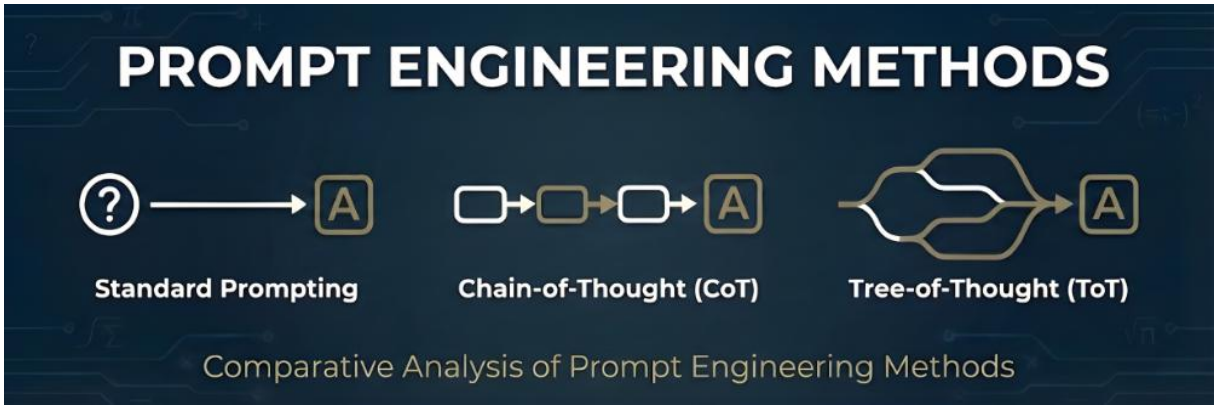
ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΠΡΟΤΡΟΠΩΝ: STANDARD, CHAIN-OF-THOUGHT, TREE-OF-THOUGHT»



Των φοιτητών
Μπαλτήρα Αγοραστού-Κωνσταντίνου
Αρ. Μητρώου:21/2024
Χαροκόπου Μιχαήλ
Αρ. Μητρώου:32/2024

Επιβλέπων
Διαμαντάρας Κωνσταντίνος
Καθηγητής

`

Θεσσαλονίκη, 11 Φεβρουαρίου 2026

Τίτλος Δ.Ε: Συγκριτική ανάλυση μεθόδων μηχανικής προτροπών: Standard, Chain-of-Thought, Tree-of-Thought

Κωδικός Δ.Ε.: 25296

Ονοματεπώνυμο φοιτητών: Μπαλτήρας Αγοραστός-Κωνσταντίνος, Χαροκόπος Μιχαήλ

Ονοματεπώνυμο εισηγητή: Κωνσταντίνος Διαμαντάρας

Ημερομηνία ανάληψης Δ.Ε. 13 Σεπτεμβρίου 2025

Ημερομηνία περάτωσης Δ.Ε. 11 Φεβρουαρίου 2026

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

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του Προγράμματος Μεταπτυχιακών Σπουδών «Ευφυείς Τεχνολογίες Διαδικτύου» (αγγλ. MSc in Web Intelligence) και πραγματεύεται το δυναμικά εξελισσόμενο πεδίο της μηχανικής προτροπών (αγγλ. Prompt Engineering) στα Μεγάλα Γλωσσικά Μοντέλα (αγγλ. Large Language Models, LLMs). Η επιλογή του συγκεκριμένου θέματος προέκυψε κατόπιν προτροπής του επιβλέποντα καθηγητή κ. Διαμαντάρα Κωνσταντίνου, καθώς και από το έντονο επιστημονικό ενδιαφέρον των φοιτητών για τις σύγχρονες εξελίξεις στον χώρο της τεχνητής νοημοσύνης. Ειδικότερος στόχος αποτέλεσε η διερεύνηση των μεθόδων που ενισχύουν τη συλλογιστική ικανότητα και την επίλυση σύνθετων προβλημάτων από τα γλωσσικά μοντέλα, αξιοποιώντας τις εγγενείς τους δυνατότητες χωρίς την απαίτηση περαιτέρω εκπαίδευσης.

Κατά την εκπόνηση της μελέτης, πραγματοποιήθηκε εις βάθος επισκόπηση και πρακτική εφαρμογή σύγχρονων ερευνητικών προσεγγίσεων, όπως η Τυπική Προτροπή (αγγλ. Standard Prompting), η Αλυσίδα Συλλογισμού (αγγλ. Chain-of-Thought) και η Δενδρική Συλλογιστική (αγγλ. Tree-of-Thought). Μέσω της διαδικασίας αυτής, καλλιεργήθηκαν ουσιαστικές δεξιότητες που αφορούν την κριτική ανάλυση της διεθνούς βιβλιογραφίας, τον σχεδιασμό πειραματικών διαδικασιών και την ερμηνεία ποσοτικών και ποιοτικών αποτελεσμάτων.

Η παρούσα εργασία φιλοδοξεί να συμβάλει στη βαθύτερη κατανόηση των δυνατοτήτων και των περιορισμών των τεχνικών προτροπής, προσφέροντας ένα συστηματικό και τεκμηριωμένο πλαίσιο μελέτης που μπορεί να αξιοποιηθεί σε ακαδημαϊκό και ερευνητικό επίπεδο.

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στη συγκριτική ανάλυση και αξιολόγηση τριών θεμελιωδών μεθόδων μηχανικής προτροπών (αγγλ. Prompt Engineering) στα Μεγάλα Γλωσσικά Μοντέλα (αγγλ. Large Language Models, LLMs): της Τυπικής Προτροπής (αγγλ. Standard Prompting), της Αλυσίδας Συλλογισμού (αγγλ. Chain-of-Thought, CoT) και της Δενδρικής Συλλογιστικής (αγγλ. Tree-of-Thought, ToT). Κύριος στόχος της έρευνας αποτελεί η διερεύνηση της επίδρασης που ασκεί η μετάβαση από τις απλές, άμεσες οδηγίες σε δομημένες στρατηγικές συλλογισμού, ως προς την ικανότητα των μοντέλων να επιλύουν προβλήματα αυξημένης πολυπλοκότητας.

Για την πειραματική αποτίμηση των μεθόδων στο καθιερωμένο σύνολο δεδομένων μαθηματικών προβλημάτων GSM8K, υιοθετήθηκε ένα σύνολο ποσοτικών μετρικών, το οποίο περιλαμβάνει την ακρίβεια (αγγλ. Accuracy) των παραγόμενων απαντήσεων, τη συνέπεια (αγγλ. Consistency), τον χρόνο απόκρισης (αγγλ. Latency) και το υπολογιστικό κόστος (αγγλ. Tokens). Παράλληλα, εξετάστηκαν διεξοδικά τα φαινόμενα αποχής (αγγλ. Abstention) των μοντέλων, συνδέοντας τα ποσοστά αποτυχίας με τους περιορισμούς του παραθύρου κειμένου (αγγλ. context window) και των διαθέσιμων πόρων. Για την ποιοτική επικύρωση της συλλογιστικής πορείας, εφαρμόστηκε συμπληρωματικά η μεθοδολογία «LLM-as-a-Judge».

Τα ευρήματα της μελέτης τεκμηριώνουν ότι η μέθοδος CoT επιφέρει σαφή βελτίωση στην επίλυση λογικών και πολυσταδιακών προβλημάτων έναντι της Standard Prompting. Παράλληλα, η μέθοδος ToT επιτυγχάνει τα βέλτιστα επίπεδα ακρίβειας σε σενάρια που απαιτούν τη διερεύνηση πολλαπλών εναλλακτικών λύσεων, συνεπαγόμενη ωστόσο σημαντική επιβάρυνση σε χρόνο εκτέλεσης και κατανάλωση tokens. Επιπροσθέτως, η ανάλυση κατέδειξε ότι η επάρκεια του παραθύρου κειμένου (αγγλ. context window) αποτελεί κατά περίπτωση κρίσιμη παράμετρο, καθώς σε συγκεκριμένα μοντέλα οι αυστηροί περιορισμοί μνήμης οδήγησαν σε αδυναμία ολοκλήρωσης του συλλογισμού, και κατ' επέκταση, σε υψηλά ποσοστά αποχής.

Συμπερασματικά, η εργασία αναδεικνύει τη σημασία της επιλογής κατάλληλης μεθοδολογίας προτροπής για τη βελτιστοποίηση της απόδοσης των LLMs, καταδεικνύοντας ότι οι δομημένες στρατηγικές συλλογισμού υπερέχουν σταθερά σε σύνθετες εργασίες.

«Comparative Analysis of Prompt Engineering Methods: Standard, Chain-of-Thought, and Tree-of-Thought»

« Baltiras Agorastos-Konstantinos, Charokopos Michail»

Abstract

This thesis focuses on the comparative analysis and evaluation of three fundamental Prompt Engineering methodologies in Large Language Models (LLMs): Standard Prompting, Chain-of-Thought (CoT), and Tree-of-Thought (ToT). The primary objective of the research is to investigate the impact of transitioning from simple, direct instructions to structured reasoning strategies on the models' ability to solve problems of increased complexity.

For the experimental assessment of these methods on the standard GSM8K math word problem benchmark, a set of quantitative metrics was adopted, including the Accuracy of the generated responses, Consistency, response time (Latency), and computational cost (Tokens). Concurrently, model abstention phenomena were thoroughly examined, linking failure rates to context window limitations and available resources. To qualitatively validate the reasoning process, the "LLM-as-a-Judge" methodology was complementarily applied.

The study's findings substantiate that the Chain-of-Thought method offers a clear improvement in solving logical and multi-step problems compared to Standard Prompting. At the same time, the Tree-of-Thought method achieves optimal accuracy levels in scenarios requiring the exploration of multiple alternative solutions, albeit with a significant burden on execution time and token consumption. Additionally, the analysis demonstrated that context window sufficiency constitutes, in certain cases, a critical parameter, as in specific models strict memory constraints led to an inability to complete the reasoning process and, by extension, to high abstention rates.

In conclusion, this thesis highlights the importance of selecting the appropriate prompting methodology for optimizing LLM performance, demonstrating that structured reasoning strategies consistently outperform standard methods in complex tasks.

Ευχαριστίες

Θα θέλαμε να εκφράσουμε τις θερμές μας ευχαριστίες προς τον επιβλέποντα καθηγητή μας κ. Κωνσταντίνο Διαμαντάρα για την πολύτιμη καθοδήγηση, τις εύστοχες παρατηρήσεις και τη συνεχή υποστήριξή του καθ' όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας. Η συμβολή του υπήρξε καθοριστική τόσο σε επιστημονικό όσο και σε μεθοδολογικό επίπεδο.

Επιπλέον, θα θέλαμε να ευχαριστήσουμε θερμά τον κ. Καζλάρη Ιωάννη για τη βοήθεια, τις χρήσιμες συμβουλές και την υποστήριξή του, οι οποίες συνέβαλαν ουσιαστικά στην ολοκλήρωση της εργασίας.

Τέλος, θα θέλαμε να εκφράσουμε την ευγνωμοσύνη μας προς τις οικογένειές μας για την αμέριστη στήριξη, την υπομονή και την ενθάρρυνση που μας παρείχαν καθ' όλη τη διάρκεια των σπουδών μας και ιδιαίτερα κατά την εκπόνηση της παρούσας εργασίας. Η ηθική τους υποστήριξη υπήρξε πολύτιμη και καθοριστική.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Συντομογραφίες και ορισμοί όρων

Abstention Rate	Ρυθμός Αποχής
Accuracy	Ακρίβεια
Aggregation techniques	Τεχνικές συνάθροισης
Application Programming Interface (API)	Διεπαφή Προγραμματισμού Εφαρμογών
Autoregressive	Αυτοπαλίνδρομος
Backtracking	Οπισθοδρόμηση
Baseline	Μέθοδος αναφοράς
Benchmark	Πρότυπο σύνολο αναφοράς ή διαδικασία συγκριτικής αξιολόγησης
BERT	Bidirectional Encoder Representations from Transformers
Big data	Μεγάλα δεδομένα
Branch	Κλαδί
Breadth-First Search (BFS)	Αναζήτηση κατά πλάτος
Chain-of-Thought (CoT)	Αλυσίδα Συλλογισμού
Consistency	Συνέπεια
Context	Πλαίσιο
Context window	Παράθυρο συμφραζομένων
Deep Learning	Βαθιά Μάθηση
Depth-First Search (DFS)	Αναζήτηση κατά βάθος
Declarative Self-improving Language Programs (DSPy)	Δηλωτικό πλαίσιο αυτο-βελτιούμενων γλωσσικών προγραμμάτων
Emergent Abilities	Αναδυόμενες ικανότητες
Ensemble Refinement	Βελτίωση μέσω συνόλου μοντέλων
Expected Calibration Error (ECE)	Αναμενόμενο Σφάλμα Βαθμονόμησης
Exact Match (EM)	Ακριβής Ταύτιση
F1 Score (F1)	Αρμονικός μέσος Ακρίβειας και Ανάκλησης
Few-Shot Prompting	Προτροπή με λίγα παραδείγματα
Fine-tuning	Μικρορύθμιση

Gemini	Οικογένεια πολυτροπικών γλωσσικών μοντέλων της Google
Gold Answer	Χρυσή απάντηση / απάντηση αναφοράς
GPT	Generative Pre-trained Transformer
Grade School Math 8K (GSM8K)	Σύνολο δεδομένων μαθηματικού συλλογισμού
Hallucinations	Παραισθήσεις
In-Context Learning (ICL)	Μάθηση Εντός Πλαισίου (ικανότητα εκτέλεσης εργασιών μέσω παραδειγμάτων εντός της προτροπής)
Inference	Εξαγωγή συμπερασμάτων / φάση πρόβλεψης
LangChain	Πλαίσιο ανάπτυξης εφαρμογών με LLMs
LangGraph	Πλαίσιο γραφημάτων συλλογισμού για LLMs
LangSmith	Πλατφόρμα παρακολούθησης και αξιολόγησης LLM pipelines
Latency	Χρόνος Απόκρισης
Large Language Model(s) (LLM / LLMs)	Μεγάλο(α) Γλωσσικό(ά) Μοντέλο(α)
Linear chain	Γραμμική αλυσίδα
LLM-as-a-Judge	Μέθοδος αξιολόγησης όπου το LLM λειτουργεί ως αξιολογητής απαντήσεων
LLaMA	Large Language Model Meta AI
Long Short-Term Memory (LSTM)	Δίκτυο Μακράς-Βραχείας Μνήμης
Majority voting	Πλειοψηφική ψηφοφορία
Mixture of Experts (MoE)	Αρχιτεκτονική Μείγματος Ειδικών
Multimodality	Πολυτροπικότητα
Multi-Head Attention	Προσοχή πολλαπλών κεφαλών
Multi-Step Reasoning	Συλλογιστική πολλαπλών βημάτων
Multi-Task Learning	Μάθηση πολλαπλών εργασιών
Natural Language Processing (NLP)	Επεξεργασία Φυσικής Γλώσσας
Node	Κόμβος
Ollama	Πλατφόρμα τοπικής εκτέλεσης Μεγάλων Γλωσσικών Μοντέλων
OpenAI-compatible client	Πελάτης συμβατός με OpenAI
PaLM	Pathways Language Model

Pattern	Πρότυπο / μοτίβο
Pipeline	Αγωγός επεξεργασίας
Pipeline Chain-of-Thought (Pipeline-CoT)	Αλυσιδωτή Συλλογιστική σε αγωγό επεξεργασίας
Pre-training	Προεκπαίδευση
Prompt(s)	Προτροπή(ές)
Prompt Engineering (PE)	Μηχανική Προτροπών
Reasoning Path	Διαδρομή συλλογισμού
Reasoning Traces	Ίχνη συλλογισμού
Recurrent Neural Network (RNN)	Επαναλαμβανόμενο Νευρωνικό Δίκτυο
Reinforcement Learning from Human Feedback (RLHF)	Ενισχυτική Μάθηση από Ανθρώπινη Ανατροφοδότηση
Robustness	Ανθεκτικότητα
Search tree	Δέντρο αναζήτησης
Self-Attention	Αυτό-προσοχή
Self-Consistency (SC)	Αυτο-Συνέπεια
Self-Evaluation	Αυτο-αξιολόγηση
Self-Consistency Chain-of-Thought (SC-CoT)	Αλυσιδωτή Συλλογιστική με Αυτο-συνέπεια
Self-Supervised Learning	Αυτο-επιβλεπόμενη μάθηση
Standard Prompting	Τυπική Προτροπή
Step-by-step reasoning	Συλλογιστική βήμα προς βήμα
State Space Search	Αναζήτηση σε χώρο καταστάσεων
Test set	Σύνολο ελέγχου
Thought evaluation	Αξιολόγηση σκέψεων
Thought generation	Παραγωγή σκέψεων
Token	Μονάδα διακριτοποίησης
Training set	Σύνολο εκπαίδευσης
Tree-of-Thought(s) (ToT)	Δενδρική Συλλογιστική
Unsupervised Learning	Μη επιβλεπόμενη μάθηση

Validation set	Σύνολο επικύρωσης
Word embeddings	Ενσωματώσεις λέξεων
Wrappers	Περιβλήματα λογισμικού
Zero-Shot	Προτροπή χωρίς παραδείγματα

Κεφάλαιο 1ο: Εισαγωγή

Η ραγδαία πρόοδος των Μεγάλων Γλωσσικών Μοντέλων (αγγλ. Large Language Models, LLMs) η οποία έχει σημειωθεί τα τελευταία χρόνια, έχει επιφέρει ριζικές αλλαγές στον τρόπο με τον οποίο προσεγγίζονται προβλήματα επεξεργασίας φυσικής γλώσσας, συλλογιστικής και λήψης αποφάσεων. Αρχιτεκτονικές αιχμής, όπως τα μοντέλα GPT, PaLM, LLaMA και Gemini έχουν επιδείξει αξιοσημείωτες επιδόσεις σε ένα ευρύ φάσμα εργασιών, από απλές ερωταπαντήσεις έως σύνθετα προβλήματα μαθηματικού και λογικού συλλογισμού [1]. Εντούτοις, παρά τη γενική αυτή πρόοδο, έχει διαπιστωθεί ότι η απόδοση των LLMs εξαρτάται σε μεγάλο βαθμό από τον τρόπο με τον οποίο διατυπώνονται οι οδηγίες εισόδου, δηλαδή οι προτροπές (αγγλ. prompts), γεγονός που έχει οδηγήσει στην ανάδειξη της μηχανικής προτροπών (αγγλ. Prompt Engineering) ως κρίσιμου ερευνητικού και πρακτικού πεδίου [2].

