ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία

Τίτλος Πτυχιακής Εργασίας	Μελέτη της αποτελεσματικότητας των LLMs σε διαφορετικά επιστημονικά πεδία *Study of the effectiveness of LLMs across different scientific fields*
Ονοματεπώνυμο Φοιτητή	Αντώνιος Μπαρδάνης
Πατρώνυμο	Νικόλαος
Αριθμός Μητρώου	Π21110
Επιβλέπων	Κωνσταντίνος Λιαγκούρας, Επίκουρος Καθηγητής

Ημερομηνία Παράδοσης: Ιούνιος 2026

Copyright ©

Ευχαριστίες

Με την ολοκλήρωση της παρούσας πτυχιακής εργασίας, θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες σε όσους συνέβαλαν στην εκπόνησή της.

Καταρχάς, ευχαριστώ θερμά τον επιβλέποντα καθηγητή μου, κ. Κωνσταντίνο Λιαγκούρα, για την πολύτιμη καθοδήγηση, τις εύστοχες παρατηρήσεις και τη συνεχή υποστήριξή του καθ' όλη τη διάρκεια της εργασίας. Η εμπιστοσύνη και οι συμβουλές του υπήρξαν καθοριστικές για την ολοκλήρωσή της.

Ένα μεγάλο ευχαριστώ οφείλω στην οικογένειά μου, για την αμέριστη στήριξη, την υπομονή και την ενθάρρυνση σε όλη τη διάρκεια των σπουδών μου, καθώς και στους φίλους μου, που έκαναν αυτή την πορεία πιο ευχάριστη.

Σε όλους όσοι συνέβαλαν, με οποιονδήποτε τρόπο, στην ολοκλήρωση αυτής της προσπάθειας, εκφράζω την ειλικρινή μου ευγνωμοσύνη.

Περίληψη

Η παρούσα πτυχιακή εργασία διερευνά την αποτελεσματικότητα των Μεγάλων Γλωσσικών Μοντέλων (LLMs) σε διαφορετικά επιστημονικά πεδία, εξετάζοντας πώς διαφοροποιείται η απόδοσή τους ανάλογα με το εκάστοτε γνωστικό αντικείμενο. Καθώς η παραγωγική τεχνητή νοημοσύνη ενσωματώνεται ταχύτατα σε ερευνητικά και επαγγελματικά περιβάλλοντα, η κριτική αξιολόγηση των δυνατοτήτων αλλά και των ορίων της αποτελεί επιτακτική ανάγκη.

Για τον σκοπό αυτό, εκπονήθηκε Συστηματική Βιβλιογραφική Ανασκόπηση (SLR) βασισμένη στο πρωτόκολλο PRISMA και το πλαίσιο επιλεξιμότητας PICOC. Μέσα από την ανάλυση 43 πρόσφατων επιστημονικών μελετών, χαρτογραφείται η απόδοση των μοντέλων σε πέντε βασικούς τομείς: Θετικές Επιστήμες και Μαθηματικά, Πληροφορική και Προγραμματισμό, Ιατρική και Βιοεπιστήμες, Νομική, καθώς και Κοινωνικές και Ανθρωπιστικές Επιστήμες.

Τα ευρήματα υποδεικνύουν ότι η αποτελεσματικότητα των LLMs δεν είναι ομοιόμορφη, αλλά εξαρτάται άμεσα από τον βαθμό τυποποίησης του πεδίου, τον όγκο των διαθέσιμων δεδομένων εκπαίδευσης και την πολυπλοκότητα του απαιτούμενου συλλογισμού. Ενώ τα μοντέλα επιδεικνύουν κορυφαίες επιδόσεις σε αυστηρά δομημένες εργασίες, όπως η παραγωγή μεμονωμένου κώδικα ή η επίλυση βασικών μαθηματικών προβλημάτων, αντιμετωπίζουν σημαντικές προκλήσεις σε καθήκοντα που απαιτούν πολυβηματική λογική ή λεπτή κλινική και νομική κρίση. Ακόμη και όταν σημειώνουν υψηλές βαθμολογίες σε τυποποιημένα benchmarks, παραμένουν επιρρεπή σε ψευδαισθήσεις (hallucinations). Παράλληλα, καταγράφεται σημαντική υστέρηση της απόδοσής τους σε γλώσσες με λιγότερους πόρους (όπως τα ελληνικά) και σε μη αγγλοσαξονικά θεσμικά πλαίσια.

Συμπερασματικά, αν και τεχνικές όπως το prompt engineering, το fine-tuning και οι πολυπρακτορικές μέθοδοι (agentic workflows) ενισχύουν κατακόρυφα την ακρίβεια των απαντήσεων, η ανθρώπινη εποπτεία παραμένει αναντικατάστατη. Τα LLMs αποδεικνύεται πως λειτουργούν βέλτιστα ως ισχυρά εργαλεία υποστήριξης και όχι ως αυτόνομοι λήπτες αποφάσεων, ιδιαίτερα σε περιβάλλοντα υψηλού ρίσκου.

Λέξεις Κλειδιά: Μεγάλα Γλωσσικά Μοντέλα, Αξιολόγηση, Benchmarks, Επιστημονικά Πεδία, Συστηματική Βιβλιογραφική Ανασκόπηση

Abstract

This thesis investigates the effectiveness of Large Language Models (LLMs) across various scientific disciplines, examining how their performance fluctuates depending on the specific domain. As generative artificial intelligence is rapidly being integrated into both research and professional environments, a critical evaluation of its capabilities and limitations is highly necessary.

To this end, a Systematic Literature Review (SLR) was conducted, adhering to the PRISMA guidelines and the PICOC framework. By analyzing 43 recent peer-reviewed studies, this research maps the performance of LLMs across five core areas: STEM and Mathematics, Computer Science and Programming, Medicine and Biosciences, Law, and Social Sciences and Humanities.

The findings indicate that the effectiveness of LLMs is far from uniform. It heavily depends on the degree of standardization within a field, the volume of available training data, and the complexity of the required reasoning. While the models demonstrate exceptional performance in highly structured tasks, such as generating isolated code snippets or solving basic mathematical problems, they struggle significantly with multi-step logical reasoning and tasks requiring nuanced clinical or legal judgment. Even when achieving high scores on standardized benchmarks, they remain prone to hallucinations. Furthermore, a substantial performance gap is evident when operating in low-resource languages or outside Anglo-centric cultural and institutional contexts.

Ultimately, although optimization techniques—such as prompt engineering, fine-tuning, and agentic workflows—drastically enhance model output, human oversight remains irreplaceable. LLMs are best utilized as advanced assistive tools rather than autonomous decision-makers, particularly in high-stakes environments.

Key Words: Large Language Models, Evaluation, Benchmarks, Scientific Fields, Systematic Literature Review

Περιεχόμενα

Κατάλογος Εικόνων

1 Εισαγωγή

Τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models, LLMs) έχουν αναδειχθεί σε μία από τις σημαντικότερες τεχνολογικές εξελίξεις της τρέχουσας δεκαετίας, με ταχεία διάχυση σε ένα ευρύ φάσμα επιστημονικών και επαγγελματικών πεδίων. Καθώς η χρήση τους επεκτείνεται από τη γενική επεξεργασία κειμένου σε εξειδικευμένες εφαρμογές —από την ιατρική διάγνωση έως τη νομική ανάλυση— προκύπτει ένα κρίσιμο ερώτημα: η αποτελεσματικότητά τους είναι ομοιόμορφη σε όλα τα πεδία, ή διαφοροποιείται ανάλογα με τη φύση του εκάστοτε γνωστικού αντικειμένου; Η παρούσα εργασία επιχειρεί να απαντήσει σε αυτό το ερώτημα μέσω μιας συστηματικής βιβλιογραφικής ανασκόπησης της σχετικής επιστημονικής βιβλιογραφίας.

1.1 Πλαίσιο και Κίνητρο

Η ραγδαία εξέλιξη των LLMs τα τελευταία χρόνια έχει μεταβάλει το τοπίο της τεχνητής νοημοσύνης. Μοντέλα όπως το GPT-4 [1], η οικογένεια Llama [2] και η οικογένεια Claude [3] επιδεικνύουν ικανότητες που, σε ορισμένες περιπτώσεις, προσεγγίζουν ή υπερβαίνουν την ανθρώπινη απόδοση σε τυποποιημένες δοκιμασίες. Η διάχυση αυτή έχει οδηγήσει σε αυξανόμενο ενδιαφέρον για την αξιοποίησή τους σε εξειδικευμένα πεδία με υψηλές απαιτήσεις ακρίβειας.

Ωστόσο, η ίδια η ευρύτητα των εφαρμογών αναδεικνύει ένα θεμελιώδες ζήτημα. Η απόδοση ενός LLM σε μια γενική δοκιμασία γνώσης δεν συνεπάγεται αναγκαστικά αντίστοιχη απόδοση σε ένα εξειδικευμένο πεδίο που απαιτεί βαθιά κατανόηση, πολυβηματικό συλλογισμό, ή λειτουργία σε μη-αγγλόφωνο περιβάλλον. Η αυξανόμενη χρήση των μοντέλων αυτών σε κρίσιμες αποφάσεις καθιστά την κατανόηση των ορίων τους όχι απλώς ακαδημαϊκό, αλλά πρακτικό ζήτημα με ουσιαστικές συνέπειες.

Παρά τον μεγάλο όγκο μεμονωμένων μελετών που αξιολογούν τα LLMs σε επιμέρους πεδία, παρατηρείται έλλειψη συστηματικής σύνθεσης που να συγκρίνει την απόδοση μεταξύ διαφορετικών επιστημονικών κλάδων και να εντοπίζει τους παράγοντες που ερμηνεύουν τις διαφορές [4]. Το κενό αυτό είναι το κίνητρο της παρούσας εργασίας: η συγκέντρωση, οργάνωση και κριτική σύνθεση των διάσπαρτων ευρημάτων σε μια ενιαία, συγκριτική εικόνα.

1.2 Σκοπός και Ερευνητικά Ερωτήματα

Σκοπός της εργασίας είναι η συστηματική χαρτογράφηση της αποτελεσματικότητας των LLMs σε διαφορετικά επιστημονικά πεδία, καθώς και ο εντοπισμός των παραγόντων που ερμηνεύουν τις παρατηρούμενες διαφορές. Για την επίτευξη του σκοπού αυτού, διατυπώνονται πέντε ερευνητικά ερωτήματα (Research Questions, RQs) που καθοδηγούν την ανασκόπηση:

- RQ1: Σε ποια επιστημονικά πεδία έχει αξιολογηθεί η απόδοση των LLMs και με ποια benchmarks;
- RQ2: Πώς διαφοροποιείται η απόδοση των LLMs μεταξύ διαφορετικών επιστημονικών πεδίων;
- RQ3: Ποιοι παράγοντες (δεδομένα εκπαίδευσης, τυποποίηση του πεδίου, γλώσσα, τύπος εργασίας) εξηγούν τις διαφορές αυτές;
- RQ4: Ποιες τεχνικές (prompt engineering, fine-tuning, RAG, agentic methods) βελτιώνουν την απόδοση ανά πεδίο;
- RQ5: Ποιοι μεθοδολογικοί περιορισμοί (data contamination, prompt sensitivity, benchmark saturation) επηρεάζουν την αξιοπιστία των αποτελεσμάτων;

Τα ερωτήματα αυτά καλύπτουν ολόκληρο το φάσμα της ανάλυσης: από την καταγραφή του πεδίου (RQ1) και τη συγκριτική αποτίμηση (RQ2), στην ερμηνεία των διαφορών (RQ3), τις τεχνικές βελτίωσης (RQ4), και τέλος την κριτική αξιολόγηση της αξιοπιστίας των ίδιων των ευρημάτων (RQ5).

1.3 Δομή της Εργασίας

Η εργασία διαρθρώνεται σε επτά κεφάλαια. Το παρόν πρώτο κεφάλαιο εισάγει το πρόβλημα, το κίνητρο και τα ερευνητικά ερωτήματα. Το δεύτερο κεφάλαιο περιγράφει τη μεθοδολογία της συστηματικής βιβλιογραφικής ανασκόπησης, ακολουθώντας το πρωτόκολλο PRISMA,

και τεκμηριώνει τη διαδικασία επιλογής των 43 μελετών. Το τρίτο κεφάλαιο αναλύει τις θεμελιώδεις τεχνολογίες στις οποίες βασίζονται τα LLMs — διανύσματα λέξεων, παραμέτρους, την αρχιτεκτονική Transformer και τον μηχανισμό προσοχής, την παράλληλη επεξεργασία, τις κάρτες γραφικών, την προ-εκπαίδευση και την ενισχυτική μάθηση από ανθρώπινη ανατροφοδότηση — καθώς και τους βασικούς τρόπους χρήσης και προσαρμογής τους στην πράξη. Το τέταρτο κεφάλαιο εξετάζει τις μεθοδολογίες αξιολόγησης και τα κύρια benchmarks. Το πέμπτο κεφάλαιο, που αποτελεί τον πυρήνα της εργασίας, αναλύει την απόδοση ανά επιστημονικό πεδίο. Το έκτο κεφάλαιο συνθέτει τα ευρήματα και συζητά τις απειλές εγκυρότητας. Τέλος, το έβδομο κεφάλαιο διατυπώνει τα συμπεράσματα, απαντά στα ερευνητικά ερωτήματα, και προτείνει κατευθύνσεις για μελλοντική έρευνα.

2 Μέθοδοι Βιβλιογραφικής Έρευνας

Στο παρόν κεφάλαιο περιγράφεται το ερευνητικό πρωτόκολλο που ακολουθήθηκε για τη συστηματική μελέτη της βιβλιογραφίας. Η διαδικασία σχεδιάστηκε ώστε να είναι διαφανής και αναπαραγώγιμη, ακολουθώντας τις αρχές της μεθοδολογίας PRISMA 2020 [5] και το πλαίσιο PICOC για τον ορισμό των κριτηρίων επιλεξιμότητας.

2.1 Συστηματική Βιβλιογραφική Ανασκόπηση (SLR)

2.1.1 Σκοπιμότητα της SLR

Σε αντίθεση με μια παραδοσιακή (αφηγηματική) ανασκόπηση, η Συστηματική Βιβλιογραφική Ανασκόπηση (Systematic Literature Review) βασίζεται σε ένα προκαθορισμένο, τεκμηριωμένο πρωτόκολλο αναζήτησης, επιλογής και αξιολόγησης των μελετών. Έτσι περιορίζεται η μεροληψία του ερευνητή, διασφαλίζεται η αναπαραγωγιμότητα των αποτελεσμάτων και επιτυγχάνεται πληρέστερη κάλυψη του πεδίου.

Δεδομένου ότι η αξιολόγηση των Μεγάλων Γλωσσικών Μοντέλων (LLMs) αποτελεί ταχύτατα εξελισσόμενο και διεπιστημονικό αντικείμενο, η συστηματική προσέγγιση κρίνεται κατάλληλη για τη χαρτογράφηση της διάσπαρτης βιβλιογραφίας και τη συγκριτική σύνθεση των ευρημάτων ανά επιστημονικό πεδίο.

2.1.2 Στόχοι και Ερευνητικά Ερωτήματα (RQs)

Στόχος της εργασίας είναι η συστηματική αποτύπωση και σύγκριση της αποτελεσματικότητας των LLMs σε διαφορετικά επιστημονικά πεδία. Για τον σκοπό αυτό διατυπώνονται τα ακόλουθα ερευνητικά ερωτήματα:

Κωδικός	Ερευνητικό Ερώτημα
RQ1	Σε ποια επιστημονικά πεδία έχει αξιολογηθεί η απόδοση των LLMs και με ποια benchmarks ή σύνολα δεδομένων;
RQ2	Πώς διαφοροποιείται η απόδοση των LLMs μεταξύ διαφορετικών επιστημονικών πεδίων;
RQ3	Ποιοι παράγοντες (φύση δεδομένων εκπαίδευσης, βαθμός τυποποίησης του πεδίου, γλώσσα, τύπος εργασίας) εξηγούν τις διαφορές απόδοσης;
RQ4	Ποιες τεχνικές (prompt engineering, fine-tuning, retrieval-augmented generation) βελτιώνουν την απόδοση σε συγκεκριμένα πεδία;
RQ5	Ποιοι μεθοδολογικοί περιορισμοί (π.χ. data contamination, ευαισθησία στο prompt) επηρεάζουν την αξιοπιστία των συγκρίσεων;

Πίνακας 2.1: Ερευνητικά Ερωτήματα (RQs) της εργασίας.

2.2 Υλικά και Μέθοδοι

2.2.1 Κριτήρια Επιλεξιμότητας (PICOC)

Τα κριτήρια επιλεξιμότητας ορίστηκαν με βάση το πλαίσιο PICOC (Population, Intervention, Comparison, Outcome, Context):

Στοιχείο	Περιγραφή
Population (Πληθυσμός)	Μεγάλα Γλωσσικά Μοντέλα (LLMs) — γενικού σκοπού και εξειδικευμένα.
Intervention (Παρέμβαση)	Η εφαρμογή ή αξιολόγηση των LLMs σε εργασίες ενός συγκεκριμένου επιστημονικού πεδίου (ερωτήσεις, εξετάσεις, benchmarks).
Comparison (Σύγκριση)	Σύγκριση απόδοσης μεταξύ πεδίων, μεταξύ μοντέλων, ή έναντι ανθρώπων-ειδικών και κλασικών μεθόδων.

Στοιχείο	Περιγραφή
Outcome (Έκβαση)	Μετρήσιμη απόδοση: ακρίβεια, pass@k, F1, ποιότητα/εγκυρότητα συλλογισμού, ποσοστό ψευδαισθήσεων.
Context (Πλαίσιο)	Ακαδημαϊκή ή πειραματική αξιολόγηση μέσω δημοσιευμένων μελετών και αναγνωρισμένων benchmarks.

Πίνακας 2.2: Πλαίσιο επιλεξιμότητας PICOC.

2.2.2 Κριτήρια Εισαγωγής και Αποκλεισμού

Κριτήρια εισαγωγής:

- Δημοσιεύσεις της περιόδου 2019–2026 (μετά την καθιέρωση των μοντέλων τύπου Transformer).
- Άρθρα με κριτές (peer-reviewed) ή αξιόπιστα preprints (π.χ. arXiv) που περιλαμβάνουν εμπειρική αξιολόγηση απόδοσης.
- Μελέτες που αναφέρουν ποσοτικά ή ποιοτικά αποτελέσματα απόδοσης σε τουλάχιστον ένα επιστημονικό πεδίο.
- Γλώσσα δημοσίευσης: αγγλικά ή ελληνικά.
- Διαθέσιμο πλήρες κείμενο.

Κριτήρια αποκλεισμού:

- Άρθρα γνώμης, εκλαϊκευτικά κείμενα ή αναρτήσεις χωρίς μεθοδολογία.
- Μελέτες που δεν αξιολογούν απόδοση (π.χ. αμιγώς θεωρητικές, χωρίς benchmark).
- Διπλότυπα ή προγενέστερες εκδόσεις της ίδιας μελέτης.
- Μη προσβάσιμο πλήρες κείμενο.

2.2.3 Πηγές Πληροφόρησης

Η αναζήτηση πραγματοποιήθηκε στις ακόλουθες βάσεις δεδομένων και ψηφιακές βιβλιοθήκες:

- Scopus
- IEEE Xplore
- ACM Digital Library
- ACL Anthology
- arXiv (cs.CL, cs.AI)
- SpringerLink και ScienceDirect
- PubMed (για το ιατρικό/βιοϊατρικό πεδίο)
- Google Scholar (συμπληρωματικά, για εντοπισμό grey literature)

Συμπληρωματικά αξιοποιήθηκε η τεχνική του snowballing (forward και backward citation) για τον εντοπισμό σχετικών μελετών που δεν επιστράφηκαν από τις αρχικές αναζητήσεις.

2.2.4 Στρατηγική Αναζήτησης

Η στρατηγική αναζήτησης δομήθηκε σε τρεις εννοιολογικές ομάδες όρων, που συνδυάστηκαν με τελεστές Boolean:

- Ομάδα Α — Μοντέλα: "large language model*" OR "LLM*" OR "GPT" OR "generative AI" OR "foundation model*"
- Ομάδα Β — Αξιολόγηση: evaluat* OR benchmark* OR performance OR accuracy OR assess*
- Ομάδα Γ — Πεδίο/Διακύμανση: domain OR discipline OR field OR "subject area" OR cross-domain

Γενική μορφή ερωτήματος: (Α) AND (Β) AND (Γ). Ενδεικτικό ερώτημα σε σύνταξη Scopus:

```
TITLE-ABS-KEY(("large language model*" OR "LLM*" OR GPT) AND (evaluat* OR benchmark*
OR performance OR accuracy) AND (domain OR discipline OR field)) AND PUBYEAR > 2018
```

Για στοχευμένη αναζήτηση ανά πεδίο, η Ομάδα Γ αντικαθίσταται από τους αντίστοιχους όρους:

Επιστημονικό Πεδίο	Όροι αναζήτησης (Ομάδα Γ)
Θετικές Επιστήμες / Μαθηματικά	mathemat* OR "STEM" OR physics OR chemistry OR "scientific reasoning"
Πληροφορική / Προγραμματισμός	"code generation" OR programming OR "software engineering" OR coding
Ιατρική / Βιοεπιστήμες	medical OR clinical OR medicine OR biomedical OR diagnosis
Νομική	legal OR law OR "bar exam" OR jurisprudence
Κοινωνικές / Ανθρωπιστικές	humanities OR "social science*" OR history OR philosophy

Πίνακας 2.3: Όροι αναζήτησης ανά επιστημονικό πεδίο.

2.2.5 Αξιολόγηση Ποιότητας

Κάθε μελέτη που πέρασε τον έλεγχο επιλεξιμότητας αξιολογήθηκε ως προς την ποιότητά της με βάση τον ακόλουθο κατάλογο κριτηρίων. Κάθε κριτήριο βαθμολογήθηκε με 1 (Ναι), 0,5 (Εν μέρει) ή 0 (Όχι):

#	Κριτήριο Ποιότητας
Q1	Είναι σαφώς διατυπωμένοι οι στόχοι ή τα ερευνητικά ερωτήματα της μελέτης;
Q2	Περιγράφονται με σαφήνεια τα μοντέλα και οι εκδόσεις τους;
Q3	Χρησιμοποιείται κατάλληλο και αναγνωρισμένο benchmark ή σύνολο δεδομένων;
Q4	Είναι κατάλληλες οι μετρικές αξιολόγησης για το εξεταζόμενο πεδίο;
Q5	Είναι αναπαραγώγιμη η μελέτη (διαθέσιμα prompts, κώδικας ή δεδομένα);
Q6	Συζητούνται οι περιορισμοί (π.χ. data contamination, μεροληψία);

Πίνακας 2.4: Κατάλογος αξιολόγησης ποιότητας μελετών.

Ως κατώφλι αποδοχής ορίστηκε η συνολική βαθμολογία ≥ 3 στα 6. Μελέτες με χαμηλότερη βαθμολογία αποκλείστηκαν από την τελική σύνθεση.

2.2.6 Ανάλυση Επιλογής (Διάγραμμα PRISMA)

Η διαδικασία επιλογής των μελετών ακολούθησε με ακρίβεια τα τέσσερα στάδια που υπαγορεύει το αναθεωρημένο πρωτόκολλο PRISMA 2020: Ταυτοποίηση (Identification), Έλεγχος (Screening), Επιλεξιμότητα (Eligibility) και Ένταξη (Included). Τα αριθμητικά αποτελέσματα σε κάθε επιμέρους φάση συνοψίζονται στον Πίνακα 2.5 και αποτυπώνονται σχηματικά στο αντίστοιχο διάγραμμα ροής (Εικόνα 1).

Αρχικά, η στοχευμένη αναζήτηση με τους συγκεκριμένους όρους και φίλτρα (ημερομηνία, γλώσσα, είδος δημοσίευσης και θεματική εστίαση) επέστρεψε έναν περιορισμένο αλλά διαχειρίσιμο αριθμό εγγραφών. Ήδη σε αυτό το στάδιο εφαρμόστηκαν αρκετά αυστηρά κριτήρια, ώστε να αποκλειστούν γενικές εργασίες για τα LLMs που δεν σχετίζονταν με αξιολόγηση ανά επιστημονικό πεδίο.

