPWORDS: εξάγουν λέξεις-κλειδιά (αγνοώντας απλά stopwords) για να υπολογιστεί επικάλυψη tokens.
- _best_support_sentence(context_text, question_text, min_overlap=1): βρίσκει την πρόταση μέσα στα συμφραζόμενα με το μεγαλύτερο token overlap σε σχέση με την ερώτηση, σε περιπτώσεις που το μοντέλο δεν έδωσε evidence.
- _question_type(q): εντοπίζει είδος ερώτησης (yes/no, age, method, which, other) με απλούς κανόνες.
- _extract_yesno,_extract_age,_extract_method,_extract_which: κανόνες εξαγωγής σύντομης απάντησης από μια πρόταση evidence, ανάλογα με τον τύπο της ερώτησης

Σημαντικά να αναφερθούν σε αυτή την φάση είναι δύο πράγματα. Πρώτον, στο συγκεκριμένο σύνολο δεδομένων, δεν υπάρχουν προτάσεις με την κλασσική έννοια (απουσία σημείων στήξης). Οι συναρτήσεις που αφορούν τον χωρισμό σε προτάσεις, λοιπόν, δεν έχουν ιδιαίτερο όφελος για αυτό το σύνολο δεδομένων. Παρόλαυτα, θα μπορούσαν να έχουν σημαντική χρήση για άλλη συλλογή κειμένων, και αποφασίστηκε να συμπεριληφθούν. Σε κάθε περίπτωση, στα επόμενα βήματα, χωρίζεται το κάθε κείμενο με άλλον τρόπο. Δεύτερον, σε αυτή την φάση δεν αποδέχονται απλές Yes/No απαντήσεις αφού μετά απο δοκιμές, οι περισσότερες τέτοιες απαντήσεις ήταν λάθος ή ελληπείς. Αν το μοντέλο επιστρέψει Yes/No, αντικαθιστάται απο κομμάτι του evidence. Σε επόμενο στάδιο επιχειρείται να υπάρχουν και κάποιες τέτοιες μονολεκτικές απαντήσεις, με μεγαλύτερη επιτυχία και μικρότερη συχνότητα.

Ροή Main:

1. Διαβάζει τα records απο την είσοδο. Αν χρειαστεί, κόβει σε maxctx και ctx_max_chars (σε αύτο το σύνολο δεδομένων, με τα δεδομένα που τρέξαμε δεν συμβαίνει). Φτιάχνει string για το contexts και το συνδέει με το draft question στο USER_TEMPLATE. Παράλληλα συλλέγει default used_doc_id

2. Καλεί το LLM με call_ollama(...), και γίνεται parsing της απάντησης με parse_loose_json(raw). Αν είναι αυστηρό JSON το κρατάει, αλλίως κάνει τεμαχισμό (slicing) και προσπαθεί json.loads. Αν αποτύχει, παράμενει η ερώτηση με κενό το evidence

3. Χρησιμοποιείται ο validator. Συνενώνει τα τεκμήρια σε ctx_text. Αν λείπει evidence/answer από το LLM, καλεί _best_support_sentence για να βρει υποψήφια πρόταση-τεκμήριο. Στην συνέχεια, υπάρχει η αρχική απόφαση απαντησιμότητας. Ξεκινά από την αρχική απάντηση του μοντέλου και αν αυτό είπε false αλλά υπάρχει καλή πρόταση, γυρίζει σε true. Αν είναι true και λείπει evidence/answer, γίνεται false. Σε περίπτωση που λείπουν πεδία, συμπληρώνονται με βάση την καλύτερη πρόταση. Όσον αφορά το rationale, συμπληρώνεται αυτόματα για answerable = true ή false.

Η έξοδος γράφεται στο αρχείο **cf_grounded.jsonl** και κάθε entry του έχει την παρακάτω μορφή:

```
{
  "source_id": <id>,
  "question_ix": <0|1>,
  "draft_question": "<draft question>",
  "grounded_question": "<grounded question>",
  "answer": "<answer>",
  "rationale": "<Supported by..",
  "used_doc_ids": [...],
  "answerable": true/false,
  "evidence": "<section from passage>",
  "generated_at": <timestamp>,
  "model": "llama3.1"
}
```

Συνολικά, στο στάδιο αυτό, οι ερωτήσεις προσαρμόζωνται έτσι ώστε να είναι σχετικές με τα ανεκτημένα τεκμήρια. Στην ουσία γίνεται μια επέκταση ερωτήματος καθοδηγούμενη απο ανάκτηση (εφαρμόζεται RAG). Επιτυγχάνεται, με αυτή την διαδικασία γείωσης, να περιοριστούν σημαντικά οι παραισθήσεις που μπορεί να προκαλούνται απο το μοντέλο στο κομμάτι παραγωγής των ερωτήσεων, αφού κάθε ερώτηση ελέγχεται και αν χρειαστεί εμπλουτίζεται είτε σε μικρό είτε σε μεγαλύτερο βαθμό. Δίνεται ιδιαίτερη έμφαση στην ακρίβεια και στην ύπαρξη τεκμηριών για την κάθε ερώτηση έτσι ώστε να έχει αξία σε οποιοδήποτε τομέα χρήσης, όχι μόνο στον τομέα της εκπαίδευσης που έχει δωθεί παραπάνω εστίαση. Η προσπάθεια για τον εντοπισμό της σωστής απάντησης στην κάθε ερώτηση, συνεχίζεται στα επόμενα στάδια του pipeline, αφού σε αυτή την φάση υπάρχουν απαντήσεις που μπορεί να είναι και ολόκληρο το κείμενο, λόγω απουσίας στήξης στην συλλογή κειμένων. Ενσωματώθηκαν, όμως, οι κατάλληλες συναρτήσεις στον κώδικα έτσι ώστε να έχει καλύτερη συμπεριφορά σε αυτό το στάδιο και για άλλες συλλογές κειμένων που μπορεί να χρησιμοποιηθούν για την παραγωγή και επέκταση ερωτήσεων, στο αντικείμενο που πραγματεύονται.

5.2.4 Φάση περικοπής/επικύρωσης απαντησεων - prune_answers.py

Το script αυτό, παίρνει ως είσοδο κάθε εγγραφή από το **cf_grounded.jsonl**, δηλαδή την grounded ερώτηση σε συνδιασμό με την αρχική απάντηση και το evidence που παρήχθησαν στο προηγούμενο βήμα. Στόχος είναι η παραγόμενη απάντηση να είναι σύντομη και τεκμηριωμένη, μήκους μικρότερο ή ίσο με ένα όριο χαρακτήρων (προεπιλογή 140). Αν δεν μπορέσει να βρει τέτοια φράση με ασφάλεια/τεκμηρίωση, γυρνάει σε μη-απαντίσημη (unanswerable). Έχει μεγάλη σημασία, η κάθε ερώτηση να έχει μια σύντομη και πιστή στα κείμενα απάντηση έτσι ώστε το σύστημα να έχει πρακτική χρήση. Αυτό το script, πετυχαίνει ακριβώς αυτό. Παρακάτω θα αναλυθούν τα κύρια σημεία του κώδικα:

Prompting:

- SYSTEM_PROMPT: Δίνονται αυστηρές οδηγίες στο μοντέλο να ξαναγράψει την απάντηση με σύντομο τρόπο, περιορισμένη σε όριο χαρακτήρων, να κρατήσει μόνο πράγματα που υπάρχουν αυτούσια στο κείμενο/evidence. Για yes/no ερωτήσεις, να το απαντάει μόνο εαν φαίνεται ξεκάθαρα απο το κείμενο, και αν δεν μπορεί να σχηματίσει κάποια απάντηση, να επιστρέψει κενό. Τελικά, επιστρέφει αυστηρό JSON.
- USER_TEMPLATE: Δίνει στο μοντέλο την γειωμένη ερώτηση, τον μέγιστο αριθμό χαρακτήρων, την τωρινή απάντηση, το evidence και το κείμενο.

Κύριες Βοηθητικές Συναρτήσεις:

- _norm, _normalize_spaces, _token_set, _key_tokens: κανονικοποιούν κείμενο, και φτιάχνουν σύνολα tokens
- _sentences(text): σπαει σε προτάσεις (στην συγκεκριμένη συλλογή κειμένων που δεν υπάρχουν προτάσεις, συνήθως επιστρέφει το passage)
- _best_support_sentence(context_text, question_text): εντοπίζει την καλύτερη υποστηρικτική πρόταση με overlap tokens
- _question_type(q), _extract_yesno(sentence), _extract_age, _extract_method, _extract_number, _extract_which: εξαγωγή τύπου ερώτησης
- _shorten_to_phrase(s, max_chars): κόβει σε μικρή φράση
- _is_prefix_clip_answer(ans, evidence, contexts, max_chars), _raw_prefix_clip, _is_bare_prefix, _has_punctuation_before_cap) : με αυτές τις συναρτήσεις ελέγχεται αν η απάντηση είναι απλά ένα πρόθεμα (κομμένο στο όριο χαρακτήρων) του passage/evidence. Προφανώς δεν είναι επιθυμητό να υπάρχουν τέτοιες απαντήσεις, που δεν έχουν φυσικό τέλος ή προέκυψαν λόγω αποτυχίας εντοπισμού κατάλληλης φράσης-τεκμηρίου.

- supported(text): γίνεται έλεγχος υποστήριξης. Επιστρέφει True, αν το text εμπεριέχεται στο evidence ή αν υπάρχει πρόταση στα συμφραζόμενα με Jaccard μεγαλύτερο ή ίσο του 0.7(και αυτό στο συγκεκριμένο σύνολο δεδομένων ίσως δεν βοηθάει πολύ) ως προς τα tokens του text. Έτσι αποφεύγονται παραφράσεις που δεν πατάνε ρητά στο τεκμήριο.

Ροή Main:

Για κάθε record της εισόδου που διαβάζει:

1. Αν answerable = False, περνάει ως έχει
2. Ελέγχεται αν χρειάζεται pruning. Δηλαδή, ελέγχει αν answerable = true και το μήκος της απάντησης είναι μεγαλύτερο από το όριο χαρακτήρων ή αν μιλάμε για απλό πρόθεμα του κειμένου (μέσω _is_prefix_clip_answer()). Σε περίπτωση που αναγνωριστεί ως πρόθεμα, επιχειρείται να διορθωθεί, αλλίως σημειώνεται ως μη απαντήσιμη.
3. Καλείται το LLM (μέσω rec = prune_one(rec, cfg, args.max_chars, args.ctx_max_chars) και γίνεται η διαδικασία της περικοπής. Αν το pruned_answer είναι μικρότερο από το όριο χαρακτήρων και υποστηρίζεται απο το τεκμήριο, δέχεται την απάντηση, ενημερώνει το evidence και σημειώνει στο rationale το μήκος της απάντησης πρίν και μετά την διαδικασία. Αν το pruned_answer είναι κάτι άλλο (στην περίπτωση εδώ, yes/no) πάει σε fallback
4. Γράφεται το νέο record με όλα τα ενημερωμένα πεδία

Το script παράγει έξοδο το αρχείο **cf_grounded_pruned.jsonl** το οποίο έχει την ίδια δομή με το **cf_grounded.jsonl**, με το επιπλέον πεδίο pruned_at που δίνει το timestamp που έγινε η περικοπή και επιπλέον λεπτομέρειες σχετικά με το pruning στα γνωστά πεδία (ένδειξη για το παλιό και το νέο μήκος απάντησης και ένδειξεις για prefix, fallback στο rationale)

Σε αυτό το σημείο, να αναφέρουμε ότι πλεόν υπάρχει υποστήριξη για Yes/No απαντήσεις. Θεωρήθηκε καλύτερη πρακτική να γίνει εδώ ο έλεγχος για τέτοιου είδους απαντήσεις, και αποδείχθηκε και στην πράξη καλύτερο αφού πλεόν δεν εντοπίζονται πολλές και λανθασμένες απαντήσεις αυτού του τύπου. Επιστρέφεται τέτοια απάντηση, μόνο αν υπάρχει ρητά στο κείμενο. Παράλληλα, να σημειωθεί ότι ο έλεγχος για το κομμάτι προθέματος (prefix clip) στην συγκεκριμένη συλλογή κειμένων μπορεί να θεωρηθεί λίγο αυστηρός λόγω απουσίας σημείων στήξης, αφού μπορεί κάποιες πιθανές απαντήσεις να απορριφθούν. Αυτό δεν αποτελεί πρόβλημα στην συνέχεια καθώς έχει υλοποιηθεί ένα επιπλεόν βήμα επανα-Ανάκτησης για περιπτώσεις όπου η ερώτηση έχει χαρακτηριστεί ως μη-απαντήσιμη. Συνολικά, το script αυτό, αποτελεί το στάδιο που ¨καθαρίζει' το σύνολο ερωταπαντήσεων, και το φέρνει ένα βήμα πιο κοντά

στην στην απόλυτη συμφωνία κειμένου-ερωτήσεων-απαντήσεων, γεγονός που εξαλείφει και τον κίνδυνο των παραισθήσεων.