1.1 Αντικείμενο και σκοπός της εργασίας

Αντικείμενο της παρούσας εργασίας είναι η συστηματική μελέτη και συγκριτική αξιολόγηση προηγμένων τεχνικών της μηχανικής προτροπών, με στόχο τη βελτίωση της συλλογιστικής ικανότητας των LLMs. Η εργασία εστιάζει ειδικότερα στις μεθόδους τυπικής προτροπής (αγγλ. Standard Prompting), αλυσίδας συλλογισμού (αγγλ. Chain-of-Thought, CoT) και δενδρικής συλλογιστικής (αγγλ. Tree-of-Thought, ToT), οι οποίες αντιπροσωπεύουν διαφορετικά επίπεδα ρητής καθοδήγησης της διαδικασίας συλλογισμού του μοντέλου.

Σκοπός της εργασίας είναι αφενός να αναλυθεί θεωρητικά η λειτουργία και οι βασικές αρχές κάθε προσέγγισης και αφετέρου να αποτιμηθεί πειραματικά η αποτελεσματικότητά τους σε προβλήματα πολυσταδιακού συλλογισμού. Για τον σκοπό αυτό, αξιοποιείται το πρότυπο σύνολο δεδομένων (αγγλ. Benchmark) GSM8K, το οποίο έχει καθιερωθεί στη διεθνή βιβλιογραφία ως σημείο αναφοράς για την αξιολόγηση μαθηματικού συλλογισμού σε φυσική γλώσσα [3]. Μέσω της συγκριτικής ανάλυσης επιδιώκεται η εξαγωγή τεκμηριωμένων συμπερασμάτων σχετικά με τα πλεονεκτήματα, την αποδοτικότητα και τους περιορισμούς κάθε μεθόδου.

1.2 Κίνητρο και σημασία της Μηχανικής Προτροπών στα LLMs

Η μηχανική προτροπών αναδείχθηκε ως απάντηση στο γεγονός ότι τα LLMs, παρά τον τεράστιο όγκο κωδικοποιημένης γνώσης που διαθέτουν, δεν αξιοποιούν πάντα αποτελεσματικά τις δυνατότητές τους όταν λαμβάνουν ασαφείς ή ελάχιστα δομημένες οδηγίες. Σε αντίθεση με παραδοσιακές προσεγγίσεις βελτίωσης της απόδοσης, όπως η μικρορύθμιση (αγγλ. fine-tuning), η μηχανική προτροπών επιτρέπει την προσαρμογή της συμπεριφοράς των μοντέλων χωρίς αλλαγή των παραμέτρων τους, αξιοποιώντας αποκλειστικά τη διατύπωση της προτροπής.

Η κρισιμότητα της μεθοδολογίας αυτής καθίσταται ιδιαίτερα εμφανής σε διεργασίες πολύπλοκου συλλογισμού, όπου τα LLMs τείνουν να παρουσιάζουν ασυνέχειες στα ενδιάμεσα λογικά βήματα. Η ενσωμάτωση τεχνικών όπως το CoT έχει αποδειχθεί ότι επιτρέπει στα μοντέλα να αποσυνθέτουν σύνθετα προβλήματα σε διακριτά και απλούστερα υποπροβλήματα, βελτιστοποιώντας δραστικά την ακρίβεια των τελικών απαντήσεων [3]. Συνακόλουθα, πιο προηγμένες αρχιτεκτονικές προτροπών, όπως το ToT, επεκτείνουν τη λογική αυτή, επιτρέποντας τη δυναμική εξερεύνηση και αξιολόγηση πολλαπλών εναλλακτικών πορειών συλλογισμού [4].

1.3 Προβλήματα στη συλλογιστική των LLMs

Παρά την αδιαμφισβήτητη τεχνολογική πρόοδο, τα LLMs εξακολουθούν να παρουσιάζουν σημαντικές αδυναμίες στη συλλογιστική. Ένα από τα βασικότερα προβλήματα είναι η παραγωγή εσφαλμένων απαντήσεων με υψηλό βαθμό βεβαιότητας, φαινόμενο που έχει καθιερωθεί στη βιβλιογραφία με τον όρο «παραισθήσεις» (αγγλ. hallucinations) [5]. Το πρόβλημα αυτό καθιστά τα μοντέλα λιγότερο αξιόπιστα, ιδιαίτερα σε εφαρμογές υψηλού ρίσκου.

Επιπλέον, τα μοντέλα συχνά αποτυγχάνουν στη διατήρηση λογικής συνέπειας κατά τη διάρκεια πολυσταδιακού συλλογισμού, οδηγούμενα σε σφάλματα στα ενδιάμεσα βήματα, ακόμη και σε περιπτώσεις όπου η τελική απάντηση τυγχάνει ορθή. Παράλληλα, οι παραδοσιακές μετρικές αξιολόγησης, όπως το Exact Match ή το F1 Score, αδυνατούν να αποτυπώσουν την ποιότητα της συλλογιστικής διαδικασίας [5]. Τα παραπάνω προβλήματα αναδεικνύουν την ανάγκη για τεχνικές που ενισχύουν τόσο την ακρίβεια όσο και τη διαφάνεια του συλλογισμού.

1.4 Ερευνητικά ερωτήματα και στόχοι

Λαμβάνοντας υπόψη το προαναφερθέν θεωρητικό και τεχνολογικό πλαίσιο, η παρούσα εργασία επιχειρεί να απαντήσει στα ακόλουθα ερευνητικά ερωτήματα:

- Σε ποιο βαθμό οι διαφορετικές τεχνικές μηχανικής προτροπών επηρεάζουν την απόδοση των LLMs σε προβλήματα πολυσταδιακού συλλογισμού;
- Πώς συγκρίνονται οι προσεγγίσεις Standard Prompting, CoT και ToT ως προς την ακρίβεια, τη συνέπεια και την ερμηνευσιμότητα των αποτελεσμάτων;
- Ποιο είναι το κόστος, υπολογιστικό και μεθοδολογικό, που συνεπάγεται η χρήση πιο σύνθετων τεχνικών συλλογισμού;

Απώτερος στόχος της παρούσας εργασίας είναι η εξαγωγή τεκμηριωμένων και εμπειρικά επαληθευμένων συμπερασμάτων, τα οποία θα συμβάλουν στην εμβάθυνση της κατανόησης του ρόλου που διαδραματίζει η μηχανική προτροπών στις γνωσιακές διεργασίες των LLMs. Παράλληλα, επιδιώκεται η παροχή πρακτικών κατευθυντήριων γραμμών για τη βέλτιστη επιλογή τεχνικών σε ρεαλιστικά σενάρια εφαρμογής και αξιολόγησης.

1.5 Συνεισφορά της παρούσας εργασίας

Η παρούσα διπλωματική εργασία εστιάζει στη συστηματική μελέτη και αξιολόγηση τεχνικών της μηχανικής προτροπών σε LLMs, με έμφαση στις μεθόδους που στοχεύουν στη βελτίωση των ικανοτήτων συλλογιστικής και επίλυσης προβλημάτων. Σε αντίθεση με μεγάλο μέρος της υπάρχουσας βιβλιογραφίας, η οποία συχνά περιορίζεται στην παρουσίαση μεμονωμένων τεχνικών ή στην αξιολόγησή τους σε αποσπασματικά πειραματικά σενάρια, η παρούσα εργασία υιοθετεί μια ενιαία και συγκριτική προσέγγιση.

Η κύρια συνεισφορά της εργασίας συνοψίζεται στα εξής:

- Παρέχεται μια οργανωμένη και επικαιροποιημένη επισκόπηση βασικών και προηγμένων τεχνικών μηχανικής προτροπών, όπως το Standard Prompting, το CoT, το ToT, η Αυτο-Συνέπεια (αγγλ. Self-Consistency, SC) και συναφείς επεκτάσεις, αναδεικνύοντας τις θεωρητικές τους βάσεις και τις διαφορές τους ως προς τη λογική λειτουργίας τους.
- Σχεδιάζεται και υλοποιείται ένα ενιαίο πειραματικό πλαίσιο αξιολόγησης, το οποίο επιτρέπει τη δίκαιη και αναπαραγώγιμη σύγκριση διαφορετικών τεχνικών και μοντέλων, αξιοποιώντας κοινό σύνολο δεδομένων και ομοιόμορφα κριτήρια αξιολόγησης.
- Πραγματοποιείται εμπειρική ανάλυση της επίδρασης των τεχνικών της μηχανικής προτροπών στην ακρίβεια, τη σταθερότητα και τη συμπεριφορά των μοντέλων κατά την επίλυση

προβλημάτων συλλογιστικής, αναδεικνύοντας τόσο τα πλεονεκτήματα όσο και τους περιορισμούς κάθε προσέγγισης.
- Διερευνάται η σχέση μεταξύ της πολυπλοκότητας της προτροπής και της απόδοσης των μοντέλων, συμβάλλοντας στην κατανόηση του κατά πόσο πιο σύνθετες τεχνικές οδηγούν πράγματι σε ουσιαστική βελτίωση ή εισάγουν επιπλέον αστάθεια και υπολογιστικό κόστος.

Γενικά, η εργασία φιλοδοξεί να συμβάλει τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο στη μελέτη της μηχανικής προτροπών, προσφέροντας χρήσιμα συμπεράσματα για ερευνητές και επαγγελματίες που επιθυμούν να αξιοποιήσουν αποδοτικά τα LLMs σε εφαρμογές που απαιτούν αξιόπιστη συλλογιστική και ερμηνεύσιμα αποτελέσματα.

1.6 Δομή της εργασίας

Η παρούσα διπλωματική εργασία είναι οργανωμένη σε οκτώ κεφάλαια.

- Στο **Κεφάλαιο 1** παρουσιάζεται το αντικείμενο, οι στόχοι, τα ερευνητικά ερωτήματα και η συνεισφορά της εργασίας.
- Το **Κεφάλαιο 2** αναλύει το θεωρητικό υπόβαθρο των Μεγάλων Γλωσσικών Μοντέλων και της μηχανικής προτροπών.
- Στο **Κεφάλαιο 3** γίνεται επισκόπηση της σχετικής βιβλιογραφίας και σύνδεσή της με την παρούσα έρευνα.
- Το **Κεφάλαιο 4** περιγράφει τη μεθοδολογία, το πειραματικό πλαίσιο και τα εργαλεία που χρησιμοποιήθηκαν.
- Στο **Κεφάλαιο 5** παρουσιάζονται τα κριτήρια και οι μετρικές αξιολόγησης.
- Το **Κεφάλαιο 6** περιλαμβάνει τα πειραματικά αποτελέσματα και τη συγκριτική ανάλυσή τους.
- Στο **Κεφάλαιο 7** ακολουθεί συζήτηση και ερμηνεία των αποτελεσμάτων.
- Τέλος, το **Κεφάλαιο 8** συνοψίζει τα βασικά συμπεράσματα και προτείνει κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2ο: Θεωρητικό υπόβαθρο

2.1 Εξέλιξη και αρχιτεκτονική των Μεγάλων Γλωσσικών Μοντέλων (LLMs)

Τα τελευταία έτη, τα Μεγάλα Γλωσσικά Μοντέλα (αγγλ. Large Language Models, LLMs) έχουν αναδειχθεί σε ένα από τα πλέον αξιοσημείωτα επιτεύγματα στον τομέα της τεχνητής νοημοσύνης. Η υπεροχή τους έγκειται στην ικανότητά τους να συνθέτουν προηγμένες τεχνικές βαθιάς μάθησης (αγγλ. Deep Learning), κατανόηση φυσικής γλώσσας και στοιχειώδεις μορφές συλλογισμού σε ένα ενιαίο υπολογιστικό σύστημα, ικανό να διαχειρίζεται πολύπλοκες γλωσσικές διεργασίες [1]. Η λειτουργία τους βασίζεται στην εκμάθηση στατιστικών και σημασιολογικών συσχετίσεων μεταξύ λέξεων και προτάσεων μέσω της επεξεργασίας τεράστιου όγκου δεδομένων κειμένου. Η διαδικασία αυτή τους επιτρέπει να παράγουν απαντήσεις που χαρακτηρίζονται από συνοχή, νοηματική αλληλουχία και ορθή δομή [6]. Σύμφωνα με πρόσφατα ερευνητικά δεδομένα, η επιτυχία των LLMs οφείλεται κυρίως στον συνδυασμό νευρωνικών δικτύων μεγάλης κλίμακας και σύγχρονων τεχνικών προ-εκπαίδευσης (αγγλ. pre-training), οι οποίες επιτρέπουν στα μοντέλα να μαθαίνουν γενικές αναπαραστάσεις της γλώσσας και στη συνέχεια να προσαρμόζονται εύκολα σε πιο εξειδικευμένες εργασίες (αγγλ. fine-tuning) [7]. Επίσης, η ανάπτυξη των LLMs συνδέεται άμεσα και με την εξέλιξη της υπολογιστικής ισχύος και τη διαθεσιμότητα μεγάλων συνόλων δεδομένων (αγγλ. Big Data), τα οποία επιτρέπουν τη δημιουργία ολοένα μεγαλύτερων και πιο ικανών μοντέλων [8].

Η ιστορική εξέλιξη των LLMs ξεκινά από τις πρώτες απόπειρες επεξεργασίας φυσικής γλώσσας, όπου τα συστήματα βασίζονταν πρωτίστως σε σύνολα χειροποίητων κανόνων και απλουστευμένες στατιστικές μεθόδους, γεγονός που περιόριζε δραστικά την ικανότητά τους για γενίκευση και κατανόηση του φυσικού, μη δομημένου λόγου [6]. Κατά τις δεκαετίες του 1980 και 1990, τα στατιστικά μοντέλα n-gram άρχισαν να χρησιμοποιούνται ευρέως, επειδή μπορούσαν να προβλέπουν την επόμενη λέξη με βάση τα κοντινά συμφραζόμενα, παρόλο που δυσκολεύονταν να χειριστούν μακρινές εξαρτήσεις ή πιο σύνθετες γλωσσικές έννοιες [1]. Η εμφάνιση των νευρωνικών δικτύων, και ιδιαίτερα των επαναλαμβανόμενων νευρωνικών δικτύων (αγγλ. Recurrent Neural Networks, RNNs), αποτέλεσε σημαντική εξέλιξη, αφού για πρώτη φορά έδωσε τη δυνατότητα στα μοντέλα να επεξεργάζονται ακολουθίες λέξεων με πιο δυναμικό τρόπο αντί για απλές στατικές μεθόδους [6]. Τα δίκτυα μακράς - βραχείας μνήμης (αγγλ. Long Short-Term Memory, LSTM) βελτίωσαν ακόμη περισσότερο αυτή τη δυνατότητα, βοηθώντας στην καλύτερη διατήρηση πληροφορίας και λύνοντας προβλήματα όπως η εξαφάνιση της βαθμίδας (αγγλ. vanishing gradient), γεγονός που βελτίωσε σημαντικά την κατανόηση της γλώσσας που μπορούσαν να πετύχουν [6]. Παρά την πρόοδό τους, οι αρχιτεκτονικές αυτές είχαν σημαντικούς περιορισμούς, καθώς δεν μπορούσαν να αξιοποιήσουν καλά τον παραλληλισμό και δυσκολεύονταν να χειριστούν πολύ μεγάλες προτάσεις ή εκτενή κείμενα, κάτι που τελικά εμπόδιζε την κλιμάκωσή τους σε μεγαλύτερα και πιο απαιτητικά συστήματα [1].

Σημαντικό βήμα στην εξέλιξη των LLMs ήταν η εμφάνιση των διανυσματικών αναπαραστάσεων λέξεων (αγγλ. word embeddings) το 2013, μέσα από τις μεθόδους εκμάθησης Word2Vec και GloVe, οι οποίες έδωσαν τη δυνατότητα στις λέξεις να αναπαρίστανται ως συνεχόμενα διανύσματα, τα οποία αποτυπώνουν τις νοηματικές και σημασιολογικές σχέσεις μεταξύ τους [6]. Με τις μεθόδους αυτές, τα συστήματα μπορούσαν πλέον να καταλαβαίνουν ότι ορισμένες λέξεις σχετίζονται μεταξύ τους, κάτι που αποτέλεσε σημαντική διασύνδεση ανάμεσα στις απλές στατιστικές τεχνικές και στα πιο εξελιγμένα νευρωνικά μοντέλα που αναπτύχθηκαν αργότερα [1]. Παρ' όλα αυτά, ακόμη και με αυτή τη σημαντική πρόοδο, οι αρχιτεκτονικές RNN και LSTM δεν κατάφερναν να αξιοποιήσουν πλήρως τα τεράστια

σύνολα δεδομένων που είχαν αρχίσει να διατίθενται, κυρίως επειδή η σειριακή επεξεργασία που απαιτούσαν έκανε την εκπαίδευση αργή και δύσκολο να κλιμακωθεί σε πολύ μεγάλα μοντέλα [6].