Στη φάση του ελέγχου (screening), οι τίτλοι και οι περιλήψεις ελέγχθηκαν χειροκίνητα με βάση τα κριτήρια εισαγωγής και αποκλεισμού. Από τη διαδικασία αυτή αποκλείστηκαν μόνο ελάχιστες εργασίες, κυρίως επειδή αφορούσαν θεωρητικές προσεγγίσεις χωρίς εμπειρική αξιολόγηση ή επειδή δεν παρείχαν διακριτά αποτελέσματα ανά πεδίο.

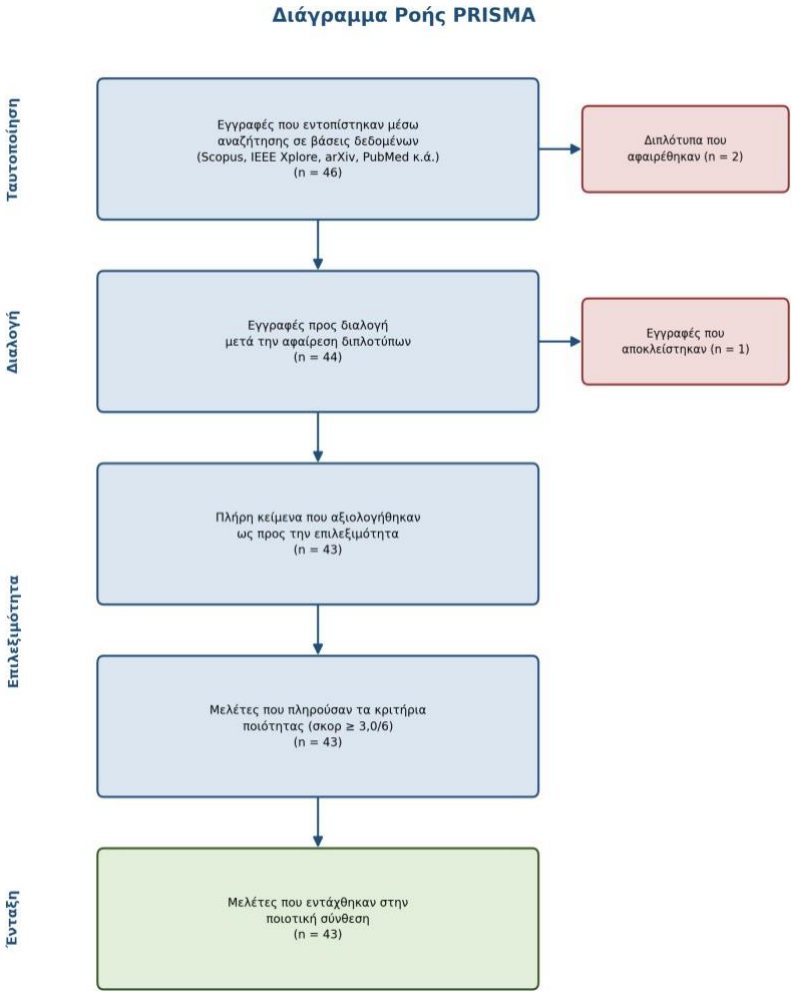
Στη συνέχεια, τα πλήρη κείμενα των υπολοίπων μελετών εξετάστηκαν διεξοδικά ως προς την επιλεξιμότητα και την ποιότητα. Όσες δεν πληρούσαν το κατώφλι του ελέγχου ποιότητας ή δεν παρείχαν επαρκώς τεκμηριωμένα αποτελέσματα αποκλείστηκαν, ενώ οι 43 τελικά επιλεγμένες μελέτες εντάχθηκαν στην ποιοτική σύνθεση της ανασκόπησης.

Είναι σημαντικό να σημειωθεί ότι ο σχετικά μικρός αριθμός τελικών μελετών δεν οφείλεται σε περιορισμένη αναζήτηση, αλλά στην αυστηρή εστίαση σε εργασίες που αξιολογούν ρητά την απόδοση των LLMs σε συγκεκριμένα επιστημονικά πεδία, με σαφείς μετρικές και συγκρίσεις.

Στάδιο	**Πλήθος εγγραφών**
Εγγραφές από βάσεις δεδομένων (Identification)	46
Εγγραφές μετά την αφαίρεση διπλότυπων	44
Έλεγχος τίτλου/περίληψης (Screening)	44 (αποκλείστηκε 1)

Στάδιο	Πλήθος εγγραφών
Αξιολόγηση πλήρους κειμένου (Eligibility)	43 (αποκλείστηκαν 0)
Προστέθηκαν μέσω snowballing	0
Τελικά ενταγμένες μελέτες (Included)	43

Πίνακας 2.5: Πλήθη εγγραφών ανά στάδιο της διαδικασίας επιλογής.



Εικόνα 1: Διάγραμμα ροής PRISMA

3 Τεχνολογικό Υπόβαθρο: Θεμελιώδεις Τεχνολογίες και Λειτουργία των LLMs

Τα Μεγάλα Γλωσσικά Μοντέλα δεν αποτελούν μια μεμονωμένη εφεύρεση, αλλά το αποτέλεσμα της σύγκλισης μιας σειράς θεμελιωδών τεχνολογιών που αναπτύχθηκαν σταδιακά τις τελευταίες δεκαετίες. Η εντυπωσιακή απόδοση που παρουσιάζουν τα σύγχρονα μοντέλα δεν εξηγείται από έναν και μόνο παράγοντα, αλλά από τον συνδυασμό κατάλληλων αναπαραστάσεων της γλώσσας, μιας ευέλικτης νευρωνικής αρχιτεκτονικής, τεράστιας υπολογιστικής ισχύος και προηγμένων μεθόδων εκπαίδευσης. Σκοπός του παρόντος κεφαλαίου είναι να εξηγήσει, με τρόπο συστηματικό και κατά το δυνατόν προσιτό, τις τεχνολογίες αυτές, ώστε να καταστούν κατανοητοί τόσο οι δυνατότητες όσο και τα όρια των μοντέλων που αναλύονται στα επόμενα κεφάλαια.

Συγκεκριμένα, εξετάζονται διαδοχικά οκτώ δομικά στοιχεία: τα διανύσματα λέξεων (word embeddings), οι παράμετροι (parameters), η αρχιτεκτονική Transformer, ο μηχανισμός της αυτο-προσοχής (self-attention), η παράλληλη επεξεργασία (parallel processing), οι κάρτες γραφικών (GPUs), η προ-εκπαίδευση (pre-training) και η ενισχυτική μάθηση από ανθρώπινη ανατροφοδότηση (RLHF). Τα στοιχεία αυτά δεν λειτουργούν ανεξάρτητα, αλλά συνθέτουν μια ενιαία αλυσίδα: από τη μετατροπή της γλώσσας σε αριθμούς, έως την ευθυγράμμιση του τελικού μοντέλου με τις ανθρώπινες προτιμήσεις.

Το κεφάλαιο ξεκινά από τα δομικά στοιχεία των μοντέλων και προχωρά σταδιακά προς τη λειτουργία τους: αφού αναλυθούν οι θεμελιώδεις τεχνολογίες, η τελευταία ενότητα εξετάζει τους βασικούς τρόπους με τους οποίους τα μοντέλα αξιοποιούνται και προσαρμόζονται στην πράξη. Η κατανόηση του «πώς» λειτουργούν τα μοντέλα σε τεχνικό επίπεδο αποτελεί απαραίτητη προϋπόθεση για την κριτική ερμηνεία των ευρημάτων απόδοσης που παρουσιάζονται στα επόμενα κεφάλαια.

Για να εκτιμηθεί η σημασία των τεχνολογιών αυτών, είναι χρήσιμη μια σύντομη ιστορική αναδρομή. Πριν από την επικράτηση της αρχιτεκτονικής Transformer, η αυτόματη επεξεργασία γλώσσας στηριζόταν σε στατιστικά μοντέλα n-gram και, αργότερα, σε αναδρομικά νευρωνικά δίκτυα. Οι προσεγγίσεις αυτές, αν και χρήσιμες για την εποχή τους, αντιμετώπιζαν σοβαρές δυσκολίες στη διαχείριση μακρινών εξαρτήσεων μέσα στο κείμενο και δεν κλιμακώνονταν αποδοτικά. Η μετάβαση στα σύγχρονα LLMs δεν ήταν, επομένως, αποτέλεσμα μίας και μόνο ιδέας, αλλά της συνέργειας όλων των τεχνολογιών που αναλύονται στη συνέχεια.

3.1 Διανύσματα Λέξεων (Word Embeddings)

Η πρώτη θεμελιώδης πρόκληση που πρέπει να αντιμετωπίσει ένα γλωσσικό μοντέλο είναι ότι οι υπολογιστές δεν επεξεργάζονται λέξεις, αλλά αριθμούς. Επομένως, πριν από οποιαδήποτε επεξεργασία, το κείμενο πρέπει να μετατραπεί σε μια αριθμητική μορφή. Το πρώτο βήμα αυτής της διαδικασίας είναι η τμηματοποίηση (tokenization), κατά την οποία το κείμενο διασπάται σε μικρότερες μονάδες που ονομάζονται tokens. Ένα token μπορεί να αντιστοιχεί σε μια ολόκληρη λέξη, σε ένα τμήμα λέξης (subword) ή ακόμη και σε έναν μεμονωμένο χαρακτήρα, ανάλογα με τη μέθοδο τμηματοποίησης που χρησιμοποιείται.

Οι μέθοδοι τμηματοποίησης που επικρατούν σήμερα, όπως ο αλγόριθμος Byte-Pair Encoding (BPE) και η παραλλαγή WordPiece, ακολουθούν μια ενδιάμεση προσέγγιση: αντί να μεταχειρίζονται ως μονάδα είτε ολόκληρες λέξεις είτε μεμονωμένους χαρακτήρες, μαθαίνουν ένα λεξιλόγιο από συχνά εμφανιζόμενα υπο-τμήματα λέξεων. Με τον τρόπο αυτό, σπάνιες ή άγνωστες λέξεις αναλύονται σε γνωστά μικρότερα κομμάτια, αποφεύγοντας το πρόβλημα των λέξεων εκτός λεξιλογίου (out-of-vocabulary) και διατηρώντας παράλληλα διαχειρίσιμο το μέγεθος του λεξιλογίου, που συνήθως κυμαίνεται σε μερικές δεκάδες χιλιάδες tokens.

Στη συνέχεια, κάθε token αντιστοιχίζεται σε ένα διάνυσμα (embedding), δηλαδή σε μια λίστα πραγματικών αριθμών — συνήθως εκατοντάδων ή και χιλιάδων διαστάσεων. Το ουσιώδες χαρακτηριστικό αυτής της αναπαράστασης είναι ότι δεν είναι τυχαία: τα διανύσματα οργανώνονται σε έναν πολυδιάστατο χώρο κατά τέτοιο τρόπο ώστε λέξεις με παρόμοια σημασία να βρίσκονται κοντά μεταξύ τους. Έτσι, η γεωμετρική απόσταση μεταξύ δύο διανυσμάτων αποτυπώνει, σε σημαντικό βαθμό, τη σημασιολογική τους συγγένεια.

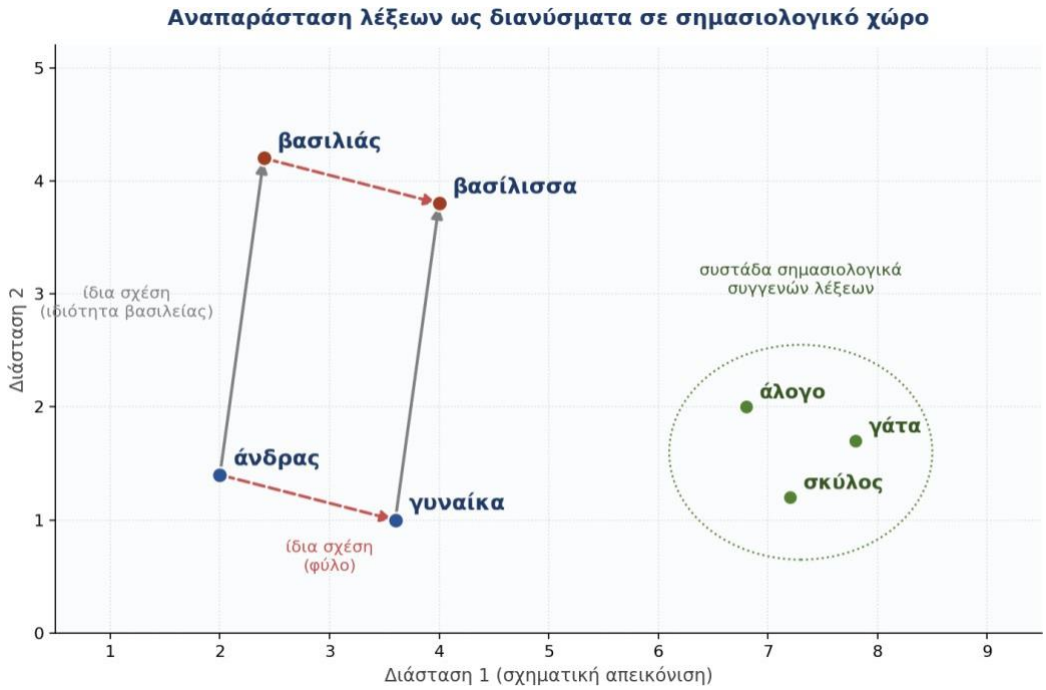
Το πλήθος των διαστάσεων κάθε διανύσματος (embedding dimension) αποτελεί βασική σχεδιαστική επιλογή και κυμαίνεται τυπικά από μερικές εκατοντάδες έως αρκετές χιλιάδες. Το σύνολο των διανυσμάτων για όλα τα tokens του λεξιλογίου συγκροτεί έναν μεγάλο πίνακα ενσωμάτωσης (embedding matrix), ο οποίος αποτελεί μέρος των συνολικών παραμέτρων του μοντέλου και μαθαίνεται κι αυτός κατά την εκπαίδευση. Μεγαλύτερες διαστάσεις προσφέρουν πλουσιότερη αναπαράσταση, αυξάνουν όμως το υπολογιστικό κόστος και τις απαιτήσεις σε μνήμη.

Η ιδιότητα αυτή έγινε ευρέως γνωστή μέσα από τα πρώιμα μοντέλα διανυσματικών αναπαραστάσεων, όπως τα word2vec [6] και GloVe [7], τα οποία ανέδειξαν ότι ο χώρος των διανυσμάτων μπορεί να κωδικοποιεί ακόμη και αναλογικές σχέσεις. Το κλασικό παράδειγμα είναι ότι το διάνυσμα της λέξης «βασιλιάς», αν αφαιρεθεί από αυτό η έννοια «άνδρας» και προστεθεί η έννοια «γυναίκα», προσεγγίζει το διάνυσμα της λέξης «βασίλισσα». Παρόμοιες κανονικότητες υποδηλώνουν ότι η σημασιολογική δομή της γλώσσας μπορεί, ως έναν βαθμό, να αναπαρασταθεί αριθμητικά.

Τα πρώιμα αυτά διανύσματα ήταν ωστόσο στατικά: κάθε λέξη αντιστοιχούσε σε ένα και μόνο διάνυσμα, ανεξάρτητα από τα συμφραζόμενα. Αυτό δημιουργούσε προβλήματα με τις πολύσημες λέξεις· για παράδειγμα, η αγγλική λέξη «bank» έχει την ίδια αναπαράσταση είτε αναφέρεται σε τράπεζα είτε σε όχθη ποταμού. Τα σύγχρονα LLMs ξεπερνούν αυτόν τον περιορισμό παράγοντας συμφραζόμενα διανύσματα (contextual embeddings): η αναπαράσταση κάθε λέξης υπολογίζεται δυναμικά με βάση το σύνολο της πρότασης, ώστε η ίδια λέξη να λαμβάνει διαφορετικό διάνυσμα ανάλογα με τη χρήση της.

Αξίζει να σημειωθεί ότι τα διανύσματα δεν χρησιμοποιούνται μόνο στην είσοδο, αλλά και στην έξοδο του μοντέλου. Όταν το μοντέλο παράγει κείμενο, η τελική εσωτερική του αναπαράσταση προβάλλεται εκ νέου στον χώρο του λεξιλογίου, ώστε να υπολογιστεί, μέσω της συνάρτησης softmax, μια κατανομή πιθανότητας για το ποιο θα είναι το επόμενο token. Συχνά, μάλιστα, ο πίνακας που χρησιμοποιείται για την προβολή αυτή στην έξοδο είναι κοινός με τον πίνακα ενσωμάτωσης της εισόδου, μια τεχνική γνωστή ως δέσμευση βαρών (weight tying).

Επειδή η σειρά των λέξεων είναι κρίσιμη για τη σημασία, στα διανύσματα προστίθεται και πληροφορία θέσης μέσω των λεγόμενων κωδικοποιήσεων θέσης (positional encodings). Χωρίς αυτές, το μοντέλο θα αντιμετώπιζε μια πρόταση ως απλό σύνολο λέξεων, χωρίς να γνωρίζει τη σειρά τους. Συνολικά, τα διανύσματα λέξεων αποτελούν τη «γλώσσα εισόδου» κάθε LLM: είναι το σημείο όπου το ανθρώπινο κείμενο μετατρέπεται σε μια μορφή που το νευρωνικό δίκτυο μπορεί να επεξεργαστεί.



Εικόνα 2: Διανυσματικός χώρος ενσωματώσεων λέξεων (σχηματική απεικόνιση)

3.2 Παράμετροι (Parameters)

Ο όρος «παράμετροι» αναφέρεται στους εσωτερικούς αριθμούς ενός νευρωνικού δικτύου που καθορίζουν τη συμπεριφορά του. Τεχνικά, πρόκειται κυρίως για τα βάρη (weights) και τις πολώσεις (biases) που συνδέουν τους τεχνητούς νευρώνες μεταξύ τους. Κάθε παράμετρος είναι ένας πραγματικός αριθμός που πολλαπλασιάζεται ή προστίθεται κατά τη ροή της πληροφορίας μέσα στο δίκτυο, διαμορφώνοντας έτσι τον τρόπο με τον οποίο το μοντέλο μετασχηματίζει την είσοδο σε έξοδο [8].

Μια χρήσιμη διαίσθηση είναι να θεωρήσει κανείς τις παραμέτρους ως εκατομμύρια ή δισεκατομμύρια μικροσκοπικά «κουμπιά ρύθμισης». Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος ρυθμίζει σταδιακά κάθε ένα από αυτά τα κουμπιά, ώστε το μοντέλο να κάνει όλο και καλύτερες προβλέψεις. Η «γνώση» ενός LLM δεν αποθηκεύεται, επομένως, σε κάποια ρητή βάση δεδομένων, αλλά είναι κατανεμημένη υπόρρητα μέσα στις τιμές αυτών των παραμέτρων.

Το πλήθος των παραμέτρων αποτελεί έναν από τους βασικότερους δείκτες του μεγέθους ενός μοντέλου. Η εξέλιξη είναι χαρακτηριστική: το GPT-2 διέθετε περίπου 1,5 δισεκατομμύρια

παραμέτρους, ενώ το GPT-3 έφθασε τα 175 δισεκατομμύρια [9]. Δεδομένου ότι κάθε παράμετρος καταλαμβάνει χώρο στη μνήμη, τα πολύ μεγάλα μοντέλα απαιτούν δεκάδες ή εκατοντάδες gigabytes μνήμης μόνο για να αποθηκευτούν, γεγονός που συνδέει άμεσα το πλήθος των παραμέτρων με τις απαιτήσεις σε υπολογιστική υποδομή.

Οι παράμετροι ενός μοντέλου δεν είναι συγκεντρωμένες σε ένα σημείο, αλλά κατανέμονται σε διαφορετικά τμήματα της αρχιτεκτονικής: στους πίνακες προβολής του μηχανισμού προσοχής, στα δίκτυα εμπρόσθιας τροφοδότησης κάθε στρώματος και στον πίνακα ενσωμάτωσης. Στα τυπικά μοντέλα, το μεγαλύτερο μέρος των παραμέτρων εντοπίζεται στα δίκτυα εμπρόσθιας τροφοδότησης. Η κατανόηση αυτής της κατανομής είναι χρήσιμη, καθώς εξηγεί γιατί τεχνικές συμπίεσης ή βελτιστοποίησης στοχεύουν συχνά συγκεκριμένα τμήματα του δικτύου.

Γενικά, περισσότερες παράμετροι προσφέρουν στο μοντέλο μεγαλύτερη ικανότητα (capacity) να αποτυπώνει σύνθετα μοτίβα, αλλά με κόστος σε υπολογιστικούς πόρους τόσο κατά την εκπαίδευση όσο και κατά τη χρήση. Η σχέση αυτή μεταξύ μεγέθους και απόδοσης δεν είναι τυχαία, αλλά διέπεται από εμπειρικές κανονικότητες γνωστές ως νόμοι κλιμάκωσης (scaling laws) [10], οι οποίες εξετάζονται αναλυτικότερα σε επόμενη ενότητα.

Πρακτικής σημασίας είναι και ο τρόπος αποθήκευσης των παραμέτρων. Αντί για πλήρη ακρίβεια 32 bits, τα σύγχρονα μοντέλα χρησιμοποιούν συχνά μειωμένη ακρίβεια (π.χ. 16 bits) ή ακόμη και τεχνικές κβαντισμού (quantization) σε 8 ή και λιγότερα bits. Οι τεχνικές αυτές μειώνουν δραστικά τις απαιτήσεις μνήμης και επιταχύνουν τους υπολογισμούς, με μικρή συνήθως απώλεια ακρίβειας, καθιστώντας εφικτή τη χρήση μεγάλων μοντέλων σε λιγότερο ισχυρό υλικό.

Ένα απλό παράδειγμα καθιστά σαφή τη σχέση μεγέθους και μνήμης: ένα μοντέλο 175 δισεκατομμυρίων παραμέτρων, αποθηκευμένο με ακρίβεια 16 bits (2 bytes ανά παράμετρο), απαιτεί από μόνο του περίπου 350 gigabytes μνήμης — ποσότητα που υπερβαίνει κατά πολύ τη χωρητικότητα μίας μεμονωμένης κάρτας γραφικών. Το παράδειγμα αυτό αναδεικνύει γιατί η ανάπτυξη και η λειτουργία των μεγαλύτερων μοντέλων προϋποθέτει εξειδικευμένες υποδομές με πολλαπλές, διασυνδεδεμένες κάρτες.

Είναι σημαντικό να διακρίνονται οι παράμετροι από τις υπερπαραμέτρους (hyperparameters). Οι πρώτες μαθαίνονται αυτόματα από τα δεδομένα κατά την εκπαίδευση, ενώ οι δεύτερες — όπως ο αριθμός των στρωμάτων, το μέγεθος των διανυσμάτων ή ο ρυθμός μάθησης — ορίζονται εκ των προτέρων από τους σχεδιαστές του μοντέλου και καθορίζουν το πλαίσιο μέσα στο οποίο πραγματοποιείται η μάθηση.

3.3 Η Αρχιτεκτονική Transformer

Η αρχιτεκτονική Transformer, που παρουσιάστηκε το 2017 στην ιστορική εργασία των Vaswani και συνεργατών με τίτλο «Attention Is All You Need» [11], αποτελεί τη βάση όλων των σύγχρονων LLMs. Πριν από αυτήν, οι κυρίαρχες αρχιτεκτονικές για την επεξεργασία γλώσσας ήταν τα αναδρομικά νευρωνικά δίκτυα (RNNs) και οι παραλλαγές τους, όπως τα LSTM, τα οποία επεξεργάζονταν το κείμενο σειριακά, λέξη προς λέξη. Η Transformer εγκατέλειψε αυτή τη σειριακή λογική και εισήγαγε έναν μηχανισμό που επιτρέπει την ταυτόχρονη εξέταση όλων των λέξεων μιας ακολουθίας.

Σε υψηλό επίπεδο, μια Transformer αποτελείται από μια στοίβα πανομοιότυπων δομικών μονάδων (blocks ή layers), η μία πάνω στην άλλη. Κάθε μονάδα δέχεται ως είσοδο τις αναπαραστάσεις των λέξεων, τις επεξεργάζεται και παράγει εμπλουτισμένες αναπαραστάσεις που τροφοδοτούν την επόμενη μονάδα. Με αυτόν τον τρόπο, καθώς η πληροφορία ρέει προς τα πάνω, το μοντέλο χτίζει σταδιακά ολοένα πιο αφηρημένες και σύνθετες αναπαραστάσεις του κειμένου.