5.2.5 Φάση τοπικής επανα-Ανάκτησης και διάσωσης ερωταπαντήσεων - salvage_unanswerablev4.py

Το script αυτό, παίρνει ως είσοδο το **cf_grounded_pruned.jsonl** και σκοπός είναι να διασωθούν μερικές απο τις απαντήσεις που σημάνθηκαν ως μη απαντήσιμες στο προηγούμενο στάδιο. Οι λόγοι για αυτή την λανθασμένη σήμανση, μπορεί να είναι απουσία κατάλληλης φράσης-τεκμήριου που την επιβεβαιώνει εξαιτίας μη ολοκληρωμένης ανάκτησης στο αντίστοιχο στάδιο, ή λόγω σήμανσης του evidence ως πρόθεμα, στο στάδιο περικοπής απάντησης. Σκοπός είναι να περιοριστούν οι λανθασμένες σημάνσεις όσο το δυνατόν περισσότερο. Δεν επιδιώκεται, όμως, η μετατροπή όλων των μη απαντήσιμων περιπτώσεων σε απαντήσιμες, καθώς στόχος είναι οι ερωταπαντήσεις να παραμένουν επαρκώς τεκμηριωμένες. Ακολουθήθηκε, λοιπόν, μια προσέγγιση η οποία διασφαλίζει ότι κάθε ερώτηση ελέγχεται ξανά, για πιθανά τεκμήρια και αν αυτά υπάρχουν, σημειώνεται ως απαντήσιμη. Συγκεκριμένα το script:

1. Παίρνει την σημασμένη ως μη-απαντήσιμη γειωμένη ερώτηση και το πλήρες, αντίστοιχο κείμενο. Συγκεκριμένα παίρνει το κείμενο για το οποίο παράχθηκε αρχικά η ερώτηση απο το LLM. Η λογική για αυτή την επιλογή είναι πως αφού δεν απαντήθηκε η ερώτηση στα προηγούμενα βήματα, και δεν βοήθησε η καθολική ανάκτηση στην συλλογή κειμένων, ο καλύτερος και πιο εύκολος τρόπος να εντοπιστεί απάντηση για την ερώτηση, είναι να την αναζητήσουμε στο original passage. Έχει εξετάσει και την υπόλοιπη συλλογή, και πάλι δεν βρέθηκε κάτι σχετικό, επομένως θα ακολουθήθεί η ¨απλή' πρακτική και θα ερευνηθεί το αρχικό κείμενο.

2. Τεμαχίζει το passage σε μικρά αποσπάσματα (είτε προτάσεις, είτε κινούμενα παράθυρα (sliding windows) αν δεν υπάρχουν τελείες, όπως στην συγκεκριμένη συλλογή).

3. Βαθμολογεί τα αποσπάσματα με συνδυασμό token overlap + in-passage BM25, για μεγαλύτερη κάλυψη και ακρίβεια, και κρατά τα top-k (προεπιλογή 4).

4. Στέλνει τα αποσπάσματα στο LLM και ζητά να επιστρέψει σύντομη φράση που απαντά την εκάστοτε ερώτηση, που βρίσκεται σε αυτά

5. Αν βρεθεί απάντηση, την δέχεται και η σήμανση της ερώτησης αλλάζει σε απαντίσιμη. Αν δεν βρεθεί, δοκιμάζει νέο πέρασμα με μεγαλύτερο window. Αν αποτύχει και αυτό, παραμένει μη-απαντήσιμη η ερώτηση, αφου όυτε με αυτή την διαδικασία βρέθηκε απάντηση.

Μετά απο και αυτό το στάδιο, οι σημασμένες ως μη-απαντίσημες ερωτήσεις, είναι κατα

πάσα πιθανότητα προϊόν παραισθήσεων του μοντέλου. Παρακάτω θα αναφερθούν τα κύρια σημεία του script:

Επανα-Ανάκτηση:

Για την διαδικασία ανάκτησης του Retrieval-Augmented Generation που εφαρμόστηκε σε αυτό το στάδιο, χρησιμοποιήθηκαν:

- Token overlap: Εξάγει key tokens από την ερώτηση και από κάθε απόσπασμα. Υπολογίζει επικάλυψη (overlap) $|Q \cap S|$. Μεγαλύτερο overlap συνεπάγεται μεγαλύτερη συνάφεια.

- In-passage BM25: Φτιάχνει BM25 μοντέλο μόνο πάνω στα αποσπάσματα του ίδιου passage, υπολογίζει idf (_bm25_fit(docs)), tf(_bm25_score(model, qtok, stok, k1=1.5, b=0.75)) και βαθμολογείται κάθε απόσπασμα

- Συνδυασμός: Η συνάρτηση _rank_inside_passage(...) συνδυάζει τις δύο παραπάνω μεθόδους, και υπολογίζει τον γραμμικό συνδιασμό τους (rows.append((alpha * overlap + beta * bm25, overlap, bm25, snippet)) όπου alpha, beta = 1). Επιστρέφονται τα top-k αποσπάσματα.

Βλέπουμε, πως η επανα-ανάκτηση είναι source-restricted, και γίνεται νέο ranking ανάμεσα στα κομμάτια του κειμένου. Έτσι λοιπόν, διορθώνονται περιπτώσεις όπου στο pruning δεν βρέθηκε μικρή φράση, που το κείμενο μπορεί να την περιέχει, αλλά και περιπτώσεις που η αρχική ανάκτηση δεν ήταν αρκετή για εντοπισμό κατάλληλης απάντησης.

Prompting:

- SYSTEM_PROMPT: Δίνεται ρόλος προσεκτικού 'answer extractor' στο μοντέλο και του ζητά να απαντήσει στις ερωτήσεις, ρητά με βάση τα τεκμήρια, να επιστρέψει την συντομότερη φράση, να μην απαντά με ναι/όχι(δημιουργούσε προβλήματα) και να επιστρέψει αυστηρό JSON

- USER_TEMPLATE : Δίνει στο μοντέλο την επεκταμένη ερώτηση, το όριο χαρακτήρων, τα τεκμήρια και ζητά JSON

Κύριες Βοηθητικές Συναρτήσεις:

- _shorten_wordsafe: Κόβει σε λέξη, σε περίπτωση που είναι κοντά στο όριο
- _supported(span, evidence_text): Ελέγχει για supported evidence (όπως και σε προηγούμενα scripts)
- _window_slices_by_tokens(...), _sentences_or_window(...): Χωρίζουν σε windows τα κείμενα
- _rank_inside_passage(...): Πραγματοποιεί την τοπική ανάκτηση μέσα στο ίδιο passage, συνδυάζοντας token overlap και BM25 για την κατάταξη των αποσπασμάτων. Επιστρέφει τα top-k πιο σχετικά για το εκ νέου extraction.
- _try_extract_with_llm(...): Καλεί το LLM δίνοντας του τα top αποσπάσματα και την ερώτηση, ζητώντας την συντομότερη απάντηση που υπάρχει μέσα στα αποσπάσματα.

Ροή Main:

1. Φορτώνονται τα αρχεία εισόδου, με τις, εως τώρα, ερωταπαντήσεις και την συλλογή κειμένων.
2. Για κάθε record ελέγχεται αν answerable = false ή αν η απάντηση είναι κενή. Αν ισχύει ένα απο τα δύο, παίρνει το αρχικό κείμενο και κάνει την διαδικασία ανάκτησης που αναφέραμε παραπάνω (μέσω salvage_record(..)). Αρχικά σπάει το κέιμενο σε windows 50/25 και αν αποτύχει, δοκιμάζει 80/30. Αν βρεθεί κατάλληλη φράση, το record ενημερώνεται με την νέα απάντηση. Αν δεν βρεθεί, γράφεται ένδειξη αποτυχίας.
3. Γράφεται η νέα έγγραφη, με κατάλληλο timestamp.

Το script παράγει έξοδο το αρχείο **cf_grounded_pruned_salvagedv4.jsonl** το οποίο έχει την ίδια δομή με τα προηγούμενα, με το επιπλέον πεδίο salvaged_at που δίνει το timestamp που έγινε η επανα-ανάκτηση και επιπλέον λεπτομέριες σχετικά με το salvage στα γνωστά πεδία (πχ ένδειξη για salvage failed).

Συνολικά, αυτό το στάδιο ολοκληρώνει την διαδικασία παραγωγής πιστοποιημένων ερωταπαντήσεων με χρήση LLM χωρίς τον κίνδυνο των παραισθήσεων. Γίνεται ο τελευταίος έλεγχος, στο πλαίσιο της επικύρωσης απαντήσεων, και στην σήμανση ερωτήσεων ως μη απαντήσιμες. Το τελικό σύνολο δεδομένων έχει εξεταστεί σε μεγάλο βαθμό, και γίνεται με βεβαιότητα να ειπωθεί ότι τα σετ που έχουν απομείνει, δεν αποτελούν παραισθήσεις του μοντέλου, ούτε στο πλαίσιο της ερώτησης ούτε στο πλαίσιο της απάντησης της. Συγκεκριμένα,

σε αυτό το στάδιο διορθώνεται ό,τι ερώτηση μπορεί να είχε λανθασμένα σημανθεί ως μη απαντίσημη στο στάδιο της περικοπής, είτε λόγω της αυστηρής λογικής επικύρωσης, είτε λόγω ανάκτησης στο πρώτο στάδιο (bm2_retrieval.py). Με αυτή την μεθοδολογία, έγινε εστίαση στην μείωση των unanswerables, χωρίς να εισαχθεί επιπλέον κίνδυνο παραισθήσεων.

Είναι σημαντικό να σημειωθεί, το ότι δοκιμάστηκαν μερικές ακόμη προσεγγίσεις στην διαδικασία επανα-ανάκτησης κυρίως με εφαρμογή απλού token overlap, αλλά ο συνδυασμός του με BM25 επέφερε καλύτερα αποτελέσματα, και συγκεκριμένα οι 34 διασωθέντες ερωταπαντήσει με την πρώτη προσέγγιση, έγιναν 50 με αυτήν που επιλέχθηκε εν τέλη (εξού και η ονομασία v4 στο αρχείο)

5.2.6 GUI εκπαιδευτικής εφαρμογής - quiz_gen.py

Στο σημείο αυτό, η κύρια ροή του pipeline έχει τελειώσει. Το τελικό αρχείο που παράγεται μετά και απο το στάδιο της επανεξέτασης των μη απαντήσιμων περιπτώσεων, είναι ανθεκτικό σε σετ ερωταπαντήσεων που αποτελούν παραισθήσεις ή δεν έχουν ξεκάθαρη απάντηση στα κείμενα. Ένα τέτοιο σύνολο δεδομένων μπορεί να έχει αρκετές χρήσεις και η ύπαρξή του έχει μεγάλη αξία. Για αυτό λοιπόν και υλοποιήθηκε η εφαρμογή που θα περιγραφεί παρακάτω. Συγκεκριμένα, πρόκειται για μια απλή αλλά λειτουργική διεπαφή χρήστη που εστιάζει στον τομέα της εκπαίδευσης και αξιοποιεί αποτελεσματικά τα σετ ερωταπαντήσεων που παρήχθησαν απο το pipeline που σχεδιάστηκε. Η εφαρμογή υλοποιήθηκε σε Python χρησιμοποιώντας το framework Streamlit, το οποίο επιτρέπει τη γρήγορη ανάπτυξη διαδραστικών web εφαρμογών χωρίς ανάγκη επιπλέον backend.

Η εφαρμογή φορτώνει το τελικό σύνολο δεδομένων (cf_grounded_pruned_salvagedv4.jsonl), το οποίο περιέχει μόνο επικυρωμένες και απαντήσιμες ερωτήσεις. Δίνει την δυνατότητα στον χρήστη να απαντήσει σε ένα τυχάιο σύνολο ερωτήσεων από το σύνολο (10 σε αριθμό) και στην συνέχεια να δει την απόδοσή του με βάση μετρικές ομοιότητας μεταξύ της απάντησης που έδωσε ο ίδιος και της σωστής απάντησης που έδωσε το μοντέλο, για κάθε ερώτηση που απαντήθηκε.