Το σημείο καμπής στην εξέλιξη των LLMs σημειώθηκε το 2017, όταν παρουσιάστηκε η αρχιτεκτονική Transformer, η οποία εισήγαγε τον επαναστατικό μηχανισμό αυτό-προσοχής (αγγλ. Self-Attention). Ο μηχανισμός αυτός επέτρεψε στα μοντέλα να συσχετίζουν κάθε λέξη μιας πρότασης με όλες τις υπόλοιπες ταυτόχρονα ανεξαρτήτως της θέσης τους, κάτι που μέχρι τότε δεν ήταν εφικτό με τόσο αποδοτικό τρόπο [2]. Με την αρχιτεκτονική Transformer δεν υπήρχε πλέον ανάγκη για αυστηρά σειριακή επεξεργασία, γεγονός που άνοιξε τον δρόμο για πολύ πιο γρήγορη και παράλληλη εκπαίδευση, στοιχείο καθοριστικό για την ανάπτυξη των σύγχρονων LLMs [1]. Επιπλέον, η χρήση του μηχανισμού «προσοχής πολλαπλών κεφαλών» (αγγλ. Multi-Head Attention) στον πυρήνα της αρχιτεκτονικής Transformers, επέτρεψε στο μοντέλο να αναλύει ταυτόχρονα διαφορετικές σημασιολογικές σχέσεις μέσα στο ίδιο κείμενο, δημιουργώντας πιο πλούσιες και ακριβείς γλωσσικές αναπαραστάσεις [9]. Χάρη σε αυτές τις καινοτομίες, η αρχιτεκτονική Transformer εξελίχθηκε γρήγορα σε βασικό θεμέλιο της σύγχρονης γενιάς LLMs, επηρεάζοντας τόσο την απόδοση όσο και τη δυνατότητα συνεχούς κλιμάκωσης των μοντέλων [6].

Από το 2018 και μετά, η εξάπλωση των LLMs υπήρξε ραγδαία, καθώς μοντέλα όπως τα BERT, GPT-2 και T5 έδειξαν ξεκάθαρα πόσο ισχυρή μπορεί να γίνει η αρχιτεκτονική Transformer όταν εφαρμόζεται σε πολλές διαφορετικές γλωσσικές εργασίες [7]. Η επιτυχία αυτών των μοντέλων οφείλεται κυρίως στο ότι μπορούν να αξιοποιούν τεράστιες ποσότητες δεδομένων χωρίς να χρειάζεται να τα επισημάνει κάποιος χειροκίνητα, μέσα από τεχνικές μη Επιβλεπόμενης Μάθησης (αγγλ. Unsupervised Learning) ή Αυτο-Επιβλεπόμενης Μάθησης (αγγλ. Self-supervised learning), κάτι που τους επιτρέπει να μαθαίνουν γενικά γλωσσικά μοτίβα και δομές [1]. Το 2020, η παρουσίαση του GPT-3, με 175 δισεκατομμύρια παραμέτρους, ανέδειξε στην πράξη πόσο σημαντικό είναι το μέγεθος ενός μοντέλου, αφού όσο αυτό αυξάνεται, τόσο εμφανίζονται νέες και πιο εξελιγμένες ικανότητες, όπως καλύτερη συλλογιστική με λίγα παραδείγματα (αγγλ. Few-Shot) και πιο αποτελεσματική γενίκευση γνώσης [8]. Το GPT-3 θεωρήθηκε σημείο καμπής, επειδή απέδειξε ότι η αύξηση των παραμέτρων δεν βελτιώνει μόνο τη μνήμη του μοντέλου, αλλά οδηγεί και σε σημαντική ποιοτική πρόοδο στη λογική σκέψη και στην κατανόηση της γλώσσας [7].

Τα επόμενα έτη, η εμφάνιση νέων μοντέλων όπως τα PaLM, LLaMA και Gemini διεύρυνε ακόμη περισσότερο τις δυνατότητες των LLMs, αφού αυτά τα συστήματα κατάφεραν να συνδυάσουν μάθηση πολλαπλών εργασιών (αγγλ. Multi-Task Learning) και εκπαίδευση σε πολλούς διαφορετικούς τομείς γνώσης, γεγονός που τα έκανε πιο ευέλικτα και πιο ικανά να χειριστούν διαφορετικού τύπου πληροφορίες [10]. Επιπλέον, η υιοθέτηση της αρχιτεκτονικής Mixture-of-Experts (MoE) συνέβαλε σημαντικά στη βελτίωση της αποδοτικότητας. Με αυτή την αρχιτεκτονική, το μοντέλο ενεργοποιεί μόνο τις ειδικές υπομονάδες που χρειάζονται κάθε φορά, προσφέροντας μεγαλύτερη εξειδίκευση χωρίς να αυξάνεται το υπολογιστικό κόστος [4]. Οι εξελίξεις αυτές βελτίωσαν θεαματικά τη δυνατότητα κλιμάκωσης των μοντέλων, επιτρέποντας την ανάπτυξη μοντέλων πολύ μεγάλου μεγέθους, τα οποία όμως καταναλώνουν λιγότερους πόρους σε σχέση με το παρελθόν [1].

Η εξέλιξη των LLMs μπορεί να γίνει πιο κατανοητή αν τη δούμε μέσα από τρεις γενιές ανάπτυξης, όπως προτείνεται στη σχετική βιβλιογραφία [6]:

- Η πρώτη γενιά (2018–2020) περιλαμβάνει μοντέλα όπως το BERT και το GPT-2, τα οποία απέδειξαν την αποτελεσματικότητα της αρχιτεκτονικής Transformer, πετυχαίνοντας εξαιρετικές επιδόσεις χωρίς να χρειάζονται επιβλεπόμενη εκπαίδευση [7].

- Η δεύτερη γενιά (2020–2022) χαρακτηρίζεται από την εμφάνιση πολύ μεγαλύτερων μοντέλων, με πιο γνωστό παράδειγμα το GPT-3, που εισήγαγε νέες ικανότητες, όπως η συλλογιστική πολλαπλών βημάτων (αγγλ. Multi-Step Reasoning) και καλύτερη γενίκευση γνώσης, αποδεικνύοντας ότι όσο αυξάνεται το μέγεθος ενός μοντέλου, τόσο πιο προηγμένες δυνατότητες μπορεί να αναπτύξει [8].
- Η τρίτη γενιά (2022–σήμερα) στρέφεται προς ακόμη πιο σύνθετες και πρακτικές εφαρμογές, δίνοντας έμφαση στην πολυτροπικότητα (αγγλ. multimodality), στη χρήση των MoE αρχιτεκτονικών και στη βελτίωση της ευθυγράμμισης των μοντέλων με ανθρώπινες αξίες, με συστήματα όπως τα PaLM, Gemini και LLaMA να θέτουν πλέον νέα πρότυπα ακρίβειας, ασφάλειας και αποτελεσματικότητας [10].

Η διαδικασία εκπαίδευσης ενός LLM ξεκινά με το στάδιο της προ-εκπαίδευσης (αγγλ. pre-training), όπου το μοντέλο διαβάζει τεράστιους όγκους κειμένου μαθαίνοντας να προβλέπει το επόμενο token, αποκτώντας έτσι θεμελιώδη κατανόηση της γλώσσας [1]. Με αυτό τον τρόπο το μοντέλο αναπτύσσει γλωσσικές, συντακτικές και σημασιολογικές γνώσεις χωρίς να χρειάζεται άμεση ανθρώπινη επίβλεψη, χτίζοντας έτσι μια πολύ ισχυρή βάση πάνω στην οποία μπορεί αργότερα να εξειδικευτεί [6]. Ακολουθεί η διαδικασία της μικρορύθμισης μέσω οδηγιών, όπου εκπαιδεύεται με παραδείγματα που έχουν επιμεληθεί άνθρωποι, ώστε να μάθει να ακολουθεί οδηγίες με πιο φυσικό, ξεκάθαρο και συνεργάσιμο τρόπο [11]. Στη συνέχεια ακολουθεί το στάδιο της ενισχυτικής μάθησης από ανθρώπινη ανατροφοδότηση (αγγλ. Reinforcement Learning from Human Feedback - RLHF), όπου το μοντέλο βελτιώνει τη συμπεριφορά του αξιοποιώντας ανθρώπινες προτιμήσεις σε συνδυασμό με τεχνικές ενισχυτικής μάθησης, με στόχο να γίνει πιο ασφαλές και περισσότερο ευθυγραμμισμένο με ανθρώπινες αξίες [1].

Σε επίπεδο αρχιτεκτονικής, τα LLMs λειτουργούν με αυτοπαλινδρομικό τρόπο (αγγλ. autoregressive), δηλαδή παράγουν κάθε νέο token βασισμένα σε όσα έχουν ήδη παραχθεί, κάτι που τους επιτρέπει να χτίζουν μια σταθερή και συνεκτική αλυσίδα σκέψης και επιχειρημάτων [3]. Αυτή η μορφή παραγωγής, σε συνδυασμό με το εύρος συμφραζομένων που διαθέτουν τα σύγχρονα μοντέλα, τους δίνει τη δυνατότητα να διατηρούν τη συνοχή σε μεγάλους διαλόγους, να ανατρέχουν σε προηγούμενα σημεία και να αναπτύσσουν πιο φυσικό και ρεαλιστικό διάλογο [1]. Παράλληλα, ο μηχανισμός Multi-Head Attention επιτρέπει στο μοντέλο να εντοπίζει ταυτόχρονα πολλές διαφορετικές σημασιολογικές σχέσεις μέσα στο ίδιο κείμενο, βελτιώνοντας έτσι τόσο την κατανόηση όσο και τη λογική επεξεργασία των πληροφοριών [1].

Τα σύγχρονα LLMs έχουν εξελιχθεί σε πραγματικά πολυδιάστατα συστήματα, ικανά να επεξεργάζονται κείμενο, εικόνα, ήχο και άλλους τύπους δεδομένων, κάτι που τα κάνει ιδιαίτερα ευέλικτα και χρήσιμα σε ένα πολύ ευρύ φάσμα εφαρμογών [1]. Η μετάβαση από τα παλαιότερα μοντέλα στατιστικής πρόβλεψης σε πιο προηγμένα μοντέλα που βασίζονται στη συλλογιστική τεχνητή νοημοσύνη σηματοδοτεί μια νέα εποχή στην πορεία των LLMs, αφού αυτά τα συστήματα μπορούν πλέον όχι μόνο να δίνουν απαντήσεις, αλλά και να εξηγούν τα βήματα που ακολούθησαν για να καταλήξουν σε αυτές [4]. Οι τεχνικές όπως η CoT και η ToT ενισχύουν ακόμη περισσότερο τη διαφάνεια και την αξιοπιστία των μοντέλων, επιτρέποντας μια βαθύτερη κατανόηση του τρόπου με τον οποίο λαμβάνονται οι αποφάσεις μέσα στο σύστημα [3][4].

Η συνεχώς αυξανόμενη πολυπλοκότητα των LLMs καθιστά αναγκαία την εφαρμογή νέων τρόπων αξιολόγησης και επεξήγησης της συμπεριφοράς τους, αφού οι παραδοσιακές διαδικασίες αξιολόγησης δεν αρκούν πλέον για να αποτυπώσουν πλήρως τις πραγματικές τους δυνατότητες [12]. Για τον λόγο αυτό δημιουργήθηκε η μέθοδος LLM-as-a-Judge, μέσω της οποίας το LLM λειτουργεί ως αξιολογητής των απαντήσεων άλλων LLMs, προσφέροντας μεγαλύτερη συνέπεια, πιο αντικειμενική κρίση και τη δυνατότητα να αξιολογούνται πολλά δεδομένα σε μεγάλη κλίμακα [12]. Την ίδια στιγμή, η ενσωμάτωση

ανθρώπινης ανάδρασης μέσα στα ενδιάμεσα βήματα του συλλογισμού μπορεί να βελτιώσει ουσιαστικά τόσο την ακρίβεια όσο και τη σταθερότητα των ακολουθιών σκέψης που παράγει ένα LLM [8].

Συνοψίζοντας, η τρέχουσα κατάσταση των LLMs δείχνει μια διαρκή πορεία εξέλιξης, όπου η πρόοδος στην αρχιτεκτονική, στα δεδομένα, στις μεθόδους εκπαίδευσης και στους μηχανισμούς συλλογισμού συνδυάζεται για να δημιουργήσει συστήματα όλο και πιο ικανά, πιο ασφαλή και πιο κοντά στον τρόπο που σκέφτεται ο άνθρωπος. Η προσθήκη συστημάτων ή μοντέλων που μπορούν να επεξεργάζονται και να συνδυάζουν πολλαπλούς τύπους δεδομένων (αγγλ. multimodal), η προσπάθεια για μεγαλύτερη ερμηνευσιμότητα και η χρήση πιο προηγμένων τεχνικών συλλογισμού αποτελούν βασικά στοιχεία της νέας γενιάς LLMs, η οποία μεταφέρει την τεχνητή νοημοσύνη από την απλή πρόβλεψη προς μια πιο τεκμηριωμένη και βαθιά κατανόηση [1]. Με τη συνεχή ερευνητική πρόοδο και την αύξηση της υπολογιστικής ισχύος, τα LLMs διαμορφώνουν πλέον ένα δυναμικό τεχνολογικό τοπίο που έχει σημαντικό επιστημονικό, κοινωνικό και οικονομικό αντίκτυπο.

2.2 Μηχανική Προτροπών (Prompt Engineering): έννοιες και κατηγορίες

Η αξιοποίηση των LLMs, όπως αναλύθηκε στο προηγούμενο κεφάλαιο, δεν εξαρτάται μόνο από το πώς εκπαιδεύονται ή από το πώς είναι σχεδιασμένη η αρχιτεκτονική τους. Η αποδοτική χρήση τους προϋποθέτει την κατανόηση και την εφαρμογή κατάλληλων μεθόδων αλληλεπίδρασης. Στο πλαίσιο αυτό, η μηχανική προτροπών (αγγλ. Prompt Engineering) έχει αναδειχθεί τα τελευταία έτη ως ένας από τους σημαντικότερους παράγοντες για την εκμετάλλευση της υπολογιστικής ισχύος των LLMs. Η μηχανική προτροπών περιγράφεται ως μια μορφή «γνωσιακού προγραμματισμού» [11], όπου, αντί της συγγραφής παραδοσιακού κώδικα, η φυσική γλώσσα διαμορφώνεται στρατηγικά ώστε να κατευθύνει το μοντέλο να σκεφτεί και να απαντήσει με τον τρόπο που επιθυμούμε. Ουσιαστικά, η σύνταξη της ερώτησης δύναται να επηρεάσει καθοριστικά τη συλλογιστική διαδικασία του μοντέλου, ενεργοποιώντας συγκεκριμένα μονοπάτια σκέψης. Συνεπώς, η μηχανική προτροπών δεν συνιστά απλώς μια τεχνική λεπτομέρεια, αλλά μια κρίσιμη δεξιότητα που καθορίζει την αποτελεσματικότητα του LLM σε ένα ευρύ φάσμα εργασιών, από απλές εφαρμογές περίληψης και μετάφρασης, έως σύνθετες διαδικασίες ανάλυσης δεδομένων, επίλυσης προβλημάτων και υποστήριξης λήψης αποφάσεων.

Η σχετική βιβλιογραφία επισημαίνει ότι, καθώς αυξάνεται η κλίμακα των LLMs, αναδύονται νέες ικανότητες οι οποίες απουσίαζαν από τα παλαιότερα και μικρότερα μοντέλα [1]. Πρόκειται για ιδιότητες όπως ο προηγμένος λογικός συλλογισμός, η αναλογική σκέψη και η ικανότητα επίλυσης εργασιών χωρίς ειδική εκπαίδευση (αγγλ. Zero-Shot). Ωστόσο, οι δυνατότητες αυτές παραμένουν συχνά λανθάνουσες και απαιτούν τον κατάλληλο τρόπο αλληλεπίδρασης για να ενεργοποιηθούν. Εδώ έγκειται η σημασία της μηχανικής προτροπών, καθώς μέσω καλά δομημένων εντολών καθίσταται δυνατή η ανάκληση αυτών των αναδυόμενων ικανοτήτων και η μεγιστοποίηση της απόδοσης των μοντέλων. Η ακρίβεια, η αναλυτική ικανότητα και η ποιότητα των απαντήσεων του LLM εξαρτώνται άμεσα από τη διατύπωση της προτροπής. Για τον λόγο αυτό η μηχανική προτροπών είναι πλέον κρίσιμη, αφού μας επιτρέπει να αξιοποιήσουμε πραγματικά τις ικανότητες των σύγχρονων μοντέλων και να λάβουμε πιο αξιόπιστες, ακριβείς και χρήσιμες απαντήσεις.

Αξίζει να σημειωθεί ότι η επίδραση της ποιότητας της προτροπής στην απόδοση των μοντέλων είναι καταλυτική, με ορισμένες εμπειρικές μελέτες να αποδίδουν έως και το 80% της τελικής απόδοσης στον ορθό σχεδιασμό της [13]. Το γεγονός αυτό καταδεικνύει ότι η μηχανική προτροπών δεν αποτελεί μια απλή τεχνική επιτήδευση, αλλά μια ουσιαστική μεθοδολογία που ορίζει τον τρόπο λειτουργίας του μοντέλου. Με λίγα λόγια, οι προτροπές λειτουργούν ως τα δομικά στοιχεία της συλλογιστικής του LLM, επηρεάζοντας άμεσα τόσο την ακρίβεια όσο και τη συνέπεια των αποτελεσμάτων. Θεμελιωδώς, η

μηχανική προτροπών κατηγοριοποιείται σε τρεις βασικές προσεγγίσεις: την προτροπή χωρίς παραδείγματα (αγγλ. Zero-Shot Prompting), την προτροπή με λίγα παραδείγματα (αγγλ. Few-Shot Prompting) και τη μάθηση εντός πλαισίου (αγγλ. In-Context Learning - ICL). Οι κατηγορίες αυτές αποτελούν το θεωρητικό υπόβαθρο πάνω στο οποίο αναπτύχθηκαν μεταγενέστερα πιο εξελιγμένες τεχνικές, όπως η CoT, η SC και η ToT, οι οποίες αναλύονται διεξοδικά στις ενότητες που ακολουθούν.