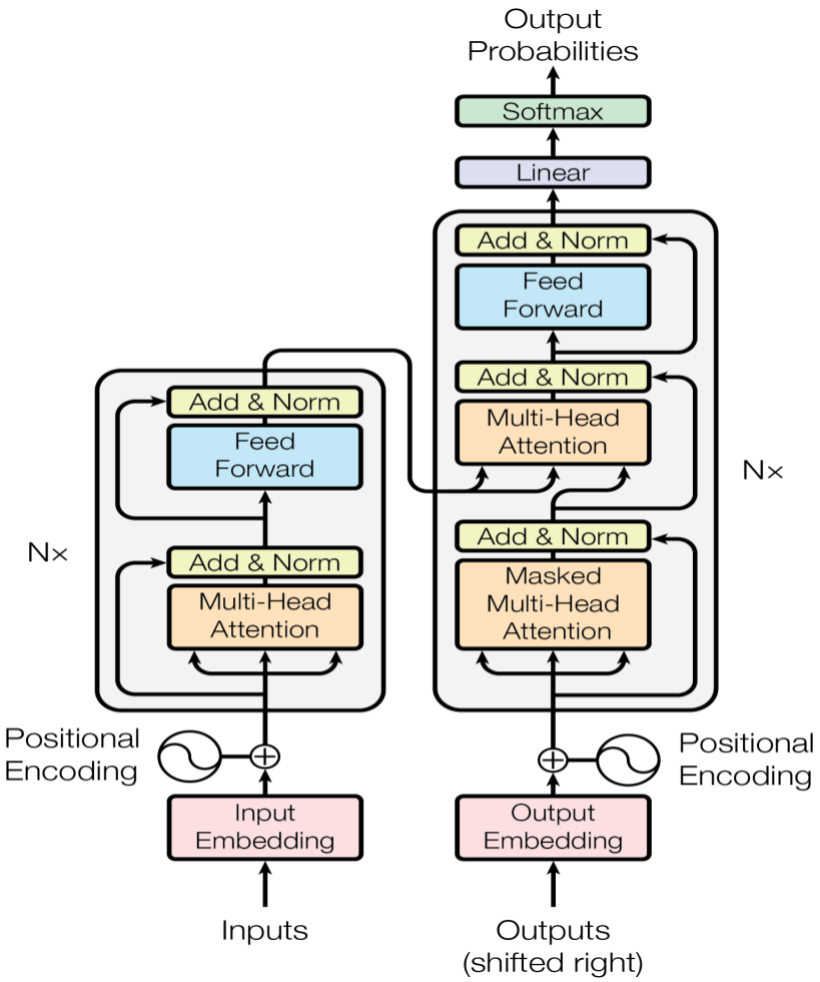
Κάθε δομική μονάδα περιλαμβάνει δύο κύρια υποστρώματα: ένα υπόστρωμα αυτο-προσοχής (self-attention), που εξετάζεται αναλυτικά στην επόμενη ενότητα, και ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης (feed-forward network). Τα δύο αυτά υποστρώματα συνοδεύονται από δύο τεχνικές που είναι κρίσιμες για τη σταθερότητα της εκπαίδευσης βαθιών δικτύων: τις παρακαμπτήριες συνδέσεις (residual connections), που επιτρέπουν στην πληροφορία να «προσπερνά» στρώματα, και την κανονικοποίηση στρώματος (layer normalization), που σταθεροποιεί τις αριθμητικές τιμές.

Το υπόστρωμα εμπρόσθιας τροφοδότησης, αν και συχνά υποτιμάται έναντι του μηχανισμού προσοχής, επιτελεί καθοριστικό ρόλο. Εφαρμόζει στην αναπαράσταση κάθε λέξης έναν μη γραμμικό μετασχηματισμό, μέσω μιας συνάρτησης ενεργοποίησης, επιτρέποντας στο δίκτυο να μοντελοποιεί σύνθετες, μη γραμμικές σχέσεις που δεν θα μπορούσαν να αποτυπωθούν με απλούς γραμμικούς συνδυασμούς. Αρκετές πρόσφατες μελέτες υποστηρίζουν ότι σημαντικό μέρος της «αποθηκευμένης γνώσης» ενός μοντέλου εδράζεται ακριβώς σε αυτά τα υποστρώματα.

Ανάλογα με τον σκοπό τους, τα μοντέλα τύπου Transformer διακρίνονται σε τρεις βασικές κατηγορίες. Τα μοντέλα μόνο-κωδικοποιητή (encoder-only), όπως το BERT [12], είναι κατάλληλα για εργασίες κατανόησης, όπως η ταξινόμηση κειμένου. Τα μοντέλα μόνο-αποκωδικοποιητή (decoder-only), όπως η οικογένεια GPT, εξειδικεύονται στην παραγωγή κειμένου, προβλέποντας την επόμενη λέξη. Τέλος, τα μοντέλα κωδικοποιητή-αποκωδικοποιητή (encoder-decoder) είναι κατάλληλα για εργασίες όπως η μετάφραση. Τα περισσότερα σύγχρονα παραγωγικά LLMs ανήκουν στην κατηγορία decoder-only.

Ένα κρίσιμο χαρακτηριστικό κάθε μοντέλου Transformer είναι το παράθυρο συμφραζομένων (context window), δηλαδή το μέγιστο πλήθος tokens που μπορεί να επεξεργαστεί ταυτόχρονα. Το μέγεθος αυτό καθορίζει πόσο εκτενές κείμενο μπορεί να «δει» το μοντέλο σε μια δεδομένη στιγμή και επηρεάζει άμεσα την ικανότητά του να διαχειρίζεται μεγάλα έγγραφα ή εκτενείς συνομιλίες. Η επέκταση του παραθύρου συμφραζομένων αποτελεί ενεργό πεδίο έρευνας, καθώς συνδέεται με σημαντικά αυξημένο υπολογιστικό κόστος.

Η σημασία της αρχιτεκτονικής Transformer δεν έγκειται μόνο στην αποτελεσματικότητά της, αλλά και στην κλιμακωσιμότητά της. Επειδή επεξεργάζεται τις λέξεις παράλληλα και όχι σειριακά, επιτρέπει την αποδοτική εκπαίδευση πάνω σε τεράστια σύνολα δεδομένων, αξιοποιώντας πλήρως τη σύγχρονη υπολογιστική υποδομή — μια ιδιότητα που, όπως θα φανεί, υπήρξε καθοριστική για την έκρηξη των δυνατοτήτων των LLMs.



Εικόνα 3: Αρχιτεκτονική Transformer

3.4 Ο Μηχανισμός Προσοχής (Self-Attention)

Ο μηχανισμός της αυτο-προσοχής (self-attention) αποτελεί τον πυρήνα της αρχιτεκτονικής Transformer και την πιο καθοριστική της καινοτομία. Η βασική ιδέα είναι ότι, για να κατανοήσει τη σημασία μιας λέξης, το μοντέλο πρέπει να λάβει υπόψη τις υπόλοιπες λέξεις της πρότασης, αποδίδοντας σε καθεμία διαφορετικό βαθμό σημασίας. Με άλλα λόγια, ο μηχανισμός επιτρέπει σε κάθε λέξη να «προσέχει» επιλεκτικά τις λέξεις εκείνες που είναι πιο σχετικές για την ερμηνεία της.

Τεχνικά, για κάθε λέξη υπολογίζονται τρία διανύσματα: ένα διάνυσμα ερωτήματος (query), ένα διάνυσμα κλειδιού (key) και ένα διάνυσμα τιμής (value). Η συνάφεια μεταξύ δύο λέξεων εκτιμάται μέσω του εσωτερικού γινομένου του ερωτήματος της μίας με το κλειδί της άλλης·

όσο μεγαλύτερο το γινόμενο, τόσο ισχυρότερη η σχέση. Οι τιμές αυτές κανονικοποιούνται μέσω της συνάρτησης softmax ώστε να μετατραπούν σε βάρη προσοχής που αθροίζουν στη μονάδα, και στη συνέχεια χρησιμοποιούνται για τον υπολογισμό ενός σταθμισμένου αθροίσματος των διανυσμάτων τιμής.

Για λόγους αριθμητικής σταθερότητας, το εσωτερικό γινόμενο διαιρείται με έναν παράγοντα κλίμακας πριν την εφαρμογή της softmax· εξ ου και η ονομασία «κλιμακωμένη προσοχή εσωτερικού γινομένου» (scaled dot-product attention). Το αποτέλεσμα είναι ότι κάθε λέξη αποκτά μια νέα αναπαράσταση που ενσωματώνει, με σταθμισμένο τρόπο, την πληροφορία από όλες τις σχετικές λέξεις του συμφραζομένου.

Τυπικά, για έναν πίνακα ερωτημάτων Q, έναν πίνακα κλειδιών K και έναν πίνακα τιμών V, ο μηχανισμός της προσοχής ορίζεται από την ακόλουθη σχέση:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Στη σχέση αυτή, το γινόμενο QK^T υπολογίζει τη συνάφεια κάθε ερωτήματος με κάθε κλειδί· ο παράγοντας √(d_k), όπου d_k είναι η διάσταση των διανυσμάτων κλειδιού, λειτουργεί ως συντελεστής κανονικοποίησης για αριθμητική σταθερότητα· η συνάρτηση softmax μετατρέπει τις τιμές συνάφειας σε βάρη που αθροίζουν στη μονάδα· και το τελικό γινόμενο με τον πίνακα V παράγει έναν σταθμισμένο συνδυασμό των διανυσμάτων τιμής. Η ίδια πράξη εκτελείται παράλληλα για όλες τις λέξεις της ακολουθίας.

Στην πράξη, τα μοντέλα δεν χρησιμοποιούν έναν μόνο μηχανισμό προσοχής, αλλά πολλούς παράλληλους, σε μια διάταξη που ονομάζεται προσοχή πολλαπλών κεφαλών (multi-head attention). Κάθε «κεφαλή» μαθαίνει να εστιάζει σε διαφορετικό είδος σχέσεων — για παράδειγμα, η μία μπορεί να αποτυπώνει συντακτικές εξαρτήσεις και η άλλη σημασιολογικές συσχετίσεις. Ο συνδυασμός πολλαπλών κεφαλών εμπλουτίζει σημαντικά την εκφραστική ικανότητα του μοντέλου.

Μια απλουστευμένη διαίσθηση για τα τρία διανύσματα έχει ως εξής: το ερώτημα (query) εκφράζει «τι αναζητά» η τρέχουσα λέξη, το κλειδί (key) εκφράζει «τι προσφέρει» κάθε άλλη λέξη, και η τιμή (value) είναι η πληροφορία που τελικά μεταφέρεται. Όταν το ερώτημα μιας λέξης ταιριάζει με το κλειδί μιας άλλης, η δεύτερη συνεισφέρει περισσότερο στη νέα αναπαράσταση της πρώτης. Έτσι, το μοντέλο μαθαίνει δυναμικά ποιες λέξεις είναι αμοιβαία

σχετικές σε κάθε συγκεκριμένο συμφραζόμενο, χωρίς αυτές οι σχέσεις να έχουν οριστεί εκ των προτέρων.

Η αξία του μηχανισμού γίνεται σαφής με ένα απλό παράδειγμα. Στην πρόταση «περπάτησε κατά μήκος της όχθης του ποταμού», η λέξη που στα αγγλικά αποδίδεται ως «bank» πρέπει να συσχετιστεί ισχυρά με τη λέξη «ποταμός». Αντιθέτως, στην πρόταση «πήγε στην τράπεζα για να καταθέσει χρήματα», η ίδια λέξη πρέπει να συνδεθεί με τη «κατάθεση». Ο μηχανισμός αυτο-προσοχής μαθαίνει αυτόματα να αποδίδει τα κατάλληλα βάρη σε κάθε περίπτωση, επιτυγχάνοντας έτσι την αποσαφήνιση της σημασίας με βάση τα συμφραζόμενα.

Στα μοντέλα παραγωγής κειμένου (decoder-only), ο μηχανισμός εφαρμόζεται σε μια ειδική, «καλυμμένη» μορφή (masked self-attention): κάθε λέξη επιτρέπεται να προσέχει μόνο τις προηγούμενες λέξεις και όχι τις επόμενες. Αυτός ο περιορισμός είναι απαραίτητος, ώστε το μοντέλο να μαθαίνει να προβλέπει την επόμενη λέξη χωρίς να «βλέπει» εκ των προτέρων την απάντηση που καλείται να παραγάγει.

Ο μηχανισμός της αυτο-προσοχής έχει, ωστόσο, ένα σημαντικό υπολογιστικό κόστος: ο αριθμός των απαιτούμενων πράξεων αυξάνεται τετραγωνικά ως προς το μήκος της ακολουθίας, καθώς κάθε token πρέπει να συσχετιστεί με κάθε άλλο. Η τετραγωνική αυτή πολυπλοκότητα εξηγεί γιατί η επεξεργασία πολύ μεγάλων κειμένων είναι ιδιαίτερα απαιτητική και αποτελεί έναν από τους βασικούς λόγους για τους οποίους το παράθυρο συμφραζομένων παραμένει περιορισμένο. Σημαντικό μέρος της σύγχρονης έρευνας αφιερώνεται στην ανάπτυξη πιο αποδοτικών παραλλαγών του μηχανισμού προσοχής.

3.5 Παράλληλη Επεξεργασία (Parallel Processing)

Ένα από τα σημαντικότερα πλεονεκτήματα της αρχιτεκτονικής Transformer έναντι των προγενέστερων αναδρομικών δικτύων είναι η δυνατότητα παράλληλης επεξεργασίας. Τα RNNs επεξεργάζονταν το κείμενο σειριακά, καθώς ο υπολογισμός για κάθε λέξη εξαρτιόταν από το αποτέλεσμα της προηγούμενης. Αυτή η εγγενώς ακολουθιακή φύση αποτελούσε σοβαρό υπολογιστικό περιορισμό, καθιστώντας την εκπαίδευση σε μεγάλα σύνολα δεδομένων εξαιρετικά αργή.

Αντιθέτως, ο μηχανισμός της αυτο-προσοχής επιτρέπει την ταυτόχρονη επεξεργασία όλων των λέξεων μιας ακολουθίας. Επειδή οι υπολογισμοί που απαιτούνται μπορούν να εκφραστούν ως πράξεις πινάκων (matrix multiplications), είναι δυνατόν να εκτελεστούν παράλληλα από εξειδικευμένο υλικό. Η ικανότητα αυτή αξιοποιεί πλήρως τις σύγχρονες μονάδες επεξεργασίας, οι οποίες έχουν σχεδιαστεί ακριβώς για τη μαζική παράλληλη εκτέλεση τέτοιων πράξεων.

Η παραλληλία ενισχύεται περαιτέρω μέσω της ομαδοποίησης (batching), όπου πολλαπλά παραδείγματα κειμένου επεξεργάζονται ταυτόχρονα. Ο συνδυασμός παράλληλης επεξεργασίας λέξεων και ομαδοποίησης παραδειγμάτων αύξησε δραματικά την ταχύτητα εκπαίδευσης, καθιστώντας εφικτή την αξιοποίηση συνόλων δεδομένων με εκατοντάδες δισεκατομμύρια λέξεις σε εύλογο χρόνο. Χωρίς αυτή τη δυνατότητα, η εκπαίδευση μοντέλων της κλίμακας των σημερινών LLMs θα ήταν πρακτικά ανέφικτη.

Η δυνατότητα παράλληλης επεξεργασίας δεν αποτελεί απλώς μια τεχνική λεπτομέρεια, αλλά έναν από τους βασικότερους λόγους για τους οποίους τα LLMs γνώρισαν τόσο ραγδαία εξέλιξη μετά το 2017. Καθιστώντας εφικτή την πλήρη αξιοποίηση του εξειδικευμένου υλικού, επέτρεψε στους ερευνητές να εκπαιδεύσουν μοντέλα πρωτοφανούς κλίμακας μέσα σε ρεαλιστικά χρονικά πλαίσια, μετατρέποντας τη θεωρητική δυνατότητα της κλιμάκωσης σε πρακτική πραγματικότητα.

Αξίζει, ωστόσο, να σημειωθεί μια λεπτή διάκριση. Ενώ η εκπαίδευση και η επεξεργασία ενός δοθέντος κειμένου είναι σε μεγάλο βαθμό παράλληλες, η παραγωγή νέου κειμένου παραμένει εν μέρει σειριακή: το μοντέλο παράγει μία λέξη τη φορά, και κάθε νέα λέξη εξαρτάται από τις προηγούμενες. Η ασυμμετρία αυτή μεταξύ της μαζικά παράλληλης εκπαίδευσης και της ακολουθιακής παραγωγής εξηγεί γιατί η δημιουργία μεγάλων κειμένων μπορεί να είναι υπολογιστικά απαιτητική ακόμη και σε ένα ήδη εκπαιδευμένο μοντέλο.

3.6 Κάρτες Γραφικών (GPUs) και Υπολογιστική Υποδομή

Η παράλληλη φύση των υπολογισμών των Transformer θα είχε περιορισμένη πρακτική αξία χωρίς το κατάλληλο υλικό για την εκτέλεσή τους. Τον ρόλο αυτό επιτελούν κατά κύριο λόγο οι κάρτες γραφικών (Graphics Processing Units, GPUs). Σε αντίθεση με τις κεντρικές μονάδες

επεξεργασίας (CPUs), που διαθέτουν λίγους αλλά ισχυρούς πυρήνες βελτιστοποιημένους για σειριακές εργασίες, οι GPUs διαθέτουν χιλιάδες απλούστερους πυρήνες, ικανούς να εκτελούν ταυτόχρονα έναν τεράστιο αριθμό απλών πράξεων.

Η αρχιτεκτονική αυτή ταιριάζει ιδανικά με τις απαιτήσεις των νευρωνικών δικτύων, οι υπολογισμοί των οποίων συνίστανται κατά βάση σε πολλαπλασιασμούς πινάκων. Σύγχρονες κάρτες διαθέτουν εξειδικευμένες μονάδες, γνωστές ως πυρήνες τανυστών (tensor cores), σχεδιασμένες αποκλειστικά για την επιτάχυνση τέτοιων πράξεων. Πέρα από τις GPUs, ορισμένες εταιρείες αναπτύσσουν και ειδικά ολοκληρωμένα κυκλώματα, όπως οι Tensor Processing Units (TPUs), για τον ίδιο σκοπό.

Μια χρήσιμη αναλογία είναι η εξής: μια CPU μοιάζει με λίγους έμπειρους ειδικούς που εκτελούν διαδοχικά σύνθετες εργασίες, ενώ μια GPU μοιάζει με χιλιάδες εργάτες που εκτελούν ταυτόχρονα ο καθένας μια απλή εργασία. Για τις επαναλαμβανόμενες και ομοιόμορφες πράξεις που κυριαρχούν στα νευρωνικά δίκτυα, η δεύτερη προσέγγιση είναι ασύγκριτα αποδοτικότερη· ακριβώς αυτή η αντιστοιχία εξηγεί τον κεντρικό ρόλο των GPUs στην ανάπτυξη των LLMs.

Ένας κρίσιμος περιορισμός είναι η διαθέσιμη μνήμη της κάρτας (VRAM). Επειδή τα μεγάλα μοντέλα διαθέτουν δισεκατομμύρια παραμέτρους, συχνά δεν χωρούν σε μία μόνο κάρτα. Για τον λόγο αυτό, η εκπαίδευσή τους κατανέμεται σε δεκάδες, εκατοντάδες ή και χιλιάδες κάρτες που λειτουργούν συντονισμένα, μέσω τεχνικών κατανεμημένης εκπαίδευσης. Οι τεχνικές αυτές περιλαμβάνουν τόσο τον επιμερισμό των δεδομένων (data parallelism) όσο και τον επιμερισμό του ίδιου του μοντέλου (model parallelism) σε πολλαπλές συσκευές.

Όταν η εκπαίδευση κατανέμεται σε πολλές κάρτες, η ταχύτητα της μεταξύ τους επικοινωνίας γίνεται καθοριστικός παράγοντας. Για τον λόγο αυτό, οι σύγχρονες υπολογιστικές συστοιχίες αξιοποιούν διασυνδέσεις πολύ υψηλής ταχύτητας, ώστε οι κάρτες να ανταλλάσσουν αποδοτικά τις ενδιάμεσες τιμές και τις κλίσεις (gradients) που απαιτούνται για τη συντονισμένη ενημέρωση των παραμέτρων. Χωρίς τέτοιες διασυνδέσεις, η επιβάρυνση της επικοινωνίας θα ακύρωνε σε μεγάλο βαθμό τα οφέλη της παραλληλίας.

Η εξάρτηση αυτή από εκτεταμένη υπολογιστική υποδομή έχει σημαντικές συνέπειες ως προς το κόστος και την ενεργειακή κατανάλωση. Εκτιμάται ότι η εκπαίδευση ενός μοντέλου της κλίμακας του GPT-3 απαιτούσε ενέργεια της τάξης άνω των 1.000 MWh, με αντίστοιχο περιβαλλοντικό αποτύπωμα [13]. Πέραν του οικονομικού και περιβαλλοντικού κόστους, η ανάγκη για ακριβή και σπάνια υλικό εγείρει και ζητήματα προσβασιμότητας, καθώς περιορίζει

τη δυνατότητα ανάπτυξης τέτοιων μοντέλων σε λίγους μεγάλους οργανισμούς με επαρκείς πόρους.

Αξίζει να σημειωθεί ότι οι GPUs δεν είναι απαραίτητες μόνο κατά την εκπαίδευση, αλλά και κατά τη φάση της χρήσης (inference), όταν δηλαδή ένα εκπαιδευμένο μοντέλο εξυπηρετεί αιτήματα χρηστών. Η εξυπηρέτηση ενός μεγάλου μοντέλου σε πραγματικό χρόνο, και μάλιστα σε πολλούς χρήστες ταυτόχρονα, απαιτεί επίσης σημαντικούς υπολογιστικούς πόρους. Το γεγονός αυτό εξηγεί γιατί το λειτουργικό κόστος των υπηρεσιών που βασίζονται σε LLMs παραμένει υψηλό, ακόμη και μετά την ολοκλήρωση της δαπανηρής φάσης της εκπαίδευσης.

3.7 Προ-εκπαίδευση (Pre-training)

Η προ-εκπαίδευση (pre-training) αποτελεί την πρώτη και πιο δαπανηρή φάση δημιουργίας ενός LLM. Κατά τη φάση αυτή, στο μοντέλο παρέχεται ένα τεράστιο σώμα κειμένου — που περιλαμβάνει ιστοσελίδες, βιβλία, άρθρα και άλλες πηγές — και εκπαιδεύεται σε μια φαινομενικά απλή εργασία: την πρόβλεψη της επόμενης λέξης (ή, ακριβέστερα, του επόμενου token) με βάση τις προηγούμενες. Η εργασία αυτή είναι γνωστή ως πρόβλεψη επόμενου token (next-token prediction).

Το ιδιαίτερο χαρακτηριστικό αυτής της προσέγγισης είναι ότι δεν απαιτεί ανθρώπινη επισημείωση των δεδομένων· για τον λόγο αυτό χαρακτηρίζεται ως αυτο-επιβλεπόμενη μάθηση (self-supervised learning). Το ίδιο το κείμενο παρέχει τη «σωστή απάντηση», αφού η επόμενη λέξη κάθε ακολουθίας είναι ήδη γνωστή. Αυτή η ιδιότητα είναι καθοριστική, διότι επιτρέπει την αξιοποίηση πρακτικά απεριόριστων ποσοτήτων μη επισημειωμένου κειμένου, κάτι που θα ήταν αδύνατο με μεθόδους που απαιτούν χειροκίνητη επισήμανση.

Η ποιότητα και η σύνθεση των δεδομένων προ-εκπαίδευσης είναι εξίσου καθοριστικές με τον όγκο τους. Τα σύγχρονα μοντέλα εκπαιδεύονται σε σώματα κειμένου που περιλαμβάνουν εκατοντάδες δισεκατομμύρια ή και τρισεκατομμύρια tokens, τα οποία υποβάλλονται σε εκτενή επεξεργασία: αφαίρεση διπλότυπων, φιλτράρισμα περιεχομένου χαμηλής ποιότητας ή ακατάλληλου, και εξισορρόπηση των πηγών. Η διαδικασία αυτή είναι κρίσιμη, διότι τυχόν προκαταλήψεις, ανακρίβειες ή κενά στα δεδομένα μεταφέρονται αναπόφευκτα στο τελικό μοντέλο.