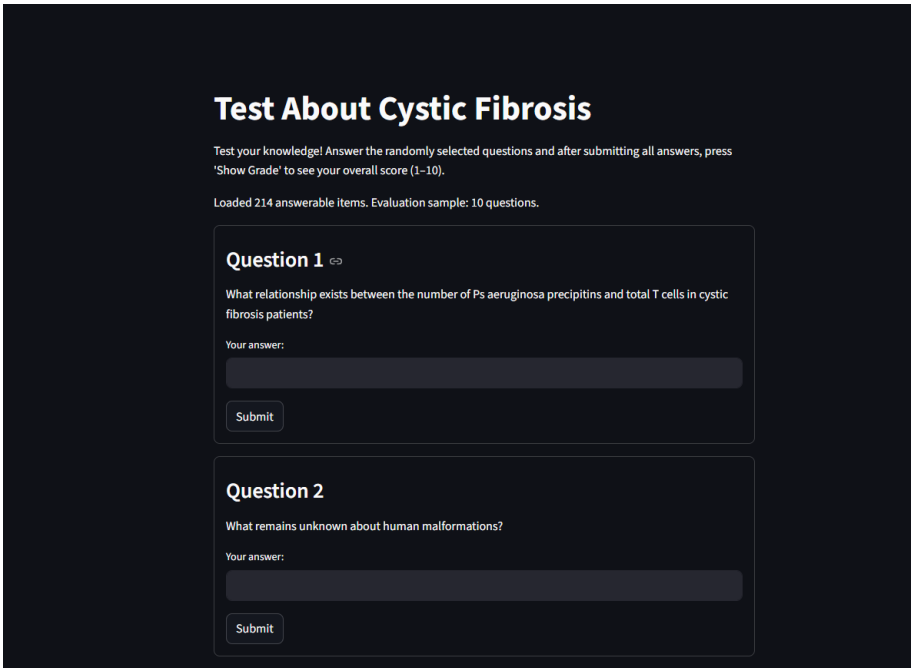
Βασική Λειτουργία

- Φορτώνεται το αρχείο και κρατούνται μόνο τα πεδία grounded_question και answer, για όσες ερωτήσεις έχουν answerable = True. Δημιουργείται μοναδικό qid για κάθε ερώτηση.

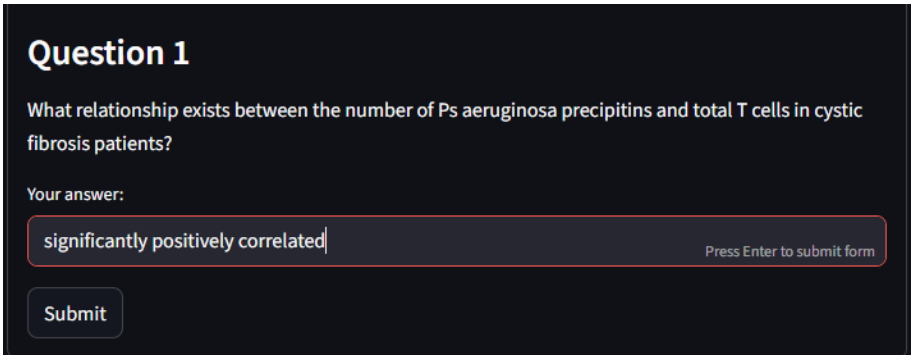
- Επιλέγεται ένα τυχαίο σύνολο 10 ερωτήσεων, έτσι ώστε να ενθαρρύνεται η επαναληψιμότητα της εφαρμογής.
- Για κάθε ερώτηση, ο χρήστης γράφει την απάντηση του στο πεδίο text_input και πατάει το κουμπί "Submit" για να κλειδώσει την απάντηση του και να ελεγχθεί η ομοιότητα της με την απάντηση του μοντέλου.
- Για κάθε απάντηση, χρησιμοποιούνται δύο μετρικές. Η ROUGE-1 Precision, που είναι και η πιο σημαντική για την ομοιότητα των απαντήσεων και η BERTScore Precision, αν είναι διαθέσμη. Η τελική βαθμολογία είναι ο μέσος όρος ROUGE-1, σε κλίμακα 1–10, στρογγυλοποιημένος στο πλησιέστερο 0.5.
- Όταν απαντηθούν όλες οι ερωτήσεις, ο χρήστης μπορεί να δεί την βαθμολογία του πατώντας "Show Grade". Υπάρχει και η δυνατότητα για εγκατάσταση CSV αρχείου που υπάρχουν αναλυτικά οι απαντήσεις και οι μετρικές, αλλά αυτό είναι περισσότερο για να το εξετάσουμε εμείς.

Συνολικά η εφαρμογή αυτή μετατρέπει το παραγόμενο σύνολο δεδομένων, μετά απο την διαδικασία που ακολουθήθηκε, σε ένα εργαλείο αξιολόγησης. Η χρήση του μπορεί να είναι η αυτοαξιολόγηση αλλά και αξιολόγηση απο κάποιον άλλον φορέα (αν για παράδειγμα τα αποτελέσματα του quiz αποσταλούν σε κάποιον εκπαιδευτικό). Ουσιαστικά, η εφαρμογή λειτουργεί ως απόδειξη της χρησιμότητας του συστήματος που υλοποιήθηκε, αφού το pipeline παράγει έγκυρο και απόλυτα σύμφωνο με τα δοθέντα κείμενα υλικό το οποίο μπορεί να χρησιμοποιηθεί χωρίς ανθρώπινη επιμέλεια. Επιτυγχάνεται, συνεπώς, ο αρχικός στόχος της διπλωματικής δηλαδή ο περιορισμός παραισθήσεων σε QA συστήματα μέσω επέκτασης ερωτημάτων και προσεκτικής εξαγωγής απαντήσεων, πάντα σύμφωνα με τεκμήρια.

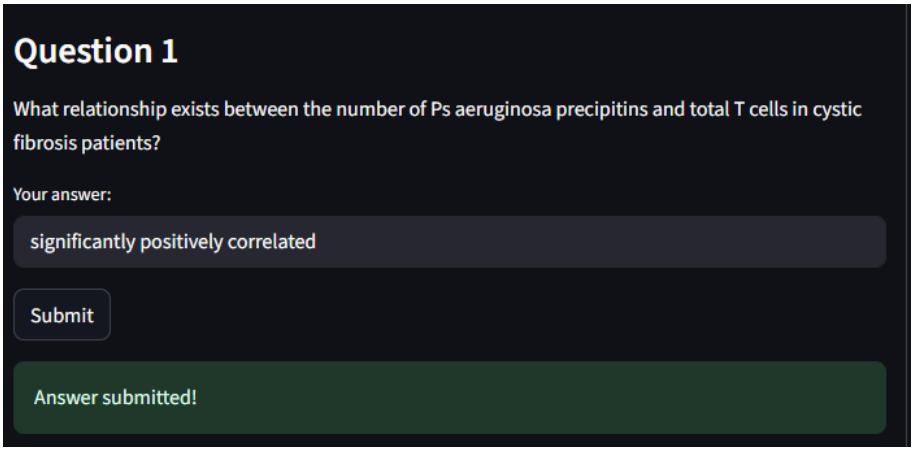
Παρακάτω ακολουθούν εικόνες που δείχνουν τα στοιχεία της εφαρμογής, όπως περιγράφηκαν προηγουμένως:



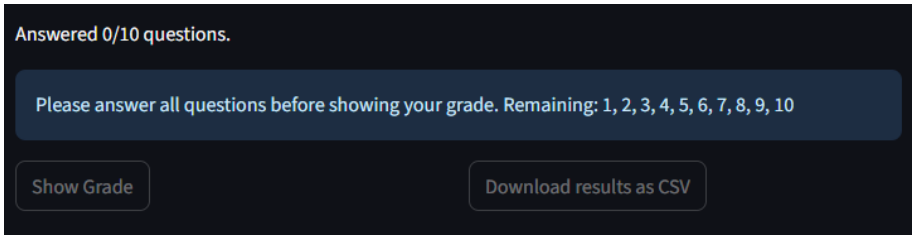
Σχήμα 5.2: Προεπισκόπιση της εφαρμογής. Εμφανίζονται οι ερωτήσεις στον χρήστη



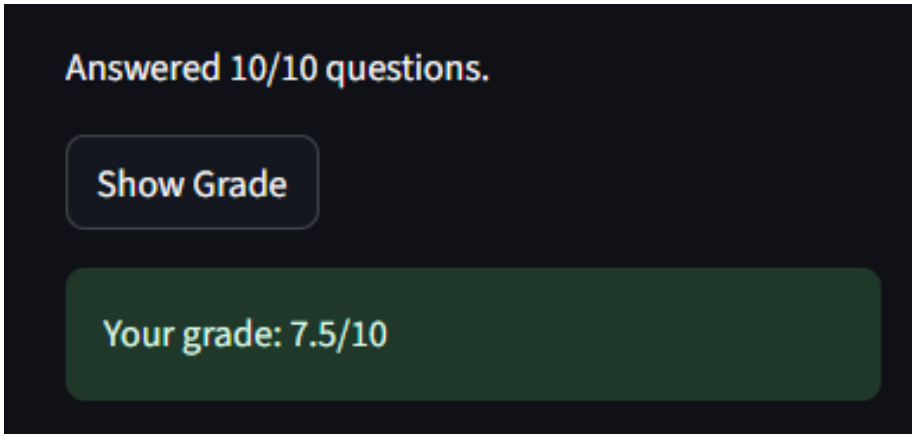
Σχήμα 5.3: Ο χρήστης γράφει την απάντηση του, μέχρι να είναι σίγουρος για αυτήν



Σχήμα 5.4: Ο χρήστης πατάει "Submit" και η απάντηση κατοχυρώνεται



Σχήμα 5.5: Σε περίπτωση που δεν έχουν απαντηθεί όλες οι ερωτήσεις, ο χρήστης βλέπει ποιες απομένουν για να τις απαντήσει, και να έχει την δυνατότητα να βαθμολογηθεί



Σχήμα 5.6: Ο χρήστης πατάει "Show Grade" και βλέπει την βαθμολογία του σε κλίμακα 1-10. Επίσης, υπάρχει η δυνατότητα ΄κατεβάσματος΄ των αποτελεσμάτων μέσω του "Download results as CSV"

5.2.7 Κώδικας αξιολόγησης απαντήσεων

Για να γίνει αποτελεσματικά η αξιολόγηση του κώδικα(θα γίνει αναλυτική αναφορά της διαδικασίας και των αποτελεσμάτων στο επόμενο κεφάλαιο), υλοποιήθηκε ένα επιπλέον script που επιτυγχάνει ακριβώς αυτό. Είναι το **evaluate_qas.py**, το οποίο συνδέει τις ανθρώπινες αξιολογήσεις που πραγματοποιήθηκαν χειρονακτικά, με τις απαντήσεις που έδωσε το μοντέλο (δηλαδή τα στοιχεία του **cf_grounded_pruned_v4.jsonl**). Στην ουσία, το script αυτό υπολογίζει τις τιμές ROUGE-1 (P/R/F1), ROUGE-L (F1) και BERTScore (P/R/F1) αν είναι εγκατεστημένο. Στην συνέχεια, γράφει ένα CSV αρχείο με μία γραμμή για κάθε σύγκριση (μία ανθρώπινη έναντι της απάντησης του μοντέλου). Το αρχείο εισόδου που περίεχει χειρόγραφες απαντήσεις είναι το **my-evaluations.txt**.

Κύρια Στοιχεία

1. **BERTScore import**: Προσπαθεί να κάνει import bertscore.score. Αν αποτύχει, θέτει BERTSCORE_AVAILABLE=False και γεμίζει None στα πεδία του BERTScore, ώστε το script να τρέχει ακόμη κι αν δεν υπάρχει εγκατεστημένο.
2. **Υπολογισμός ROUGE:** Έχουν υλοποιηθεί μερικές συναρτήσεις, για τον υπολογισμό της τιμής ROUGE. Η normalize_text() κανονικοποιεί τις απαντήσεις (lowercaseing, αφαίρεση μη αλφαριθμητικών κτλ, η tokenize() κάνει απλή διάσπαση μετά την κανονικοποίηση, η rouge_1_p_r_f1() υπολογίζει το unigram overlap και οι _lcs(), rouge_l_f1() υπολογίζουν την LCS-based ROUGE-L.
3. **Υπολογισμός BERTScore:** Η run_bertscore(refs, hyps, lang) τρέχει bertscore_score(hyps, refs, ...) και επιστρέφει λίστες P/R/F. Αν το πακέτο λείπει ή υπάρξει εξαίρεση, επιστρέφει λίστες από None.
4. **Parser για το my-evaluations.txt:** Η parse_human_file(txt) εντοπίζει την ερώτηση που έχουμε απαντήσει (μέσω source_id και question_ix). Εντοπίζει, παράλληλα, την απάντηση που δώσαμε (ή παραλλαγές απαντήσεων αν έχουν δωθεί). Τέλος, εντοπίζει πιθανά σχόλια που κάναμε για κάποια συγκεκριμένη ερώτηση ή απάντηση.