2.2.1 Μηχανική Προτροπών ως γνωσιακή διεπαφή

Η μηχανική προτροπών δεν συνιστά απλώς μια μέθοδο αλληλεπίδρασης ανθρωπου-μηχανής, αλλά λειτουργεί ως μια γνωσιακή διεπαφή, δηλαδή ως ένας μηχανισμός μέσω του οποίου ο χρήστης διαμορφώνει ενεργά τη διαδικασία συλλογισμού του LLM [11]. Επί της ουσίας, η προτροπή δρα ως ένα στρατηγικό πλαίσιο αναφοράς που καθοδηγεί το μοντέλο σχετικά με τον τρόπο επεξεργασίας της πληροφορίας, την οργάνωση της σκέψης του και, τελικά, την παραγωγή της βέλτιστης λύσης (απάντησης). Στο πλαίσιο αυτό, η φυσική γλώσσα μετεξελίσσεται σε μια νέα μορφή υψηλού επιπέδου προγραμματισμού. Μέσω των προτροπών, ο χρήστης δύναται να ορίσει κανόνες, παραδείγματα, δομικά πρότυπα, ακόμη και λογικούς περιορισμούς, χρησιμοποιώντας αποκλειστικά εντολές μέσω κειμένου [14]. Η διάκριση αυτή είναι θεμελιώδης, καθώς τα LLMs δεν λειτουργούν βάσει ντετερμινιστικών αλγορίθμων όπως τα συμβατικά λογισμικά, αλλά βασίζονται στην πρόβλεψη της επόμενης λέξης και στην ανάλυση των συμφραζομένων. Συνεπώς, η προτροπή διαδραματίζει καθοριστικό ρόλο στην κατανόηση και εκτέλεση μιας εργασίας, επιτρέποντας τη μετάβαση από τη γενική χρήση ενός LLM στη στοχευμένη αξιοποίηση των πιο εξελιγμένων δυνατοτήτων του [2]. Μέσω ορθά σχεδιασμένων προτροπών, καθίσταται εφικτό για το μοντέλο:

- Να παράγει πιο συνεκτικές και λογικά δομημένες αλυσίδες συλλογισμού.
- Να εκτελεί πολύπλοκες εργασίες χωρίς την ανάγκη προηγούμενης εξειδικευμένης εκπαίδευσης (αγγλ. fine-tuning).
- Να προσαρμόζει το ύφος, τον τόνο και το επίπεδο λεπτομέρειας της απόκρισης σύμφωνα με τις εκάστοτε απαιτήσεις του χρήστη.

Η βασική ιδέα είναι ότι τα LLMs διαθέτουν ήδη την απαραίτητη γνώση, αλλά αυτό που συχνά απουσιάζει είναι ο κατάλληλος μηχανισμός ανάκλησης και εφαρμογής της, κενό το οποίο καλείται να καλύψει η μηχανική προτροπών.

2.2.2 Προτροπή χωρίς παραδείγματα (Zero-Shot Prompting)

Η τεχνική προτροπής χωρίς παραδείγματα (αγγλ. Zero-Shot Prompting) συνιστά την πλέον θεμελιώδη και απλή μορφή προτροπής, κατά την οποία το μοντέλο λαμβάνει αποκλειστικά μια εντολή, χωρίς την παροχή συνοδευτικών παραδειγμάτων, και καλείται να συμπεράνει αυτόνομα το ζητούμενο [11]. Υπό τη συνθήκη αυτή, το LLM εδράζεται αποκλειστικά στη γνώση που απέκτησε κατά το στάδιο της προ-εκπαίδευσης, αξιοποιώντας στατιστικές συσχετίσεις για την ερμηνεία και την εκτέλεση της εργασίας. Ο Naveed και συν. (2024) επισημαίνουν ότι τα σύγχρονα LLMs, χάρη στην ευρεία κλίμακα των παραμέτρων τους και την ποικιλομορφία των δεδομένων εκπαίδευσης, επιδεικνύουν πλέον την ικανότητα επίλυσης ενός ευρύτατου φάσματος εργασιών με τη μέθοδο Zero-Shot. Εφαρμογές όπως η μετάφραση, η περίληψη κειμένων και τα απλά συστήματα ερωταποκρίσεων εκτελούνται συχνά με εντυπωσιακή ακρίβεια, καθώς το μοντέλο έχει ήδη εσωτερικεύσει πλήθος γλωσσικών μοτίβων και δομών.

Εντούτοις, τονίζεται ότι τα LLMs παρουσιάζουν συχνά αστάθεια όταν η μέθοδος Zero-Shot εφαρμόζεται σε εργασίες που απαιτούν σύνθετο συλλογισμό ή υψηλή ακρίβεια [11]. Χωρίς παραδείγματα που θα λειτουργούσαν ως οδηγοί, τα μοντέλα συχνά καταφεύγουν σε πιθανοτικές

εικασίες (αγγλ. hallucinations), βασιζόμενα επιφανειακά στις στατιστικές συσχετίσεις του κειμένου, γεγονός που οδηγεί σε σφάλματα. Ο Wang και συν. (2022) υποστηρίζουν ότι οι προτροπές Zero-Shot αδυνατούν να ενεργοποιήσουν μηχανισμούς συλλογισμού υψηλού επιπέδου. Η αδυναμία αυτή απορρέει από την απουσία πλαισίου (αγγλ. context), καθώς χωρίς παραδείγματα το μοντέλο δυσκολεύεται να αντιληφθεί το ακριβές λογικό πρότυπο που απαιτεί η εκάστοτε εργασία. Ως εκ τούτου, η μέθοδος συχνά αποδεικνύεται ανεπαρκής σε προβλήματα μαθηματικών, λογικής, γρίφων (αγγλ. puzzles) και συμβολικού συλλογισμού, όπου απαιτείται αυστηρή και μεθοδική διεργασία.

Συνοψίζοντας, η Zero-Shot παραμένει εξαιρετικά πολύτιμη στην πράξη, καθώς είναι ταχεία, απλή στην εφαρμογή και επιτρέπει την άμεση αλληλεπίδραση με το μοντέλο χωρίς την προαπαιτούμενη προετοιμασία παραδειγμάτων. Επιπροσθέτως, αποδίδει εξαιρετικά σε εργασίες όπου το μοντέλο διαθέτει ήδη ισχυρή αναπαράσταση γνώσης από την προ-εκπαίδευση, καθιστώντας την ως τη βασική τεχνική για την πλειονότητα των καθημερινών σεναρίων χρήσης.

Πίνακας 1 Παράδειγμα Zero-Shot

Κατηγορία	**Αναλυτική Περιγραφή**
Τεχνική Προτροπής	**Zero-Shot Prompting**
Σκοπός	Να ζητηθεί από το LLM να εκτελέσει μία εργασία (περίληψη κειμένου) **χωρίς να του δοθεί κανένα παράδειγμα** ή ενδιάμεσο βήμα συλλογισμού.
Χαρακτηριστικό Zero-Shot	Το μοντέλο βασίζεται **αποκλειστικά στη γνώση που απέκτησε κατά την προεκπαίδευσή του**, χωρίς βοήθεια από δείγματα στη προτροπή.
Prompt προς το LLM	«Γράψε μια σύντομη και ακριβή περίληψη σε **μία πρόταση**: Από τη δεκαετία του 1960, οι ανθρώπινες διαστημικές αποστολές έχουν επεκτείνει συνεχώς τα όρια της τεχνολογίας και της επιστήμης. Η εξερεύνηση του διαστήματος οδήγησε σε σημαντικές ανακαλύψεις, ανέπτυξε νέες τεχνολογίες επικοινωνιών και επέτρεψε την κατανόηση φαινομένων που δεν μπορούν να μελετηθούν από τη Γη. Επιπλέον, η πρόοδος αυτή ώθησε την ανάπτυξη υλικών υψηλής αντοχής και βελτίωσε συστήματα πλοήγησης που χρησιμοποιούμε σήμερα στην καθημερινότητά μας.»
Ενδεικτική Απάντηση του LLM	«Από τη δεκαετία του 1960, η διαστημική εξερεύνηση έχει επιταχύνει την τεχνολογική και επιστημονική πρόοδο, οδηγώντας σε σημαντικές ανακαλύψεις και εφαρμογές που επηρεάζουν την καθημερινή ζωή.»
Απαίτηση από το LLM	● Κατανόηση του πλαισίου του αρχικού κειμένου. ● Εντοπισμός και εξαγωγή των κεντρικών ιδεών. ● Παραγωγή σύντομης και συνεκτικής περίληψης χωρίς να του προσφερθεί παράδειγμα ή καθοδήγηση
Αναμενόμενος Μηχανισμός Λειτουργίας	Το LLM ανακαλεί τις ενσωματωμένες παραστάσεις του για: ● Την αναγνώριση θεματικών σχέσεων ● Την εξαγωγή βασικών σημείων ● Την σύνθεση νέου συνοπτικού κειμένου
Πλεονεκτήματα Zero-Shot	● Μηδενικό κόστος προετοιμασίας παραδειγμάτων ● Γρήγορη εκτέλεση ● Ιδανικό για εργασίες που απαιτούν ευρεία, γενική γνώση
Μειονεκτήματα	● Πιθανότερα λάθη λόγω έλλειψης παραδειγμάτων ● Μικρότερος έλεγχος στον τρόπο απάντησης ● Χειρότερα αποτελέσματα σε εξεζητημένες εργασίες συλλογισμού

Κριτήρια Αξιολόγησης Απάντησης	● Συντομία και περιεκτικότητα ● Ακρίβεια νοήματος ● Διατήρηση κύριων πληροφοριών ● Γλωσσική και συντακτική ορθότητα
Εφαρμογές Zero-Shot Summarization	● Αναλύσεις κειμένων ● Αυτόματη δημιουργία σημειώσεων ● Δημοσιογραφία & ενημέρωση ● Εκπαιδευτικό υλικό
Στο παράδειγμα αυτό, η απουσία παραδειγμάτων αναγκάζει το μοντέλο να βασιστεί αποκλειστικά στην προ-εκπαιδευμένη γνώση του και στην ικανότητά του να αναγνωρίζει το γενικό γλωσσικό μοτίβο της εντολής «περίληψη».	

2.2.3 Προτροπή με λίγα παραδείγματα (Few-Shot Prompting)

Η τεχνική προτροπή με λίγα παραδείγματα (αγγλ. Few-Shot Prompting) αναγνωρίζεται ως μία από τις σημαντικότερες εξελίξεις στο πεδίο της σύγχρονης τεχνητής νοημοσύνης, καθώς καταδεικνύει την ικανότητα των LLMs να προσαρμόζονται σε νέες εργασίες μέσω της έκθεσης σε περιορισμένο αριθμό παραδειγμάτων εντός της προτροπής [7]. Η συγκεκριμένη μεθοδολογία έχει καθιερωθεί ως μία από τις πλέον αποδοτικές στρατηγικές στη μηχανική προτροπών.

Όπως επισημαίνει ο Gao (2023), στην προσέγγιση Few-Shot ο χρήστης παρέχει συνήθως από 2 έως 8 παραδείγματα (αγγλ. shots), τα οποία υποδεικνύουν στο μοντέλο τη φύση και τη δομή της ζητούμενης εργασίας. Το LLM αναγνωρίζει το σημασιολογικό μοτίβο (αγγλ. pattern) που συνδέει την είσοδο με την έξοδο και, εν συνεχεία, εφαρμόζει την ίδια λογική για να παραγάγει την απάντηση στη νέα ερώτηση. Η διαδικασία αυτή μειώνει δραστικά την ασάφεια που χαρακτηρίζει τη μέθοδο Zero-Shot και βελτιστοποιεί την ακρίβεια των αποτελεσμάτων [2]. Το φαινόμενο αυτό ερμηνεύεται από το γεγονός ότι τα παραδείγματα λειτουργούν ως ένα άμεσο σύνολο εκπαίδευσης, το οποίο καθοδηγεί το μοντέλο. Μέσω της χρήσης κατάλληλων δειγμάτων, το LLM δύναται να μιμηθεί με ακρίβεια το επιθυμητό ύφος, τη δομή, το επίπεδο λεπτομέρειας, καθώς και το συνολικό στυλ της απόκρισης.

Ο Nguyen και συν. (2025) τονίζουν ότι η Few-Shot είναι ιδιαίτερα αποτελεσματική σε κατηγορίες εργασιών που απαιτούν λογικές συσχετίσεις, μετασχηματισμό ή αναδόμηση κειμένου, ανάλυση συναισθήματος, καθώς και σε περιπτώσεις όπου απαιτείται αυστηρά προτυποποιημένη μορφή εξόδου. Επιπροσθέτως, προηγμένες τεχνικές συλλογισμού, όπως η CoT, βασίζονται σε μεγάλο βαθμό στη μέθοδο Few-Shot, δεδομένου ότι ο χρήστης καλείται να παρέχει δείγματα που αποτυπώνουν όχι μόνο το τελικό αποτέλεσμα, αλλά και τη δομή του συλλογισμού [3].

Ωστόσο, η μέθοδος υπόκειται σε συγκεκριμένους περιορισμούς. Όπως αναφέρει ο Obuchowski (n.d.), η επιλογή των παραδειγμάτων καθίσταται εξαιρετικά κρίσιμη παράμετρος, καθώς λανθασμένα, ασαφή ή κακοδιατυπωμένα παραδείγματα ενδέχεται να παραπλανήσουν το μοντέλο, οδηγώντας σε εσφαλμένα συμπεράσματα. Επιπλέον, το πεπερασμένο μέγεθος των προτροπών, θέτει όρια στο πλήθος των παραδειγμάτων που μπορούν να συμπεριληφθούν, γεγονός που δύναται να περιορίσει την αποτελεσματικότητα της τεχνικής σε πολύπλοκα προβλήματα.

Συμπερασματικά, η Few-Shot θεωρείται συχνά ως η χρυσή τομή ανάμεσα στην απλότητα εφαρμογής και την αποτελεσματική καθοδήγηση, καθώς προσφέρει σημαντική βελτίωση στην απόδοση χωρίς την πολυπλοκότητα και το κόστος που απαιτεί η διαδικασία της πλήρους μικρορύθμισης.

Πίνακας 2 Παράδειγμα με προτροπή Few-Shot

Κατηγορία	Αναλυτική Περιγραφή
Τεχνική Prompting	**Few-Shot Prompting** – το LLM λαμβάνει λίγα παραδείγματα (shots) πριν από την τελική ερώτηση, ώστε να «μάθει» το μοτίβο απάντησης.
Σκοπός	Κατηγοριοποίηση συναισθήματος (Sentiment Classification) με βάση τα παραδείγματα που δίνονται.
Χαρακτηριστικό Few-Shot	Το μοντέλο δεν βασίζεται μόνο στη δική του προεκπαιδευμένη γνώση, αλλά μαθαίνει το επιθυμητό μοτίβο από τα παραδείγματα (input → output) μέσα στην ίδια την προτροπή.
Prompt προς το LLM	**Παράδειγμα 1**: Ερώτηση: «Μου άρεσε πολύ η ταινία» – Θετική **Παράδειγμα 2**: Ερώτηση: «Ήταν απαίσιο φαγητό» – Αρνητική **Τελική Ερώτηση:** «Δεν ήταν κακό, αλλά περίμενα περισσότερα» –
Ενδεικτική Απάντηση του LLM	«Αρνητική» ή «Ουδέτερη»
Απαίτηση από το LLM	● Να αναγνωρίσει το μοτίβο ταξινόμισης: κάθε ερώτηση πρέπει να απαντηθεί μόνο με «Θετική» ή «Αρνητική». ● Να υιοθετήσει το ύφος και τη δομή της απάντησης (π.χ. μονολεκτική απάντηση). ● Να ταξινομήσει σωστά το συναίσθημα της νέας φράσης.
Αναμενόμενος Μηχανισμός Λειτουργίας	1. **Αναλύει τα παραδείγματα** και εξάγει το μοτίβο απάντησης. 2. **Εντοπίζει τη συναισθηματική πολικότητα** στη νέα φράση. 3. **Παράγει απάντηση στο ίδιο ύφος** και με τον ίδιο τύπο κατηγοριοποίησης.
Πλεονεκτήματα Few-Shot	● Μεγαλύτερη ακρίβεια από το Zero-Shot. ● Το LLM μιμείται τη μορφή απάντησης που ορίζει ο χρήστης. ● Όχι ανάγκη για εκπαίδευση ή fine-tuning. ● Καλή απόδοση σε ποικιλία εργασιών.
Μειονεκτήματα	● Κίνδυνος υιοθέτησης λανθασμένων συσχετίσεων από τα παραδείγματα. ● Απαιτεί προσεκτική επιλογή ποιοτικών παραδειγμάτων. ● Μικρή αλλαγή στη διατύπωση των παραδειγμάτων μπορεί να αλλάξει την ποιότητα της απάντησης.
Κριτήρια Αξιολόγησης Απάντησης	● Συνέπεια με το μοτίβο των παραδειγμάτων. ● Σωστή ταξινόμηση συναισθήματος. ● Σαφήνεια και σύντομη απόδοση (μία λέξη ή σύντομη φράση). ● Συμμόρφωση με τη ζητούμενη μορφή (format).
Εφαρμογές Few-Shot	● Συναισθηματική ανάλυση κειμένων. ● Κατηγοριοποίηση κειμένων (email, σχολίων, reviews). ● Ερωτήσεις–απαντήσεις με συγκεκριμένο μοτίβο. ● Εξαγωγή θεμάτων / κατηγοριών. ● Ορισμός νέων κανόνων ταξινόμησης εντός της προτροπής.