Σε κάθε βήμα, το μοντέλο παράγει μια πρόβλεψη, η οποία συγκρίνεται με την πραγματική επόμενη λέξη. Η απόκλιση μεταξύ πρόβλεψης και πραγματικότητας ποσοτικοποιείται μέσω μιας συνάρτησης σφάλματος (loss). Το σφάλμα αυτό αξιοποιείται από τον αλγόριθμο της ανάστροφης μετάδοσης (backpropagation) και της καθοδικής κλίσης (gradient descent), προκειμένου να ρυθμιστούν ελαφρώς οι παράμετροι του δικτύου προς την κατεύθυνση που μειώνει το σφάλμα. Επαναλαμβάνοντας τη διαδικασία αυτή δισεκατομμύρια φορές, το μοντέλο βελτιώνει σταδιακά τις προβλέψεις του.

Σε όλη τη διάρκεια της εκπαίδευσης, η τιμή της συνάρτησης σφάλματος παρακολουθείται προσεκτικά: η σταδιακή της μείωση υποδεικνύει ότι το μοντέλο μαθαίνει, ενώ η μορφή της καμπύλης σφάλματος προσφέρει ενδείξεις για την πρόοδο και τη σταθερότητα της διαδικασίας. Είναι αξιοσημείωτο ότι ορισμένες ικανότητες δεν εμφανίζονται σταδιακά, αλλά αναδύονται σχετικά απότομα μόλις το μοντέλο ξεπεράσει ένα ορισμένο κατώφλι μεγέθους ή όγκου εκπαίδευσης — ένα φαινόμενο που έχει προσελκύσει ιδιαίτερο ερευνητικό ενδιαφέρον.

Δύο είναι οι κυρίαρχες παραλλαγές προ-εκπαίδευσης. Στην αυτοπαλίνδρομη (autoregressive) προσέγγιση, που χρησιμοποιείται από τα μοντέλα τύπου GPT [9], το μοντέλο προβλέπει την επόμενη λέξη με βάση μόνο τις προηγούμενες. Στην προσέγγιση καλυμμένης γλωσσικής μοντελοποίησης (masked language modeling), που εισήγαγε το BERT [12], ορισμένες λέξεις της πρότασης αποκρύπτονται και το μοντέλο καλείται να τις ανακτήσει αξιοποιώντας τόσο τα προηγούμενα όσο και τα επόμενα συμφραζόμενα.

Ένα από τα σημαντικότερα εμπειρικά ευρήματα της προ-εκπαίδευσης είναι οι νόμοι κλιμάκωσης (scaling laws) [10]: η απόδοση του μοντέλου βελτιώνεται προβλέψιμα καθώς αυξάνονται το μέγεθός του, ο όγκος των δεδομένων και η διαθέσιμη υπολογιστική ισχύς. Ακόμη πιο εντυπωσιακό είναι ότι, πέρα από ένα ορισμένο μέγεθος, εμφανίζονται αναδυόμενες ικανότητες, όπως η μάθηση εντός συμφραζομένων (in-context learning), δηλαδή η ικανότητα του μοντέλου να προσαρμόζεται σε νέες εργασίες με βάση λίγα μόνο παραδείγματα στο prompt, χωρίς περαιτέρω εκπαίδευση [9].

Μεταγενέστερες μελέτες εμπλούτισαν την κατανόηση των νόμων κλιμάκωσης, αναδεικνύοντας ότι η βέλτιστη αξιοποίηση ενός δεδομένου υπολογιστικού προϋπολογισμού δεν επιτυγχάνεται απαραίτητα με τη μεγιστοποίηση μόνο του αριθμού των παραμέτρων, αλλά με την ισορροπημένη αύξηση τόσο του μεγέθους του μοντέλου όσο και του όγκου των δεδομένων εκπαίδευσης. Το εύρημα αυτό μετατόπισε εν μέρει την έμφαση από τα απλώς μεγαλύτερα μοντέλα προς μοντέλα εκπαιδευμένα σε περισσότερα και ποιοτικότερα δεδομένα.

Η φάση της προ-εκπαίδευσης διαρκεί τυπικά εβδομάδες ή και μήνες, απαιτώντας χιλιάδες GPUs που λειτουργούν αδιάλειπτα. Το αποτέλεσμα είναι ένα μοντέλο βάσης (base model) με ευρεία γλωσσική ικανότητα, το οποίο όμως δεν είναι ακόμη βελτιστοποιημένο για να ακολουθεί οδηγίες ή να παράγει χρήσιμες απαντήσεις. Η μετάβαση από αυτό το γενικό μοντέλο σε ένα χρηστικό βοηθό απαιτεί μια επιπλέον φάση προσαρμογής, η οποία εξετάζεται στη συνέχεια.

3.8 Ενισχυτική Μάθηση από Ανθρώπινη Ανατροφοδότηση (RLHF)

Ένα μοντέλο που έχει υποστεί μόνο προ-εκπαίδευση μαθαίνει να παράγει στατιστικά εύλογο κείμενο, όχι όμως απαραίτητα κείμενο που είναι χρήσιμο, ασφαλές ή ευθυγραμμισμένο με τις προθέσεις του χρήστη. Για να γεφυρωθεί αυτό το χάσμα ευθυγράμμισης, αναπτύχθηκε η τεχνική της Ενισχυτικής Μάθησης από Ανθρώπινη Ανατροφοδότηση (Reinforcement Learning from Human Feedback, RLHF) [14], η οποία προσαρμόζει το μοντέλο ώστε οι απαντήσεις του να ανταποκρίνονται στις ανθρώπινες προτιμήσεις.

Η ανάγκη για ευθυγράμμιση γίνεται εύκολα κατανοητή με ένα παράδειγμα. Ένα μοντέλο που έχει υποστεί μόνο προ-εκπαίδευση, αν δεχθεί μια ερώτηση, ενδέχεται να μην απαντήσει ευθέως, αλλά να συνεχίσει το κείμενο παράγοντας, για παράδειγμα, μια λίστα παρόμοιων ερωτήσεων — διότι αυτό είναι στατιστικά εύλογο με βάση τα δεδομένα εκπαίδευσης. Η RLHF διδάσκει στο μοντέλο ότι, στο πλαίσιο μιας συνομιλίας, η επιθυμητή συμπεριφορά είναι να δοθεί μια χρήσιμη και άμεση απάντηση, ευθυγραμμίζοντας έτσι την έξοδό του με τις πραγματικές προσδοκίες του χρήστη.

Η διαδικασία RLHF περιλαμβάνει τρία διαδοχικά στάδια. Στο πρώτο στάδιο, της επιβλεπόμενης μικρορύθμισης (supervised fine-tuning), το μοντέλο εκπαιδεύεται σε ένα σύνολο υψηλής ποιότητας παραδειγμάτων, όπου άνθρωποι έχουν συντάξει τις επιθυμητές απαντήσεις σε διάφορες προτροπές. Με τον τρόπο αυτό, το μοντέλο μαθαίνει το γενικό ύφος και τη μορφή των επιθυμητών αποκρίσεων.

Στο δεύτερο στάδιο εκπαιδεύεται ένα ξεχωριστό μοντέλο ανταμοιβής (reward model). Για τον σκοπό αυτό, το βασικό μοντέλο παράγει πολλαπλές εναλλακτικές απαντήσεις στην ίδια

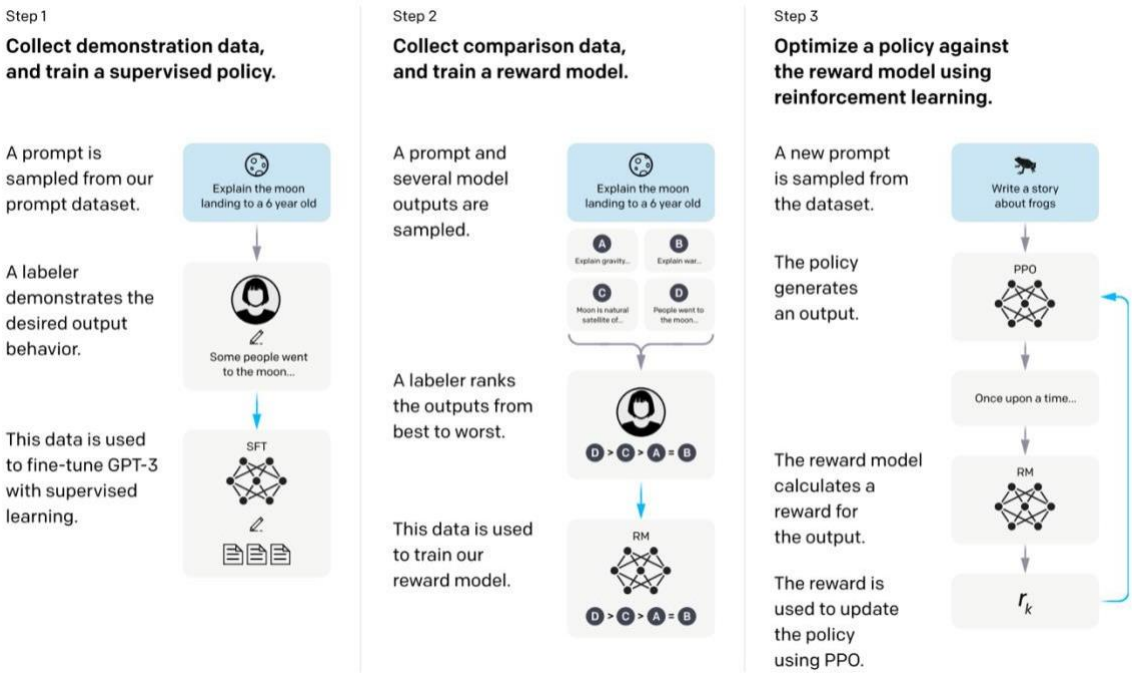
προτροπή, και ανθρώπινοι αξιολογητές τις κατατάσσουν από την καλύτερη προς τη χειρότερη. Το μοντέλο ανταμοιβής μαθαίνει, από αυτές τις συγκρίσεις, να προβλέπει ποια απάντηση θα προτιμούσε ένας άνθρωπος, αποδίδοντας σε κάθε απάντηση μια αριθμητική βαθμολογία.

Στο τρίτο στάδιο, το μοντέλο ανταμοιβής χρησιμοποιείται για την περαιτέρω εκπαίδευση του γλωσσικού μοντέλου μέσω ενός αλγορίθμου ενισχυτικής μάθησης — συνήθως του Proximal Policy Optimization (PPO). Το γλωσσικό μοντέλο «ανταμείβεται» όταν παράγει απαντήσεις που λαμβάνουν υψηλή βαθμολογία από το μοντέλο ανταμοιβής, και έτσι προσαρμόζει σταδιακά τη συμπεριφορά του προς την κατεύθυνση των ανθρώπινων προτιμήσεων.

Το αποτέλεσμα της διαδικασίας αυτής είναι μοντέλα σημαντικά πιο χρήσιμα, πιο ειλικρινή και λιγότερο επιβλαβή, ικανά να ακολουθούν οδηγίες με φυσικό τρόπο. Η RLHF ήταν, σε μεγάλο βαθμό, ο παράγοντας που μετέτρεψε τα ισχυρά αλλά «ακατέργαστα» μοντέλα βάσης σε προσιτούς συνομιλητικούς βοηθούς ευρείας χρήσης.

Παρά την αποτελεσματικότητά της, η RLHF δεν είναι απαλλαγμένη από περιορισμούς. Το μοντέλο ενδέχεται να «εκμεταλλευτεί» αδυναμίες του μοντέλου ανταμοιβής (reward hacking), παράγοντας απαντήσεις που βαθμολογούνται υψηλά χωρίς να είναι πραγματικά ποιοτικές. Επιπλέον, οι ανθρώπινες αξιολογήσεις μπορεί να ενσωματώνουν μεροληψίες, ενώ η όλη διαδικασία απαιτεί σημαντικό κόστος και ανθρώπινη εργασία. Τα ζητήματα αυτά παραμένουν ενεργά πεδία έρευνας.

Λόγω της πολυπλοκότητας και του κόστους της κλασικής RLHF, έχουν προταθεί και εναλλακτικές ή συμπληρωματικές προσεγγίσεις. Ορισμένες μέθοδοι, όπως η άμεση βελτιστοποίηση προτιμήσεων (Direct Preference Optimization), επιχειρούν να επιτύχουν παρόμοιο αποτέλεσμα χωρίς την ανάγκη ενός ξεχωριστού σταδίου ενισχυτικής μάθησης, ενώ άλλες αξιοποιούν ανατροφοδότηση που παράγεται από το ίδιο το μοντέλο με βάση ένα προκαθορισμένο σύνολο αρχών. Οι προσεγγίσεις αυτές στοχεύουν στη μείωση της εξάρτησης από εκτενή ανθρώπινη επισημείωση, διατηρώντας παράλληλα τα οφέλη της ευθυγράμμισης.



Εικόνα 4: RLHF (Reinforcement Learning from Human Feedback).

3.9 Χρήση, Προσαρμογή και Λήψη Αποφάσεων

3.9.1 Zero-Shot, Few-Shot Learning και In-Context Learning

Μόλις εκπαιδευτεί ένα LLM, μπορεί να χρησιμοποιηθεί με τρεις κύριους τρόπους. Στο zero-shot learning, δίνεται στο μοντέλο μια εργασία χωρίς κανένα παράδειγμα. Στο few-shot learning, του δίνονται λίγα παραδείγματα της εργασίας, τα οποία βελτιώνουν σημαντικά την απόδοση. Τέλος, υπάρχει το fine-tuning, όπου το μοντέλο εκπαιδεύεται περαιτέρω σε εξειδικευμένα δεδομένα.

Ένα συναφές και αρκετά ενδιαφέρον φαινόμενο ονομάζεται in-context learning: όσο μεγαλύτερο είναι το μοντέλο, τόσο καλύτερα 'μαθαίνει' από παραδείγματα που παρέχονται στο prompt χωρίς να ενημερώνονται τα βάρη του δικτύου. Αυτό σημαίνει ότι το μοντέλο με κάποιο τρόπο 'αναγνωρίζει' τα παραδείγματα και προσαρμόζει τη συμπεριφορά του ανάλογα. Ωστόσο, αυτό το φαινόμενο εξαρτάται κατά κύριο λόγο από το μέγεθος του μοντέλου — τα πολύ μικρότερα μοντέλα συχνά αποτυγχάνουν στο in-context learning και απαιτούν ρητό fine-tuning για να μάθουν νέες εργασίες.

3.9.2 Prompt Engineering: Τέχνη και Επιστήμη

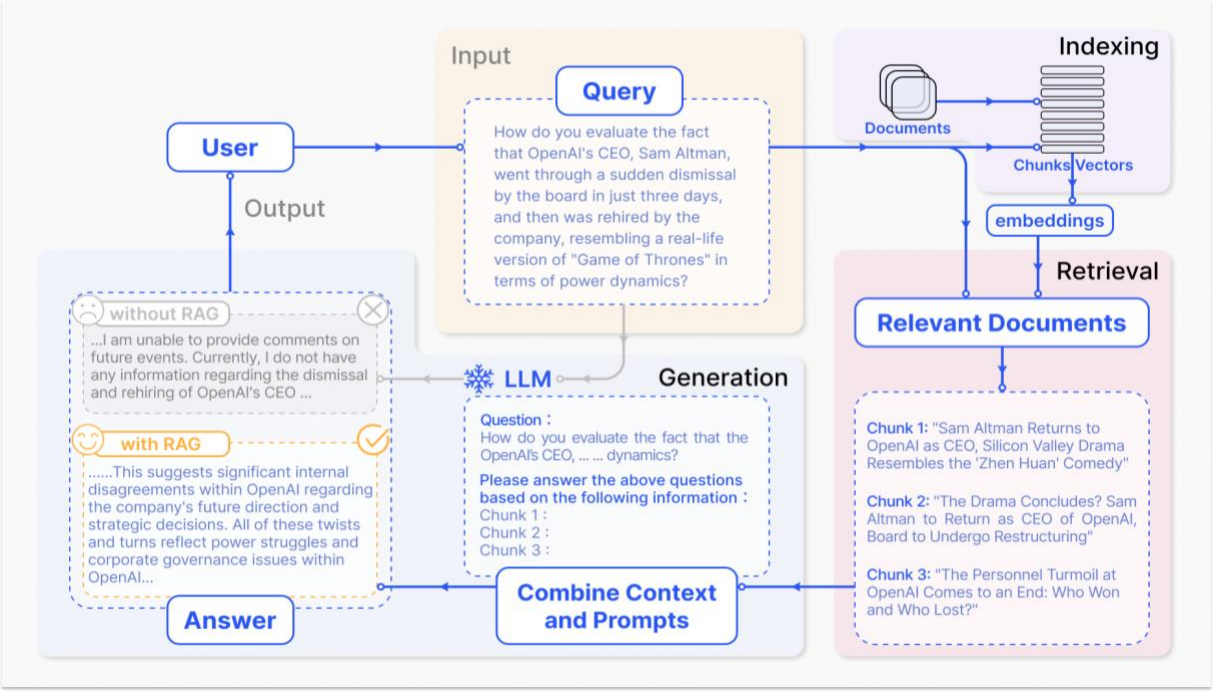
Η τέχνη του prompt engineering — το σχεδίασμα ακριβών, σαφών και αποτελεσματικών οδηγιών — αποδείχθηκε ότι είναι καίρια για την επίτευξη καλής απόδοσης από ένα LLM. Μικρές αλλαγές στη διατύπωση, τη δομή ή το περιεχόμενο ενός prompt μπορούν να έχουν σημαντικές επιπτώσεις στην ποιότητα και την αξιοπιστία της απάντησης. Παράδειγμα: 'Λύσε αυτό το μαθηματικό πρόβλημα' συχνά δίνει χαμηλότερης ποιότητας απαντήσεις από 'Λύσε αυτό το μαθηματικό πρόβλημα βήμα προς βήμα, δείχνοντας όλη τη δουλειά σου'.

Ένας από τους πιο αποτελεσματικούς τρόπους prompt engineering είναι το chain-of-thought (CoT) prompting [15]. Αντί να ζητούμε απλώς μια τελική απάντηση, ζητούμε από το μοντέλο να δείξει τη σκέψη του — δηλαδή τα ενδιάμεσα βήματα που οδηγούν στην απάντηση. Παράδειγμα CoT: 'Σκέψου βήμα-προς-βήμα: Αν υπάρχουν 3 μήλα και προσθέσω 5 ακόμα, πόσα υπάρχουν συνολικά;' Τα CoT prompts έχουν αποδειχθεί ότι βελτιώνουν σημαντικά την ικανότητα του μοντέλου να λύνει σύνθετα προβλήματα, ιδιαίτερα σε μαθηματικά και λογικό συλλογισμό. Άλλες τεχνικές περιλαμβάνουν το self-consistency (δημιουργία πολλαπλών διαδρομών συλλογισμού και λήψη της πλειοψηφικής απόφασης) και το instruction prompting (ξεκάθαρες, δομημένες οδηγίες).

3.9.3 Retrieval-Augmented Generation (RAG) και Εξωτερικές Γνώσεις

Μία άλλη αναδυόμενη και σημαντική προσέγγιση είναι η Retrieval-Augmented Generation (RAG) [16]. Σε ένα σύστημα RAG, το LLM δεν λειτουργεί μόνο βασιζόμενο στη γνώση του που έχει απορροφήσει κατά την εκπαίδευση, αλλά έχει πρόσβαση σε μια εξωτερική βάση δεδομένων ή μηχανή αναζήτησης. Πριν να δημιουργήσει μια απάντηση, το σύστημα RAG ανακτά τα πιο σχετικά κομμάτια πληροφορίας από τη βάση δεδομένων (retrieval) και τα εισάγει ως πρόσφατο πλαίσιο πριν ζητήσει από το μοντέλο να δημιουργήσει την απάντηση (generation). Αυτό είναι ιδιαίτερα χρήσιμο για εργασίες που απαιτούν πρόσφατες πληροφορίες

(π.χ. τρέχοντα γεγονότα), εξειδικευμένη γνώση από κάποιο πεδίο, ή ευαίσθητες πληροφορίες που ενδέχεται να μην περιλαμβάνονται πλήρως ή κατάλληλα στα δεδομένα εκπαίδευσης του μοντέλου. Η RAG μπορεί να βελτιώσει σημαντικά την αξιοπιστία ενός συστήματος LLM, ειδικά όταν η ακρίβεια γεγονών είναι σημαντική.



Εικόνα 5: Αρχιτεκτονική συστήματος RAG

3.10 Σύνοψη

Συνοψίζοντας, οι τεχνολογίες που εξετάστηκαν στο παρόν κεφάλαιο συνθέτουν μια συνεκτική αλυσίδα. Τα διανύσματα λέξεων μετατρέπουν τη γλώσσα σε αριθμητική μορφή· οι παράμετροι αποθηκεύουν, υπόρρητα, τη γνώση του μοντέλου· η αρχιτεκτονική Transformer και ο μηχανισμός της αυτο-προσοχής επιτρέπουν την κατανόηση των λέξεων μέσα στα συμφραζόμενά τους· η παράλληλη επεξεργασία, σε συνδυασμό με τις GPUs, καθιστά εφικτή την εκπαίδευση σε τεράστια κλίμακα· η προ-εκπαίδευση παρέχει την ευρεία γλωσσική ικανότητα· και, τέλος, η RLHF ευθυγραμμίζει το μοντέλο με τις ανθρώπινες προτιμήσεις.

Η κατανόηση των δομικών αυτών στοιχείων δεν έχει μόνο θεωρητική αξία. Όπως θα φανεί στα επόμενα κεφάλαια, πολλά από τα ισχυρά σημεία αλλά και τους περιορισμούς των LLMs — από την εξάρτηση της απόδοσης από τον όγκο των δεδομένων εκπαίδευσης έως την τάση

τους για ψευδαισθήσεις [17] — μπορούν να ερμηνευθούν ακριβώς υπό το πρίσμα του τρόπου με τον οποίο λειτουργούν οι θεμελιώδεις αυτές τεχνολογίες.

4 Αξιολόγηση και Benchmarks των LLMs

Το κεφάλαιο αυτό παρουσιάζει μια εκτενή ανάλυση των μεθοδολογιών και εργαλείων που χρησιμοποιούνται για την αξιολόγηση των LLMs. Βάσει της ανασκόπησης 43 εργασιών, εντοπίστηκε ένα εκτεταμένο σύνολο benchmarks και μετρικών που χρησιμοποιούνται σε διάφορα επιστημονικά πεδία. Η κατανόηση αυτών των εργαλείων είναι κρίσιμη για την ερμηνεία των αποτελεσμάτων απόδοσης που παρουσιάζονται στο επόμενο κεφάλαιο.

4.1 Μεθοδολογίες Αξιολόγησης

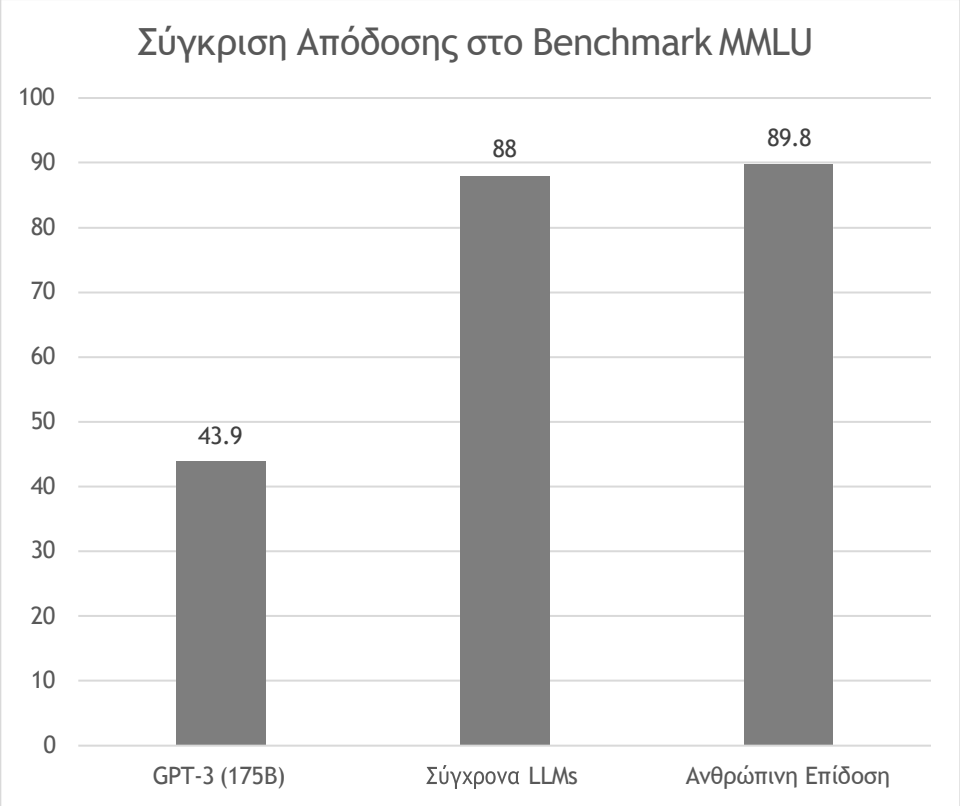
Με βάση τη βιβλιογραφία, διακρίνονται τρεις κύριες προσεγγίσεις αξιολόγησης [4]. Πρώτον, τα αυτόματα benchmarks χρησιμοποιούν τυποποιημένα σύνολα δεδομένων με γνωστές σωστές απαντήσεις, επιτρέποντας επαναλήψιμη αξιολόγηση. Δεύτερον, η ανθρώπινη αξιολόγηση (human evaluation) χρησιμοποιεί ειδικούς για την αποτίμηση της ποιότητας των απαντήσεων, μέθοδος ακριβέστερη αλλά με υψηλότερο κόστος και χρόνο. Τρίτον, η προσέγγιση LLM-as-a-judge χρησιμοποιεί ένα ισχυρό μοντέλο για την αξιολόγηση των απαντήσεων άλλων μοντέλων, μέθοδος αποδοτική αλλά με τεκμηριωμένο κίνδυνο συστηματικής μεροληψίας [18].