Κύρια Ροή Main

1. Διάβαζει τα κατάλληλα αρχεία και ορισμάτα. Συγκεκριμένα, φορτώνει το my-evaluations.txt και το περνά από parse_human_file. Φορτώνει επίσης το JSONL των απαντήσεων του μοντέλου και χτίζει ένα λεξικό model_map[(source_id, question_ix)] = record.
2. Ξεκινάει η διαδικασία των συγκρίσεων ανθρώπινης απάντησης και απάντησης μοντέλου. Για κάθε ανθρώπινη απάντηση, ψάχνει την αντιστοιχία μέσω (source_id, question_ix).

Αν υπάρχει αντιστοίχιση, παίρνει answer και σημειώνεται model_found=True, αλλιώς σημειώνει το ζεύγος στα missing_pairs για debugging αργότερα. Αν υπάρχει κάποιο variant απάντησης, δημιουργείται ξεχωριστή γραμμή έτσι ώστε να αναγνωριστεί (το συγκεκριμένο υλοποιήθηκε σε πλαίσιο δοκιμών).

3. Υπολογίζει τις τιμές των μετρικών μεταξύ απάντησης μοντέλου και απάντησης αναφοράς. Συγκεκριμένα, υπολογίζονται ROUGE-1 P/R/F1 με rouge_1_p_r_f1(), ROUGE-L F1 με rouge_l_f1() και BERTScore με run_bertscore().

4. Γράφει το CSV εξόδου και τυπώνει πόσες γραμμές αποθήκευσε. Αν υπάρχουν "missing pairs", τυπώνει τα πρώτα 5 για έλεγχο.

Το τελικό αποτέλεσμα είναι ένα CSV αρχείο, το **qa_eval_metrics.csv**, στο οποίο περιέχονται οι τιμές των μετρικών σύγκρισης μεταξύ των ανθρώπινων απάντησεων και των απάντησεων του μοντέλου. Στο σημείο αυτό, είναι σημαντικό να αναφερθεί ότι ο εντοπισμός των απαντήσεων γίνεται με βάση το αναγνωριστικό source_id χωρίς αυτό να σημαίνει ότι η ερώτηση απαντάται αναγκαστικά απο το αντίστοιχο κείμενο. Υπάρχει πάντα το ενδεχόμενο, σύμφωνα με τον σχεδιασμό του συστήματος, η απάντηση στην ερώτηση να βρίσκεται σε άλλο κείμενο το οποίο σημάνθηκε ως σχετικότερο με αυτήν. Για να απαντηθούν οι ερωτήσεις πραγματοποιήθηκε έλεγχος στο rationale έτσι ώστε να εντοπιστεί ποιό κείμενο σημάνθηκε ως το πιο σχετικό (ένδειξη Supported by doc_id(s)[...] και χρησιμοποιήθηκε αυτό για άντληση της απάντησης. Για να διαπιστωθεί και η καλή λειτουργία της ανάκτησης σχετικών κειμένων, έγινε προσπάθεια εξαγωγής απάντησης και με βάση το source_id-κείμενο. Κάποιες φορές η απάντηση έβγαζε νοήμα αλλά πάντα το σημασμένο ως σχετικότερο κείμενο έδινε πιο ξεκάθαρη και ολοκληρωμένη απάντηση για την ίδια ερώτηση, γεγονός που τονίζει την αξιοπιστία και εγκυρότητα του συστήματος.

Κεφάλαιο 6

Έλεγχος - Αξιολόγηση

Στο κεφάλαιο αυτό θα γίνει μια συνολική αξιολόγηση των αποτελεσμάτων του pipeline. Θα εξεταστεί τόσο το κομμάτι της επέκτασης και γείωσης των ερωτήσεων, όσο και το κομμάτι της εγκυρότητας της απάντησης. Για το πρώτο θα αναφερθούν μερικά παραδείγματα και για το δεύτερο έχει πραγματοποιηθεί μια αξιολόγηση με βάση μετρικές, οι οποίες θα αναλυθούν και θα σχολιαστούν τα αποτελέσματα. Παράλληλα, θα γίνει μια αναφορά και στην συλλογή κειμένων που χρησιμοποιήθηκε για το πείραμα που πραγματοποιήθηκε.

6.1 Περιγραφή της Συλλογής Κειμένων

Για την ανάπτυξη του συστήματος και την αξιολόγηση του, χρησιμοποιήθηκε μια συλλογή ιατρικών κειμένων που αφορούν την Κυστική Ίνωση (Cystic Fibrosis). Στην αρχική της μορφή, ήταν ένα σύνολο απο αρχεία απλού κειμένου (plain text). Η ουσιαστική αλλαγή που πραγματοποιήθηκε για τις ανάγκες της συγκεκριμένης εργασίας, ήταν η μετατροπή των πολλών αυτών αρχείων (1239 σε αριθμό) σε ένα ενιαίο αρχείο CSV. Το αρχείο αυτό, περιέχει όλα τα επι μέρους διαθέσιμα κείμενα, με μοναδικό αναγνωριστικό (id) για το κάθε ένα. Η μετατροπή έγινε για λόγους πρακτικότητας και ευκολίας ενσωμάτωσης στον κώδικα. Με αυτόν τον τρόπο, όλα τα αποσπάσματα συγκεντρώνονται σε ένα αρχείο, διευκολύνοντας την επεξεργασία τους και επιτρέποντας στο σύστημα να τα χειρίζεται αποδοτικά, χωρίς να χρειάζεται να ανοίγει εκατοντάδες επιμέρους αρχεία κατά την εκτέλεση του pipeline. Παράλληλα, επιλέχθηκε μια τέτοια θεματική συλλογή για δύο κύριους λόγους. Αρχικά, το εξειδικευμένο πεδίο γνώσης που υπάρχει σε αυτήν οδηγεί το σύστημα RAG στο να παρουσιάζει καλύτερη απόδοση, όπως έχει αναφερθεί και παραπάνω [44]. Παράλληλα, η εκπαιδευτική αξία μιας τέτοιας συλλογής κειμένων εξυπηρετεί τον τελικό στόχο της διπλωματικής, ο οποίος είναι η παραγωγή ενός συστήματος ερωταπαντήσεων, με χρήση Μεγάλων Γλωσσικών Μοντέλων, απαλλαγμένου απο τον κίνδυνο των παραισθήσεων.

6.2 Περιγραφή Πειράματος

Το πείραμα που πραγματοποιήθηκε εστιάζει στη βελτίωση της αξιοπιστίας ενός συστήματος ερωταπαντήσεων με χρήση RAG. Η διαδικασία αξιολογήθηκε και σε κομμάτι Ανάκτησης Πληροφορίας (Information Retrieval-IR) αλλά και σε κομμάτι περιλήψεων (Συμμαριζατιον). Το κομμάτι των περιλήψεων αφορά κυρίως την παραγωγή των απαντήσεων(η σύνδεση αυτή θα επεξηγηθεί στην συνέχεια) και το κομμάτι του IR απασχόλησε παραπάνω στην αξιολόγηση των επεκταμένων ερωτήσεων(μέσω χειρονακτικού ελέγχου και αξιολόγησης με βάση αρχικές-τελικές ερωτήσεις και σχετικό κείμενο). Τα αποτελέσματα του πειράματος και συνολικά όλου του συστήματος, θα παρουσιαστούν παρακάτω.

6.3 Αποτελέσματα διαδικασίας Γείωσης/Επέκτασης ερωτήσεων

Ένα απο τα κομμάτια που δώθηκε ιδιαίτερη έμφαση κατά την εκπόνηση της διπλωματικής αυτής εργασίας είναι η γείωση/επέκταση των παραγώμενων ερωτήσεων σύμφωνα με ανακτόμενα, σχετικά με αυτές, κείμενα. Εκτώς απο πιθανότητα αναδιατύπωσης της εκάστοτε ερώτησης, αυτή η διαδικασία εξασφαλίζει και την σχετικότητα της ερώτησης συγκριτικά με τα κείμενα, κάτι που είναι πολύ σημαντικό για να επιτευχθεί ο στόχος της εργασίας, δηλαδή ο περιορισμός παραισθήσεων σε συστήματα ερωταπαντήσεων. Για το συγκεκριμένο σύνολο δεδομένων, γίνεται προσπάθεια στο στο αρχικό στάδιο (prompting στο gen_questions_ollama.py) να παραχθούν ερωτήσεις οι οποίες είναι ήδη συναφείς με τα κείμενα. Δηλαδή πρίν γίνει η διαδικασία γείωσης, έχει γίνει μια προσπάθεια η κάθε ερώτηση να έχει σχέση με το κείμενο. Παρόλα αυτά, δεν γίνεται να επιστευθεί τυφλά το μοντέλο και συνεπώς εφαρμόζεται την διαδικασία RAG στα επόμενα βήματα. Με βάση τα παραπάνω, και για την συγκεκριμένη συλλογή κειμένων και διαδικασία prompting, είναι σημαντικό να σημειώθει πως το grounding δεν αλλάζει τρομερά την διατύπωση όλων των ερωτήσεων, αφού αρκετές απο αυτές παραμένουν ίδιες ή αλλάζουν μικρά στοιχεία τους, κάνοντας τες περισσότερο σύμφωνες με τα περιεγχόμενα των κειμένων και αφαιρώντας περιττά στοιχεία. Η διαδικασία της γείωσης, συνεπώς, δεν επιδιώκει ριζική αναδιατύπωση αλλά επιβεβαίωση ή ποιοτική αναδιατύπωση κάθε ερώτησης σύμφωνα με τα ανακτημένα τεκμήρια.

Παρακάτω θα αναφερθούν και θα σχολιαστούν μερικά παραδείγματα που εντοπίστηκαν, έτσι ώστε να φανούν στην πράξη τα αποτελέσματα της διαδικασίας:

Αρχική Ερώτηση	**Γειωμένη Ερώτηση**	**Σχόλια**
What were the concentrations of two specific serum proteins significantly changed in patients with pseudomonas aeruginosa infection compared to control persons?	What serum proteins had significantly changed concentrations in patients with pseudomonas aeruginosa infection compared to control persons?	Αλλαγή στον τρόπο έκφρασης της ερώτησης, τώρα ρωτάει για τις πρωτεΐνες και όχι την συγκέντρωση τους, που είναι πιο σχετική ερώτηση με το κείμενο
What condition also affected a cousin of the index case?	What other condition affected a cousin of the index case?	Προσθήκη της λέξης "other" που μπορεί να καθοδηγήσει αυτόν που διαβάζει στην ερώτηση στην εξαγωγή της απάντησης.
What is the effect of pancreatic insufficiency on trypsin activity and proteolytic activity in duodenal juice?	What effect does pancreatic insufficiency have on trypsin and proteolytic activity in duodenal juice?	Η ερώτηση στην ουσία είναι ίδια, με λίγο διαφορετική διατύπωση σύμφωνα με αυτήν του κειμένου.
Which immunoglobulins were detected in normal levels in CF patients?	What immunoglobulins were found at normal levels in CF patients?	Αντικατάσταση λέξης "detected" με "found", που υπάρχει αυτούσια στο κείμενο.
What is often not present in chronic sinusitis in young children?	What is often absent in chronic sinusitis in young children?	Αντικατάσταση φράσης "not present" με "absent" για καλύτερη αντιστοίχηση με το κέιμενο.
What is one of the methods used to induce intracisternal inclusion bodies?	What method induces intracisternal inclusion bodies?	Αφαιρείται το κομμάτι "one of the methods", αφού αναφέρεται μια μέθοδος - η ερώτηση είναι πιο ξεκάθαρη.

Πίνακας 6.1: Παραδείγματα αρχικών και γειωμένων ερωτήσεων με σχολιασμό.