2.2.4 Μάθηση εντός πλαισίου (In-Context Learning)

Ένα από τα πλέον διακριτά και επαναστατικά χαρακτηριστικά των σύγχρονων LLMs συνιστά η ικανότητά τους να αποκτούν γνώση και να προσαρμόζουν τη συμπεριφορά τους μέσω των συμφραζομένων, διαδικασία γνωστή ως μάθηση εντός πλαισίου (αγγλ. In-Context Learning, ICL) [7]. Η ICL ορίζεται ως μια αναδυόμενη ιδιότητα των μοντέλων μεγάλης κλίμακας, η οποία τους επιτρέπει να εκτελούν νέες, άγνωστες εργασίες αξιοποιώντας αποκλειστικά τα παραδείγματα που παρέχονται εντός της προτροπής.

Λειτουργικά, η ICL δρα ως μια μορφή «προσωρινής εκπαίδευσης», η οποία δεν απαιτεί καμία τροποποίηση των εσωτερικών παραμέτρων του δικτύου, ούτε τη διαδικασία της μικρο-ρύθμισης [8]. Αποτελεί, ουσιαστικά, τη θεμελιώδη αρχή που καθιστά τα LLMs ικανά να γενικεύουν σε πραγματικό χρόνο. Παράλληλα, η ICL αποτελεί το λειτουργικό υπόβαθρο επί του οποίου θεμελιώνονται οι προηγμένες τεχνικές προτροπών, όπως η CoT, η SC και η ToT Μέσω του μηχανισμού αυτού, καθίσταται εφικτή η καθοδήγηση του μοντέλου σε συγκεκριμένα γνωσιακά μοτίβα συλλογισμού, αποκλειστικά μέσω της παράθεσης στοχευμένων παραδειγμάτων εντός της προτροπής.

Σύμφωνα με τον Obuchowski (n.d.), η ICL προσφέρει το απαραίτητο ερμηνευτικό πλαίσιο για την κατανόηση της ικανότητας των LLMs να ανταποκρίνονται επιτυχώς σε εργασίες που δεν περιλαμβάνονταν ρητά στο σύνολο δεδομένων εκπαίδευσής τους. Η δυναμική αυτή, κατά την οποία το μοντέλο υιοθετεί νέα συμπεριφορά αποκλειστικά βάσει του περιεχομένου της προτροπής, συνιστά ένα από τα πλέον αξιοσημείωτα χαρακτηριστικά των σύγχρονων LLMs.

Στον Πίνακα που ακολουθεί, παρουσιάζεται ένα χαρακτηριστικό παράδειγμα όπου το μοντέλο μαθαίνει προσωρινά τη μεθοδολογία επίλυσης προβλημάτων μαθηματικού συλλογισμού πολλαπλών βημάτων, απλώς αναλύοντας τα παρεχόμενα παραδείγματα

Πίνακας 3 Παράδειγμα με ICL

Κατηγορία	**Αναλυτική Περιγραφή**
Τεχνική Prompting	**In-Context Learning (ICL)** – το LLM μαθαίνει εντός της προτροπής το μοτίβο λύσης, χωρίς ενημέρωση παραμέτρων ή εκπαίδευση.
Σκοπός	Αριθμητικό πρόβλημα πολλαπλασιασμού μέσω παραδειγμάτων που λειτουργούν ως επεξήγηση της διαδικασίας.
Χαρακτηριστικό ICL	Το μοντέλο μαθαίνει προσωρινά τη λογική επίλυσης από τα παραδείγματα (patterns) που παρέχονται και την αναπαράγει στην τελική ερώτηση, **χωρίς fine-tuning**.
Prompt προς το LLM	**Παράδειγμα 1:** Ερώτηση: «Ο Γιάννης έχει 3 κουτιά με 5 μολύβια το καθένα. Πόσα μολύβια έχει συνολικά;» Απάντηση: «3 × 5 = 15. Άρα έχει 15 μολύβια.» **Παράδειγμα 2:** Ερώτηση: «Η Μαρία αγόρασε 4 τσάντες και κάθε τσάντα έχει μέσα 2 βιβλία. Πόσα βιβλία έχει συνολικά;» Απάντηση: «4 × 2 = 8. Άρα έχει 8 βιβλία.» **Τελική Ερώτηση:** «Ο Πέτρος έχει 6 κουτιά και κάθε κουτί περιέχει 3 μπάλες. Πόσες μπάλες έχει συνολικά;»
Ενδεικτική Απάντηση του LLM	«6 × 3 = 18. Άρα έχει 18 μπάλες.»
Απαίτηση από το LLM	● Να εντοπίσει ότι τα παραδείγματα ακολουθούν την ίδια λογική: *πλήθος κουτιών × περιεχόμενο ανά κουτί*. ● Να εφαρμόσει ακριβώς το ίδιο μοτίβο στην τελική ερώτηση. ● Να παράγει απάντηση με παρόμοια μορφή (εξίσωση → συμπέρασμα).
Αναμενόμενος Μηχανισμός Λειτουργίας	1. Αναγνωρίζει κοινή αριθμητική δομή στα παραδείγματα. 2. Εξάγει το μοτίβο πολλαπλασιασμού ως «κανόνα» επίλυσης. 3. Αναπαράγει τη μέθοδο στη νέα άσκηση. 4. Παράγει απάντηση στο ίδιο πρότυπο (π.χ., «6 × 3 = 18. Άρα…»).
Πλεονεκτήματα ICL	● Πολύ υψηλή προσαρμοστικότητα σε νέα εργασία. ● Δεν απαιτεί fine-tuning ή αλλαγή παραμέτρων του μοντέλου. ● Το μοντέλο μιμείται με ακρίβεια την επιθυμητή μορφή λύσης. ● Κατάλληλο για εργασία με σαφές μοτίβο (patterns).

Μειονεκτήματα	● Η ποιότητα της απάντησης εξαρτάται από την ποιότητα/σαφήνεια των παραδειγμάτων. ● Μπορεί να παρερμηνεύσει το μοτίβο αν τα παραδείγματα είναι ασυνεπή. ● Η μάθηση είναι προσωρινή και δεν διατηρείται πέρα από τη προτροπή.
Κριτήρια Αξιολόγησης Απάντησης	● Τήρηση του μοτίβου που δόθηκε στα παραδείγματα. ● Ορθός πολλαπλασιασμός. ● Παρόμοια μορφή παρουσίασης της λύσης (εξίσωση και συμπέρασμα). ● Λογική συνέπεια.
Εφαρμογές ICL	● Μαθηματικά προβλήματα (π.χ., πολλαπλασιασμοί, ποσοστά). ● Γλωσσικά μοτίβα (μεταφράσεις, διορθώσεις). ● Λογικά προβλήματα και υπολογισμοί. ● Δομημένες εργασίες όπου η μορφή απάντησης είναι κρίσιμη. ● Προσωρινή εκμάθηση κανόνων για νέα εργασία εντός προτροπής.

Πίνακας 4 Συγκριτική αξιολόγηση Zero-Shot, Few-Shot και ICL

Τεχνική	**Κύρια χαρακτηριστικά**	**Πλεονεκτήματα**	**Περιορισμοί**
Zero-Shot	Μόνο εντολή	Ταχύτητα, απλότητα	Αστάθεια, χαμηλή ακρίβεια
Few-Shot	Παραδείγματα εισόδου-εξόδου	Υψηλή ακρίβεια, σταθερότητα	Απαιτεί προσεκτικά παραδείγματα
ICL	Προσωρινή μάθηση εντός προτροπής	Εξαιρετικό για σύνθετες εργασίες	Ευαισθησία στη δομή της προτροπής

2.2.5 Η μέθοδος Τυπικής Προτροπής (Standard Prompting)

Η μέθοδος τυπικής προτροπής (αγγλ. Standard Prompting) συνιστά την βασική και διαδεδομένη προσέγγιση αλληλεπίδρασης χρηστών με τα LLMs. Στο πλαίσιο της σύγχρονης βιβλιογραφίας, η μέθοδος ορίζεται ως η διαδικασία κατά την οποία το μοντέλο λαμβάνει μία λιτή, άμεση και σαφώς διατυπωμένη εντολή σε φυσική γλώσσα, χωρίς παροχή συμπληρωματικών παραδειγμάτων (few-shot), ενδιάμεσων βημάτων ή εξειδικευμένων μηχανισμών συλλογισμού όπως η CoT ή ToT. Η παραγόμενη απόκριση βασίζεται αποκλειστικά στην παγιωμένη γνώση που απέκτησε το μοντέλο κατά τη φάση της προ-εκπαίδευσης, καθώς και στην ικανότητά του να αναγνωρίζει λεκτικά και σημασιολογικά πρότυπα [7].

Σύμφωνα με τον Gao (2023), η Standard Prompting αποτελεί τη θεμελιώδη αρχή επί της οποίας αναπτύχθηκαν όλες οι μεταγενέστερες, εξελιγμένες προσεγγίσεις. Η αποτελεσματικότητά της συναντάται άμεσα με τη σαφήνεια της εντολής, τη συντακτική δομή, την ακρίβεια στη διατύπωση του στόχου και την απουσία μη ξεκάθαρων διατυπώσεων στην προτροπή. Αξίζει να τονιστεί ότι η απέριττη μορφή της Standard Prompting εκμεταλλεύεται στο έπακρο τις εγγενείς ιδιότητες των LLMs, όπως η στατιστική μοντελοποίηση, η συμπλήρωση προτύπων, ο τεράστιος όγκος κωδικοποιημένης γνώσης που έχει λάβει κατά την προ-εκπαίδευση και η ικανότητα γενίκευσης σε ποικίλες εργασίες [1].

Επιπροσθέτως, ο Liu και συν. (2021) επισημαίνουν ότι η Standard Prompting λειτουργεί ως η βασική μέθοδος αναφοράς (αγγλ. Baseline), δηλαδή ως το μέτρο σύγκρισης για την αξιολόγηση κάθε άλλης

προηγμένης τεχνικής. Ακριβώς επειδή δεν ενσωματώνει πρόσθετη πληροφορία ή καθοδήγηση συλλογισμού, συνιστά τον αντικειμενικότερο δείκτη των θεμελιωδών ικανοτήτων κατανόησης, ανάκλησης πληροφοριών και παραγωγής κειμένου ενός LLM.

Παρότι η μέθοδος χαρακτηρίζεται από απλότητα, εμφανίζει σημαντικούς περιορισμούς όταν η εργασία απαιτεί λογική πολλαπλών βημάτων, επαγωγικό συλλογισμό, αναλυτική αιτιολόγηση ή αριθμητική επεξεργασία. Τα LLMs τείνουν να αποτυγχάνουν σε προβλήματα που προϋποθέτουν δομημένη σκέψη, διότι η Standard Prompting δεν εξωθεί το μοντέλο στην εξωτερίκευση της συλλογιστικής του πορείας (αγγλ. step-by-step reasoning) ούτε στην παραγωγή ενδιάμεσων σταδίων [3].

Παρόμοια ευρήματα παρουσιάζουν η Reynolds και η McDonell (2021), σημειώνοντας ότι οι απαντήσεις ενός LLM σε σύνθετες εργασίες υπό καθεστώς Standard Prompting παραμένουν συχνά επιφανειακές ή εσφαλμένες. Οι ανωτέρω αδυναμίες αποτέλεσαν το έναυσμα για την ανάπτυξη εξελιγμένων τεχνικών, όπως η CoT, η SC και η ToT.

Παρά τους περιορισμούς της, η προσέγγιση παραμένει κρίσιμη για την επιστημονική αξιολόγηση των μοντέλων, καθώς επιτρέπει την ακριβή μέτρηση:

- της γλωσσικής ικανότητας,
- της σημασιολογικής κατανόησης,
- της πληρότητας της ανάλυσης,
- της ακρίβειας χωρίς μεθόδους ενίσχυσης συλλογισμού.

Συμπερασματικά, η Standard Prompting αποτελεί τη βάση της μηχανικής προτροπών και την πλέον ουδέτερη προσέγγιση. Λειτουργεί ως μηχανισμός αποκάλυψης των πραγματικών, ανεπεξέργαστων δυνατοτήτων ενός LLM, προσφέροντας το καθαρότερο δυνατό πλαίσιο αξιολόγησης πριν την εφαρμογή σύνθετων δομημένων ακολουθιών προτροπών.

2.3 Η μέθοδος Αλυσίδας Συλλογισμού (Chain-of-Thought)

Η μέθοδος αλυσίδας συλλογισμού (αγγλ. Chain-of-Thought, CoT) συνιστά μια θεμελιώδη τεχνική στη σύγχρονη μηχανική προτροπών, καθώς καθιστά εφικτή την επεξήγηση της συλλογιστικής πορείας των LLMs και τη βέλτιστη αντιμετώπιση προβλημάτων που απαιτούν πολυσταδιακή λογική επεξεργασία. Σε αντιδιαστολή με τη μέθοδο Standard Prompting, όπου το μοντέλο παράγει άμεσα την τελική απάντηση, η προσέγγιση CoT απαιτεί από το μοντέλο να εξωτερικεύσει ρητά τα ενδιάμεσα βήματα του συλλογισμού του πριν καταλήξει στο τελικό συμπέρασμα [3].

Επί της ουσίας, η CoT δεν περιορίζεται στην παρουσίαση της εισόδου (ερώτηση) και της εξόδου (απάντηση), αλλά αποτυπώνει το διακριτό «μονοπάτι σκέψης» (αγγλ. Reasoning Path) που συνδέει τα δύο άκρα. Η διαδικασία αυτή προσομοιάζει τον ανθρώπινο τρόπο επίλυσης σύνθετων προβλημάτων: αντί για την άμεση εξαγωγή συμπεράσματος, το σύστημα επεξεργάζεται σταδιακά κάθε τμήμα της πληροφορίας, δομώντας μια συνεκτική λογική αλυσίδα. Η προσέγγιση αυτή καθιστά τη γνωσιακή διαδικασία του μοντέλου διαφανή και ερμηνεύσιμη μέσω της φυσικής γλώσσας [11].

Θεμελιώδης παραδοχή της μεθόδου είναι ότι τα LLMs διαθέτουν εσωτερικές λογικές δυνατότητες, οι οποίες ωστόσο παραμένουν ανενεργές ελλείψει κατάλληλης καθοδήγησης. Όταν η προτροπή εμπλουτίζεται με παραδείγματα που επιδεικνύουν βήμα-προς-βήμα τον συλλογισμό, τα μοντέλα τείνουν να μιμούνται το συγκεκριμένο πρότυπο σκέψης, παράγοντας εντέλει ακριβέστερα και πιο σταθερά αποτελέσματα [7].

Η λογική αυτή είναι πλήρως εναρμονισμένη με όσα παρουσιάστηκαν σχετικά με τη χρήση της Few-Shot. Για να γίνει πιο απτή η διαφορά ανάμεσα στη Standard Prompting και στην CoT, ακολουθεί ένα απλό αριθμητικό παράδειγμα. Έστω η ερώτηση:

Πίνακας 5 Παράδειγμα βασισμένο στη Standard Prompting και στη CoT

«Ένα τρένο κινείται με 90 km/h. Πόση απόσταση θα καλύψει σε 3 ώρες;»
Με Standard Prompting, το μοντέλο θα έδινε άμεσα την τελική απάντηση:
«270 km.»
Με CoT, το μοντέλο παράγει τα ενδιάμεσα βήματα συλλογισμού:
«Το τρένο κινείται με 90 km/h για 3 ώρες. Άρα η απόσταση είναι 3 × 90 = 270 km.
Η τελική απάντηση είναι 270 km.»

Η φαινομενικά απλή προσθήκη ενδιάμεσων βημάτων αποδεικνύεται καθοριστική για την αποτελεσματικότητα της μεθόδου, ιδίως σε προβλήματα υψηλής πολυπλοκότητας. Πλήθος ερευνητικών μελετών επιβεβαιώνει ότι η υποχρεωτική διαδικασία βήμα-προς-βήμα συλλογισμού οδηγεί σε ποιοτικά ανώτερες απαντήσεις [3] και συγκεκριμένα:

- αποφεύγονται βιαστικά λάθη, γιατί η σκέψη σπάει σε μικρά και ελέγξιμα στάδια
- βελτιώνεται η συνέπεια της απάντησης
- μειώνονται τα φαινόμενα παραισθήσεων, καθώς το μοντέλο ευθυγραμμίζεται με τη δική του λογική ακολουθία
- η λύση γίνεται πιο ερμηνεύσιμη, κάτι ιδιαίτερα σημαντικό σε εφαρμογές υψηλής ευθύνης.

Συμπερασματικά, η μέθοδος CoT αξιοποιεί στο έπακρο τις αναδυόμενες ικανότητες (αγγλ. Emergent Abilities) των LLMs. Τα αποτελέσματα είναι ιδιαίτερα εμφανή σε πρότυπα αξιολόγησης μαθηματικού συλλογισμού, όπως το GSM8K. Στο συγκεκριμένο πρότυπο σύνολο αναφοράς, η εφαρμογή της CoT σε μοντέλα όπως το PaLM 540B οδήγησε σε κορυφαίες επιδόσεις, ξεπερνώντας ακόμη και συστήματα που είχαν εκπαιδευτεί ειδικά για την επίλυση τέτοιων προβλημάτων [3]. Το γεγονός αυτό καταδεικνύει ότι μέσω της μεθόδου CoT η βελτίωση επιτυγχάνεται αποκλειστικά μέσω του αποτελεσματικού σχεδιασμού προτροπών, χωρίς την απαίτηση πρόσθετης εκπαίδευσης.