4.2 Βασικά Benchmarks ανά Κατηγορία

4.2.1 Γενικής Γνώσης / Πολυτομεακά (π.χ. MMLU)

Το MMLU (Massive Multitask Language Understanding, όπως ορίζεται στο paper 'Measuring Massive Multitask Language Understanding') αποτελείται από 15.908 ερωτήσεις πολλαπλής επιλογής σε 57 θέματα [19]. Η εξέλιξη της απόδοσης είναι ενδεικτική: το αρχικό GPT-3 175B σημείωσε 43,9%, ενώ η ανθρώπινη απόδοση εκτιμάται περίπου στο 89,8% [19]. Παρόμοιας λογικής, αλλά με έμφαση σε ερωτήσεις ανθρώπινων εξετάσεων (εισαγωγικές, νομικές, ιατρικές), είναι το AGIEval [20]. Μεταγενέστερα μοντέλα όπως τα Claude 3.5 Sonnet, GPT-4o και Llama 3.1 405B υπερβαίνουν το 88%, προσεγγίζοντας ή υπερβαίνοντας τη μέση

ανθρώπινη επίδοση [19], [2], [3]. Ένα ακόμη πιο απαιτητικό benchmark πολλαπλών πεδίων, σχεδιασμένο να εξετάζει τα όρια της γνώσης των LLMs, αποτελεί το Humanity's Last Exam [21]. Το BIG-Bench, με 204 εργασίες, καταδεικνύει ότι η απόδοση βελτιώνεται με την κλιμάκωση των παραμέτρων, αλλά παραμένει περιορισμένη σε απόλυτους όρους σε αρκετές κατηγορίες εργασιών [22].



Εικόνα 6: Σύγκριση απόδοσης στο MMLU

4.2.2 Επιστημών (π.χ. GPQA)

Στις φυσικές επιστήμες, το GPQA (Graduate-Level Google-Proof QA) περιλαμβάνει 448 ερωτήσεις σε βιολογία, φυσική και χημεία, σχεδιασμένες ώστε να μην απαντώνται μέσω απλής αναζήτησης [23]. Τα αποτελέσματα δείχνουν ότι οι κάτοχοι διδακτορικού σημειώνουν 65%, ενώ το GPT-4 περιορίζεται στο 39% και το Claude 3 Opus προσεγγίζει το 60% [23], [3]. Συμπληρωματικά, η αξιολόγηση SciAssess διαπιστώνει ότι τα μοντέλα παρουσιάζουν ισχυρή ανάκληση γεγονότων αλλά περιορισμένη ικανότητα σύνθεσης πληροφοριών από πολλαπλές πηγές [24].

4.2.3 Μαθηματικών (π.χ. GSM8K, MATH)

Στα μαθηματικά παρατηρείται σημαντική βελτίωση της απόδοσης με την εξέλιξη των μοντέλων. Στο benchmark MATH (12.500 προβλήματα), η απόδοση αυξήθηκε από 6-8% στα αρχικά μοντέλα σε περίπου 94% για το GPT-4o [25]. Αντίστοιχη πορεία καταγράφεται στο GSM8K (8.500 μαθηματικά επιπέδου δημοτικού), όπου το GPT-4o προσεγγίζει το 94% [26]. Ωστόσο, στα προβλήματα ολυμπιακού επιπέδου η απόδοση μειώνεται σημαντικά: το benchmark CHAMP καταγράφει βέλτιστη επίδοση 58,1% [27], ενώ το OlympicArena (11.163 προβλήματα) δίνει 39,97% για το GPT-4o και κάτω του 20% για τα open-source μοντέλα [28]. Ανάλογη πρόκληση πολυβηματικού συλλογισμού παρουσιάζει και το EconLogicQA [29], το οποίο απαιτεί από το μοντέλο να εκτελέσει αλληλουχία λογικών βημάτων για την εξαγωγή συμπερασμάτων – μια ικανότητα που δεν ταυτίζεται με την απομνημόνευση γεγονότων.

4.2.4 Προγραμματισμού (π.χ. HumanEval, SWE-bench)

Στην πληροφορική, το benchmark HumanEval (164 προβλήματα κώδικα) καταγράφει για το μοντέλο Codex απόδοση pass@1 ίση με 28,8%, έναντι 0% του GPT-3 και 11,4% του GPT-J, με την pass@100 να φθάνει το 70,2% [30]. Το SWE-bench (2.294 πραγματικά GitHub issues) παρουσιάζει διαφορετική εικόνα: το Claude 2 σημειώνει μόλις 1,96% χωρίς πρόσθετες τεχνικές, ενώ με τη χρήση agentic methods η απόδοση αυξάνεται σημαντικά, από 4,4% σε 71,7% στο επαληθευμένο υποσύνολο [31]. Το εύρημα αυτό υποδεικνύει ότι η δημιουργία κώδικα διαφοροποιείται ουσιωδώς από τη διόρθωση πραγματικού κώδικα.

4.2.5 Ιατρικής / Εξειδικευμένων (π.χ. MedQA)

Στην ιατρική, τα εξειδικευμένα (fine-tuned) μοντέλα παρουσιάζουν αξιοσημείωτη απόδοση: το Med-PaLM πέτυχε περίπου 67% στο MedQA (USMLE-style), υπερβαίνοντας το ενδεικτικό

κατώφλι επιτυχίας (~60%) για το USMLE στις εξετάσεις USMLE [32], ενώ το Med-PaLM 2 σημείωσε 86,5% στο MedQA, με τους αξιολογητές ιατρούς να προτιμούν στατιστικά σημαντικά τις απαντήσεις του έναντι ανθρώπινων ($p<0,001$) [33]. Τα αποτελέσματα ωστόσο διαφοροποιούνται ανά εργασία: στην καρδιολογία το GPT-4 υπερτερεί στο 52,6% των περιπτώσεων χωρίς στατιστική σημαντικότητα [34], στο NEJM Image Challenge το Claude 3 Haiku σημειώνει 59,8% έναντι μέσου ανθρώπινου 49,4% [35], και το μοντέλο o1 υπερβαίνει το GPT-4 κατά 6,2% κατά μέσο όρο [36]. Πέρα από τα αγγλόφωνα ιατρικά benchmarks, υπάρχουν και γλωσσικά εξειδικευμένα σύνολα δεδομένων. Για παράδειγμα, οι εξετάσεις για την Ιατρική Ειδίκευση στην Ισπανία (MIR) [37] αποτελούν ένα τέτοιο benchmark, αναδεικνύοντας τη σημασία της γλωσσικής και πολιτισμικής προσαρμογής στην αξιολόγηση των LLMs. Σημαντικά, στο MEDEC οι ιατροί υπερτερούν όλων των μοντέλων στην ανίχνευση κλινικών σφαλμάτων [38].

4.3 Μετρικές Απόδοσης

Διαφορετικές μετρικές χρησιμοποιούνται ανάλογα με τη φύση της εργασίας [4]. Για εργασίες πολλαπλής επιλογής εφαρμόζεται η ακρίβεια (accuracy). Για τη δημιουργία κώδικα χρησιμοποιείται η μετρική pass@k, που εκφράζει το ποσοστό προβλημάτων που επιλύονται σε k παραγόμενα αποτελέσματα [30]. Για την παραγωγή κειμένου εφαρμόζονται οι μετρικές BLEU και ROUGE, ενώ για εργασίες με πολλαπλές έγκυρες απαντήσεις προτιμάται το F1 score. Αρκετές μελέτες επισημαίνουν ότι οι αυτόματες μετρικές ενδέχεται να είναι ανεπαρκείς για την αποτίμηση της ποιότητας, ιδίως σε κλινικό πλαίσιο [38].

4.4 Παγίδες και Περιορισμοί Αξιολόγησης

Η βιβλιογραφία αναδεικνύει αρκετές μεθοδολογικές προκλήσεις. Πρώτον, η μόλυνση δεδομένων (data contamination) προκύπτει όταν δεδομένα ελέγχου περιλαμβάνονται στα δεδομένα εκπαίδευσης, οδηγώντας σε τεχνητά υψηλές επιδόσεις που είναι δύσκολο να εντοπιστούν [39]. Δεύτερον, ο κορεσμός των benchmarks (saturation) μειώνει τη διακριτική

τους ικανότητα όταν τα μοντέλα συγκλίνουν σε παρόμοιες επιδόσεις [39]. Τρίτον, η ευαισθησία στη διατύπωση των prompts μπορεί να επιφέρει σημαντικές μεταβολές στα αποτελέσματα [40]. Τέταρτον, η γλωσσική επίδραση οδηγεί σε χαμηλότερη απόδοση σε γλώσσες χαμηλών πόρων (π.χ. 72% στα αγγλικά έναντι 44% στα αραβικά), ζήτημα ιδιαίτερα σχετικό για γλώσσες όπως η ελληνική [41].

5 Απόδοση των LLMs ανά Επιστημονικό Πεδίο

Το κεφάλαιο αυτό αποτελεί τον πυρήνα της εργασίας. Με βάση τη συστηματική ανασκόπηση 43 επιστημονικών εργασιών (peer-reviewed και αξιόπιστα preprints), εξετάζεται αναλυτικά η διαφοροποίηση της απόδοσης των LLMs ανά επιστημονικό πεδίο. Για κάθε πεδίο αναλύονται τα μετρούμενα μεγέθη, τα χρησιμοποιούμενα benchmarks, τα κύρια αποτελέσματα, και η σύγκριση με την ανθρώπινη απόδοση. Τα δεδομένα αναδεικνύουν συστηματικά μοτίβα που ερμηνεύονται από τη φύση κάθε πεδίου, τη δομή των δεδομένων εκπαίδευσης, και τις απαιτήσεις συλλογισμού.

5.1 Θετικές Επιστήμες και Μαθηματικά

Τα μαθηματικά και οι θετικές επιστήμες συγκαταλέγονται στα πλέον μελετημένα πεδία αξιολόγησης των LLMs, εν μέρει επειδή προσφέρουν σαφείς, μονοσήμαντες σωστές απαντήσεις που διευκολύνουν την αυτόματη αξιολόγηση. Τα ευρήματα της ανασκόπησης καταδεικνύουν σημαντική βελτίωση της απόδοσης διαχρονικά, καθώς και σαφή όρια που παραμένουν.

Σύμφωνα με τα ευρήματα των σχετικών μελετών, στα προβλήματα μαθηματικής λογικής χαμηλής και μεσαίας δυσκολίας τα σύγχρονα μοντέλα σημειώνουν αξιοσημείωτη απόδοση. Στο benchmark GSM8K (8.500 μαθηματικά επιπέδου δημοτικού (grade-school math)), η απόδοση αυξήθηκε από τιμές κοντά στο μηδέν σε περίπου 94% για το GPT-4o [26]. Αντίστοιχη πορεία καταγράφεται στο MATH (12.500 προβλήματα ποικίλης δυσκολίας), όπου η απόδοση μεταβλήθηκε από 6-8% στα αρχικά μοντέλα σε περίπου 94% για το GPT-4o [25]. Επιπλέον, το μοντέλο Gemini 2.0 Flash Thinking, που ενσωματώνει ρητό συλλογισμό, σημείωσε 73,6% στο απαιτητικότερο U-MATH [25].

Ωστόσο, με την αύξηση της δυσκολίας και της πολυπλοκότητας, η απόδοση μειώνεται σημαντικά. Στο benchmark CHAMP, που περιλαμβάνει προβλήματα μαθηματικού διαγωνισμού ολυμπιακού επιπέδου, η βέλτιστη καταγεγραμμένη επίδοση περιορίζεται στο 58,1% [27]. Η μείωση αυτή σε σχέση με τα απλούστερα προβλήματα υποδεικνύει ότι ο σύνθετος μαθηματικός συλλογισμός, ο οποίος απαιτεί δημιουργική επίλυση και πολλαπλά βήματα λογικής, παραμένει ουσιώδης πρόκληση για τα LLMs.

Το benchmark OlympicArena ενισχύει το συμπέρασμα αυτό. Με 11.163 προβλήματα από διαφορετικούς κλάδους των φυσικών επιστημών, απαιτεί διασύνδεση γνώσης από πολλαπλά πεδία. Το GPT-4o σημειώνει 39,97%, ενώ τα open-source μοντέλα παρουσιάζουν απόδοση κάτω του 20% [28]. Το αποτέλεσμα υπογραμμίζει ότι η ικανότητα σύνθεσης γνώσης από πολλαπλά πεδία (multi-domain reasoning) είναι σημαντικά πιο περιορισμένη από την απόδοση σε μεμονωμένα, εστιασμένα προβλήματα.

Στις φυσικές επιστήμες ειδικότερα, το GPQA (Graduate-Level Google-Proof Q&A) προσφέρει αυστηρή αξιολόγηση. Αποτελείται από 448 ερωτήσεις σε βιολογία, φυσική και χημεία, σχεδιασμένες ώστε να μην απαντώνται μέσω απλής αναζήτησης και να απαιτούν εξειδικευμένη γνώση. Τα αποτελέσματα δείχνουν ότι οι κάτοχοι διδακτορικού σημειώνουν 65% ακρίβεια, ενώ το GPT-4 περιορίζεται στο 39% [23]. Το Claude 3 Opus προσεγγίζει το 60%, χωρίς να φθάνει το επίπεδο των ειδικών [3], [23].

Συμπληρωματικά, η αξιολόγηση SciAssess, που αφορά την ανάλυση επιστημονικής βιβλιογραφίας, αναδεικνύει ένα συστηματικό μοτίβο: τα μοντέλα παρουσιάζουν ισχυρή ανάκληση γεγονότων (factual recall) αλλά περιορισμένη ικανότητα σύνθεσης πληροφοριών από πολλαπλές εργασίες (cross-paper synthesis) [24]. Το εύρημα υποδεικνύει ότι τα LLMs ανακαλούν επιστημονικά δεδομένα αλλά δυσχεραίνονται στον συνδυασμό γνώσης από διαφορετικές πηγές κατά τρόπο ανάλογο του έμπειρου ερευνητή.

Συνολικά, η εικόνα στις θετικές επιστήμες και τα μαθηματικά είναι διττή. Αφενός, τα LLMs έχουν αναπτύξει αξιοσημείωτη ικανότητα στα τυποποιημένα, σαφώς ορισμένα προβλήματα, προσεγγίζοντας ή υπερβαίνοντας τα ανθρώπινα επίπεδα σε benchmarks όπως το GSM8K και το MATH [25], [26]. Αφετέρου, ο σύνθετος και διεπιστημονικός συλλογισμός που απαιτείται για προβλήματα ολυμπιακού επιπέδου ή ερωτήσεις επιπέδου διδακτορικού παραμένει σαφής περιοχή υστέρησης [27], [28], [23].

Σε επίπεδο εξετάσεων φυσικών επιστημών, η μελέτη των Farhat et al. αξιολόγησε τα GPT-3.5, GPT-4 και Bard σε ερωτήσεις φυσικής, χημείας και βιολογίας, καταγράφοντας για το GPT-4 ακρίβεια 73% στη φυσική, 44% στη χημεία και 51% στη βιολογία, με το GPT-3.5 να σημειώνει χαμηλότερες επιδόσεις [42]. Η σημαντική διακύμανση μεταξύ κλάδων ενισχύει το συμπέρασμα ότι η απόδοση δεν είναι ομοιόμορφη ούτε εντός των ίδιων των θετικών επιστημών, με τη χημεία να αναδεικνύεται ως ιδιαίτερα απαιτητική.

5.2 Πληροφορική και Προγραμματισμός

Η πληροφορική και ο προγραμματισμός συνιστούν ένα ιδιαίτερα ενδιαφέρον πεδίο αξιολόγησης, καθώς συνδυάζουν αυστηρή λογική δομή με πρακτική εφαρμοσιμότητα. Τα ευρήματα αναδεικνύουν ένα κρίσιμο μοτίβο: σημαντική διαφοροποίηση μεταξύ της απόδοσης σε απομονωμένα προβλήματα κωδικοποίησης και της απόδοσης σε πραγματικές προκλήσεις ανάπτυξης λογισμικού.

Ως προς τη δημιουργία μεμονωμένων τμημάτων κώδικα, το benchmark HumanEval αποτελεί σημείο αναφοράς. Περιλαμβάνει 164 προβλήματα προγραμματισμού σε Python, όπου το μοντέλο καλείται να συμπληρώσει μια συνάρτηση βάσει της περιγραφής της. Το μοντέλο Codex, ειδικά εκπαιδευμένο σε κώδικα, σημείωσε pass@1 ίσο με 28,8%, έναντι 0% του γενικού σκοπού GPT-3 και 11,4% του GPT-J [30]. Η διαφορά αυτή υπογραμμίζει τη σημασία της εξειδικευμένης εκπαίδευσης σε κώδικα.

Ιδιαίτερα ενδεικτική είναι η μετρική pass@100, που εκφράζει αν τουλάχιστον ένα από 100 παραγόμενα αποτελέσματα είναι ορθό. Στη μετρική αυτή ο Codex σημειώνει 70,2%, ποσοστό σημαντικά υψηλότερο της pass@1 [30]. Η διαφορά υποδεικνύει ότι το μοντέλο διαθέτει την ικανότητα παραγωγής ορθών λύσεων, αλλά με περιορισμένη αξιοπιστία στην πρώτη προσπάθεια. Σε πρακτικό επίπεδο, αυτό συνεπάγεται ότι ο προγραμματιστής ενδέχεται να χρειαστεί να εξετάσει πολλαπλές προτεινόμενες λύσεις.

Η εικόνα διαφοροποιείται σημαντικά στις πραγματικές προκλήσεις ανάπτυξης λογισμικού. Το benchmark SWE-bench περιλαμβάνει 2.294 πραγματικά ζητήματα (issues) από αποθετήρια GitHub, όπου το μοντέλο πρέπει να κατανοήσει υπάρχον σύστημα κώδικα, να εντοπίσει το πρόβλημα και να παραγάγει διόρθωση. Το Claude 2 σημειώνει αρχικά μόλις 1,96% επιτυχία [31]. Η περιορισμένη αυτή απόδοση αντανακλά την πολυπλοκότητα της πραγματικής μηχανικής λογισμικού σε σχέση με μεμονωμένα προβλήματα.

Το σημαντικότερο εύρημα του πεδίου αφορά την επίδραση των agentic methods. Όταν το μοντέλο αποκτά τη δυνατότητα να χρησιμοποιεί εργαλεία, να εκτελεί κώδικα, να παρατηρεί τα αποτελέσματα και να επαναλαμβάνει τη διαδικασία, η απόδοση αυξάνεται σημαντικά. Από το αρχικό 1,96%, οι agentic προσεγγίσεις ανέβασαν την επίδοση σε 4,4% και εν συνεχεία σε 71,7% στο επαληθευμένο υποσύνολο (SWE-bench Verified) [31]. Η αύξηση αυτή, άνω των 35 φορών, συνιστά ένα από τα σημαντικότερα ευρήματα της ανασκόπησης.

Το συμπέρασμα από το πεδίο της πληροφορικής είναι διττό. Πρώτον, η ικανότητα δημιουργίας νέου κώδικα διαφοροποιείται ποιοτικά από την ικανότητα κατανόησης, διόρθωσης και βελτίωσης υπάρχοντος κώδικα σε πραγματικά συστήματα. Δεύτερον, ο τρόπος αξιοποίησης ενός LLM (απλή προτροπή έναντι agentic πλαισίου) ενδέχεται να επιδρά στην απόδοση περισσότερο από την εγγενή ικανότητα του ίδιου του μοντέλου [31]. Το γεγονός αυτό έχει σημαντικές πρακτικές συνέπειες για την ανάπτυξη εργαλείων βασισμένων σε LLMs.

5.3 Ιατρική και Βιοεπιστήμες

Η ιατρική συνιστά ένα από τα πλέον εκτενώς μελετημένα πεδία εφαρμογής των LLMs, γεγονός που αποτυπώνεται στον μεγάλο αριθμό σχετικών εργασιών της ανασκόπησης. Οι δυνητικές εφαρμογές, από την υποστήριξη διάγνωσης έως την απάντηση ερωτήσεων ασθενών, έχουν ουσιώδη πρακτική σημασία. Τα ευρήματα καταδεικνύουν ότι τα LLMs επιτυγχάνουν αξιοσημείωτη απόδοση σε δοκιμές ιατρικής γνώσης, αλλά παρουσιάζουν περιορισμούς σε εργασίες που απαιτούν λεπτή κλινική κρίση.

Όπως καταγράφεται στη βιβλιογραφία, η σημαντικότερη επίδοση στο ιατρικό πεδίο προέρχεται από τα εξειδικευμένα μοντέλα της οικογένειας Med-PaLM. Το αρχικό Med-PaLM, μοντέλο PaLM προσαρμοσμένο (fine-tuned) σε ιατρικά δεδομένα, πέτυχε περίπου 67% στο MedQA (USMLE-style), υπερβαίνοντας το ενδεικτικό κατώφλι επιτυχίας (~60%) για το USMLE στις εξετάσεις USMLE [32]. Επιπλέον, το fine-tuning μείωσε τις δυνητικά επιβλαβείς απαντήσεις από 29,7% σε 5,9%, γεγονός που υπογραμμίζει τη σημασία της εξειδικευμένης εκπαίδευσης για την ασφάλεια [32].

Το μεταγενέστερο μοντέλο Med-PaLM 2 σημείωσε περαιτέρω βελτίωση, φθάνοντας το 86,5% στο MedQA, αύξηση περίπου 19 ποσοστιαίων μονάδων [33]. Αξιοσημείωτο είναι ότι, σε μελέτη δημοσιευμένη στο περιοδικό Nature, οι αξιολογητές ιατροί προτίμησαν στατιστικά σημαντικά τις απαντήσεις του Med-PaLM 2 έναντι των ανθρώπινων ($p<0,001$) [33]. Το εύρημα υποδεικνύει ότι, για ορισμένες κατηγορίες ιατρικών ερωτήσεων, τα εξειδικευμένα μοντέλα παράγουν απαντήσεις ποιότητας ισοδύναμης ή ανώτερης των ειδικών.

Η εικόνα διαφοροποιείται ανά τύπο ιατρικής εργασίας. Σε αξιολόγηση του GPT-4 σε 251 πραγματικές ερωτήσεις καρδιολογίας, το μοντέλο σημείωσε ανώτερες απαντήσεις στο 52,6%

των περιπτώσεων έναντι καρδιολόγων, χωρίς όμως στατιστική σημαντικότητα [34]. Επιπλέον, μόλις το 12,6% των απαντήσεών του χαρακτηρίστηκαν υψηλής ακρίβειας, έναντι 21,1% των ανθρώπινων [34]. Το αποτέλεσμα δείχνει ότι σε εξειδικευμένα κλινικά πεδία το πλεονέκτημα των LLMs περιορίζεται.