Αρχική Ερώτηση	**Γειωμένη Ερώτηση**	**Σχόλια**
"According to the authors, what may replace the fibroblast technique in detecting cystic fibrosis heterozygotes?	What may replace the fibroblast technique for detecting cystic fibrosis heterozygotes?	Αφαιρείται το κομμάτι "According to the authors και η ερώτηση αποκτά πιο γενική φύση.
What was the result of observing the kinetics of ouabain binding in both normal and CF fibroblasts?	What were the results of observing the kinetics of ouabain binding in both normal and CF fibroblasts?	Αναφορά σε πολλά results, αντί για ένα, έτσι ώστε να οδηγείται ο χρήστης σε καλύτερη απάντηση, με βάση το κείμενο.
What is suggested as the reason for the anemia in the twins with cystic fibrosis?	What is suggested as a possible reason for anemia in twins with cystic fibrosis?	Αναφορά σε "possible reason" αντί για "the reason", που δεν είναι τόσο απόλυτο και είναι πιο σύμφωνο με το περιεγχόμενο του αποσπάσματος.
What is the main product of anisylalanine metabolism in normal adult subjects?	What are the main products of anisylalanine metabolism in normal adult subjects?	Χρήση πληθυντικού για τα products, που καθοδηγεί τον χρήστη στο να απαντήσει καλύτερα, αφού το κείμενο αναφέρει διάφορα main products.
What caused the wrinkle appearance in infants with cystic fibrosis?	What causes skin wrinkling in infants with cystic fibrosis?	Καλύτερη έκφραση της ερώτησης αφού αντί για παρελθοντικό χρόνο (caused) χρησιμοποιείται το παρον (causes)

Πίνακας 6.2: Επιπλέον παραδείγματα αρχικών και γειωμένων ερωτήσεων με σχολιασμό.

Εξετάζοντας τα παραπάνω παραδείγματα, γίνεται εμφανές ότι το στάδιο της επέκτασης των ερωτήσεων, παρότι οι αρχικές ερωτήσεις είναι αρκετά καλές και σχετικές σε έναν βαθμό με το κείμενο, βοηθάει στην παραιτέρω ενίσχυση της πληροφορίας που παρέχουν. Μερικά ανακριβή στοιχεία αφαιρούνται ή αντικαθιστούνται με στοιχεία απο τα κείμενα. Συνεπώς, το αποτέλεσμα που βλέπει ο χρήστης, δηλαδή το τελικό σετ ερωταπαντήσεων είναι σε απόλυτη συμφωνία με την διαθέσιμη συλλογή κειμένων.

6.4 Αξιολόγηση απαντήσεων

Πέρα απο την αξιολόγηση των ερωτήσεων, έχει ιδιαίτερη σημασία και η αξιολόγηση των αντίστοιχων απαντήσεων. Το σύστημα που αναπτύχθηκε, είναι ένα ολοκληρωμένο και αξιόπιστο σύστημα παραγωγής ερωτήσεων και απαντήσεων, πιστό σε συλλογή κειμένων. Είναι σημαντικό, λοιπόν, να υπάρχει σιγουριά για την υψηλή ποιότητα των παραγόμενων απαντήσεων. Η αξιολόγηση τους επιτεύχθηκε με χρήση κατάλληλων μετρικών, οι οποίες θα αναλυθούν στην συνέχεια.

6.4.1 Ομοιότητα με περιλήψεις κειμένων και διαδικασία αξιολόγησης

Κατά την εκπόνηση της διπλωματικής αυτής εργασίας, ένα κομμάτι που απασχόλησε ιδιαίτερα ήταν αυτό της αξιολόγησης των, παραγόμενων απο το μοντέλο, απαντήσεων στις ερωτήσεις. Στην συγκεκριμένη εργασία, η ακρίβεια της απάντησης έχει μεγάλη βαρύτητα και είναι σημαντικό αυτή να εξασφαλίζεται όσο το δυνατόν περισσότερο. Το πρόβλημα είναι το πως γίνεται να ποσοτικοποιηθεί και να εξεταστεί η ακρίβεια και η εγκυρότητα κάθε ερώτησης. Κάθε απάντηση του συστήματος είναι αυτούσιο κομμάτι του κειμένου αλλά πρέπει κάπως να εξασφαλιστεί ότι η διαδικασία απαντά ικανοποιητικά, αν όχι όλες, την μεγάλη πλειοψηφία των ερωτήσεων.

Υπάρχουν αρκετές γνωστές μέθοδοι για την αξιολόγηση QA συστημάτων. Βέβαια, αυτές οι μέθοδοι δεν εξυπηρέτησαν την παρούσα αξιολόγηση, επομένως επιλέχθηκε κάτι άλλο. Παρακάτω θα αναφερθούν μερικές απο τις μεθόδους που εντοπίστηκαν, αλλά εν τέλη δεν επιλέχθηκαν [54]:

- Ακριβή Αντιστοίχηση (Exact Match): Είναι η απλούστερη μορφή αξιολόγησης, αφού απλά ελέγχεται αν η απάντηση του μοντέλου είναι αρκιβώς ίδια με την ανθρώπινη απάντηση. Αυτή η μέθοδος προφανώς δεν λαμβάνει υπόψην συνόνυμα, παραφράσεις και σημεία στήξης. Μετράει μόνο ταυτοσημία, κάτι που δεν είναι κατάλληλο για ένα σύστημα QA όπου ο χρήστης απαντά ερωτήσεις.

- Αξιολόγηση με βάση κανόνων (Rule Based Evaluation): Εφαρμογή κανόνων όπως "αν περιέχεται η λέξη-κλειδί X, η απάντηση θεωρείται σωστή". Προφανώς και αυτή η μέθοδος είναι ακατάλληλη για την αξιολόγηση της εργασίας αυτής, εφόσον δεν λειτουργεί σε πλαίσιο απάντησης σύνθετων ερωτήσεων και δεν αξιολογείται καθόλου το νόημα της πρότασης.

- Σκόρ επιπέδου λεκτικών μονάδων (Token-Level F1 Score): Υπολογίζει το ποσοστό κοινών tokens μεταξύ της απάντησης του μοντέλου και της σωστής απάντησης, εξισορροπώντας ακρίβεια (precision) και ανάκληση (recall.) Το πρόβλημα με αυτή την περίπτωση είναι οτι στηρίζεται στην επιφανειακή σύγκριση λέξεων και μόνο αυτών. Αυ-

τό μπορεί να τιμωρήσει απαντήσεις που είναι πιο αναλυτικές και περιγραφικές από την απάντηση του μοντέλου, χωρίς να περιέχουν λανθασμένα στοιχεία.

Απο τα παραπάνω γίνεται εμφανές πως μερικές κλασσικές προσεγγίσεις δεν ταιριάζουν απόλυτα με τον σκοπό της εργασίας. Για αυτό και επιλέχθηκε κάτι λίγο διαφορετικό. Οι απαντήσεις αντιμετωπίζονται ως περιλήψεις των αντίστοιχων κειμένων και, συνεπώς, γίνεται χρήση γνωστών μετρικών που εφαρμόζονται συνήθως στην αξιολόγηση περιλήψεων κειμένων (text summarization). Η προσέγγιση αυτή βασίζεται στην παρατήρηση ότι, στο πλαίσιο ενός συστήματος ερωταπαντήσεων, η απάντηση που παράγει το μοντέλο μπορεί να θεωρηθεί ως μια σύνοψη του σχετικού αποσπάσματος της βάσης γνώσης, επικεντρωμένη γύρω από την πληροφορία που ζητά η ερώτηση. Γενικά, στην βιβλιογραφία, υπάρχουν προσεγγίσεις που συνδέουν την έννοια του QA με αυτή των περιλήψεων και συγκεκριμένα την αξιολόγηση τους [12, 11]. Επομένως, η σύνδεση των δύο έχει επιστημονική βάση και προσφέρει οφέλη στην διαδικασία αξιολόγησης.

Στο πλαίσιο αυτό κάθε απάντηση του συστήματος αντιμετωπίστηκε ως μια σύνοψη της πληροφορίας που βρίσκεται στο σχετικό κείμενο, και συγκρίθηκε με μια αντίστοιχη απάντηση αναφοράς, δηλαδή με την περίληψη/απάντηση που θα παρήγαγε ένας άνθρωπος για την ίδια ερώτηση. Η σύγκριση αυτή επιτρέπει να εκτιμηθεί σε ποιό βαθμό η απάντηση του μοντέλου παρέχει τις ίδιες βασικές πληροφορίες, είναι πιστή στο κείμενο και αποφεύγει παραλείψεις ή παραισθήσεις. Παρόλο που η διαδικασία υλοποίησης εξασφαλίζει μερικά απο αυτά, είναι σημαντικό να φανούν και στην πράξη.

Συγκεκριμένα, η αξιολόγηση μετατρέπεται σε πρόβλημα υπολογισμού ομοιότητας μεταξύ δύο απαντήσεων/περιλήψεων: της ανθρώπινης και της παραγόμενης από το σύστημα. Για να επιτευχθεί αυτό με σωστό τρόπο, επιλέχθηκε τυχαία ένας αριθμός ερωτήσεων και απαντήθηκαν με βάση το σχετικότερο κείμενο, το οποίο χρησιμοποίησε και το μοντέλο για να αντλείσει και να παράξει την απάντηση. Έγινε προσπάθεια η κάθε δωσμένη απάντηση να είναι σύντομη και σύμφωνη με το κείμενο χρησιμοποιώντας εκφράσεις και λέξεις που βρίσκονται μέσα σε αυτό. Στην συνέχεια, έγινε σύγκριση των ανθρώπινων απαντήσεων, με αυτές του μοντέλου χρησιμοποιώντας μετρικές περιλήψεων κειμένων (text summarization), οι οποίες θα αναλυθούν παρακάτω.

Σε αυτό το σημείο, να σημειωθεί ότι αξιολογήθηκαν ξεχωριστά οι ερωτήσεις που είχαν ως απάντηση Ναι/Όχι (Yes/No). Για τις 250 ερωτήσεις που παρήχθησαν, οι 9 είχαν απάντηση "No", από το μοντέλο. Έγινε χειρονακτική απάντηση των ερωτήσεων αυτών και μονάχα μια διαπιστώθηκε ως λανθασμένη (έπρεπε να είναι Yes αντί για No). Αυτό πιθανότατα οφείλεται σε αδυναμία του LLM να αναγνωρίσει την συγκεκριμένη απάντηση, αλλά είναι κάτι που δεν προκαλεί ιδιαίτερη ανησυχία αφου η λανθασμένη περίπτωση ήταν μονάχα μια. Συνεπώς, δεν επηρεάζεται η συνολική αξιοπιστία και ακρίβεια το συστήματος.

6.4.2 Μετρικές που εξετάστηκαν και χρησιμοποιήθηκαν

Για να αποφασιστεί ποιες μετρικές text summarization θα χρησιμοποιηθούν για την αξιολόγηση των απαντήσεων του μοντέλου, μελετήθηκαν αρκετές [16, 8]. Οι κυριότερες απο αυτές είναι οι εξής:

- **ROUGE**: Είναι η πιο διαδεδομένη μετρική για αξιολόγηση ποιότητας περίληψης. Μετρά την επικάλυψη λέξεων (n-grams) μεταξύ της παραγόμενης περίληψης/απάντησης και της αναφοράς. Η ROUGE διακρίνεται στις παραλλαγές ROUGE-1 που υπολογίζει επικάλυψη μεμονομένων λέξεων (unigrams), ROUGE-2 που υπολογίζει επικάλυψη ζεύγων λέξεων (bigrams), ROUGE-L που βασίζεται στη μεγαλύτερη κοινή ακολουθία λέξεων (LCS) και ROUGE-Lsum που συγκρίνει τις περιλήψεις σε επίπεδο προτάσεων. Συνολικά, η μετρική ROUGE δείχνει πόσο η παραγόμενη σύνοψη μοιάζει λεκτικά με την αναφορά.

- **BLEU**: Αναπτύχθηκε αρχικά για μηχανική μετάφραση, αλλά χρησιμοποιείται και για περίληψεις κειμένων. Συγκρίνει την παραγόμενη περίληψη με την αναφορά, μετρώντας την επικάλυψη n-grams. Όσο μεγαλύτερη επικάλυψη, τόσο καλύτερο και το σκορ (απο 0 έως 1). Η μετρική αυτή μετρά μόνο λεκτική επικάλυψη και όχι σημασιολογική.