Επεκτάσεις της CoT

Η θεμελιώδης επιτυχία της CoT αποτέλεσε το εφαλτήριο για την ανάπτυξη ισχυρότερων μεθόδων, δημιουργώντας μια νέα οικογένεια τεχνικών ενίσχυσης της λογικής. Μία από τις σημαντικότερες εξελίξεις είναι η αλυσιδωτή συλλογιστική με αυτο-συνέπεια (αγγλ. Self-Consistency Chain-of-Thought, SC-CoT). Η μέθοδος αυτή καινοτομεί προτείνοντας τη δημιουργία πολλαπλών, ανεξάρτητων αλυσίδων συλλογισμού για την ίδια είσοδο και την επιλογή της τελικής απάντησης βάσει της αρχής της πλειοψηφίας (αγγλ. majority voting) [15]. Μέσω της διαδικασίας αυτής, επιτυγχάνεται η σταθεροποίηση των απαντήσεων και ο δραστικός περιορισμός της επίδρασης τυχαίων λαθών ή εσφαλμένων συλλογιστικών μονοπατιών, αυξάνοντας σημαντικά την αξιοπιστία του τελικού αποτελέσματος.

Παράλληλα, αναπτύχθηκαν δομικά πιο σύνθετες προσεγγίσεις:

- Η ToT, η οποία επιτρέπει τη μη γραμμική εξερεύνηση πολλαπλών λογικών διαδρομών σε μορφή δέντρου αποφάσεων, ευνοώντας τη στρατηγική αναζήτηση και την αυτο-διόρθωση.
- Η Pipeline-CoT, όπου η διαδικασία συλλογισμού κατατέμνεται και εφαρμόζεται σε διακριτά, διαδοχικά στάδια επεξεργασίας για καλύτερο έλεγχο.
- Η Fact-CoT, η οποία ενισχύει την εγκυρότητα της εξαγωγής συμπερασμάτων ενσωματώνοντας τεκμηριωμένα πραγματικά δεδομένα (αγγλ. facts) στη ροή της σκέψης.

Οι εξελίξεις αυτές, σε συνδυασμό με τις αυξημένες ικανότητες σύγχρονων μοντέλων όπως το GPT-4 [16], επιβεβαιώνουν ότι η CoT έχει μετεξελιχθεί σε κεντρικό πυλώνα για τη βελτίωση και την ερμηνευσιμότητα των μηχανισμών λογικής στα LLMs.

Πίνακας 6 Παράδειγμα CoT

Ερώτηση	«Ένα κατάστημα είχε 45 τετράδια. Πούλησε 18 και μετά αγόρασε άλλα 25. Πόσα έχει τώρα;»
Βήμα	Συλλογισμός (CoT)
1	Το κατάστημα ξεκίνησε με 45 τετράδια.
2	Πούλησε 18 τετράδια, άρα υπολογίζουμε: 45 − 18 = 27.
3	Στη συνέχεια αγόρασε άλλα 25 τετράδια, οπότε: 27 + 25 = 52.
4	Άρα, στο τέλος το κατάστημα έχει 52 τετράδια.
Τελική απάντηση:	
52 τετράδια.	

2.4 Η μέθοδος Δενδρικής Συλλογιστικής (Tree-of-Thought)

Η μέθοδος Δενδρικής Συλλογιστικής (αγγλ. Tree-of-Thought, ToT) αποτελεί το επόμενο εξελικτικό βήμα στη βελτίωση των ικανοτήτων συλλογισμού των LLMs, επιχειρώντας να προσομοιώσει ακριβέστερα τις γνωσιακές διεργασίες της ανθρώπινης επίλυσης προβλημάτων. Η κεντρική καινοτομία της ToT έγκειται στην υπέρβαση της γραμμικής αλυσίδας σκέψης (αγγλ. linear chain) που χαρακτηρίζει την CoT. Αντ' αυτού, επιτρέπει στο μοντέλο να εξετάζει ταυτόχρονα πολλαπλές διαφορετικές πορείες συλλογισμού, δημιουργώντας ένα δέντρο εναλλακτικών λύσεων και πιθανών μονοπατιών [4]. Μέσω της προσέγγισης αυτής, η σκέψη δεν περιορίζεται σε μία μονοσήμαντη κατεύθυνση, αλλά επεκτείνεται στην εξέταση, σύγκριση και αξιολόγηση πολλαπλών πιθανών λύσεων πριν την επιλογή της καταλληλότερης.

Η ToT δύναται να χαρακτηριστεί ως ένα υβριδικό σύστημα που συνενώνει τις παραγωγικές δυνατότητες των LLMs με τις κλασικές αλγοριθμικές τεχνικές αναζήτησης της τεχνητής νοημοσύνης. Στο πλαίσιο αυτό, τα LLMs αναλαμβάνουν τον ρόλο της παραγωγής ιδεών και πιθανών επόμενων βημάτων, ενώ η αρχιτεκτονική της ToT οργανώνει και αξιολογεί αυτά τα βήματα ως ένα δέντρο αναζήτησης (αγγλ. search tree). Στο δέντρο αυτό, κάθε κόμβος αντιστοιχεί σε μια ενδιάμεση κατάσταση του προβλήματος και κάθε κλαδί σε μια πιθανή κατεύθυνση επίλυσης. Το σύστημα επιλέγει σταδιακά το μονοπάτι που κρίνεται ως το πλέον λογικό ή υποσχόμενο, συνδυάζοντας τη δημιουργικότητα των γλωσσικών μοντέλων με μια μεθοδική διαδικασία αξιολόγησης.

Στη δομική συγκρότηση της ToT, κάθε κόμβος (αγγλ. node) αντιπροσωπεύει μια ενδιάμεση νοητική κατάσταση, δηλαδή έναν μερικό συλλογισμό ή μια προσωρινή λύση. Κάθε κλαδί (αγγλ. branch) αντιστοιχεί σε μια εναλλακτική ιδέα ή στρατηγική επέκτασης. Το μοντέλο παράγει αρχικά πολλαπλές εκδοχές και εν συνεχεία τις αξιολογεί, αποφασίζοντας ποιες θα διερευνηθούν περαιτέρω και ποιες θα απορριφθούν ως ασθενείς ή αντιφατικές. Κατ' αυτόν τον τρόπο, το σύστημα μιμείται τους κλασικούς αλγορίθμους «Αναζήτησης σε Χώρο Καταστάσεων» (αγγλ. State Space Search), όπως η Αναζήτηση κατά Πλάτος (BFS) ή η Αναζήτηση κατά Βάθος (DFS) [4]. Η σκέψη του μοντέλου παύει να είναι μια ευθεία γραμμή και μετατρέπεται σε μια δυναμική αναζήτηση εντός ενός χώρου πολλαπλών επιλογών. Η διαδικασία αυτή ενισχύει τόσο τη δημιουργικότητα όσο και την ακρίβεια στην επίλυση σύνθετων προβλημάτων. Τα πειραματικά δεδομένα του Yao και συν. (2023) επιβεβαιώνουν την ανωτερότητα της μεθόδου, καθώς η εφαρμογή της ToT σε γνωστικά και μαθηματικά προβλήματα οδήγησε σε σημαντικά

υψηλότερη απόδοση συγκριτικά με την CoT. Η ToT προσφέρει στα LLMs μια ευέλικτη μορφή σκέψης, επιτρέποντάς τους:

- Να επανεξετάζουν και να αναθεωρούν τις ενδιάμεσες ιδέες τους.
- Να συγκρίνουν ταυτόχρονα διαφορετικές στρατηγικές κατευθύνσεις.
- Να απορρίπτουν εγκαίρως τα μονοπάτια που δεν οδηγούν σε βιώσιμη λύση.
- Να επιλέγουν τελικά το σενάριο που παρουσιάζει τη μέγιστη λογική συνέπεια.

Ένα κρίσιμο πλεονέκτημα της ToT είναι η δυνατότητα αναζήτησης με οπισθοχώρηση (αγγλ. Backtracking), δηλαδή η ικανότητα επιστροφής σε προηγούμενους κόμβους όταν μια επιλεγμένη διαδρομή αποδειχθεί ατελέσφορη. Στην παραδοσιακή CoT, ένα σφάλμα σε ενδιάμεσο βήμα οδηγεί αναπόφευκτα σε λανθασμένο τελικό αποτέλεσμα (διάδοση σφάλματος), καθώς το μοντέλο στερείται μηχανισμού αναθεώρησης. Αντιθέτως, στην ToT, ένας κλάδος μπορεί να εγκαταλειφθεί και η αναζήτηση να συνεχιστεί από εναλλακτικό σημείο του δέντρου, καθιστώντας τη συλλογιστική διαδικασία εξαιρετικά ανθεκτική σε σφάλματα (αγγλ. robustness) και πιο σταθερή σε σύνθετα προβλήματα. Θεμελιώδες θεωρητικό στοιχείο της ToT αποτελεί ο λειτουργικός διαχωρισμός μεταξύ της παραγωγής σκέψεων (αγγλ. thought generation) και της αξιολόγησης σκέψεων (αγγλ. thought evaluation). Η παραγωγή βασίζεται στα πιθανοτικά πρότυπα που έχει διδαχθεί το μοντέλο κατά την εκπαίδευσή του, ενώ η αξιολόγηση δύναται να πραγματοποιηθεί είτε από το ίδιο το μοντέλο (αγγλ. Self-Evaluation) είτε μέσω εξωτερικών μηχανισμών (π.χ. heuristics ή ελεγκτές ορθότητας). Η διάκριση αυτή επιτρέπει τον αυστηρότερο έλεγχο της συλλογιστικής πορείας και ελαχιστοποιεί τον κίνδυνο συσσώρευσης λαθών [17].

Ο συνδυασμός της προσέγγισης αυτής με τεχνικές όπως της SC και η Βελτίωση μέσω Συνόλου Λύσεων (αγγλ. Ensemble Refinement), οδηγεί στη δημιουργία μοντέλων με σταθερή, οργανωμένη και, κυρίως, αυτο-διορθούμενη πορεία σκέψης [15]. Κατ' επέκταση, η ToT αναγνωρίζεται σήμερα ως μία από τις πλέον προηγμένες μορφές δομημένων προτροπών, ενισχύοντας την ικανότητα των LLMs να παράγουν απαντήσεις υψηλής ποιότητας και ισχυρής λογικής συνοχής. Η αποτελεσματικότητά της είναι ιδιαίτερα εμφανής σε εργασίες που απαιτούν πολυβηματικό στρατηγικό σχεδιασμό, ανάλυση εναλλακτικών σεναρίων και λήψη αποφάσεων σε περιβάλλοντα υψηλής πολυπλοκότητας [18].

Πίνακας 7 Παράδειγμα ToT

Ερώτηση	«Ένας μαθητής μπορεί να διαβάσει είτε στο σπίτι, είτε στη βιβλιοθήκη, είτε σε καφετέρια. Ποιο μέρος είναι πιθανότερα πιο ήσυχο για μελέτη;»
Branch	**Συλλογισμός**
Branch 1: Σπίτι	Μπορεί να είναι ήσυχο.– Πιθανότητα θορύβου από οικογένεια ή δραστηριότητες στο σπίτι.
Branch 2: Βιβλιοθήκη	Είναι χώρος σχεδιασμένος για μελέτη.– Συνήθως υπάρχουν κανόνες ησυχίας.– Παρέχει σταθερά χαμηλό επίπεδο θορύβου.
Branch 3: Καφετέρια	Υπάρχει μουσική, συνομιλίες και γενικός θόρυβος.– Δεν αποτελεί ιδανικό περιβάλλον συγκέντρωσης.
Τελική αξιολόγηση (Root Decision): Το branch της βιβλιοθήκης παρουσιάζει τη μεγαλύτερη λογική συνοχή και τις πιο σταθερές συνθήκες ησυχίας.	
Τελική απάντηση: Η βιβλιοθήκη είναι το πιο ήσυχο και κατάλληλο μέρος για μελέτη.	

Κεφάλαιο 3ο: Συναφής έρευνα και σύνδεση με την παρούσα εργασία

3.1 Η Μηχανική Προτροπών ως ανεξάρτητος μηχανισμός βελτιστοποίησης των LLMs

Η συστηματική ανασκόπηση των Vatsal και Dubey (2024) τεκμηριώνει ότι η μηχανική προτροπών αποτελεί έναν αυτοτελή και εξαιρετικά αποδοτικό μηχανισμό βελτίωσης της απόδοσης των LLMs, χωρίς την ανάγκη επανεκπαίδευσης. Μέσα από τη μελέτη 39 τεχνικών σε 29 εργασίες NLP, η έρευνα καταδεικνύει ότι οι τεχνικές που εισάγουν δομή στον συλλογισμό επιτυγχάνουν τις μεγαλύτερες και πιο σταθερές βελτιώσεις. Συγκεκριμένα, αναφέρονται αυξήσεις ακρίβειας έως και 30 - 40% σε μαθηματικά προβλήματα με CoT, περαιτέρω βελτιώσεις 6 - 11% με SC, ενώ οι προσεγγίσεις τύπου ToT εμφανίζουν τα υψηλότερα κέρδη σε προβλήματα αυξημένης πολυπλοκότητας, με βελτιώσεις που υπερβαίνουν το 50% σε σχέση με τη μέθοδο αναφοράς.

Τα ευρήματα αυτά συνδέονται άμεσα με την παρούσα εργασία, η οποία υιοθετεί τρεις αντιπροσωπευτικές κατηγορίες prompting (Standard, CoT, ToT) και επιβεβαιώνει εμπειρικά ότι η απόδοση των LLMs επηρεάζεται καθοριστικά από τη δομή της προτροπής, ακόμη και όταν οι παράμετροι του μοντέλου παραμένουν αμετάβλητοι.

3.2 Chain-of-Thought και αναδυόμενες ικανότητες συλλογισμού

Οι θεμελιώδεις εργασίες των Wei και συν. (2023), Brown και συν. (2020) και Wang και συν. (2022) συγκροτούν έναν ενιαίο άξονα που αφορά τη φύση και τους μηχανισμούς της CoT. Στα μαθηματικά προβλήματα όπως το GSM8K, η υιοθέτηση της CoT επιφέρει δραματική βελτίωση της απόδοσης, οδηγώντας συχνά σε διπλασιασμό της ακρίβειας έναντι της Τυπικής Προτροπής (Standard Prompting), με τα μεγάλα γλωσσικά μοντέλα να επιτυγχάνουν ποσοστά που υπερβαίνουν το 55-60%, σε πλήρη αντιδιαστολή με τις επιδόσεις κάτω του 30% που καταγράφονται απουσία της μεθόδου. Παράλληλα, το φαινόμενο αυτό είναι άμεσα συνδεδεμένο με την κλίμακα του μοντέλου, καθώς μικρότερα μοντέλα αποτυγχάνουν να αξιοποιήσουν τη CoT [7]. Επιπλέον, η μελέτη των Wang και συν. (2022) συμπληρώνει τα ποσοτικά αποτελέσματα με ποιοτική ανάλυση, καταδεικνύοντας ότι η CoT δεν απαιτεί τέλεια παραδείγματα, αλλά συνεπή και σχετική δομή συλλογισμού. Το εύρημα αυτό ερμηνεύει τη λειτουργία της CoT ως μηχανισμού σταθεροποίησης, ο οποίος καθοδηγεί το μοντέλο προς τη σωστή κατεύθυνση, ακόμη και όταν μεμονωμένα ενδιάμεσα βήματα ενέχουν μικρές ανακρίβειες.

Η παρούσα εργασία επιβεβαιώνει τα παραπάνω ευρήματα, καθώς τα πειραματικά αποτελέσματα στο GSM8K δείχνουν σαφή βελτίωση της ακρίβειας με τη μετάβαση από Standard Prompting σε CoT, καθώς και αυξημένη σταθερότητα των αποκρίσεων.

3.3 Self-Consistency και μείωση στοχαστικής αστάθειας

Η ανάγκη περιορισμού της στοχαστικής αστάθειας των LLMs αναδεικνύεται στα άρθρα των Wang και συν. (2023) και Liu και συν. (2024). Τα ερευνητικά δεδομένα καταδεικνύουν ότι η μέθοδος SC αυξάνει την ακρίβεια στο GSM8K κατά 8 - 18 ποσοστιαίες μονάδες σε σχέση με την απλή CoT, μέσω της πλειοψηφικής επιλογής μεταξύ πολλαπλών συλλογιστικών διαδρομών. Επιπλέον, μειώνεται σημαντικά η διακύμανση μεταξύ επαναλήψεων, αυξάνοντας την αξιοπιστία των αποτελεσμάτων.

Η παρούσα εργασία εφαρμόζει αντίστοιχο μηχανισμό, επιβεβαιώνοντας ότι η συνάθροιση πολλαπλών απαντήσεων οδηγεί σε πιο σταθερά και αναπαραγώγιμα αποτελέσματα, ιδίως σε μοντέλα με υψηλή στοχαστικότητα.

3.4 Δενδροειδής συλλογισμός και στρατηγική εξερεύνηση μέσω Tree-of-Thought

Οι πρωτοποριακές προσεγγίσεις των Yao και συν. (2023) και Long (2023) καθιερώνουν τη μέθοδο ToT ως ένα πλαίσιο στρατηγικής εξερεύνησης, το οποίο επιτρέπει στα γλωσσικά μοντέλα να λειτουργούν πέρα από τους περιορισμούς της γραμμικής λήψης αποφάσεων. Τα εμπειρικά δεδομένα είναι αποκαλυπτικά: στο πρόβλημα «Game of 24», η ToT επιτυγχάνει ποσοστά επιτυχίας άνω του 70%, έναντι μόλις 20% για την CoT και 4% για την Zero-Shot. Αντιστοίχως, σε γρίφους Sudoku υψηλής πολυπλοκότητας, οι δενδροειδείς αρχιτεκτονικές υπερέχουν εντυπωσιακά, καταγράφοντας προβάδισμα 50 έως 60 ποσοστιαίων μονάδων έναντι των συμβατικών γραμμικών μεθόδων.