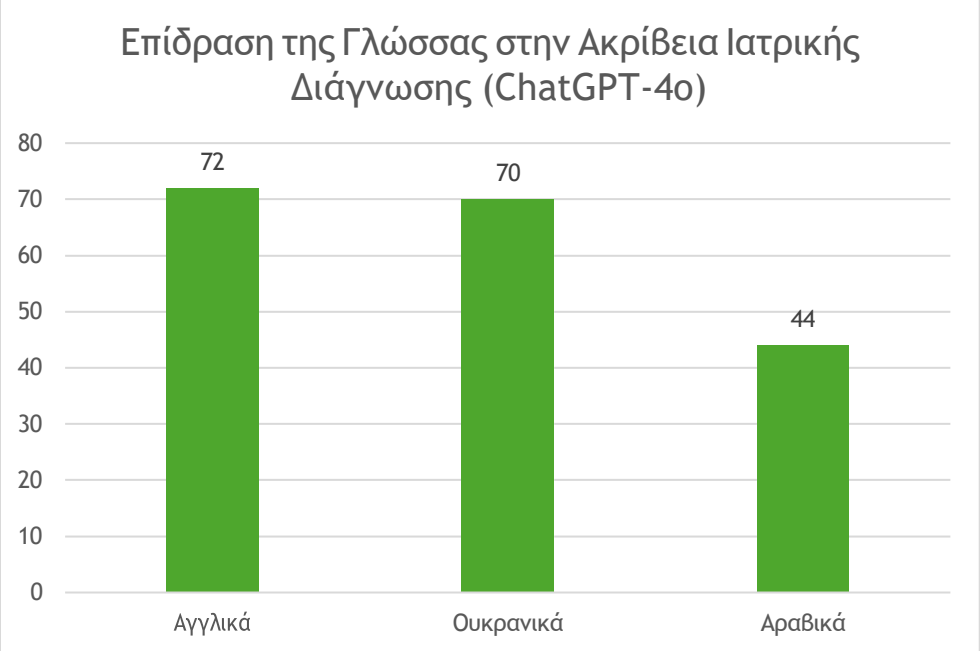
Στην πολυτροπική (multimodal) ιατρική, όπου απαιτείται ανάλυση εικόνων, τα αποτελέσματα είναι θετικά. Στο NEJM Image Challenge, συλλογή 945 περιπτώσεων διάγνωσης βάσει εικόνας, το Claude 3 Haiku σημείωσε 59,8%, υπερβαίνοντας τον μέσο όρο των μεμονωμένων ανθρώπων (49,4%) [35]. Ωστόσο, η συλλογική ανθρώπινη απόδοση (90,8%) παρέμεινε σημαντικά υψηλότερη, ενώ το GPT-4 Vision αρνήθηκε να απαντήσει στο 24% των περιπτώσεων, συμπεριφορά που περιορίζει την πρακτική χρησιμότητα [35].

Σύμφωνα με τα διαθέσιμα ευρήματα, τα νεότερα μοντέλα με ενισχυμένο συλλογισμό παρουσιάζουν περαιτέρω βελτίωση. Σε αξιολόγηση 37 συνόλων δεδομένων από έγκριτες πηγές (NEJM, Lancet), το μοντέλο o1 υπερέβη το GPT-4 κατά 6,2% κατά μέσο όρο, με μεγαλύτερα κέρδη (6,6%) στα απαιτητικότερα σύνολα [36]. Το εύρημα υποδεικνύει ότι η ενσωμάτωση ρητού συλλογισμού (reasoning) βελτιώνει την απόδοση σε σύνθετες ιατρικές εργασίες.

Παρά τα θετικά αποτελέσματα, ένα κρίσιμο εύρημα περιορίζει τη γενίκευσή τους. Το benchmark MEDEC, που αξιολογεί την ανίχνευση ιατρικών σφαλμάτων σε κλινικές σημειώσεις (3.848 κείμενα, 5 τύποι σφαλμάτων), καταδεικνύει ότι οι ιατροί υπερτερούν όλων των αξιολογημένων μοντέλων, συμπεριλαμβανομένων των o1-preview, GPT-4, Claude 3.5 Sonnet και Gemini 2.0 Flash [38]. Επιπλέον, η μελέτη επισημαίνει ότι οι αυτόματες μετρικές BLEU και ROUGE είναι ανεπαρκείς για την αξιολόγηση κλινικών κειμένων [38]. Το εύρημα υπογραμμίζει ότι η λεπτή κλινική κρίση παραμένει περιοχή ανθρώπινης υπεροχής. Η ανάγκη για ανάπτυξη τοπικών, μη-αγγλόφωνων ιατρικών benchmarks επισημαίνεται και από το AfriMed-QA [43], μια πρώτη συστηματική προσπάθεια δημιουργίας παν-αφρικανικού συνόλου δεδομένων για ιατρικές ερωτήσεις. Το εγχείρημα αυτό καταδεικνύει ότι ακόμη και ηπείρους με εκατοντάδες γλώσσες και διαφορετικές κλινικές πρακτικές υποεκπροσωπούνται στα τρέχοντα δεδομένα εκπαίδευσης των LLMs.

Τέλος, αναδεικνύεται το ζήτημα της γλωσσικής εξάρτησης της ιατρικής απόδοσης. Σε αξιολόγηση 50 κλινικών περιπτώσεων σε αγγλικά, αραβικά και ουκρανικά, το ChatGPT-4o σημείωσε συνολική ακρίβεια 62%, με σημαντική διακύμανση ανά γλώσσα: 72% στα αγγλικά, 70% στα ουκρανικά, αλλά 44% στα αραβικά [41]. Το Claude 3.5 Sonnet παρουσίασε παρόμοιο

μοτίβο (60,67% συνολικά) [41]. Η γλωσσική αυτή ανισότητα είναι ιδιαίτερα σημαντική για γλώσσες χαμηλών πόρων, καθώς υποδεικνύει ότι η ποιότητα της ιατρικής υποστήριξης ενδέχεται να διαφέρει ουσιωδώς ανάλογα με τη γλώσσα του χρήστη.



Εικόνα 7: Επίδραση γλώσσας στην ακρίβεια διάγνωσης ChatGPT-4o

Πέραν των τυποποιημένων ιατρικών εξετάσεων, η μελέτη των Stribling et al. αξιολόγησε το GPT-4 σε εννέα εξετάσεις μεταπτυχιακού επιπέδου βιοϊατρικών επιστημών, διαπιστώνοντας ότι το μοντέλο υπερέβη τη μέση επίδοση των φοιτητών σε επτά από τις εννέα εξετάσεις [44]. Το εύρημα επιβεβαιώνει την ισχυρή απόδοση των μοντέλων σε εργασίες ανάκλησης εξειδικευμένης γνώσης, σε αντιδιαστολή με την υστέρησή τους σε εργασίες κλινικής κρίσης.

Ενδιαφέρουσα κατεύθυνση βελτίωσης προσφέρει η προσέγγιση της συνέργειας μοντέλων (model synergy). Η μελέτη των Yang et al. κατέδειξε ότι ο συνδυασμός πολλαπλών μοντέλων σε σχήμα συνόλου (ensemble) για ιατρικές ερωτήσεις βελτιώνει την απόδοση, με το GPT-4 να σημειώνει 71% στο MedMCQA και εξειδικευμένα μοντέλα όπως το Vicuna-13B να φθάνουν 89,5% στο PubMedQA [45]. Το αποτέλεσμα υποδεικνύει ότι η σύνθεση εξειδικευμένων μοντέλων αποτελεί υποσχόμενη εναλλακτική έναντι ενός ενιαίου γενικού μοντέλου.

Η γλωσσική διάσταση της ιατρικής απόδοσης τεκμηριώνεται περαιτέρω από τη μελέτη των Rosoł et al. στην Ιατρική Τελική Εξέταση της Πολωνίας. Το GPT-4 πέτυχε μέση ακρίβεια 79,7% τόσο στην πολωνική όσο και στην αγγλική έκδοση, περνώντας όλες τις εκδόσεις της εξέτασης, ενώ το GPT-3.5 σημείωσε 54,8% στα πολωνικά έναντι 60,3% στα αγγλικά [46]. Η διαφορά αυτή μεταξύ των δύο γλωσσών για το ασθενέστερο μοντέλο επιβεβαιώνει ότι η

γλωσσική ανισότητα είναι εντονότερη στα λιγότερο ισχυρά μοντέλα, ζήτημα με άμεση σημασία για γλώσσες χαμηλών πόρων όπως η ελληνική.

5.4 Νομική

Η νομική συνιστά πεδίο όπου τα LLMs έχουν επιδείξει ορισμένες από τις σημαντικότερες επιδόσεις τους, αλλά και σαφή όρια συνδεόμενα με τη δικαιοδοσία και τη γλώσσα. Όπως και στην ιατρική, παρατηρείται διάκριση μεταξύ της απόδοσης σε τυποποιημένες εξετάσεις γνώσης και της απόδοσης σε σύνθετες εργασίες νομικού συλλογισμού.

Η ευρύτερα αναφερόμενη επίδοση προέρχεται από τις εξετάσεις δικηγορίας (Bar Exam) των ΗΠΑ. Σε αρχική μελέτη, το GPT-3.5 αξιολογήθηκε στο τμήμα πολλαπλής επιλογής (MBE) και σημείωσε 50,3% ορθές απαντήσεις [47]. Αν και η επίδοση αυτή δεν επαρκούσε για επιτυχία, υπερέβαινε σημαντικά το τυχαίο baseline του 25%, καταδεικνύοντας ουσιαστική νομική γνώση [47].

Σύμφωνα με τη μελέτη των Katz et al., σημαντική πρόοδος καταγράφηκε με το GPT-4. Σε αξιολόγηση στις πλήρεις ενιαίες εξετάσεις δικηγορίας (Uniform Bar Exam, που περιλαμβάνει MBE, MEE και MPT), το GPT-4 υπερέβη την ανθρώπινη απόδοση στο τμήμα πολλαπλής επιλογής, σημειώνοντας συνολικά περίπου 297 πόντους, επίδοση που αντιστοιχεί περίπου στο 90ό εκατοστημόριο των υποψηφίων [48]. Στα δοκίμια (MEE/MPT) σημείωσε μέσο όρο 4,2 στα 6 [48].

Ωστόσο, η εικόνα διαφοροποιείται σε σύνθετες νομικές εργασίες πέραν των τυποποιημένων εξετάσεων. Το benchmark LegalBench, με 162 διαφορετικές νομικές εργασίες, αξιολόγησε 20 μοντέλα [49]. Αν και το GPT-4 ηγήθηκε της κατάταξης, παρέμεινε κάτω από το επίπεδο των ειδικών, ενώ τα open-source μοντέλα σημείωσαν χαμηλότερες επιδόσεις [49]. Το εύρημα υποδεικνύει ότι η επιτυχία στις εξετάσεις πολλαπλής επιλογής δεν μεταφράζεται αυτομάτως σε ικανότητα εφαρμοσμένου νομικού συλλογισμού.

Διαφωτιστική είναι μια μελέτη που εξέτασε την ικανότητα του GPT για σημασιολογική επισήμανση (semantic annotation) νομικών κειμένων σε συνθήκες zero-shot. Τα αποτελέσματα παρουσίασαν σημαντική διακύμανση ανά τύπο εγγράφου: F1 score 0,86 σε συμβόλαια, 0,73 σε δικαστικές αποφάσεις, και 0,54 σε νομοθετικά κείμενα [50]. Η διαβάθμιση

αυτή υποδεικνύει ότι τα LLMs αποδίδουν καλύτερα σε δομημένα έγγραφα με προβλέψιμη μορφή και χαμηλότερα σε κείμενα που απαιτούν βαθύτερη ερμηνεία του νομικού πλαισίου.

Ένα κρίσιμο όριο αφορά τη δικαιοδοσία και τη γλώσσα. Μελέτες πρόβλεψης δικαστικών αποφάσεων κατέδειξαν ότι, αν και το GPT-4 ήταν το ισχυρότερο γενικό μοντέλο, υστερούσε έναντι εξειδικευμένων (fine-tuned) μοντέλων στη νομολογία γλωσσών όπως τα κινεζικά [51]. Το εύρημα αποτελεί σημαντική επιφύλαξη: η απόδοση των LLMs στις αγγλόφωνες αμερικανικές εξετάσεις δικηγορίας δεν μεταφέρεται απαραίτητα σε άλλα νομικά συστήματα και γλώσσες, συμπεριλαμβανομένου του ελληνικού νομικού πλαισίου.

5.5 Κοινωνικές και Ανθρωπιστικές Επιστήμες

Οι κοινωνικές και ανθρωπιστικές επιστήμες παρουσιάζουν ιδιαίτερη πρόκληση για την αξιολόγηση των LLMs, καθώς το ορθό κριτήριο είναι συχνά δυσχερέστερο να οριστεί αντικειμενικά. Σε αντίθεση με τα μαθηματικά ή τον προγραμματισμό, πολλές ερωτήσεις απαιτούν ερμηνεία, πολιτισμική κατανόηση και ευαισθησία στο πλαίσιο. Τα ευρήματα αναδεικνύουν τόσο αξιοσημείωτες δυνατότητες όσο και ουσιώδεις προκλήσεις, ιδίως ως προς τη γλώσσα και τον πολιτισμό.

Στον τομέα της κοινής λογικής (commonsense reasoning), τα αποτελέσματα που αναφέρονται στη βιβλιογραφία είναι θετικά. Το benchmark WinoGrande, με 44.000 ερωτήσεις κατανόησης καθημερινών καταστάσεων, αποτελεί αυστηρό κριτήριο. Ενώ τα παλαιότερα μοντέλα όπως το RoBERTa σημείωναν 59-79%, το GPT-4 με χρήση chain-of-thought προτροπής έφθασε περίπου το 95,3%, επίδοση συγκρίσιμη με την ανθρώπινη (94%) [52]. Το εύρημα υποδεικνύει ότι, για τυποποιημένες μορφές κοινής λογικής, τα σύγχρονα μοντέλα προσεγγίζουν το ανθρώπινο επίπεδο.

Ωστόσο, η βαθύτερη πολιτισμική κατανόηση παραμένει περιοχή υστέρησης. Το benchmark CMMLU, το κινεζικό αντίστοιχο του MMLU, ανέδειξε ότι τα κυρίως αγγλόφωνα (English-centric) μοντέλα παρουσιάζουν μειωμένη απόδοση σε ερωτήσεις κινεζικής πολιτισμικής γνώσης, ενώ τα εξειδικευμένα στα κινεζικά μοντέλα υπερτερούν στο συγκεκριμένο περιεχόμενο [53]. Το εύρημα δείχνει ότι η απόδοση εξαρτάται άμεσα από την αντιπροσώπευση του εκάστοτε πολιτισμού στα δεδομένα εκπαίδευσης.

Το ζήτημα της πολυγλωσσίας αναδεικνύεται ως ένα από τα σημαντικότερα. Όπως και στο ιατρικό πεδίο, η απόδοση μειώνεται σε γλώσσες χαμηλών πόρων: το ChatGPT-4o σημείωσε 72% στα αγγλικά αλλά 44% στα αραβικά για τις ίδιες κλινικές περιπτώσεις [41]. Η ανισότητα αυτή έχει συνέπειες για την ισότητα στην πρόσβαση. Για χρήστες γλωσσών με περιορισμένη παρουσία στα δεδομένα εκπαίδευσης, συμπεριλαμβανομένων ευρωπαϊκών γλωσσών όπως η ελληνική, η ποιότητα των υπηρεσιών ενδέχεται να είναι χαμηλότερη.

Στην πολυτροπική κατανόηση (multimodal), που συνδυάζει κείμενο και εικόνα σε ακαδημαϊκό επίπεδο, το benchmark MMMU προσφέρει σημαντικά δεδομένα. Με 11.500 ερωτήσεις κολεγιακού επιπέδου, τα αποτελέσματα δείχνουν ότι το GPT-4V σημείωσε 56% και το Gemini Ultra 59% [54]. Αν και αξιοσημείωτα, τα ποσοστά παραμένουν κάτω από την ανθρώπινη απόδοση, με χάσμα περίπου 8,1% έως το τέλος του 2024 [54]. Το εύρημα υποδεικνύει ότι η ολοκληρωμένη πολυτροπική κατανόηση σε ακαδημαϊκό επίπεδο παραμένει ανοιχτή πρόκληση.

Συνολικά, στις κοινωνικές και ανθρωπιστικές επιστήμες αναδεικνύεται ένα θεμελιώδες ζήτημα. Τα LLMs επιδεικνύουν αξιοσημείωτη ικανότητα σε τυποποιημένες μορφές συλλογισμού με καλή αντιπροσώπευση στα κυρίως αγγλικά δεδομένα εκπαίδευσης. Ωστόσο, η βαθιά πολιτισμική κατανόηση, η ευαισθησία σε τοπικά συμφραζόμενα και η ισότιμη απόδοση σε διαφορετικές γλώσσες παραμένουν ουσιώδεις προκλήσεις. Το ζήτημα είναι ιδιαίτερα σχετικό για την ελληνική ακαδημαϊκή και επαγγελματική κοινότητα, καθώς υπογραμμίζει την ανάγκη προσοχής κατά την εφαρμογή των εργαλείων αυτών σε μη-αγγλόφωνα πλαίσια.

5.6 Διατομεακοί Παράγοντες Διακύμανσης

Έχοντας εξετάσει την απόδοση ανά πεδίο, είναι δυνατή η σύνθεση των διατομεακών παραγόντων που ερμηνεύουν τη διακύμανση. Η ανάλυση των 43 εργασιών αναδεικνύει πέντε κύριους παράγοντες οι οποίοι, μεμονωμένα ή σε συνδυασμό, καθορίζουν την απόδοση ενός LLM σε δεδομένη εργασία.

Πρώτος παράγοντας: ο όγκος και η ποιότητα των δεδομένων εκπαίδευσης. Αναδεικνύεται ως ο πλέον καθοριστικός παράγοντας. Τα μοντέλα που έχουν υποστεί εξειδικευμένο fine-tuning

σε δεδομένα συγκεκριμένου πεδίου (όπως το Med-PaLM στην ιατρική) υπερτερούν σταθερά των γενικού σκοπού μοντέλων, με βελτίωση που μπορεί να φθάσει το 20-40% [32], [33]. Αντιστρόφως, πεδία και γλώσσες με περιορισμένη αντιπροσώπευση στα δεδομένα παρουσιάζουν σταθερά χαμηλότερη απόδοση [41].

Χαρακτηριστικό παράδειγμα της επίδρασης των εξειδικευμένων δεδομένων αποτελεί το BloombergGPT, ένα μοντέλο 50 δισεκατομμυρίων παραμέτρων εκπαιδευμένο σε χρηματοοικονομικά δεδομένα. Σύμφωνα με τους Wu et al., το μοντέλο υπερτερεί των αντίστοιχων ανοιχτών μοντέλων σε εξειδικευμένες χρηματοοικονομικές εργασίες, διατηρώντας παράλληλα ανταγωνιστική απόδοση σε γενικές εργασίες [55]. Το εύρημα αυτό, από τον τομέα των οικονομικών, ενισχύει τη διαπίστωση ότι η εκπαίδευση σε δεδομένα συγκεκριμένου πεδίου αποτελεί καθοριστικό παράγοντα απόδοσης, ανεξαρτήτως του επιστημονικού κλάδου.

Δεύτερος παράγοντας: ο βαθμός τυποποίησης του πεδίου. Πεδία με σαφείς, ρητούς κανόνες και τυποποιημένη δομή (όπως τα μαθηματικά ή ο προγραμματισμός) επιτρέπουν την αξιόπιστη αναγνώριση και εφαρμογή μοτίβων [25], [30]. Αντιθέτως, πεδία που απαιτούν ερμηνεία, κρίση ή κατανόηση ασαφών συμφραζομένων (όπως η κλινική κρίση ή η νομική ερμηνεία) παρουσιάζουν μεγαλύτερες προκλήσεις, ακόμη και όταν τα μοντέλα διαθέτουν την απαιτούμενη γνώση [38].

Τρίτος παράγοντας: η πολυπλοκότητα και ο τύπος του απαιτούμενου συλλογισμού. Τα δεδομένα καταδεικνύουν με συνέπεια ότι οι εργασίες ενός βήματος ή ανάκλησης γνώσης αποδίδουν σημαντικά καλύτερα από εργασίες που απαιτούν συλλογισμό πολλαπλών βημάτων (όπως τα ολυμπιακά μαθηματικά, με απόδοση 58%) ή σύνθεση γνώσης από πολλαπλές πηγές (όπως το OlympicArena, με 40%) [27], [28], [24]. Το εύρημα υποδεικνύει ότι ο αλυσιδωτός συλλογισμός παραμένει θεμελιώδης περιορισμός.

Τέταρτος παράγοντας: οι εφαρμοζόμενες τεχνικές βελτίωσης. Όπως καταδεικνύεται στο SWE-bench (αύξηση από 1,96% σε 71,7%), ο τρόπος αξιοποίησης του μοντέλου δύναται να είναι εξίσου ή περισσότερο καθοριστικός από την εγγενή του ικανότητα [31]. Η chain-of-thought προτροπή προσθέτει τυπικά 10-20%, το fine-tuning 20-40%, ενώ οι agentic methods πολλαπλασιάζουν την απόδοση σε ορισμένες εργασίες [31], [40]. Συνεπώς, η αξιολόγηση ενός μοντέλου αποτελεί στην πραγματικότητα αξιολόγηση ενός συστήματος μοντέλου και μεθόδου.

Πέμπτος παράγοντας: η γλώσσα και η πολιτισμική αντιπροσώπευση. Διατρέχοντας πολλαπλά πεδία (ιατρική, νομική, κοινωνικές επιστήμες), η γλωσσική ανισότητα αναδεικνύεται ως

σταθερός και σημαντικός παράγοντας. Η απόδοση σε γλώσσες χαμηλών πόρων μπορεί να είναι 25-30 ποσοστιαίες μονάδες χαμηλότερη από αυτήν στα αγγλικά (π.χ. 72% έναντι 44%) [41]. Ο παράγοντας αυτός είναι ιδιαίτερα κρίσιμος για την ελληνική πραγματικότητα και υποδεικνύει ότι τα κυρίως αγγλόφωνα διεθνή benchmarks ενδέχεται να υπερεκτιμούν την απόδοση που θα βίωνε ένας Έλληνας χρήστης.

Ένα ιδιαίτερα σημαντικό παράδειγμα εγχώριας προσπάθειας προς αυτή την κατεύθυνση αποτελεί το Meltemi, το πρώτο ανοιχτό Μεγάλο Γλωσσικό Μοντέλο για τα ελληνικά, το οποίο αναπτύχθηκε από το Ινστιτούτο Επεξεργασίας του Λόγου του Ερευνητικού Κέντρου «Αθηνά». Το Meltemi βασίζεται σε αρχιτεκτονική τύπου Mistral και έχει εκπαιδευτεί σε εκτεταμένο σώμα ελληνικού κειμένου, με στόχο τη βελτίωση της κατανόησης και παραγωγής γλώσσας σε περιβάλλον χαμηλών πόρων. Η ύπαρξη τέτοιων μοντέλων δείχνει ότι το κενό απόδοσης μεταξύ αγγλικών και ελληνικών μπορεί σταδιακά να μειωθεί, υπό την προϋπόθεση ότι θα υπάρξει συστηματική επένδυση σε δεδομένα, υπολογιστικούς πόρους και αξιολόγηση με κατάλληλα ελληνικά benchmarks.

Συμπερασματικά, η διακύμανση της απόδοσης των LLMs ανά πεδίο δεν είναι τυχαία ούτε ανερμήνευτη, αλλά προκύπτει από την αλληλεπίδραση των πέντε αυτών παραγόντων. Ένα πεδίο καλά τυποποιημένο, με επαρκή αγγλόφωνα δεδομένα εκπαίδευσης, που απαιτεί συλλογισμό ενός βήματος και επωφελείται από κατάλληλες τεχνικές προτροπής, παρουσιάζει υψηλή απόδοση. Αντιθέτως, ένα πεδίο που απαιτεί λεπτή κρίση, σε γλώσσα χαμηλών πόρων, με πολυβηματικό συλλογισμό, παρουσιάζει σημαντικές προκλήσεις. Η κατανόηση αυτή αποτελεί την κύρια συνεισφορά της παρούσας ανασκόπησης.