- **BERTScore**: Βασίζεται σε μοντέλα BERT και μετρά σημασιολογική ομοιότητα μεταξύ παραγόμενης περίληψης/απάντησης και αναφοράς. Ο υπολογισμός γίνεται μετρώντας διανυσματικές αναπαραστάσεις (embeddings) κάθε λέξης. Προσφέρει την δυνατότητα διαχείρησης συνώνυμων και γενικά σημασιολογικής κάλυψης, γεγονός που την καθιστά αρκετά χρήσιμη σε πλαίσιο αξιολόγησης.

- **METEOR**: Η μετρική αρχικά αναπτύχθηκε για μετάφραση, αλλά εφαρμόζεται πλέον και σε πλαίσιο αξιολόγησης περίληψεων. Εκτιμά τη γραμματική και σημασιολογική ορθότητα, λαμβάνοντας υπόψιν την σειρά των λέξεων, συνώνυμα, παραλλαγές καθώς και την φυσικότητα του κειμένου. Ωστόσο, δημιουργούνται σημαντικοί περιορισμοί όσον αφορά την χρήση της. Αφερνός απαιτεί προεπεξεργασία των δεδομένων και αφετέρου δεν καθίσταται βέλτιστο να χρησιμοποιηθεί μόνη της, μιας και πρέπει να συνδυαστεί με κάποια άλλη μετρική για ολοκληρωμένη αξιολόγηση.

Μελετώντας τις παραπάνω μετρικές, λήφηκε η απόφαση πως οι καλύτερες για την συγκεκριμένη αξιολόγηση ήταν οι ROUGE-1, ROUGE-L παραλλαγές της ROUGE και η BERTSCORE. Συγκεκριμένα, η ROUGE-1 είναι η καλύτερη, αφού οι απαντήσεις που δίνει το μοντέλο (και θα απαντήσει και ο χρήστης) είναι σύντομες και περιέχουν σημαντικές 'λέξεις-κλειδιά' που πρέπει και ο χρήστης να αναφέρει στην απάντηση του, ώστε αυτή να είναι πλήρως σωστή. Η unigram επικάλυψη είναι προτιμότερη σε τέτοιες μικρές απαντήσεις, επειδή το να πάρουμε τις λέξεις ως ζεύγη μπορεί να έχει ακόμη και αρνητικό αποτέλεσμα. Θεωρήθηκε οφέλιμη και η μετρική ROUGE-L, γιατί θα μπρούσε να θεωρηθεί πως οι 2 απαντήσεις(μοντέλου και ανθρώπινη) μπορούν να συγκριθούν και αν έχουν πολύ παρόμοια Ελάχιστη Κοινή Ακολουθία, τότε η απάντηση είναι καλή. Προφανώς, αυτό δεν ισχύει για όλες τις περιπτώσεις αλλά η εφαρμογή της

μετρικής θεωρήθηκε ότι μπορεί να προσφέρει στην αξιολόγηση τους συστήματος. Παράλληλα, η BERTScore, προσφέρει σημασιολογική κάλυψη σε περίπτωση που ο χρήστης απαντήσει με συνώνυμα ή παραλλαγές των περιεχομένων του κειμένου. Είναι σημαντικό να υπάρχει και μια τέτοια αξιολόγηση, έτσι ώστε να υπάρχει ευελιξία στην έκφραση της απάντησης. Η εστίαση έγινε κυρίως στο PRECISION τμήμα των μετρικών για την παρούσα αξιολόγηση, διότι έχει περισσότερο αξία να εξεταστεί η ακρίβεια των απαντήσεων παρά η ανάκληση, στο πλαίσιο της εργασίας. Αυτό επιλέχθηκε, επειδή ο κάθε χρήστης μπορεί να εκφράσει την απάντηση του με διαφορετική έκταση, αλλά να περιέχονται σε αυτή οι σωστές πληροφορίες.

Στο σημείο αυτό, είναι σημαντικό να σημειωθεί πως η αξιολόγηση του χρήστη στην εκπαιδευτική εφαρμογή (συγκεκριμένα η βαθμολογία που βλέπει ο χρήστης όταν πατήσει "Show Grade") είναι ο μέσος όρος της μετρικής ROUGE-1 PRECISION, μιας και θεωρήθηκε πως αυτή ποσοτικοποιεί καλύτερα την βαθμολογία του χρήστη.

6.4.3 Αποτελέσματα αξιολόγησης απαντήσεων

Παρακάτω θα παρατεθούν σε μορφή πίνακα τα αποτελέσματα της εκτέλεσης του κώδικα αξιολόγησης(evaluation code) της παρούσας εργασίας:

Πίνακας 6.3: Μετρικές Αξιολόγησης

entry	rouge1_p	rouge1_r	rouge1_f1	rougeL_f1	bertscore_p	bertscore_r	bertscore_f1
1)	1.0000	0.5000	0.6667	0.6667	0.6303	0.4436	0.5361
2)	1.0000	0.3750	0.5455	0.5455	0.7844	0.4653	0.6208
3)	0.6364	1.0000	0.7778	0.7778	0.6673	0.7758	0.7215
4)	1.0000	0.4000	0.5714	0.5714	0.7178	0.3869	0.5481
5)	1.0000	1.0000	1.0000	1.0000	0.6145	0.9043	0.7561
6)	1.0000	0.5000	0.6667	0.6667	0.3837	0.1921	0.2873
7)	0.5000	0.4000	0.4444	0.4444	0.6286	0.4718	0.5498
8)	1.0000	0.2143	0.3529	0.3529	0.5439	-0.0514	0.2304
9)	0.8000	0.3077	0.4444	0.4444	0.5192	0.2099	0.3611
10)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11)	1.0000	0.2632	0.4167	0.4167	0.6079	0.1840	0.3885
12)	1.0000	0.5000	0.6667	0.6667	0.7191	0.3859	0.5482
13)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14)	1.0000	0.6250	0.7692	0.7692	0.5561	0.3157	0.4341
15)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
16)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
17)	0.9231	0.8571	0.8889	0.8889	0.8629	0.8578	0.8606
18)	1.0000	0.2500	0.4000	0.4000	0.4499	0.1032	0.2720
19)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20)	1.0000	0.7500	0.8571	0.8571	0.8381	0.6571	0.7466
21)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
22)	1.0000	0.5000	0.6667	0.6667	0.6735	0.5458	0.6096
23)	1.0000	0.6250	0.7692	0.7692	0.6785	0.4799	0.5781
24)	0.2500	0.1000	0.1429	0.1429	0.3124	0.1438	0.2280
25)	1.0000	1.0000	1.0000	1.0000	0.9857	0.9857	0.9857
26)	1.0000	1.0000	1.0000	1.0000	0.9831	0.9831	0.9832
27)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
28)	1.0000	0.3750	0.5455	0.5455	0.5640	0.2670	0.4123
29)	0.6667	0.6667	0.6667	0.5556	0.6846	0.6090	0.6471
30)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
31)	0.9375	0.9375	0.9375	0.9375	0.9806	0.9807	0.9807
Mean	0.9051	0.6655	0.7307	0.7272	0.7329	0.6047	0.6671

Μελετώντας τα παραπάνω, γίνεται εμφανές πως οι τιμές των μετρικών είναι αρκετά καλές, ειδικά όσο αφορά το τμήμα της Ακρίβειας (PRECISION) που είναι και το πιο σημαντικό στην αξιολόγηση των σύντομων απαντήσεων που παράγει το σύστημα. Συγκεκριμένα, το ROUGE-1 PRECISION έχει κατά μέσο όρο τιμή κοντά στο 0.91, γεγονός που αποδεικνύει την ακρίβεια των απαντήσεων του μοντέλου. Η υψηλή αυτή τιμή σημαίνει πως οι ίδιες ΄λέξεις κλειδιά' που περιέχονται στην απάντηση αναφοράς(την ανθρώπινη απάντηση), υπάρχουν και στην απάντηση του μοντέλου. Διαβάζοντας το κείμενο, λοιπόν, άνθρωπος και μοντέλο εντόπισαν τα ίδια ή πολύ συναφή κομμάτια, ως απάντηση. Η τιμή της ανάκλησης (RECALL τμήματος των μετρικών) είναι μικρότερη, γεγονός απολύτως φυσιολογικό, αφού οι απαντήσεις του μοντέλου είναι σύντομα αποσπάσματα του κειμένου. Η απάντηση του ανθρώπου μπορεί να περιέχει παραπάνω λέξεις, διαφορετικές προτάσεις και διαφορετική σύνταξη απο το κομμάτι του κειμένου που εντόπισε το μοντέλο ως απάντηση, ακόμη και αν το νόημα είναι ακριβώς το ίδιο.

Παρατηρείται επίσης ότι η τιμή της BERTScore PRECISION είναι μικρότερη απο την αντίστοιχη της ROUGE-1 PRECISION. Αυτό δεν σημαίνει πως οι απαντήσεις δεν έχουν σημασιολογική κάλυψη, αλλά προκύπτει απο τον τρόπο που λειτουργεί η μετρική BERTSCORE. Θα αναφερθεί ένα παράδειγμα, για να επεξηγηθεί το παραπάνω:

Υπάρχει η ερώτηση: **"What relationship exists between the number of Ps aeruginosa precipitins and total T cells in cystic fibrosis patients?"** (entry 1 στον πίνακα).

Η ανθρώπινη απάντηση, μετά απο ανάγνωση του αντίστοιχου κειμένου ήταν η εξής:**"the number was significantly positively correlated"**, ενώ το μοντέλο είχε ώς απάντηση: **"Significantly positively correlated"**.

Η ROUGE-1 P έδειξε πως η απάντηση είναι ολόσωστη, γεγονός που συμπίπτει με την πραγματικότητα. Η BERTScore P όμως έδειξε τιμή 0.6303, πιθανότητα λόγω απουσίας της λέξης "number" στην απάντηση του μοντέλου που μάλλον θεωρήθηκε σημαντικό σημασιολογικό πλαίσιο απο το BERT.

Αν ληφθούν υπόψην τέτοιες περιπτώσεις(υπάρχουν και άλλες παρόμοιες μέσα στις αξιολογήσεις), γίνεται εμφανές πως οι απαντήσεις του μοντέλου συμπίπτουν απόλυτα με το ζητούμενο. Είναι σύντομες, έγκυρες, περιεκτικές και βγαλμένες αυτούσιες απο την συλλογή κειμένων. Η εξαγωγή τους δεν είναι δύσκολη ή περίπλοκη αφού διαβάζοντας τα κείμενα, εντοπίστηκαν άνετα. Το μοντέλο, λοιπόν, εκτώς απο την σωστή λειτουργία που παρουσιάζει, παράγει και τα θεμιτά αποτελέσματα στο πλαίσιο της ακρίβειας των παραγόμενων απαντήσεων.

Κεφάλαιο 7

Επίλογος

Στο κεφάλαιο αυτό, θα αναφερθούν και θα αναλυθούν τα συμπεράσματα και τα πορίσματα που προέκυψαν απο την έρευνα που πραγματοποιήθηκε και γενικά απο την υλοποίηση της διπλωματικής αυτής εργασίας. Θα αναφερθούν, επίσης, μερικοί περιορισμοί της υλοποίησης και της έρευνας καθώς και προτάσεις για μελλοντική έρευνα και επέκταση του συστήματος και τις ιδέες που πραγματεύεται η εργασία.

7.1 Συμπεράσματα

Η παρούσα διπλωματική εργασία είχε ως κύριο στόχο τη δημιουργία ενός αυτόνομου συστήματος παραγωγής και επικύρωσης ερωταπαντήσεων βασισμένου σε Μεγάλα Γλωσσικά Μοντέλα, ενισχυμένα με την τεχνική RAG. Αναδείχθηκαν, παράλληλα, τόσο η δυναμική όσο και τα όρια των σύγχρονων μοντέλων τεχνητής νοημοσύνης σε εκπαιδευτικές εφαρμογές, κυρίως όσον αφορά τα ζητήματα της ακρίβειας και της αξιοπιστίας της παραγόμενης απάντησης ή πληροφορίας. Εξάλλου, σε ένα αυτόματο εκπαιδευτικό πλαίσιο είναι σημαντικότερη, σε πρώτο στάδιο, η εξασφάλιση της ακρίβειας και εγκυρότητας του συστήματος από την ποσότητα της πληροφορίας που παράγει. Σε γενικότερο πλάισιο, η ενσωμάτωση των LLMs στην εκπαιδευτική διαδικασία προσφέρει σημαντικά πλεονεκτήματα όπως ταχύτερη παραγωγή περιεχομένου, δυνατότητα εξατομίκευσης της μάθησης, αλλά και αυτοματοποίηση εργασιών που έως τώρα απαιτούσαν ανθρώπινη παρέμβαση. Ωστόσο, μέσα απο την υλοποίηση αυτής της εργασίας, αναδεικνύεται η πραγματική αξία των Μεγάλων Γλωσσικών Μοντέλων όταν αυτά χρησιμοποιούνται σε συνδυασμό με τεχνικές όπως το RAG και η Επέκταση Ερωτημάτων, οι οποίες επιτρέπουν στα μοντέλα να βασίζονται σε τεκμηριωμένα δεδομένα αντί να εμπιστεύονται τυφλά την παραμετρική τους γνώση. Μέσω της εφαρμογής τέτοιων τεχνικών περιορίζεται το φαινόμενο των παραισθήσεων, επιβλαβές σε οποιοδήποτε πλαίσιο εφαρμογής, και συνεπώς ενισχύεται η αξιοπιστία των παραγόμενων αποτελεσμάτων.