Η παρούσα εργασία ενσωματώνει την ToT σε ένα αυστηρό συγκριτικό πλαίσιο αξιολόγησης, επαληθεύοντας ότι η μη γραμμική εξερεύνηση δύναται να αποφέρει σημαντικά περιθώρια βελτίωσης σε προβλήματα υψηλής δυσκολίας. Ωστόσο, τα ευρήματα της εργασίας επιβεβαιώνουν παράλληλα τις βιβλιογραφικές επισημάνσεις αναφορικά με το τίμημα αυτής της βελτίωσης, το οποίο μεταφράζεται σε δραματική αύξηση του υπολογιστικού κόστους σε tokens και του χρόνου απόκρισης.

3.5 Αξιολόγηση ποιότητας συλλογισμού και LLM-as-a-Judge

Οι ερευνητικές εργασίες των Gu και συν. (2025) και Aithal και συν. (2025) αναδεικνύουν ότι η απλή μέτρηση της ακρίβειας δεν επαρκεί για την αξιολόγηση των συλλογιστικών αποκρίσεων των μοντέλων. Τα ερευνητικά δεδομένα καταδεικνύουν ότι τα LLMs μπορούν να λειτουργήσουν ως αξιόπιστοι κριτές, επιδεικνύοντας υψηλό βαθμό συσχέτισης με την ανθρώπινη κρίση, παρουσιάζοντας ωστόσο συχνά εγγενείς μεροληψίες (π.χ. προτίμηση σε εκτενέστερες απαντήσεις). Ως εκ τούτου, η σύγχρονη βιβλιογραφία προκρίνει τη χρήση τεχνικών συνάθροισης και τον αυστηρό σχεδιασμό των κριτηρίων αξιολόγησης για τη διασφάλιση της αντικειμενικότητας.

Σε εναρμόνιση με τις τάσεις αυτές, η παρούσα διατριβή ενσωματώνει τη μεθοδολογία «LLM-as-a-Judge» για την ποιοτική αποτίμηση της λογικής συνοχής και της σαφήνειας των παραγόμενων αποκρίσεων. Η προσέγγιση αυτή επιβεβαιώνει ότι οι μέθοδοι CoT και ToT υπερέχουν ποιοτικά έναντι της Τυπικής Προτροπής (Standard Prompting), ακόμη και σε περιπτώσεις όπου το τελικό αριθμητικό αποτέλεσμα είναι ορθό.

3.6 Όρια γενίκευσης και εξάρτηση από το είδος της εργασίας

Οι πρόσφατες μελέτες των Chochlakis και συν. (2024), Nguyen και συν. (2025), Joshi (2025), καθώς και των Dai και Qin (2023), οριοθετούν με σαφήνεια το πεδίο εφαρμογής των τεχνικών προτροπής, καταδεικνύοντας ότι η αποτελεσματικότητά τους τελεί σε άμεση συνάρτηση με τη γνωσιακή φύση του εκάστοτε προβλήματος. Τα ερευνητικά δεδομένα υποδεικνύουν ότι σε εργασίες υποκειμενικού χαρακτήρα, η εφαρμογή της CoT δεν προσφέρει μετρήσιμα οφέλη, ενώ σε τομείς όπως η μετάφραση και η χρηματοοικονομική ανάλυση, τα αποτελέσματα κρίνονται ως μικτά, εξαρτώμενα σε μεγάλο βαθμό από τη δομική πολυπλοκότητα της εισόδου. Αντίθετα, εξειδικευμένες παραλλαγές όπως η Fact-CoT επιτυγχάνουν βελτιώσεις της τάξης του 8-12% σε σενάρια συλλογισμού κοινής λογικής, συμβάλλοντας δραστικά στον περιορισμό των λογικών σφαλμάτων.

Τα ευρήματα αυτά επικυρώνουν την εγκυρότητα των αποτελεσμάτων της παρούσας διατριβής για αντικειμενικά, πολυσταδιακά προβλήματα όπως αυτά του συνόλου δεδομένων GSM8K,

υπογραμμίζοντας ταυτόχρονα ότι η υπεροχή των δομημένων μεθόδων δεν πρέπει να γενικεύεται άκριτα σε κάθε τύπο εργασίας, αλλά απαιτεί προσεκτική προσαρμογή στις απαιτήσεις της εκάστοτε εφαρμογής.

Κεφάλαιο 4ο: Μεθοδολογία

4.1 Σχεδιασμός συγκριτικής ανάλυσης

Σκοπός του παρόντος κεφαλαίου είναι η παρουσίαση του μεθοδολογικού πλαισίου και του συνολικού σχεδιασμού της συγκριτικής ανάλυσης που διενεργείται στην εργασία. Η ανάλυση αυτή στοχεύει στη συστηματική και σε βάθος διερεύνηση της αποτελεσματικότητας διαφορετικών τεχνικών προτροπών στα LLMs, εστιάζοντας συγκεκριμένα στη σύγκριση τριών μεθόδων: της Standard Prompting, της CoT και της ToT. Η ανάγκη για έναν αυστηρά δομημένο πειραματικό σχεδιασμό προκύπτει από το γεγονός ότι, παρά τις εντυπωσιακές ικανότητες των LLMs σε πλήθος εφαρμογών, η συμπεριφορά και η απόδοσή τους επηρεάζονται καθοριστικά από τον τρόπο διατύπωσης της προτροπής. Ειδικότερα, προηγούμενες μελέτες έχουν καταδείξει ότι τεχνικές όπως η CoT δύνανται να βελτιώσουν αισθητά την ακρίβεια σε προβλήματα που απαιτούν συλλογισμό πολλαπλών βημάτων, όπως αυτά που περιλαμβάνονται στο σύνολο δεδομένων GSM8K, έναντι της απλής, άμεσης διατύπωσης της ερώτησης [3].

Ωστόσο, η βελτίωση της ακρίβειας συνοδεύεται συχνά από σημαντικά αντισταθμίσματα, όπως το αυξημένο υπολογιστικό κόστος, οι μεγαλύτεροι χρόνοι απόκρισης (αγγλ. latency) και η μειωμένη συνέπεια μεταξύ των επαναλήψεων. Προκειμένου να αξιολογηθούν αντικειμενικά αυτές οι παράμετροι, ο σχεδιασμός της ανάλυσης βασίζεται σε ένα ενιαίο, τυποποιημένο και αναπαραγώγιμο πειραματικό πλαίσιο, όπου όλες οι μέθοδοι εξετάζονται υπό πανομοιότυπες συνθήκες. Συγκεκριμένα, χρησιμοποιείται το ίδιο σύνολο δεδομένων (GSM8K), οι ίδιες παραμετροποιήσεις στα μοντέλα και κοινά κριτήρια αξιολόγησης για κάθε πείραμα, διασφαλίζοντας ότι οι παρατηρούμενες διαφοροποιήσεις στα αποτελέσματα αποδίδονται αποκλειστικά στη μέθοδο προτροπής και όχι σε εξωγενείς παράγοντες. Η διαχείριση και προετοιμασία των δεδομένων πραγματοποιείται με τυποποιημένες διαδικασίες, εξάγοντας με συνέπεια την πρότυπη σωστή απάντηση (αγγλ. Gold Answer) για κάθε ερώτημα προς σύγκριση.

Η υλοποίηση της πειραματικής διαδικασίας πραγματοποιείται μέσω μιας αυτοματοποιημένης ροής εργασιών pipeline, η οποία έχει σχεδιαστεί για να διαχειρίζεται το σύνολο του κύκλου αξιολόγησης. Το σύστημα αυτό αναλαμβάνει την εκτέλεση των διαφορετικών στρατηγικών προτροπής, τη συστηματική συλλογή των απαντήσεων από τα μοντέλα και την καταγραφή των βασικών μετρικών απόδοσης, όπως η ακρίβεια, ο χρόνος εκτέλεσης και η συνέπεια. Επιπλέον, το pipeline παράγει αυτόματα συγκεντρωτικούς πίνακες και διαγράμματα, διευκολύνοντας την περαιτέρω στατιστική επεξεργασία. Ιδιαίτερη έμφαση δίνεται στο γεγονός ότι η αξιολόγηση δεν περιορίζεται σε μια διπολική προσέγγιση (σωστό/λάθος). Αντιθέτως, συνεκτιμώνται κρίσιμα ποιοτικά χαρακτηριστικά της συμπεριφοράς των μοντέλων, όπως η σταθερότητα των απαντήσεων σε πολλαπλές εκτελέσεις μέσω της μετρικής Consistency και η αξιοπιστία της κρίσης, υιοθετώντας στοιχεία από τη σύγχρονη προσέγγιση «LLM-as-a-Judge» [12].

4.2 Περιγραφή δεδομένων (GSM8K)

Για τη διεξαγωγή της συγκριτικής ανάλυσης της παρούσας εργασίας χρησιμοποιείται το μαθηματικό σύνολο δεδομένων GSM8K (Grade School Math 8K), το οποίο αποτελεί ένα καθιερωμένο πρότυπο αναφοράς (Benchmark) για την αξιολόγηση των ικανοτήτων συλλογισμού και αριθμητικής επίλυσης των LLMs και έχει αξιοποιηθεί εκτενώς στη σχετική βιβλιογραφία [3] [19]. Το σύνολο περιλαμβάνει

περίπου 8.500 μαθηματικά προβλήματα λεκτικής μορφής, τα οποία αντιστοιχούν σε επίπεδο δημοτικού και πρώτων τάξεων γυμνασίου. Τα προβλήματα αυτά είναι σχεδιασμένα ώστε να απαιτούν συλλογισμό πολλαπλών βημάτων, συνήθως κυμαινόμενα από 2 έως 8 διακριτά στάδια, και βασίζονται αποκλειστικά σε στοιχειώδεις αριθμητικές πράξεις, όπως πρόσθεση, αφαίρεση, πολλαπλασιασμό και διαίρεση. Σύμφωνα με τον Cobbe και συν. (2021), τα εν λόγω προβλήματα δεν απαιτούν εξειδικευμένες μαθηματικές γνώσεις και θεωρείται ότι μπορούν να επιλυθούν από έναν μέσο μαθητή γυμνασίου.

Ένα ιδιαίτερα σημαντικό χαρακτηριστικό του GSM8K είναι η μορφή με την οποία παρέχονται οι λύσεις. Οι απαντήσεις δεν διατυπώνονται ως απλές μαθηματικές εκφράσεις, αλλά σε φυσική γλώσσα, περιγράφοντας αναλυτικά τα ενδιάμεσα βήματα του συλλογισμού και καταλήγοντας στο τελικό αριθμητικό αποτέλεσμα. Η συγκεκριμένη δομή επιλέχθηκε σκόπιμα, καθώς κρίνεται ως η πλέον κατάλληλη για τη μελέτη των «εσωτερικών μονολόγων» των LLMs και της διαδικασίας σκέψης τους [19]. Στη βασική (αγγλ. main) διαμόρφωση του συνόλου δεδομένων, κάθε εγγραφή αποτελείται από δύο θεμελιώδη πεδία:

- question: το κείμενο της μαθηματικής ερώτησης σε φυσική γλώσσα,
- answer: την πλήρη λύση του προβλήματος, η οποία περιέχει τα βήματα συλλογισμού, ενδιάμεσους υπολογισμούς και στο τέλος το τελικό αποτέλεσμα, το οποίο δηλώνεται ρητά με τον δείκτη ####.

Η τυποποίηση αυτή αξιοποιείται άμεσα στον πειραματικό κώδικα της εργασίας, επιτρέποντας την αυτόματη εξαγωγή της ορθής απάντησης (αγγλ. Gold Answer) από το πεδίο answer, ώστε να καθίσταται δυνατή η αντικειμενική και αυτοματοποιημένη σύγκριση με τις απαντήσεις που παράγουν τα μοντέλα.

Αξίζει να σημειωθεί ότι το GSM8K διατίθεται και σε εναλλακτική «σωκρατική» (socratic) διαμόρφωση, όπου η λύση οργανώνεται σε υπο-ερωτήματα. Ωστόσο, στην παρούσα εργασία επιλέχθηκε η βασική διαμόρφωση, καθώς είναι η πλέον διαδεδομένη στη βιβλιογραφία και επιτρέπει την άμεση σύγκριση με προηγούμενες μελέτες, όπως αυτές του Wei και συν. (2023). Το σύνολο δεδομένων διαχωρίζεται σε δύο υποσύνολα: το σύνολο εκπαίδευσης (αγγλ. training set), που περιλαμβάνει 7.473 ερωτήσεις, και το σύνολο αξιολόγησης (αγγλ. validation/test set), που αποτελείται από 1.319 ερωτήσεις. Η παρούσα ανάλυση επικεντρώνεται αποκλειστικά στο σύνολο αξιολόγησης, καθώς ο στόχος είναι η μέτρηση της απόδοσης των τεχνικών προτροπής σε δεδομένα που το μοντέλο δεν έχει χρησιμοποιήσει για προσαρμογή ή εκπαίδευση. Τέλος, η δημιουργία του GSM8K βασίστηκε σε ανθρώπινη συγγραφή και επαλήθευση μέσω της πλατφόρμας Surge AI, διασφαλίζοντας υψηλή ποιότητα [19]. Παρότι έχουν εφαρμοστεί αυστηροί έλεγχοι, εκτιμάται ότι ένα μικρό ποσοστό (περίπου 1,7%) ενδέχεται να περιέχει ασάφειες ή σφάλματα, παράγοντας που λαμβάνεται υπόψη κατά την ερμηνεία των αποτελεσμάτων.

4.3 Παραμετροποίηση των μοντέλων

Η παραμετροποίηση των LLMs αποτελεί θεμέλιο της πειραματικής διαδικασίας, καθώς καθορίζει τον τρόπο με τον οποίο το σύστημα προσαρμόζεται στις τεχνικές συλλογισμού που εξετάζονται στο πλαίσιο της συγκριτικής ανάλυσης. Με βάση τον σχεδιασμό που έχει παρουσιαστεί στις προηγούμενες ενότητες (3.1 και 3.2), το παρόν πειραματικό περιβάλλον απαιτεί την εκτέλεση πολλαπλών μοντέλων και μεθόδων προτροπών υπό απολύτως ελεγχόμενες συνθήκες. Για τον λόγο αυτό, υλοποιήθηκε μια ενιαία αρχιτεκτονική παραμετροποίησης, η οποία επιτρέπει την απρόσκοπτη εναλλαγή μεταξύ τοπικών μοντέλων (μέσω της πλατφόρμας Ollama) και απομακρυσμένων παρόχων (OpenAI, Google Gemini, DeepSeek) χωρίς να απαιτείται τροποποίηση στον κώδικα που επιτελεί τις μεθόδους συλλογισμού. Η μετάβαση από έναν πάροχο σε άλλον πραγματοποιείται αποκλειστικά μέσω μεταβλητών

περιβάλλοντος, οι οποίες καθορίζουν τόσο την ταυτότητα του μοντέλου όσο και τις παραμέτρους λειτουργίας του.

Στο αρχείο pipeline.py υλοποιείται ένας μηχανισμός αναγνώρισης του κατάλληλου παρόχου, ο οποίος λειτουργεί ως ενδιάμεσο επίπεδο αφαίρεσης. Όταν το σύστημα εντοπίσει για παράδειγμα τη μεταβλητή OLLAMA_MODEL, μεταβαίνει σε τοπική εκτέλεση μέσω του υποσυστήματος Ollama. Αντίθετα, όταν εντοπιστούν μεταβλητές όπως OPENAI_MODEL, GOOGLE_MODEL ή DEEPSEEK_MODEL, δημιουργείται ένας OpenAI-compatible client με την αντίστοιχη διεύθυνση API. Η στρατηγική αυτή διασφαλίζει ότι το σύστημα χρησιμοποιεί τον ίδιο ακριβώς μηχανισμό κλήσης για όλα τα μοντέλα, επιτρέποντας την αυστηρή συγκρισιμότητα των αποτελεσμάτων. Στην περίπτωση τοπικής εκτέλεσης μέσω του Ollama, η παραμετροποίηση επιτρέπει πλήρη τοπικό έλεγχο. Ο ερευνητής μπορεί να καθορίσει παραμέτρους όπως το NUM_PREDICT και το παράθυρο συμφραζομένων (αγγλ. context window), οι οποίες επηρεάζουν τη δυνατότητα παραγωγής αλυσίδων σκέψης και μεγάλων ίχνων συλλογισμού (αγγλ. Reasoning Traces). Η τοπική λειτουργία διαθέτει τρία βασικά πλεονεκτήματα: την απουσία κόστους ανά token, τη μέγιστη παραμετροποιησιμότητα της διαδικασίας εξαγωγής συμπερασμάτων (αγγλ. inference) και τη σταθερή λειτουργία ακόμη και σε περιβάλλοντα εκτός σύνδεσης (αγγλ. offline), γεγονός που καθιστά την ανάλυση επαναλήψιμη και ανεξάρτητη από εξωτερικούς περιορισμούς. Από την άλλη πλευρά, στα cloud-based μοντέλα, ο ενιαίος OpenAI-compatible client επιτρέπει την ομοιόμορφη κλήση όλων των απομακρυσμένων μοντέλων. Με αυτόν τον τρόπο, το GPT-4o, το Gemini 2.5 Flash και το DeepSeek R1 μπορούν να κληθούν με την ίδια ακριβώς δομή αιτήματος. Αυτή η ενιαία δομή συμβάλλει στη σταθερή συμπεριφορά του pipeline, αφού οι ίδιες ακολουθίες προτροπών χρησιμοποιούνται για όλα τα μοντέλα με συνεπή μορφή. Έτσι διασφαλίζεται ότι τα αποτελέσματα της συγκριτικής ανάλυσης αποδίδονται αποκλειστικά στις ικανότητες των μοντέλων και όχι σε ιδιομορφίες των APIs τους.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η εσωτερική αλληλεπίδραση των προηγμένων τεχνικών συλλογισμού, όπως η SC και η ToT, με τους παρόχους API. Στη μέθοδο SC, η οποία υλοποιείται στο evaluation.py, η παράμετρος SC_K καθορίζει τον αριθμό των ανεξάρτητων κλήσεων που θα πραγματοποιηθούν για την ίδια προτροπή. Κάθε κλήση δημιουργεί μια διαφορετική εκδοχή αλυσίδας σκέψης, βασισμένη τόσο στη στοχαστικότητα του μοντέλου όσο και στη διακύμανση των πιθανών διαδρομών συλλογισμού (αγγλ. Reasoning Paths). Όταν η τιμή της SC_K είναι υψηλή, το σύστημα παράγει πολλαπλές απαντήσεις διαδοχικά, αναγκάζοντας το API (ή το Ollama) να αποδώσει όσο το δυνατόν περισσότερες διακριτές λογικές διαδρομές. Αν και ο μηχανισμός αυτός αυξάνει το υπολογιστικό φορτίο, προσφέρει βαθύτερη κατανόηση της συμπεριφοράς του μοντέλου στα προβλήματα του GSM8K. Η τελική επιλογή της απάντησης γίνεται μέσω πλειοψηφικού κανόνα (αγγλ. majority voting), διαδικασία που επιτρέπει στον ερευνητή να διακρίνει αν το μοντέλο ακολουθεί σταθερό μοτίβο συλλογισμού ή αν ο συλλογισμός του είναι εύθραυστος.