6 Συζήτηση και Σύνθεση Ευρημάτων

Το κεφάλαιο αυτό συνθέτει τα ευρήματα που παρουσιάστηκαν στα προηγούμενα κεφάλαια, επιχειρώντας μια ερμηνευτική ανάγνωση πέρα από την παράθεση επιμέρους αποτελεσμάτων. Στόχος δεν είναι η επανάληψη των δεδομένων ανά πεδίο, αλλά η ανάδειξη των διατομεακών μοτίβων, των ερμηνειών τους, και των περιορισμών της ίδιας της ανασκόπησης. Η ανάλυση οργανώνεται γύρω από τρεις άξονες: τη συγκριτική εικόνα της απόδοσης μεταξύ πεδίων, τους παράγοντες που ερμηνεύουν τις παρατηρούμενες διαφορές, και τις απειλές εγκυρότητας που οριοθετούν τα συμπεράσματα.

6.1 Συγκριτική Εικόνα ανά Πεδίο

Η συγκριτική θεώρηση των πεδίων αναδεικνύει ένα σαφές μοτίβο: η απόδοση των LLMs δεν είναι ομοιόμορφη, αλλά διαβαθμίζεται ανάλογα με τη φύση της εργασίας. Με βάση τα δεδομένα της ανασκόπησης, μπορεί να διακριθεί μια ιεραρχία τριών επιπέδων ως προς τη σχέση της απόδοσης των μοντέλων με την ανθρώπινη.

Στο πρώτο επίπεδο, τα LLMs προσεγγίζουν ή υπερβαίνουν την ανθρώπινη απόδοση σε τυποποιημένες εργασίες με σαφώς ορισμένες απαντήσεις. Χαρακτηριστικά παραδείγματα αποτελούν η γενική γνώση (MMLU >88% έναντι ανθρώπινου 89,8%) [19], τα μαθηματικά λογικής χαμηλής δυσκολίας (GSM8K και MATH περίπου 94%) [25], [26], η κοινή λογική (WinoGrande 95,3% έναντι 94%) [52], και οι τυποποιημένες νομικές εξετάσεις πολλαπλής επιλογής (GPT-4 περίπου στο 90ό εκατοστημόριο) [48]. Σε αυτό το επίπεδο, η ισοδυναμία με τον άνθρωπο είναι πραγματική αλλά περιορισμένη στη μορφή της δοκιμασίας: αφορά αναγνώριση και ανάκληση τυποποιημένης γνώσης.

Στο δεύτερο επίπεδο, τα μοντέλα επιδεικνύουν ικανοποιητική αλλά υποδεέστερη της ανθρώπινης απόδοση σε εργασίες που απαιτούν εξειδικευμένη κρίση. Στις φυσικές επιστήμες επιπέδου διδακτορικού (GPQA), το GPT-4 σημειώνει 39% έναντι 65% των ειδικών [23], ενώ στην πολυτροπική ακαδημαϊκή κατανόηση (MMMU) παραμένει χάσμα περίπου 8,1% [54]. Στο τρίτο επίπεδο, η ανθρώπινη υπεροχή είναι σαφής σε εργασίες που απαιτούν λεπτή κρίση ή λειτουργία σε πραγματικές συνθήκες: η ανίχνευση κλινικών σφαλμάτων (MEDEC), όπου οι ιατροί υπερτερούν όλων των μοντέλων [38], και η επίλυση πραγματικών ζητημάτων

λογισμικού (SWE-bench), όπου χωρίς εξειδικευμένες τεχνικές η απόδοση περιορίζεται στο 1,96% [31].

Ένα ιδιαίτερα σημαντικό συγκριτικό εύρημα είναι ότι η ίδια τεχνολογική βάση παράγει δραστικά διαφορετικά αποτελέσματα ανάλογα με τον τρόπο εφαρμογής της. Το παράδειγμα του SWE-bench, όπου η ενσωμάτωση agentic methods αυξάνει την απόδοση από 1,96% σε 71,7% [31], καταδεικνύει ότι η συγκριτική αξιολόγηση των πεδίων πρέπει να λαμβάνει υπόψη όχι μόνο το μοντέλο αλλά και το μεθοδολογικό πλαίσιο εφαρμογής του. Συνεπώς, η ιεραρχία των πεδίων δεν είναι στατική, αλλά εξαρτάται από την ωριμότητα των τεχνικών που εφαρμόζονται σε κάθε τομέα.

Τέλος, η συγκριτική εικόνα διαφοροποιείται συστηματικά ανάλογα με τη γλώσσα. Σε όλα τα πεδία όπου εξετάστηκε η πολυγλωσσική διάσταση, παρατηρείται σημαντική πτώση της απόδοσης σε γλώσσες χαμηλών πόρων, με χαρακτηριστικό παράδειγμα τη μείωση από 72% στα αγγλικά σε 44% στα αραβικά για ταυτόσημες ιατρικές περιπτώσεις [41]. Το εύρημα αυτό υποδηλώνει ότι η ιεραρχία απόδοσης που προκύπτει από τα κυρίως αγγλόφωνα διεθνή benchmarks ενδέχεται να μην αντανακλά την πραγματική εμπειρία χρηστών σε άλλα γλωσσικά περιβάλλοντα, ζήτημα με άμεση σημασία για το ελληνικό πλαίσιο.

6.2 Παράγοντες που Εξηγούν τις Διαφορές

Πέρα από την καταγραφή των παραγόντων διακύμανσης που προηγήθηκε, το ερώτημα που τίθεται εδώ είναι ερμηνευτικό: γιατί αυτοί οι παράγοντες παράγουν τα συγκεκριμένα μοτίβα. Η σύνθεση των ευρημάτων υποδεικνύει ότι οι επιμέρους παράγοντες δεν δρουν ανεξάρτητα, αλλά συγκλίνουν σε μια κοινή υποκείμενη αρχή που αφορά τη φύση της μάθησης των LLMs.

Η κεντρική ερμηνεία που προκύπτει είναι ότι τα LLMs αποδίδουν βέλτιστα όταν η εργασία μπορεί να αναχθεί σε αναγνώριση στατιστικών μοτίβων που υπάρχουν επαρκώς στα δεδομένα εκπαίδευσης. Αυτό εξηγεί γιατί η απόδοση είναι υψηλή σε τυποποιημένα προβλήματα με άφθονα παραδείγματα (όπως το GSM8K) και χαμηλή σε προβλήματα που απαιτούν πρωτότυπη σύνθεση (όπως το CHAMP, στο 58,1%) [26], [27]. Η διάκριση αυτή δεν είναι απλώς ζήτημα δυσκολίας, αλλά ζήτημα του κατά πόσον η λύση απαιτεί παρεμβολή εντός γνωστών μοτίβων ή προέκταση πέρα από αυτά.

Η ερμηνεία αυτή φωτίζει και τον ρόλο του fine-tuning. Η σημαντική βελτίωση που επιφέρει η εξειδίκευση σε δεδομένα πεδίου (όπως στο Med-PaLM 2, με 86,5% στο MedQA) [33] δεν συνιστά απόκτηση νέας συλλογιστικής ικανότητας, αλλά πυκνότερη αντιπροσώπευση των σχετικών μοτίβων στα δεδομένα εκπαίδευσης. Συνεπώς, η εξάρτηση από τα δεδομένα και η γλωσσική ανισότητα αποτελούν δύο όψεις του ίδιου φαινομένου: όπου τα μοτίβα είναι πυκνά αντιπροσωπευμένα (αγγλικά, τυποποιημένα πεδία), η απόδοση είναι υψηλή· όπου είναι αραιά (γλώσσες χαμηλών πόρων, εξειδικευμένη κρίση), η απόδοση μειώνεται.

Η ίδια αρχή ερμηνεύει και την αξιοσημείωτη επίδραση των agentic methods. Η αύξηση της απόδοσης στο SWE-bench από 1,96% σε 71,7% [31] δεν προέρχεται από βελτίωση του υποκείμενου μοντέλου, αλλά από την αναδιάρθρωση της εργασίας σε μια ακολουθία μικρότερων βημάτων, καθένα από τα οποία προσεγγίζει περισσότερο τα μοτίβα που το μοντέλο χειρίζεται αξιόπιστα. Αυτό υποδηλώνει ότι ένα μέρος των παρατηρούμενων περιορισμών δεν είναι εγγενές στα μοντέλα, αλλά συνδέεται με τον τρόπο διατύπωσης του προβλήματος. Η διάκριση αυτή έχει σημασία για την ερμηνεία της βιβλιογραφίας, καθώς αποτελέσματα που παρουσιάζονται ως όρια ικανότητας ενδέχεται στην πραγματικότητα να αντανακλούν όρια της εφαρμοζόμενης μεθόδου.

Συνολικά, η σύνθεση των παραγόντων οδηγεί σε ένα ενοποιημένο ερμηνευτικό σχήμα: η απόδοση των LLMs σε ένα δεδομένο πεδίο είναι συνάρτηση της πυκνότητας των σχετικών μοτίβων στα δεδομένα εκπαίδευσης, της δυνατότητας αναγωγής της εργασίας σε τέτοια μοτίβα, και της καταλληλότητας του μεθοδολογικού πλαισίου εφαρμογής. Το σχήμα αυτό έχει προβλεπτική αξία: επιτρέπει να εκτιμηθεί εκ των προτέρων σε ποια νέα πεδία τα LLMs είναι πιθανό να αποδώσουν ικανοποιητικά και σε ποια όχι, ζήτημα κρίσιμο για την υπεύθυνη ενσωμάτωσή τους σε επαγγελματικά και ακαδημαϊκά πλαίσια.

6.3 Απειλές Εγκυρότητας (Threats to Validity)

Η ορθή ερμηνεία των ευρημάτων προϋποθέτει την αναγνώριση των περιορισμών της ίδιας της ανασκόπησης. Ακολουθώντας την καθιερωμένη πρακτική των συστηματικών ανασκοπήσεων, οι απειλές εγκυρότητας ταξινομούνται σε τέσσερις κατηγορίες: μεροληψία επιλογής, ταχεία παλαίωση των ευρημάτων, ετερογένεια των πρωτογενών μελετών, και περιορισμοί που κληρονομούνται από τις ίδιες τις πρωτογενείς μελέτες.

Η πρώτη απειλή αφορά τη μεροληψία επιλογής (selection bias), η οποία λειτουργεί σε δύο επίπεδα. Σε επίπεδο αναζήτησης, ο περιορισμός σε συγκεκριμένες βάσεις δεδομένων και σε αγγλόφωνες εργασίες ενδέχεται να απέκλεισε σχετικές μελέτες σε άλλες γλώσσες, ενώ η τάση δημοσίευσης θετικών αποτελεσμάτων (publication bias) μπορεί να οδηγεί σε υπερεκτίμηση των δυνατοτήτων των μοντέλων. Σε ένα βαθύτερο και κρισιμότερο επίπεδο, ωστόσο, τα ίδια τα benchmarks που χρησιμοποιούνται στις πρωτογενείς μελέτες είναι στη συντριπτική τους πλειονότητα αγγλόφωνα (MMLU, MATH, HumanEval, MedQA) [19], [25], [30]. Αυτό σημαίνει ότι η συνολική εικόνα απόδοσης που προκύπτει από την ανασκόπηση είναι εγγενώς μεροληπτική υπέρ της αγγλικής γλώσσας. Όπου η πολυγλωσσική διάσταση εξετάστηκε ρητά, η απόδοση σε γλώσσες χαμηλών πόρων ήταν σημαντικά χαμηλότερη (π.χ. 44% στα αραβικά έναντι 72% στα αγγλικά) [41]· επομένως, η πραγματική απόδοση σε γλώσσες με περιορισμένη αντιπροσώπευση, όπως η ελληνική, είναι πιθανό να υπολείπεται των τιμών που καταγράφουν τα διεθνή benchmarks. Ο αναγνώστης οφείλει να ερμηνεύει τα αναφερόμενα ποσοστά υπό το πρίσμα αυτού του γλωσσικού περιορισμού.

Η δεύτερη και ίσως σοβαρότερη απειλή είναι η ταχεία παλαίωση των ευρημάτων. Το πεδίο των LLMs εξελίσσεται με ρυθμό που καθιστά πολλά αποτελέσματα παρωχημένα εντός μηνών από τη δημοσίευσή τους. Η ίδια η ιστορία του MMLU, όπου η απόδοση μεταβλήθηκε από 43,9% σε άνω του 88% εντός λίγων ετών [19], καταδεικνύει ότι κάθε στιγμιαία αποτύπωση της απόδοσης έχει περιορισμένη χρονική ισχύ. Νέες εκδόσεις μοντέλων κυκλοφορούν σε διαστήματα ολίγων μηνών, συχνά ανατρέποντας προηγούμενες κατατάξεις. Συνεπώς, τα συμπεράσματα της παρούσας ανασκόπησης θα πρέπει να ερμηνεύονται ρητά ως αποτύπωση της κατάστασης κατά την περίοδο διεξαγωγής της και όχι ως διαχρονικά σταθερές διαπιστώσεις· ορισμένα από τα επιμέρους αριθμητικά στοιχεία ενδέχεται να έχουν ήδη ξεπεραστεί κατά τον χρόνο ανάγνωσης.

Η τρίτη απειλή αφορά την ετερογένεια των πρωτογενών μελετών. Οι εργασίες που εντάχθηκαν χρησιμοποιούν διαφορετικές εκδόσεις μοντέλων, διαφορετικές παραμέτρους, και διαφορετικά πρωτόκολλα αξιολόγησης. Η ετερογένεια αυτή καθιστά τις άμεσες συγκρίσεις μεταξύ μελετών προβληματικές και απέτρεψε τη διενέργεια ποσοτικής μετα-ανάλυσης. Για τον λόγο αυτό, η σύνθεση των ευρημάτων είναι ποιοτική και τα αναφερόμενα ποσοστά θα πρέπει να θεωρούνται ενδεικτικά τάσεων παρά αυστηρά συγκρίσιμα μεγέθη.

Η τέταρτη απειλή αφορά περιορισμούς που κληρονομούνται από τις πρωτογενείς μελέτες. Η μόλυνση δεδομένων (data contamination), όπου δεδομένα ελέγχου ενδέχεται να έχουν περιληφθεί στα δεδομένα εκπαίδευσης, μπορεί να έχει οδηγήσει σε υπερεκτίμηση της

απόδοσης σε ορισμένα benchmarks [39]. Παρομοίως, η ευαισθησία των μοντέλων στη διατύπωση των prompts σημαίνει ότι τα αναφερόμενα αποτελέσματα ενδέχεται να μην είναι πλήρως αναπαραγώγιμα υπό διαφορετικές συνθήκες [40]. Η παρούσα ανασκόπηση δεν μπορεί να διορθώσει αυτούς τους περιορισμούς των πρωτογενών μελετών, αλλά τους λαμβάνει υπόψη κατά την ερμηνεία, αποφεύγοντας υπερβολικά κατηγορηματικά συμπεράσματα και επισημαίνοντας τις περιπτώσεις όπου η αβεβαιότητα είναι αυξημένη.

Η συστηματική επισκόπηση 283 benchmarks από τους Ni et al. επιβεβαιώνει ότι τα ζητήματα της μόλυνσης δεδομένων, του κορεσμού και της μεροληψίας είναι διάχυτα στο σύνολο των διαθέσιμων benchmarks, και όχι μεμονωμένα φαινόμενα [56]. Η διαπίστωση αυτή ενισχύει την ανάγκη επιφυλακτικής ερμηνείας των αναφερόμενων επιδόσεων σε ολόκληρη την παρούσα ανασκόπηση.

Η αναγνώριση των περιορισμών αυτών δεν αναιρεί την αξία των ευρημάτων, αλλά οριοθετεί την εμβέλειά τους και υποδεικνύει κατευθύνσεις για μελλοντική έρευνα. Ειδικότερα, το γλωσσικό κενό που αναδείχθηκε καθιστά σαφές ότι η μελλοντική έρευνα θα πρέπει να επικεντρωθεί στην ανάπτυξη εξειδικευμένων benchmarks στην ελληνική γλώσσα —ιδίως σε πεδία υψηλής διακινδύνευσης όπως η ιατρική και η νομική, όπου η γλωσσική και η θεσμική ιδιαιτερότητα είναι κρίσιμες— ώστε η αξιολόγηση των LLMs να αντανακλά την πραγματική τους χρησιμότητα για τους Έλληνες χρήστες και όχι μια εικόνα προερχόμενη αποκλειστικά από αγγλόφωνα δεδομένα. Οι κατευθύνσεις αυτές αναπτύσσονται περαιτέρω στο επόμενο κεφάλαιο.

7 Συμπεράσματα και Μελλοντική Έρευνα

Το τελευταίο κεφάλαιο συνοψίζει τη συνεισφορά της εργασίας, απαντά ρητά στα ερευνητικά ερωτήματα που τέθηκαν στην εισαγωγή, και προσδιορίζει κατευθύνσεις για μελλοντική έρευνα. Η δομή ακολουθεί την αντίστροφη πορεία της εισαγωγής: από τη συνολική σύνοψη, στις στοχευμένες απαντήσεις, και τέλος στα ανοιχτά ζητήματα που η ανασκόπηση ανέδειξε.

7.1 Σύνοψη

Η παρούσα εργασία διενήργησε μια συστηματική βιβλιογραφική ανασκόπηση 43 επιστημονικών μελετών (peer-reviewed και αξιόπιστα preprints), με στόχο τη χαρτογράφηση της αποτελεσματικότητας των Μεγάλων Γλωσσικών Μοντέλων σε διαφορετικά επιστημονικά πεδία. Ακολουθώντας το πρωτόκολλο PRISMA, η ανασκόπηση κάλυψε πέντε κύρια πεδία: θετικές επιστήμες και μαθηματικά, πληροφορική, ιατρική, νομική και κοινωνικές επιστήμες.

Το κεντρικό συμπέρασμα της εργασίας είναι ότι η απόδοση των LLMs δεν είναι ομοιόμορφη, αλλά διαβαθμίζεται συστηματικά ανάλογα με τη φύση της εργασίας, τη διαθεσιμότητα δεδομένων εκπαίδευσης, και το μεθοδολογικό πλαίσιο εφαρμογής. Τα μοντέλα προσεγγίζουν ή υπερβαίνουν την ανθρώπινη απόδοση σε τυποποιημένες εργασίες αναγνώρισης και ανάκλησης γνώσης, αλλά υστερούν σε εργασίες που απαιτούν λεπτή κρίση, πρωτότυπο συλλογισμό, ή λειτουργία σε πραγματικές συνθήκες. Δεν υπάρχει, συνεπώς, ένα ενιαίο μοντέλο που να κυριαρχεί καθολικά· η καταλληλότητα ενός LLM είναι συνάρτηση του συγκεκριμένου πεδίου και της εργασίας.

Δευτερευόντως, η εργασία ανέδειξε δύο διατομεακά ζητήματα με ιδιαίτερη σημασία. Το πρώτο είναι η καθοριστική επίδραση του μεθοδολογικού πλαισίου, όπως κατέδειξε η περίπτωση των agentic methods. Το δεύτερο είναι η συστηματική γλωσσική ανισότητα, με την απόδοση σε γλώσσες χαμηλών πόρων να υπολείπεται σταθερά εκείνης στα αγγλικά, εύρημα με άμεση σημασία για την αξιοποίηση των LLMs στο ελληνικό περιβάλλον.

7.2 Απαντήσεις στα Ερευνητικά Ερωτήματα

Με βάση τη σύνθεση των ευρημάτων, δίνονται στη συνέχεια ρητές απαντήσεις στα πέντε ερευνητικά ερωτήματα που τέθηκαν στην εισαγωγή.

RQ1 — Σε ποια επιστημονικά πεδία έχει αξιολογηθεί η απόδοση των LLMs και με ποια benchmarks; Η ανασκόπηση εντόπισε αξιολογήσεις σε πέντε κύρια πεδία. Τα κυριότερα benchmarks ανά πεδίο είναι: για τη γενική γνώση το MMLU και το BIG-Bench [19], [22]· για τα μαθηματικά τα MATH, GSM8K, CHAMP και OlympicArena [25], [26], [27], [28]· για την πληροφορική τα HumanEval και SWE-bench [30], [31]· για την ιατρική τα MedQA, MEDEC και το NEJM Image Challenge [33], [38]· για τη νομική το Uniform Bar Exam και το LegalBench [48], [49]· και για τις κοινωνικές επιστήμες τα WinoGrande, CMMLU και MMMU [53], [52], [54].

RQ2 — Πώς διαφοροποιείται η απόδοση μεταξύ διαφορετικών πεδίων; Η απόδοση διαβαθμίζεται σε τρία επίπεδα. Σε τυποποιημένες εργασίες τα μοντέλα προσεγγίζουν την ανθρώπινη επίδοση (MMLU άνω του 88%, GSM8K και MATH περίπου 94%) [19], [25], [26]. Σε εργασίες εξειδικευμένης κρίσης υστερούν (GPQA: 39% έναντι 65% των ειδικών) [23]. Σε εργασίες πραγματικών συνθηκών η ανθρώπινη υπεροχή είναι σαφής (MEDEC, SWE-bench με 1,96% χωρίς πρόσθετες τεχνικές) [31], [38]. Η διαφοροποίηση συνδέεται με τον βαθμό στον οποίο η εργασία απαιτεί παρεμβολή εντός γνωστών μοτίβων έναντι προέκτασης πέραν αυτών.

RQ3 — Ποιοι παράγοντες εξηγούν τις διαφορές; Εντοπίστηκαν πέντε παράγοντες: ο όγκος και η ποιότητα των δεδομένων εκπαίδευσης, ο βαθμός τυποποίησης του πεδίου, η πολυπλοκότητα και ο τύπος του απαιτούμενου συλλογισμού, οι εφαρμοζόμενες τεχνικές, και η γλωσσική αντιπροσώπευση. Επισημαίνεται ότι η παρούσα εργασία δεν αναλύει άμεσα τα ίδια τα σύνολα δεδομένων εκπαίδευσης των μοντέλων (τα οποία είναι συχνά μη δημοσιοποιημένα), αλλά συνάγει τον ρόλο τους έμμεσα, μέσα από τα ευρήματα των πρωτογενών μελετών της ανασκόπησης (όπως η σύγκριση εξειδικευμένων μοντέλων τύπου Med-PaLM και BloombergGPT έναντι γενικού σκοπού μοντέλων). Η σύνθεση των παραγόντων αυτών υποδεικνύει μια κοινή υποκείμενη αρχή: η απόδοση είναι συνάρτηση της πυκνότητας των σχετικών μοτίβων στα δεδομένα εκπαίδευσης. Το fine-tuning σε δεδομένα πεδίου βελτιώνει την απόδοση κατά 20-40% [33], ενώ η ανεπαρκής γλωσσική αντιπροσώπευση τη μειώνει αντίστοιχα [41].

RQ4 — Ποιες τεχνικές βελτιώνουν την απόδοση ανά πεδίο; Τρεις κατηγορίες τεχνικών αναδείχθηκαν ως αποτελεσματικές. Η προτροπή αλυσίδας σκέψης (chain-of-thought)

βελτιώνει τις συλλογιστικές εργασίες, όπως στην περίπτωση του WinoGrande [52]. Το fine-tuning σε εξειδικευμένα δεδομένα επιφέρει σημαντική βελτίωση, όπως κατέδειξε το Med-PaLM 2 [33]. Οι agentic methods έχουν τη μεγαλύτερη επίδραση σε εργασίες πραγματικών συνθηκών, αυξάνοντας την απόδοση στο SWE-bench από 1,96% σε 71,7% [31]. Το εύρημα αυτό υποδεικνύει ότι το μεθοδολογικό πλαίσιο εφαρμογής είναι συχνά εξίσου καθοριστικό με την επιλογή του μοντέλου.

RQ5 — Ποιοι μεθοδολογικοί περιορισμοί επηρεάζουν την αξιοπιστία; Τέσσερις περιορισμοί κρίνονται κρίσιμοι. Η μόλυνση δεδομένων (data contamination) ενδέχεται να οδηγεί σε υπερεκτίμηση της απόδοσης [39]. Ο κορεσμός των benchmarks (saturation) μειώνει τη διακριτική τους ικανότητα. Η ευαισθησία στη διατύπωση των prompts επηρεάζει τηναναπαραγωγιμότητα [40]. Τέλος, η ετερογένεια των μεθοδολογιών μεταξύ μελετών δυσχεραίνει τις άμεσες συγκρίσεις και τη συγκέντρωση των ευρημάτων σε ποσοτική μετα-ανάλυση.