Το σύστημα που αναπτύχθηκε υλοποιεί μια πολυεπίπεδη μεθοδολογία, όπου κάθε βήμα λειτουργεί συμπληρωματικά προς το προηγούμενο. Συγκεκριμένα:

- Η παραγωγή των αρχικών ερωτήσεων και την ανάκτηση τεκμηρίων.
- Η γείωση/επέκταση των ερωτήσεων με βάση τα τεκμήρια και η διαδικασία αναζήτησης, περικοπής και επικύρωσης απαντήσεων.
- Η επανεξέταση μη απαντήσιμων περιπτώσεων και η δημιουργία της εκπαιδευτικής εφαρμογής, σύμφωνα με το τελικό σύνολο ερωταπαντήσεων.

Όταν συγχωνεύονται όλα τα παραπάνω, αποδεικνύεται ότι είναι εφικτό και βέλτιστο να συνδυαστούν τεχνικές δημιουργίας φυσικής γλώσσας με μηχανισμούς ανάκτησης για να παραχθεί ποιοτικό, αξιόπιστο και τεκμηριωμένο υλικό ερωταπαντήσεων. Στην συνέχεια, αυτό το υλικό μπορεί να αξιοποιηθεί με πολλούς τρόπους, ένας εξ΄ αυτών και η εκπαίδευση. Για τις ανάγκες της εργασίας, επιλέχθηκε η εστίαση στον τομέα αυτόν έτσι ώστε να επιβεβαιωθεί με πρακτικό τρόπο η χρησιμότητα του τελικού, επικυρωμένου συνόλου. Τα παραγόμενα δεδομένα μπορούν να ενσωματωθούν εύκολα σε εκπαιδευτικά περιβάλλοντα, προσφέροντας στους χρήστες έναν άμεσο, προσβάσιμο και διαδραστικό τρόπο αξιολόγησης γνώσεων. Σε αυτή την διαδικασία δεν είναι απαραίτητη η ύπαρξη κάποιου διδάσκοντα ή φορέα εκπαίδευσης, γεγονός που ενισχύει περαιτέρω την λειτουργικότητα του συστήματος.

Η εργασία, επομένως, αποδεικνύει στην πράξη ότι:

1. Η συνδυαστική χρήση LLMs και RAG μπορεί να παράγει αξιόπιστο υλικό και συγκεκριμένα εκπαιδευτικό.
2. Η αυτοματοποίηση της επικύρωσης των απαντήσεων βελτιώνει την ποιότητα τους και περιορίζει τα σφάλματα και τις παραισθήσεις που μπορεί να προέκυψαν(ακόμη και αν στοιχίζει σε μικρό βαθμό, στην ποσότητα τους).
3. Η παρουσίαση των αποτελεσμάτων μέσω εύχρηστης και απλής εφαρμογής καθιστά την τεχνολογία προσιτή και λειτουργική για εκπαιδευτικά ιδρύματα και απλούς χρήστες.

Επιπλεόν, είναι σημαντικό να αναφερθεί πως η ανάπτυξη και δοκιμή της μεθοδολογίας σε ιατρικά κείμενα (περίπτωση Κυστικής Ίνωσης) έδειξε πως το σύστημα μπορεί να αποδώσει καλά ακόμη και σε εξειδικευμένα πεδία γνώσης, όπου η ακρίβεια είναι ζωτικής σημασίας και τυχόντα λάθη στον τρόπο παρουσίασης και μετάδωσης των πληροφοριών κοστίζουν αρκετά. Συνοψίζοντας, τα αποτελέσματα αποδεικνύουν ότι η αξιοποίηση Μεγάλων Γλωσσικών Μοντέλων στην εκπαίδευση έχει πρακτική και επιστημονική αξία, αρκεί να συνοδεύεται από μηχανισμούς επαλήθευσης και τεκμηρίωσης όπως αυτοί που προτάθηκαν στην παρούσα εργασία. Τα αποτελέσματα που παρουσιάστηκαν είναι αρκετά ενθαρρυντικά σε αυτό το κομμάτι, χωρίς αυτό να σημαίνει ότι δεν χωράνε επιπλέον βελτιώσεις και καινοτομίες πάνω στην βασική ιδέα.

7.2 Μελλοντικές Επεκτάσεις

Η παρούσα εργασία, αν και παρουσιάζει ένα καλό τελικό αποτέλεσμα, αφήνει αρκετά πεδία για περαιτέρω ανάπτυξη και βελτίωση. Η χρήστη του RAG με Μεγάλα Γλωσσικά Μοντέλα είναι ένα σύγχρονο πεδίο έρευνας και προφανώς αυτό σημαίνει ότι μπορούν να γίνουν πολλές προσθήκες και βελτιώσεις, σε ό,τι εφαρμογή σχετίζεται με αυτήν. Αυτό ισχύει και για την διπλωματική αυτή. Παρακάτω θα αναφερθούν μερικές απο τις πιο προφανείς βελτιώσεις που θα μπορούσαν να υλοποιηθούν στο μέλλον:

1. **Ενσωμάτωση μοντέλου εκτίμησης δυσκολίας ερώτησης:** Εξετάστηκε η εφαρμογή μιας τέτοιας λειτουργίας, αφού στο πρώτο στάδιο παραγωγής των αρχικών ερωτήσεων και απαντήσεων μέσω απλού LLM Prompt ζητάμε απο το μοντέλο να ενσωματώσει και πεδίο δυσκολίας ερώτησης. Τελικά, δεν πραγματοποιήθηκε η επικύρωση και η ουσιαστική εφαρμογή της συγκεκριμένης λειτουργίας γιατί θεωρήθηκε πως μπορεί να επηρεάσει αρνητικά την αξιοπιστία και εγκυρότητα του τελικού σετ ερωταπαντήσεων. Θα μπορούσε, όμως, να υλοποιηθεί η λειτουργία αν υπάρχει βεβαιότητα για το επίπεδο δυσκολίας. Για να επιτευχθεί αυτό, θα ήταν αναγκαίο ένα ακόμη στάδιο στοχευμένης ανάκτησης και χρήσης του Μεγάλου Γλωσσικού Μοντέλου για κρίση κάθε ερώτησης. Θα χρειαζόταν, επίσης, προσεκτική και αναλυτική αξιολόγηση απο ανθρώπους, αφού πολλές φορές τα Μεγάλα Γλωσσικά Μοντέλα δεν εφαρμόζουν τα αντικειμενικά κριτήρια δυσκολίας που γνωρίζουν οι άνθρωποι.

2. **Δημιουργία πλατφόρμας αυτόματης διεξαγωγής εξετάσεων και βαθμολογιών:** Η εφαρμογή που αναπτύχθηκε, λειτουργεί ως απόδειξη για την συνεισφορά που μπορεί να έχει το τελικό σύνολο δεδομένων στο πλαίσιο της εκπαίδευσης. Αυτό θα μπορούσε να πάει ένα βήμα παραπάνω, μέσα απο την ανάπτυξη μιας πλατφόρμας διεξαγωγής εξετάσεων. Το παραγόμενο σύνολο ερωταπαντήσεων θα μπορούσε να ενσωματωθεί σε ένα τέτοιο πλαίσιο, όπου οι μαθητές θα απαντούν και θα βαθμολογούνται αυτόματα. Στην συνέχεια, οι βαθμολογίες τους θα καταχωρούνταν και η εξέταση θα ολοκληρωνόταν. Μια τέτοια πλατφόρμα κάνει την διαδικασία διεξαγωγής εξετάσεων πλήρως αυτοματοποιημένη χωρίς κινδύνους ύπαρξης λανθασμένων ερωτήσεων ή απαντήσεων. Για να υλοποιηθεί κάτι τέτοιο, είναι σημαντικό να επικυρωθούν σε ακόμη μεγαλύτερο βαθμό τα αποτελέσματα καθώς και να διορθωθούν μικρά προβλήματα και λάθη που μπορεί να υπάρχουν σε μερικά απο τα σετ. Επίσης, θα πρέπει να ενσωματωθούν και κατάλληλα χρονικά όρια για την απάντηση των ερωτήσεων καθώς και διαφορετικοί τύποι ερωτήσεων (πχ συμπλήρωση κενού, πολλαπλής επιλογής) για μεγαλύτερη κάλυψη του γνωστικού πεδίου και ποικιλία θεμάτων εξέτασης.

3. **Εμπλουτισμός συλλογών κειμένων με νέες πηγές:** Παρόλο που το σύστημα έχει υλοποιηθεί έτσι ώστε να μπορούν να ενσωματωθούν διαφορετικές συλλογές κειμένων (προφανώς με το κατάλληλο csv format), θα είχε ενδιαφέρον η δοκιμή παραγωγής ερωτήσεων και απαντήσεων με χρήση μεγαλύτερων ή ετερογενών συνόλων κειμένων (π.χ. επιστημονικά άρθρα, σχολικά βιβλία, διαλέξεις, διαφάνειες). Η οπτική

αυτή, θα έδινε στο σύστημα μια νέα διάσταση και θα ενίσχυε την ικανότητα του να παράγει διαφοροποιημένες ερωτήσεις, δοκιμάζοντας το σε πιο σύνθετα σενάρια γνώσης.

4. **Αλλαγές στην διαδικασία επικύρωσης απαντήσεων:** Η μεθολογία εξαγωγής και επικύρωσης απαντήσεων (πάντα αυτολεξεί κομμάτι του κειμένου) λειτουργεί ικανοποιητικά και πετυχαίνει τέλεια τον αρχικό και βασικό στόχο της διπλωματικής, δηλαδή την ακρίβεια και τεκμηρίωση των απαντήσεων αποκλειστικά με βάση τα ανεκτημένα τεκμήρια (συλλογή κειμένων). Θα μπορούσε όμως κάποια μελλοντική έρευνα να εστιάσει σε αλλαγή αυτής της φιλοσοφίας. Συγκεκριμένα, οι απαντήσεις θα μπορούσαν να εξάγωνται απο το κείμενο, αλλά να διατυπώνονται με διαφορετικό τρόπο. Αν αυτός ο τρόπος εξασφάλιζε την εγκυρότητα και την συμφωνία με τα σχετικά κείμενα, τότε το αποτέλεσμα θα ήταν θεμιτό.

5. **Αυτοματοποίση αξιολόγησης απαντήσεων:** Όσον αφορά το πλαίσιο αξιολόγησης των απαντήσεων, στην συγκεκριμένη διπλωματική επιλέχθηκε η δειγματοληπτική ανθρώπινη αξιολόγηση και η σύγκριση με τις απαντήσεις του μοντέλου. Για τον περιορισμένο αριθμό σετ ερωταπαντήσεων που είχαν παραχθεί, η μέθοδος αυτή αποδίδει καλά (προφανώς ελέγθηκαν και μερικές ακόμη που δεν σημειώθηκαν στις αξιολογήσεις). Σε μεγαλύτερης έκτασης πειράματα, όμως, ενδέχεται η εφαρμογή μιας τέτοιας προσέγγισης να αποδειχθεί δύσκολα εφαρμόσιμη. Μια πιθανή λύση, θα μπορούσε να είναι η αυτοματοποίηση της αξιολόγησης απο κάποιο μοντέλο που λειτουργεί ως ¨κριτής¨. Με κατάλληλο prompting και περιορισμούς, κάτι τέτοιο ίσως είχε πρακτική χρήση.