7.3 Μελλοντικές Κατευθύνσεις

Πριν τη διατύπωση των μελλοντικών κατευθύνσεων, οφείλει να επισημανθεί ένας θεμελιώδης περιορισμός που πλαισιώνει το σύνολο των συμπερασμάτων: η παρούσα εργασία βασίστηκε στα διαθέσιμα benchmarks και μοντέλα της περιόδου 2024-2025, ένα πεδίο που εξελίσσεται ραγδαία. Συνεπώς, η εμφάνιση νεότερων και ισχυρότερων μοντέλων ενδέχεται να μεταβάλει ορισμένα από τα επιμέρους ευρήματα, αν και τα διαρθρωτικά μοτίβα που εντοπίστηκαν (η διαβάθμιση της απόδοσης, η σημασία του μεθοδολογικού πλαισίου, και η γλωσσική ανισότητα) αναμένεται να παραμείνουν σχετικά σταθερά. Υπό το πρίσμα αυτό, η ανασκόπηση ανέδειξε αρκετά ανοιχτά ζητήματα που οριοθετούν κατευθύνσεις για μελλοντική έρευνα, οργανωμένα σε τρεις άξονες: μεθοδολογία αξιολόγησης, γλωσσική και πολιτισμική επέκταση, και θεωρητική κατανόηση.

Στον άξονα της μεθοδολογίας αξιολόγησης, προτεραιότητα αποτελεί η ανάπτυξη ενός πιο ενιαίου και διαχρονικά σταθερού πλαισίου. Τα τρέχοντα benchmarks πάσχουν από κορεσμό, μόλυνση δεδομένων, και ετερογένεια πρωτοκόλλων [39]. Η μελλοντική έρευνα θα πρέπει να επιδιώξει benchmarks ανθεκτικά στη μόλυνση, με συνεπή καταγραφή ανθρώπινων τιμών αναφοράς, ώστε οι συγκρίσεις να είναι αξιόπιστες και επαναλήψιμες. Παράλληλα, απαιτείται

συστηματικότερη μελέτη των agentic methods, των οποίων η επίδραση αποδείχθηκε καθοριστική αλλά παραμένει ανεπαρκώς κατανοητή [31].

Στον άξονα της γλωσσικής και πολιτισμικής επέκτασης εντοπίζεται το σημαντικότερο κενό που ανέδειξε η εργασία. Η συντριπτική πλειονότητα των υφιστάμενων benchmarks εστιάζει στην αγγλόφωνη γενική γνώση, αφήνοντας ανεπαρκώς αξιολογημένη τη συλλογιστική ικανότητα σε εξειδικευμένα και μη-αγγλόφωνα περιβάλλοντα. Η παρούσα εργασία ανέδειξε, επομένως, την ανάγκη για benchmarks που να εστιάζουν λιγότερο στην αγγλόφωνη γενική γνώση και περισσότερο στη συλλογιστική ικανότητα σε εξειδικευμένα, μη-αγγλόφωνα νομικά και ιατρικά περιβάλλοντα. Η ανάπτυξη τέτοιων πόρων στην ελληνική γλώσσα είναι ιδιαίτερα κρίσιμη σε πεδία υψηλής διακινδύνευσης, όπου τόσο η γλωσσική όσο και η θεσμική ιδιαιτερότητα —για παράδειγμα, το ελληνικό νομικό σύστημα έναντι του αγγλοσαξονικού - καθιστούν τα διεθνή benchmarks ανεπαρκή. Η δημιουργία τους θα επέτρεπε μια ρεαλιστική αποτίμηση της χρησιμότητας των LLMs για τους Έλληνες χρήστες, αντί για μια εικόνα προερχόμενη αποκλειστικά από αγγλόφωνα δεδομένα γενικής γνώσης [41].

Στο ίδιο πλαίσιο επέκτασης σε νέα πεδία, ιδιαίτερο ενδιαφέρον παρουσιάζουν τα χρηματοοικονομικά και επιχειρηματικά πεδία, τόσο λόγω της σημασίας τους για την οικονομική δραστηριότητα όσο και λόγω της στενής σύνδεσής τους με το προφίλ του Πανεπιστημίου Πειραιώς. Το BloombergGPT, για παράδειγμα, είναι ένα μεγάλο γλωσσικό μοντέλο που έχει εκπαιδευτεί σε εκτεταμένο σώμα χρηματοοικονομικών δεδομένων και αξιολογείται τόσο σε γενικά όσο και σε εξειδικευμένα financial benchmarks, καταδεικνύοντας ότι η προσαρμογή σε domain-specific corpora μπορεί να βελτιώσει σημαντικά την απόδοση σε σύνθετες οικονομικές εργασίες. Παράλληλα, σύνολα δεδομένων όπως το EconLogicQA εστιάζουν στον λογικό συλλογισμό σε οικονομικά σενάρια, αναδεικνύοντας τις προκλήσεις που προκύπτουν όταν τα μοντέλα καλούνται να χειριστούν ακολουθιακές αποφάσεις και όχι απλώς ανάκληση γνώσης.

Στον άξονα της θεωρητικής κατανόησης, η ανασκόπηση υποδεικνύει την ανάγκη για βαθύτερη διερεύνηση των μηχανισμών που διέπουν τη διακύμανση της απόδοσης. Το ενοποιημένο ερμηνευτικό σχήμα που προτάθηκε —ότι η απόδοση συναρτάται με την πυκνότητα των μοτίβων στα δεδομένα εκπαίδευσης— αποτελεί υπόθεση που χρήζει εμπειρικού ελέγχου. Η εξήγηση (explainability) των LLMs, δηλαδή η κατανόηση του τι πραγματικά «γνωρίζουν» έναντι του τι αναπαράγουν, παραμένει ανοιχτό ερευνητικό πεδίο με άμεσες συνέπειες για την αξιόπιστη ενσωμάτωσή τους σε κρίσιμες εφαρμογές [4].

Συμπερασματικά, τα Μεγάλα Γλωσσικά Μοντέλα συνιστούν ένα ισχυρό αλλά ανομοιόμορφο εργαλείο, του οποίου η χρησιμότητα εξαρτάται από την κριτική κατανόηση των δυνατοτήτων και των ορίων του ανά πεδίο. Η παρούσα ανασκόπηση επιχείρησε να προσφέρει μια τέτοια χαρτογράφηση, με την επίγνωση ότι σε ένα ταχέως εξελισσόμενο πεδίο κάθε αποτύπωση είναι προσωρινή. Η υπεύθυνη αξιοποίηση των LLMs προϋποθέτει όχι την άκριτη αποδοχή ούτε την απόρριψή τους, αλλά τη συνεχή, τεκμηριωμένη αξιολόγηση της απόδοσής τους στο εκάστοτε συγκεκριμένο πλαίσιο εφαρμογής.

Κατάλογος Ακρωνυμίων και Συντομογραφιών

Ακρωνύμιο	Πλήρης Όρος (Αγγλικά / Ελληνικά)
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CNNs	Convolutional Neural Networks
CoT	Chain-of-Thought
DL	Deep Learning
GPQA	Graduate-Level Google-Proof Q&A Benchmark
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
GSM8K	Grade School Math 8K Benchmark
LLM	Large Language Model
LSTM	Long Short-Term Memory
MBE	Multistate Bar Examination

ML	Machine Learning
MMLU	Massive Multitask Language Understanding
NEJM	New England Journal of Medicine
NLP	Natural Language Processing

Ακρωνύμιο	**Πλήρης Όρος (Αγγλικά / Ελληνικά)**
PaLM	Pathways Language Model
PICOC	Population, Intervention, Comparison, Outcome, Context
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval-Augmented Generation
RL	Reinforcement Learning
RLHF	Reinforcement Learning from Human Feedback
RNNs	Recurrent Neural Networks
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RQs	Research Questions
SLR	Systematic Literature Review
TPU	Tensor Processing Unit
USMLE	United States Medical Licensing Examination

Πίνακας Ορολογίας

Επιστημονικός Όρος	Ερμηνεία / Ελληνική Απόδοση
Allocational Harms	Βλάβες Κατανομής (Άνιση ή άδικη κατανομή πόρων, ευκαιριών ή υπηρεσιών από ένα σύστημα AI)
Backpropagation	Αλγόριθμος Ανάστροφης Μετάδοσης Λάθους (Η βασική μέθοδος εκπαίδευσης των νευρωνικών δικτύων)
Benchmark	Κριτήριο Αξιολόγησης / Τυποποιημένη Δοκιμασία Αναφοράς για τη μέτρηση της απόδοσης μοντέλων
Bias	Προκατάληψη / Συστημική Μεροληψία που ενυπάρχει στα δεδομένα εκπαίδευσης του μοντέλου
Classification	Ταξινόμηση (Πρόβλημα μηχανικής μάθησης όπου η έξοδος ανήκει σε διακριτές κατηγορίες)
Clustering	Ομαδοποίηση / Συστάδωση Δεδομένων χωρίς επίβλεψη βάσει κοινών χαρακτηριστικών
Cross-Domain Analysis	Διατομεακή / Διακλαδική Ανάλυση (Η σύγκριση της απόδοσης ενός μοντέλου σε διαφορετικά πεδία γνώσης)
Data Contamination	Μόλυνση Δεδομένων (Η διαρροή δεδομένων αξιολόγησης/τεστ μέσα στο σύνολο δεδομένων εκπαίδευσης του LLM)
Explainability	Επεξηγησιμότητα / Ερμηνευσιμότητα (Η δυνατότητα κατανόησης και αιτιολόγησης των αποφάσεων ενός αλγορίθμου)

Επιστημονικός Όρος	Ερμηνεία / Ελληνική Απόδοση
Factual Hallucination	Πραγματολογική Ψευδαίσθηση (Η παραγωγή ανυπόστατων ή ψευδών γεγονότων από ένα LLM)
Faithfulness Hallucination	Ψευδαίσθηση Πιστότητας (Όταν το μοντέλο αποκλίνει από το δοθέν πλαίσιο κειμένου ή τις οδηγίες του χρήστη)
Few-Shot Prompting	Προτροπή Ολιγο-στιγμιοτύπου (Η παροχή ελάχιστων παραδειγμάτων λύσης στο prompt πριν την τελική ερώτηση)
Fine-tuning	Μικρορύθμιση / Εξειδικευμένη Προσαρμογή ενός προ-εκπαιδευμένου μοντέλου σε ένα συγκεκριμένο σύνολο δεδομένων
Gradient Descent	Καθοδική Κλίση (Μαθηματική μέθοδος βελτιστοποίησης για την ελαχιστοποίηση της συνάρτησης απώλειας (loss function))
Human Evaluation	Ανθρώπινη Αξιολόγηση (Η αξιολόγηση των εξόδων ενός LLM από ανθρώπους-ειδικούς του εκάστοτε κλάδου)
Instruction Tuning	Μικρορύθμιση βάσει Οδηγιών (Εκπαίδευση του μοντέλου ώστε να απαντά σε μορφή διαλόγου και εντολών)
LLM-as-a-judge	Μεθοδολογία αξιολόγησης όπου ένα προηγμένο LLM (π.χ. GPT-4) χρησιμοποιείται ως κριτής για τις απαντήσεις άλλων μοντέλων
Low-resource Languages	Γλώσσες Χαμηλών Πόρων (Γλώσσες όπως η Ελληνική, που έχουν περιορισμένο όγκο δεδομένων στο διαδίκτυο για εκπαίδευση AI)

Επιστημονικός Όρος	Ερμηνεία / Ελληνική Απόδοση
Multi-step Reasoning	Πολυβηματικός Συλλογισμός (Η ικανότητα ενός μοντέλου να επιλύει προβλήματα που απαιτούν μια σειρά λογικών βημάτων)
Next-Token Prediction	Πρόβλεψη Επόμενου Διακριτικού (Η στατιστική διαδικασία παραγωγής κειμένου λέξη προς λέξη από τα LLMs)
Overfitting	Υπερπροσαρμογή (Όταν ένα μοντέλο αποστηθίζει τα δεδομένα εκπαίδευσης αλλά αποτυγχάνει σε νέα, άγνωστα δεδομένα)
Pass@k metric	Μετρική Pass@k (Το ποσοστό επιτυχίας παραγωγής σωστού κώδικα, όταν επιτρέπεται στο μοντέλο να κάνει k προσπάθειες)
Pre-training	Προ-εκπαίδευση (Η αρχική φάση εκπαίδευσης ενός LLM σε τεράστιο όγκο γενικών κειμένων για την εκμάθηση της γλώσσας)
Prompt Engineering	Μηχανική Προτροπών / Μηχανική Εντολών (Η στρατηγική διαμόρφωση των κειμένων εισόδου προς ένα LLM)
Regression	Παλινδρόμηση (Πρόβλημα μηχανικής μάθησης όπου η ζητούμενη μεταβλητή-στόχος παίρνει συνεχείς τιμές)
Representational Harms	Βλάβες Αναπαράστασης (Η αναπαραγωγή και ενίσχυση κοινωνικών στερεοτύπων ή υποτιμητικών αναφορών από το AI)
Self-Attention Mechanism	Μηχανισμός Αυτο-προσοχής (Ο μηχανισμός που επιτρέπει στους Transformers να συσχετίζουν κάθε λέξη με όλες τις άλλες στο κείμενο)

Επιστημονικός Όρος	Ερμηνεία / Ελληνική Απόδοση
Supervised Learning	Επιβλεπόμενη Μάθηση (Εκπαίδευση μοντέλου με χρήση επισημειωμένων δεδομένων εισόδου-εξόδου)
Token	Διακριτικό / Λεκτική Μονάδα (Η βασική μονάδα κειμένου, λέξη ή τμήμα λέξης, που επεξεργάζεται ένα γλωσσικό μοντέλο)
Transformer	Αρχιτεκτονική Μετασχηματιστή (Η επαναστατική νευρωνική αρχιτεκτονική πάνω στην οποία βασίζονται όλα τα σύγχρονα LLMs)
Unsupervised Learning	Μη Επιβλεπόμενη Μάθηση (Εκπαίδευση μοντέλου σε δεδομένα που δεν φέρουν ετικέτες ή προκαθορισμένες εξόδους)
Zero-Shot Prompting	Προτροπή Μηδενικού Στιγμιοτύπου (Η υποβολή ερωτήματος στο μοντέλο χωρίς να του δοθεί κανένα παράδειγμα λύσης)

Βιβλιογραφία

[1] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023. doi: https://doi.org/10.48550/arXiv.2303.08774

[2] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288, 2023. doi: https://doi.org/10.48550/arXiv.2307.09288

[3] Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku," *Anthropic Technical Report*, 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-family

[4] Y. Chang et al., "A Survey on Evaluation of Large Language Models," *ACM Trans. Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024. doi: https://doi.org/10.1145/3641289

[5] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," BMJ, vol. 372, art. no. n71, 2021. doi: https://doi.org/10.1136/bmj.n71

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, 2013.

[7] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020. doi: https://doi.org/10.48550/arXiv.2005.14165

[10] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language Models," arXiv:2001.08361, 2020. doi: https://doi.org/10.48550/arXiv.2001.08361

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017. doi: https://doi.org/10.48550/arXiv.1706.03762

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), vol. 1, pp. 4171–4186, 2019. doi: https://doi.org/10.18653/v1/N19-1423

[13] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3645–3650, 2019. doi: https://doi.org/10.18653/v1/P19-1355

[14] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training Language Models to Follow Instructions with Human Feedback," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022. doi: https://doi.org/10.48550/arXiv.2203.02155

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022. doi: https://doi.org/10.48550/arXiv.2201.11903

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020. doi: https://doi.org/10.48550/arXiv.2005.11401

[17] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023. doi: https://doi.org/10.1145/3571730

[18] S.E.S. Adib, A.A. Sani, E.A. Esham, and A. Abrar, "Assessing Large Language Models for Medical QA: Zero-Shot and LLM-as-a-Judge Evaluation," in *IEEE Conference on Computer and Communications*, 2025. doi: https://doi.org/10.1109/ICCIT68739.2025.11491589

[19] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring Massive Multitask Language Understanding," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021. doi: https://doi.org/10.48550/arXiv.2009.03300

[20] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan, "AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models," in *Findings of the Association for Computational Linguistics (NAACL)*, 2024. doi: https://doi.org/10.18653/v1/2024.findings-naacl.149

[21] L. Phan et al., "Humanity's Last Exam," arXiv:2501.14249, 2025. doi: https://doi.org/10.48550/arXiv.2501.14249

[22] A. Srivastava et al., "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models," *Transactions on Machine Learning Research*, 2023. doi: https://doi.org/10.48550/arXiv.2206.04615

[23] D. Rein et al., "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," in *Proc. Conf. on Language Modeling (COLM)*, 2024. doi: https://doi.org/10.48550/arXiv.2311.12022

[24] H. Cai et al., "SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis," in *Findings of the Association for Computational Linguistics (NAACL)*, 2025. doi: https://doi.org/10.48550/arXiv.2403.01976

[25] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring Mathematical Problem Solving With the MATH Dataset," in *Proc. NeurIPS (Datasets and Benchmarks Track)*, 2021. doi: https://doi.org/10.48550/arXiv.2103.03874

[26] K. Cobbe et al., "Training Verifiers to Solve Math Word Problems," arXiv:2110.14168, 2021. doi: https://doi.org/10.48550/arXiv.2110.14168

[27] Y. Mao, Y. Kim, and Y. Zhou, "CHAMP: A Competition-level Dataset for Fine-Grained Analyses of LLMs' Mathematical Reasoning Capabilities," in *Findings of the Association for Computational Linguistics (ACL)*, pp. 13256–13274, 2024. doi: https://doi.org/10.18653/v1/2024.findings-acl.785

[28] Z. Huang et al., "OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI," in *Proc. NeurIPS*, 2024. doi: https://doi.org/10.48550/arXiv.2406.12753

[29] Y. Quan and Z. Liu, "EconLogicQA: A Question-Answering Benchmark for Evaluating Large Language Models in Economic Sequential Reasoning," in *Findings of the Association*

for Computational Linguistics (EMNLP), pp. 2273–2282, 2024. doi: https://doi.org/10.48550/arXiv.2405.07938

[30] M. Chen et al., "Evaluating Large Language Models Trained on Code," arXiv:2107.03374, 2021. doi: https://doi.org/10.48550/arXiv.2107.03374

[31] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan, "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024. doi: https://doi.org/10.48550/arXiv.2310.06770

[32] K. Singhal et al., "Large Language Models Encode Clinical Knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023. doi: https://doi.org/10.1038/s41586-023-06291-2

[33] K. Singhal et al., "Toward Expert-Level Medical Question Answering with Large Language Models," *Nature Medicine*, vol. 31, pp. 943–950, 2025. doi: https://doi.org/10.1038/s41591-024-03423-7

[34] E. Jo, S. Song, J.-H. Kim, S. Lim, J.H. Kim, J.-J. Cha, Y.-M. Kim, and H.J. Joo, "Assessing GPT-4's Performance in Delivering Medical Advice: Comparative Analysis With Human Experts," *JMIR Medical Education*, vol. 10, e51282, 2024. doi: https://doi.org/10.2196/51282

[35] R. Kaczmarczyk, T.I. Wilhelm, R. Martin, and J. Roos, "Multimodal AI in Medical Diagnostics: Evaluation on the NEJM Image Challenge," *npj Digital Medicine*, vol. 7, art. no. 215, 2024. doi: https://doi.org/10.1038/s41746-024-01208-3

[36] Y. Xie, J. Wu, H. Tu, S. Yang, B. Zhao, Y. Zong, Q. Jin, C. Xie, and Y. Zhou, "A Preliminary Study of o1 in Medicine: Are We Closer to an AI Doctor?," arXiv:2409.15277, 2024. doi: https://doi.org/10.48550/arXiv.2409.15277

[37] C. Luengo Vera et al., "Evaluating Large Language Models on the Spanish Medical Intern Resident (MIR) Examination 2024/2025," arXiv:2503.00025, 2025. doi: https://doi.org/10.48550/arXiv.2503.00025

[38] A. Ben Abacha, W.-w. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, and T. Lin, "MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes," in *Findings of the Association for Computational Linguistics (ACL)*, pp. 22539–22550, 2025. doi: https://doi.org/10.48550/arXiv.2412.19260

[39] I. Jahan, M.T.R. Laskar, C. Peng, and J.X. Huang, "A comprehensive evaluation of large language models on benchmark biomedical text processing tasks," *Computers in Biology and*

Medicine, vol. 171, art. no. 108189, 2024. doi: https://doi.org/10.1016/j.compbiomed.2024.108189

[40] J. Huang and K.C.-C. Chang, "Towards Reasoning in Large Language Models: A Survey," in *Findings of the Association for Computational Linguistics (ACL)*, 2023. doi: https://doi.org/10.18653/v1/2023.findings-acl.67

[41] E.H. El Mahdaoui and Y.O. Mykhalko, "Language-dependent performance of large language models in medical diagnostics: A comparative study of ChatGPT-4o and Claude 3.5 Sonnet," *Polski Merkuriusz Lekarski*, 2025. doi: https://doi.org/10.36740/Merkur202506117

[42] F. Farhat, B.M. Chaudhry, M. Nadeem, S.S. Sohail, and D.Ø. Madsen, "Performance of GPT-3.5, GPT-4 and Bard on Physics, Chemistry and Biology Examination Questions," 2024. doi: https://doi.org/10.2196/51523

[43] T. Olatunji et al., "AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset," in *Proc. Association for Computational Linguistics (ACL)*, 2025. doi: https://doi.org/10.48550/arXiv.2411.15640

[44] D. Stribling, Y. Xia, M.K. Amer, K.S. Graim, C.J. Mulligan, and R. Renne, "Evaluating GPT-4 on Graduate-Level Biomedical Science Examinations," *Scientific Reports*, vol. 14, art. no. 4241, 2024. doi: https://doi.org/10.1038/s41598-024-55568-7

[45] H. Yang, M. Li, H. Zhou, Y. Xiao, Q. Fang, S. Zhou, and R. Zhang, "Large Language Model Synergy for Ensemble Learning in Medical Question Answering," 2024. doi: https://doi.org/10.2196/70080

[46] M. Rosoł, J.S. Gąsior, J. Łaba, K. Korzeniewski, and M. Młyńczak, "Evaluation of the Performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination," *Scientific Reports*, vol. 13, art. no. 20512, 2023. doi: https://doi.org/10.1038/s41598-023-46995-z

[47] M. Bommarito II and D. M. Katz, "GPT Takes the Bar Exam," arXiv:2212.14402, 2022. doi: https://doi.org/10.48550/arXiv.2212.14402

[48] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 Passes the Bar Exam," *Philosophical Transactions of the Royal Society A*, vol. 382, no. 2270, 2024. doi: https://doi.org/10.1098/rsta.2023.0254

[49] N. Guha et al., "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models," in *Proc. NeurIPS (Datasets and Benchmarks Track)*, 2023. doi: https://doi.org/10.48550/arXiv.2308.11462

[50] J. Savelka, K. D. Ashley, M. A. Gray, H. Westermann, and H. Xu, "Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise?," arXiv:2306.13906, 2023. doi: https://doi.org/10.48550/arXiv.2306.13906

[51] R. Shui, Y. Cao, X. Wang, and T.-S. Chua, "Large Language Models in Legal Judgment Prediction: A Comparative Study," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. doi: https://doi.org/10.18653/v1/2023.findings-emnlp.490

[52] K. Sakaguchi, R. Le Bras, C. Bhagavatula, and Y. Choi, "WinoGrande: An Adversarial Winograd Schema Challenge at Scale," *Communications of the ACM*, vol. 64, no. 9, pp. 99–106, 2021. doi: https://doi.org/10.1145/3474381

[53] H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, Y. Gong, N. Duan, and T. Baldwin, "CMMLU: Measuring Massive Multitask Language Understanding in Chinese," in *Findings of the Association for Computational Linguistics (ACL)*, 2024. doi: https://doi.org/10.18653/v1/2024.findings-acl.671

[54] X. Yue et al., "MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: https://doi.org/10.48550/arXiv.2311.16502

[55] S. Wu et al., "BloombergGPT: A Large Language Model for Finance," arXiv:2303.17564, 2023. doi: https://doi.org/10.48550/arXiv.2303.17564

[56] S. Ni, G. Chen, S. Li, X. Chen, S. Li, B. Wang et al., "A Survey on Large Language Model Benchmarks," arXiv:2508.15361, 2025. doi: https://doi.org/10.48550/arXiv.2508.15361