7.3 Περιορισμοί της Υλοποίησης

Η μεθοδολογία που εφαρμόστηκε, στο πλαίσιο της διπλωματικής αυτής, απέδωσε ικανοποιητικά αποτελέσματα. Τα περισσότερα, αν όχι όλα, τα αποτελέσματα ελέχθηκαν και απο ανθρώπινο παράγοντα έτσι ώστε να υπάρχει σιγουρία για την σταθερά καλή ποιότητα τους. Παρόλαυτα, υπάρχουν ορισμένοι περιορισμοί που πρέπει να αναγνωριστούν:

1. **Αυστηρή λογική περικοπής και επικύρωσης:** Λόγω της αυστηρής συνθήκης ότι η απάντηση πρέπει να υπάρχει αυτούσια μέσα στο κείμενο, ορισμένα έγκυρα ζεύγη ερωτήσεων–απαντήσεων μπορεί να απορρίφθηκαν. Η συμπεριφορά αυτή, αυξάνει την αξιοπιστία του τελικού συνόλου αλλά μειώνει εν μέρει την κάλυψη. Βέβαια, στο πλαίσιο της εργασίας, κάτι τέτοιο δεν είναι ιδιαίτερα επιβλαβές εφόσον αποφασίστηκε πως θα γίνει εστίαση στην ενίσχυση της ακρίβειας.

2. **Μερικές λανθασμένες απαντήσεις:** Κατα την διάρκεια της διαδικασίας (και ιδιαίτερα αν κάποιο σετ έχει περάσει απο όλα τα στάδια της υλοποίησης) το μοντέλο μερικές φορές απέτυχε να κατανοήσει ότι ένα απόσπασμα, παρότι εμφανίζεται μέσα στο κείμενο, δεν απαντά πραγματικά την ερώτηση. Αυτό οδήγησε σε λίγες περιπτώσεις

όπου η απάντηση σημειώθηκε λανθασμένα ως έγκυρη λόγω υψηλής επικάλυψης λεκτικών μονάδων (tokens). Συγκεκριμένα εντοπίστηκε η περίπτωση του σετ: Q: Why is measuring labelled ethacrynic acid binding recommended over ouabain?, A:labelled ethacrynic acid. Εδώ φαίνεται ότι απλά επαναλαμβάνεται κομμάτι της ερώτησης, που υπάρχει αυτούσιο στο κείμενο. Αυτό πιθανότατα συνέβη γιατί το μοντέλο δεν κατανόησε ότι η φράση αυτή δεν απαντά στην ερώτηση, αλλά την θεώρησε σωστή απάντηση.

3. **Μορφολογικοί περιορισμοί της συλλογής κειμένων:** Η απουσία στίξης και η μη φυσική διαίρεση σε προτάσεις δυσκολεύει κάποιες λειτουργίες (όπως η περικοπή ή η αναγνώριση "best sentence"). Παρόλο που η διαδικασία λειτουργεί σωστά, αν η μορφή της συλλογής κειμένων ήταν πιο φυσική, ίσως να υπήρχε μια μικρή βελτίωση στα αποτελέσματα (συγκεκριμένα στο στάδιο περικοπής).

4. **Υπολογιστικό κόστος:** Η χρήση Μεγάλων Γλωσσικών Μοντέλων, ακόμη και σε τοπικό περιβάλλον όπως το Ollama, απαιτεί σημαντικούς υπολογιστικούς πόρους. Η εκτέλεση πολλαπλών σταδίων (παραγωγή, ανάκτηση, γείωση, επανεξέταση) συνεπάγεται χρονική και υπολογιστική επιβάρυνση. Συγκεκριμένα, το στάδιο της γείωσης/επέκτασης ερωτήσεων χρειαζόταν πάρα πολύ ώρα για να λειτουργήσει. Αυτό οδήγησε σε περιορισμένο αριθμό γειωμένων ερωτήσεων (250) σε σχέση με τις συνολικές που ήταν διαθέσιμες (περίπου 2400). Αυτό δεν μειώνει την αξιοπιστία του συστήματος, αλλά είναι σημαντικό να αναφερθεί

Παρά τους παραπάνω περιορισμούς, το σύστημα αποδεικνύεται λειτουργικό, αξιόπιστο και επεκτάσιμο. Οι σχεδιαστικές επιλογές προτιμούν την ποιότητα και την εγκυρότητα έναντι της ποσότητας, στοιχείο που συνάδει πλήρως με τον στόχο της εργασίας: την παραγωγή τεκμηριωμένων και ασφαλών ερωτήσεων–απαντήσεων για εκπαιδευτική χρήση.

Βιβλιογραφία

[1] Ali Alkhatlan και Jugal K. Kalita. Intelligent tutoring systems: A comprehensive historical survey with recent developments. *arXiv preprint arXiv:1812.09628*, 2018.

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil και Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

[3] Sourav Banerjee, Ayushi Agarwal και Saloni Singla. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*, 2024.

[4] Liam Barkley και Brinkvan der Merwe. Investigating the role of prompting and external tools in hallucination rates of large language models. *arXiv preprint arXiv:2410.19385*, 2024.

[5] Delphine Bernhard. Query expansion based on pseudo relevance feedback from definition clusters. Στο *Coling 2010: Posters*, σελίδες 54–62, Beijing, China, 2010. Coling 2010 Organizing Committee.

[6] Patrice Béchard και Orlando Marquez Ayala. Reducing hallucination in structured outputs via retrieval-augmented generation. *arXiv preprint arXiv:2404.08189*, 2024.

[7] Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, Daoyu Wang και Enhong Chen. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*, 2025.

[8] Ping Chen, Fei Wu, Tong Wang και Wei Ding. A semantic qa-based approach for text summarization evaluation. *arXiv preprint arXiv:1704.06259*, 2018.

[9] Deepchecks. What are the practical applications of long-context llms? , 2025.

[10] deepset. Advanced rag: Query expansion. , 2024.

[11] Daniel Deutsch και Dan Roth. Incorporating question answering-based signals into abstractive summarization via salient span selection. *arXiv preprint arXiv:2111.07935*, 2021.

[12] Daniel Deutsch και Dan Roth. Benchmarking answer verification methods for question answering-based summarization evaluation metrics. Στο *Findings of the Association for Computational Linguistics: ACL 2022*, σελίδες 3759–3765, Dublin, Ireland, 2022. Association for Computational Linguistics.

[13] Esin Durmus, He He και Mona Diab. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. Στο *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, σελίδες 5055–5070, Online, 2020. Association for Computational Linguistics.

[14] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness και Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[15] Gökhan Ergen. The danger of unresearched information: blind trust in llm models. , 2024.

[16] Fabiano Falcąo. Metrics for evaluating summarization of texts performed by transformers: how to evaluate the quality of summaries. , 2024.

[17] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang και Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

[18] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang και Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[19] GeeksforGeeks. Python programming language tutorial. , 2024.

[20] GeeksforGeeks. What is bm25 (best matching 25) algorithm? , 2024.

[21] GeeksforGeeks. Sentence transformer - geeksforgeeks. , 2025.

[22] Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart και Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? Στο *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, σελίδες 7765–7784, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[23] Hussam Ghanem και Saba Munawar. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790, 2024.

[24] Deepanway Ghosal, Somak Aditya, Sandipan Dandapat και Monojit Choudhury. Vector space interpolation for query expansion. Στο *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and*

the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), σελίδες 405–410, Online only, 2022. Association for Computational Linguistics.

[25] Shailja Gupta, Rajesh Ranjan και Surya Narayan Singh. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*, 2024.

[26] Stefan Gusenbauer. Relevance feedback in information retrieval systems for specific applications. Διπλωματική εργασία Master, Johannes Kepler Universität Linz (JKU), Linz, Austria, 2006.

[27] Desta Haileselassie Hagos, Rick Battle και Danda B. Rawat. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *arXiv preprint arXiv:2407.14962*, 2024.

[28] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin και Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[29] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu και Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

[30] Aida Kostikova, Zhipin Wang, Deidamea Bajri, Ole Pütz, Benjamin Paaßen και Steffen Eger. Lllms: A data-driven survey of evolving research on limitations of large language models. *arXiv preprint arXiv:2505.19240*, 2025.

[31] Zhengzhong Liang, Yiyun Zhao και Mihai Surdeanu. Using the hammer only on nails: A hybrid method for evidence retrieval for question answering. *arXiv preprint arXiv:2009.10791*, 2020.

[32] Siran Li, Linus Stenzel, Carsten Eickhoff και Seyed Ali Bahrainian. Enhancing retrieval-augmented generation: A study of best practices. *arXiv preprint arXiv:2501.07391*, 2025.

[33] Shengjie Liu, Jing Wu, Jingyuan Bao, Wenyi Wang, Naira Hovakimyan και Christopher G. Healey. Towards a robust retrieval-based summarization system. *arXiv preprint arXiv:2403.19889*, 2024.

[34] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei και Michael Bendersky. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*, 2024.

[35] Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav και Masoud Hashemi. Do llms know when to not answer? investigating abstention abilities of large language models. *arXiv preprint arXiv:2407.16221*, 2024.

[36] Potsawee Manakul, Adian Liusie και Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[37] Rick Merritt. What is retrieval-augmented generation (rag)? , 2025.

[38] Rohan Mistry. The truth behind hallucination in llms: what it is, why it happens and how to tackle it. , 2024.

[39] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes και Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[40] Huyen Nguyen, Haihua Chen, Lavanya Pobbathi και Junhua Ding. A comparative study of quality evaluation methods for text summarization. *arXiv preprint arXiv:2407.00747*, 2024.

[41] Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong και Chien Sheng Wu. Unanswerability evaluation for retrieval-augmented generation. *arXiv preprint arXiv:2412.12300*, 2024.

[42] PythonTutorial.net. What is python? , 2024.

[43] Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib και Most Marufatul Jannat Mim. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839 – 26874, 2024.

[44] Salman Rakin, Md. A. R. Shibly, Zahin M. Hossain, Zeeshan Khan και Md. Mostofa Akbar. Leveraging the domain adaptation of retrieval augmented generation models for question answering and reducing hallucination. *arXiv preprint arXiv:2410.17783*, 2024.

[45] Julian Risch, Timo Möller, Julian Gutsch και Malte Pietsch. Semantic answer similarity for evaluating question answering models. *arXiv preprint arXiv:2108.06130*, 2021.

[46] Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez και Zhe Feng. Delucionqa: Detecting hallucinations in domain-specific question answering. Στο *Findings of the Association for Computational Linguistics: EMNLP 2023*, σελίδες 822–835, Singapore, 2023. Association for Computational Linguistics.

[47] Samia Sahin. Query expansion in enhancing retrieval-augmented generation (rag). , 2024.

[48] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie και Christopher D. Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*, 2024.

[49] Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen και Zhaochun Ren. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. Στο *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, σελίδες 7339–7353, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[50] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana και Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023.

[51] Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu και Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*, 2024.

[52] TechTarget. Ai hallucination – definition and explanation. , 2025.

[53] S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha και Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

[54] Tuyen T. S. T. Metrics for qa: Things to note. , 2024.

[55] Rekha Vaidyanathan, Sujoy Das και Namita Srivastava. Query expansion strategy based on pseudo relevance feedback and term weight scheme for monolingual retrieval. *arXiv preprint arXiv:1502.05168*, 2015.

[56] Guidovan Rossum. An introduction to python. , 2001.

[57] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu και Qingsong Wen. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*, 2024.

[58] Wikipedia. Query expansion. , 2025.

[59] Wikipedia. Retrieval-augmented generation. , 2025.

[60] Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei Wei Kuo, Nan Guan και Chun Jason Xue. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*, 2024.

[61] Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng και Jie Zhou. Unsupervised information refinement training of large language models for retrieval-augmented generation. Στο *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, σελίδες 137–151, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[62] Abdulgaffar Abubakar Yahaya, Bashir Muhammad Ahmad, Faisal Rasheed, Usman Haruna, Abbas Sani, Babangida Salisu Muaz, Ismail Abubakar Yahaya και Aliyu Hamza Idris. Application of machine learning in education: Recent trends, challenges and future perspective. *British Journal of Computer, Networking and Information Technology (BJCNIT)*, 7(3):118–131, 2024.

[63] Borui Yang, Md Afif Al Mamun, Jie M. Zhang και Gias Uddin. Hallucination detection in large language models with metamorphic relations. *arXiv preprint arXiv:2502.15844*, 2025.

[64] Shi Qi Yan, Jia Chen Gu, Yun Zhu και Zhen Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

[65] Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond και Franziska Gouverneur. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 2019.

[66] Le Zhang, Yihong Wu, Qian Yang και Jian Yun Nie. Exploring the best practices of query expansion with large language models. Στο *Findings of EMNLP 2024*, σελίδες 1872–1883. Association for Computational Linguistics, 2024.

[67] Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica και Joseph E. Gonzalez. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*, 2024.