Αντίστοιχα, η μέθοδος ToT εισάγει μια ακόμη πιο σύνθετη αλληλεπίδραση με τα APIs. Στη μέθοδο αυτή, το pipeline δημιουργεί μια πολυεπίπεδη δομή κλήσεων, όπου κάθε επίπεδο αντιστοιχεί σε μια εκδοχή του συλλογισμού που εκτείνεται προς διαφορετικές κατευθύνσεις. Κάθε προτροπή που δημιουργείται από την ToT δεν αποτελεί απλώς επανάληψη του αρχικού ερωτήματος, αλλά παράγεται ως επέκταση της σκέψης που προηγήθηκε. Η παράμετρος TOT_BRANCHES καθορίζει τον αριθμό των μονοπατιών που θα εξερευνηθούν σε κάθε στάδιο. Αυτό σημαίνει ότι το API καλείται επανειλημμένα, όχι απλώς για την παραγωγή διαφορετικών απαντήσεων όπως στην SC, αλλά για την παραγωγή διαδοχικών βημάτων συλλογισμού, καθένα από τα οποία αποτελεί την είσοδο για το επόμενο. Το αποτέλεσμα είναι ένας δυναμικός διάλογος ανάμεσα στο pipeline και το μοντέλο, όπου το σύστημα λειτουργεί ως συντονιστής πολλαπλών υπο-συλλογισμών που πρέπει να αξιολογούνται και να

φιλτράρονται πριν συνεχιστεί η διαδικασία. Η αλληλεπίδραση αυτή είναι κρίσιμη για την ποιότητα του συλλογισμού, καθώς μοντέλα όπως το GPT-4o ή το DeepSeek R1 έχουν την ικανότητα να διατηρούν τη συνεκτικότητα σε διαδοχικά βήματα, εν αντιθέσει με άλλα μοντέλα που παρουσιάζουν απόκλιση καθώς αυξάνεται το βάθος των κλήσεων.

Τέλος, η παράμετρος NUM_PREDICT μέσω της οποίας προσδιορίζεται ο μέγιστος αριθμός των tokens που δύναται να καταναλώσει το μοντέλο σε κάθε ερώτηση, διαδραματίζει καταλυτικό ρόλο τόσο στη SC όσο και στην ToT. Με περιορισμένο αριθμό tokens, το μοντέλο ενδέχεται να παράγει αποσπασματικές αλυσίδες σκέψης ή να μην ολοκληρώνει τα βήματα της ToT, οδηγώντας σε ακρωτηριασμένα μονοπάτια και μειωμένη ακρίβεια. Η αύξηση του ορίου αυτού παρέχει τον απαραίτητο χώρο για την ανάπτυξη της σκέψης του μοντέλου, αλλά αυξάνει παράλληλα το κόστος και τον χρόνο εκτέλεσης στα Cloud APIs, αναδεικνύοντας τη σημασία της τοπικής εκτέλεσης για πειράματα μεγάλου όγκου. Συνολικά, η παραμετροποίηση των μοντέλων στο παρόν σύστημα δεν περιορίζεται στην απλή επιλογή παρόχου, αλλά επεκτείνεται σε μια βαθιά, δυναμική και αλληλοεξαρτώμενη σχέση μεταξύ του pipeline, του μοντέλου και των τεχνικών συλλογισμού, επιτρέποντας μια ολοκληρωμένη συγκριτική αξιολόγηση, προσαρμοσμένη στις απαιτήσεις του συνόλου δεδομένων GSM8K και των σύγχρονων τεχνικών συλλογισμού.

4.4 Εργαλεία ανάπτυξης: LangChain, LangGraph, LangSmith

Η ανάπτυξη της πειραματικής πλατφόρμας αυτής στηρίζεται σε ένα σύνολο εργαλείων που έχουν σχεδιαστεί ειδικά για την οργάνωση και εκτέλεση εφαρμογών μεγάλης κλίμακας που βασίζονται σε LLMs. Τα εργαλεία LangChain, LangGraph και LangSmith δεν λειτουργούν απλώς ως μεμονωμένες βιβλιοθήκες, αλλά συγκροτούν ένα ολοκληρωμένο πλαίσιο (αγγλ. framework), το οποίο υποστηρίζει τον πλήρη κύκλο ζωής μιας εφαρμογής βασισμένης σε LLMs: από τη διαχείριση των μοντέλων και τον σχεδιασμό των προτροπών, έως την παρακολούθηση, την αξιολόγηση και τη βελτιστοποίηση των παραγόμενων αποτελεσμάτων. Στην παρούσα εργασία, οι δυνατότητες των εργαλείων αυτών αξιοποιούνται τόσο άμεσα όσο και έμμεσα, καθώς η αρχιτεκτονική του κώδικα ακολουθεί πιστά τις αρχές τους και προσαρμόζεται στη λειτουργική τους φιλοσοφία.

Το LangChain λειτουργεί ως το θεμελιώδες ενδιάμεσο επίπεδο για την επικοινωνία με τα γλωσσικά μοντέλα. Η λογική αυτή αντικατοπτρίζεται στο αρχείο pipeline.py, όπου ορίζεται η κλάση LLM με τρόπο που προσομοιάζει λειτουργικά τα wrappers του LangChain. Η συγκεκριμένη κλάση ενοποιεί σε μία κοινή διεπαφή όλα τα διαφορετικά μοντέλα που χρησιμοποιεί η πλατφόρμα. Μέσω μιας ενιαίας εντολής llm.complete(), ο κώδικας δύναται να καλέσει είτε ένα τοπικό μοντέλο μέσω Ollama είτε ένα απομακρυσμένο μοντέλο μέσω OpenAI-compatible API (όπως OpenAI, Gemini, DeepSeek). Αυτό επιτυγχάνεται επειδή η δομή της κλάσης περιλαμβάνει έλεγχο τύπου παρόχου και προσαρμόζει αυτόματα τη συμπεριφορά της. Η ενότητα αυτή της κωδικοποίησης αντικατοπτρίζει τη φιλοσοφία του LangChain: το υψηλότερο επίπεδο της εφαρμογής δεν ασχολείται με το "πού" βρίσκεται το μοντέλο, αλλά μόνο με το "τι" ζητά από αυτό. Ως εκ τούτου, ο κώδικας που υλοποιεί τις μεθόδους Standard Prompting, CoT, SC και ToT παραμένει αμετάβλητος ανεξαρτήτως του παρόχου. Η επιρροή του LangChain φαίνεται και στον τρόπο που υλοποιείται η μορφοποίηση των προτροπών. Στο pipeline οι λειτουργίες make_problem_prompt και make_tot_prompt, κατασκευάζουν προτροπές σε σταθερές φόρμες όπως ακριβώς συμβαίνει και στα LangChain PromptTemplates, επιτρέποντας στο σύστημα να εφαρμόζει διαφορετικές μεθόδους προτροπών χωρίς να χρειάζεται πολλαπλές υλοποιήσεις. Επιπλέον, οι run_standard, run_cot και run_tot καλούν το μοντέλο με τον ίδιο ακριβώς τρόπο, απλώς διαφοροποιώντας τη προτροπή και τις παραμέτρους του συλλογισμού, δυνατότητα που αποτελεί άμεσο πλεονέκτημα της λογικής LangChain.

Το LangGraph αποτελεί ένα εργαλείο σχεδιασμένο για τη μοντελοποίηση ροών συλλογισμού, όπου ο χρήστης δεν περιορίζεται σε μία γραμμική απάντηση, αλλά εξερευνά πολλαπλές εναλλακτικές διαδρομές σκέψης. Παρόλο που ο κώδικάς δεν χρησιμοποιεί απευθείας το LangGraph, η φιλοσοφία του εμφανίζεται έντονα στην υλοποίηση της ToT. Στην πράξη, η μέθοδος ToT στο pipeline.py δημιουργεί πολλούς ανεξάρτητους κόμβους σκέψης, δηλαδή διαφορετικές απαντήσεις στο ίδιο πρόβλημα, κάθε μία από τις οποίες αποτελεί διαφορετική πιθανή μαθηματική πορεία. Όπως ακριβώς γίνεται και στο LangGraph, αυτές οι διαδρομές δεν εξετάζονται μεμονωμένα, αλλά συγκρίνονται, αξιολογούνται και συνδυάζονται για να καθοριστεί η πιο λογική και συνεπής. Ο μηχανισμός αυτός λειτουργεί σαν ένα απλοποιημένο γράφημα σκέψης, όπου οι διαφορετικές απαντήσεις αποτελούν τα branches και η τελική επιλογή, μέσω majority vote, παίζει τον ρόλο του κεντρικού κόμβου απόφασης. Το γεγονός ότι ο κώδικας παράγει και αποθηκεύει όλες τις ενδιάμεσες απαντήσεις αντανακλά ξεκάθαρα τη λογική LangGraph, όπου η ανάλυση πολλαπλών μονοπατιών είναι απαραίτητη για βαθύτερο συλλογισμό.

Το LangSmith αποτελεί εργαλείο παρακολούθησης και αξιολόγησης των ροών εκτέλεσης, ενώ η δομή του κώδικα έχει υλοποιηθεί ώστε να εναρμονίζεται πλήρως με τη φιλοσοφία του. Η λεπτομερής καταγραφή που γίνεται σε όλα τα στάδια, από τα ακατέργαστα αποτελέσματα (αγγλ. raw outputs) της CoT έως τους κλάδους της ToT και τις επαναλήψεις της SC, δημιουργεί ένα πλούσιο σύνολο δεδομένων που μπορεί να αναλυθεί είτε τοπικά είτε μέσω εργαλείων παρακολούθησης όπως το LangSmith. Η υλοποίηση στο judge_evaluation.py, όπου ένα δεύτερο μοντέλο χρησιμοποιείται ως κριτής αξιολόγησης, αντικατοπτρίζει άμεσα όσα προτείνει το LangSmith για οργανωμένα evaluation pipelines. Η μέτρηση ποιοτικών παραμέτρων όπως η συντομία (αγγλ. conciseness), η σαφήνεια (αγγλ. clarity), η λογική (αγγλ. logic) και η ορθότητα (αγγλ. correctness) συνιστά μια δομημένη αξιολόγηση πολλαπλών διαστάσεων, όπως ακριβώς ορίζεται και στα LangSmith evaluation frameworks.

Η ενσωμάτωση των παραπάνω εργαλείων είναι εμφανής και στην οργανωτική δομή της εφαρμογής. Το αρχείο main_parallel.py λειτουργεί ως κεντρικός εκτελεστής που διαχειρίζεται πολλαπλά μοντέλα και παραμετροποιήσεις, αξιοποιώντας την αφαίρεση (αγγλ. abstraction) που παρέχει το LangChain για να διατηρεί ενιαία εκτέλεση. Το evaluation.py υπολογίζει μετρικές που μπορούν να εισαχθούν απευθείας σε dashboards LangSmith. Τα radar_models_methods.py και plots_and_tables.py αναλύουν τις καταγεγραμμένες μετρικές και τις μετατρέπουν σε συγκριτικά γραφήματα κατάλληλα για ερευνητική παρουσίαση, ενώ ακόμη και το load_data.py έχει σχεδιαστεί με τρόπο που θυμίζει LangChain loaders, παρέχοντας τυποποιημένο χειρισμό του GSM8K. Συνοψίζοντας, η υιοθέτηση των αρχών των LangChain, LangGraph και LangSmith μετατρέπει το σύστημα από ένα απλό script εκτέλεσης σε ένα ολοκληρωμένο ερευνητικό περιβάλλον, ικανό να συγκρίνει με ακρίβεια, διαφάνεια και συνέπεια διαφορετικά μοντέλα και τεχνικές συλλογισμού.

4.5 Υλοποίηση pipelines με LangChain

Η υλοποίηση των pipelines πραγματοποιήθηκε στο αρχείο pipeline.py, όπου συγκεντρώνονται οι βασικές λειτουργίες του συστήματος. Ο σχεδιασμός ακολουθεί τη φιλοσοφία της αρχιτεκτονικής LangChain, η οποία στηρίζεται σε επιμέρους δομικά στοιχεία όπως οι προτροπές, οι περιέκτες LLM (αγγλ. LLM wrappers) και οι αλυσίδες (αγγλ. chains). Η δομή αυτή υλοποιείται μέσω ενός ενιαίου pipeline που περιλαμβάνει τη δημιουργία των προτροπών, την κλήση του LLM, την παραγωγή της απάντησης και την τελική εξαγωγή του αριθμητικού αποτελέσματος. Η υλοποίηση είναι πλήρως εναρμονισμένη με την αρχή «LLM-agnostic pipelines», επιτρέποντας την εφαρμογή των ίδιων μεθόδων προτροπής σε πλήθος διαφορετικών μοντέλων (OpenAI, Gemini, DeepSeek, Ollama) χωρίς αλλαγές στον κώδικα. Στο αρχείο pipeline.py βρίσκεται η κλάση LLM, η οποία υλοποιεί μια ενιαία διεπαφή επικοινωνίας. Κατά την αρχικοποίηση, η κλάση ανιχνεύει αυτόματα την ύπαρξη κλειδιού API (OpenAI

API key) και επιλέγει τον αντίστοιχο Openai-Compatible Client (ο οποίος καλύπτει OpenAI, Gemini, DeepSeek). Αν δεν εντοπιστεί API Key, το σύστημα μεταβαίνει σε τοπική εκτέλεση μέσω Ollama. Χάρη σε αυτή τη σχεδίαση, ο υπόλοιπος του κώδικας είτε αφορά Standard Prompting, είτε CoT, SC-K ή ToT, καλεί απλώς τη μέθοδο complete(), αγνοώντας τις λεπτομέρειες της υποκείμενης υποδομής (αγγλ. backend). Η μέθοδος αυτή επιστρέφει το κείμενο της απάντησης καθώς και μετρικές χρήσης (αγγλ. tokens), απαραίτητες για την ανάλυση κόστους και απόδοσης.

Επί της κοινής αυτής διεπαφής υλοποιούνται οι επιμέρους στρατηγικές προτροπής. Η συνάρτηση run_standard() αντιπροσωπεύει την απλούστερη μορφή pipeline, υλοποιώντας την άμεση προτροπή (αγγλ. direct prompting), στην οποία το μοντέλο δεν ενθαρρύνεται να προβεί σε συλλογισμό. Αντίθετα, λαμβάνει μια σύντομη εντολή μέσω της _prompt_standard() και καλείται να απαντήσει αποκλειστικά με το τελικό αποτέλεσμα στη μορφή:

FINAL: <number>.

Η απουσία ενδιάμεσων βημάτων οδηγεί σε χαμηλότερο υπολογιστικό κόστος και μειωμένη κατανάλωση tokens, καθιστώντας τη μέθοδο ιδανική ως σημείο αναφοράς (αγγλ, Baseline) για τη σύγκριση με πιο σύνθετες τεχνικές.

Η δεύτερη μέθοδος, run_cot(), υλοποιεί την CoT, ενθαρρύνοντας το μοντέλο να αναλύσει βήμα προς βήμα τη λύση μέσω της εντολής «Let’s solve step by step», που παράγεται από την _prompt_cot(). Η ιδιαιτερότητα της συγκεκριμένης υλοποίησης έγκειται στην ενσωμάτωση της αυτο-συνέπειας (SC-K). Η συνάρτηση run_cot() καλεί το μοντέλο k φορές, παράγοντας k διαφορετικές αλυσίδες σκέψης, και επιλέγει την τελική απάντηση βάσει πλειοψηφίας (αγγλ. majority vote). Η τεχνική αυτή βασίζεται στην τεκμηριωμένη στη βιβλιογραφία παραδοχή ότι η σύνθεση πολλαπλών λύσεων οδηγεί σε ακριβέστερα συμπεράσματα. Ο κώδικας καταγράφει σχολαστικά τις πρωτογενείς εξόδους (raw outputs) και εξάγει τις τιμές FINAL: για τον υπολογισμό της επικρατούσας λύσης με τη μέθοδο της πλειοψηφίας.

Η τρίτη και πλέον σύνθετη μέθοδος είναι η run_tot(), η οποία υλοποιεί τη ToT. Σε αντίθεση με τις ανεξάρτητες αλυσίδες της SC, η ToT δημιουργεί διακλαδώσεις σκέψης (αγγλ. branches). Το pipeline παράγει μια προτροπή μέσω της _prompt_tot_branch(), ζητώντας από το μοντέλο μια πιθανή προσέγγιση λύσης με την ένδειξη TENTATIVE: <number>. Η διαδικασία επαναλαμβάνεται b φορές, δημιουργώντας διαφορετικές πορείες, από τις οποίες εξάγονται οι ενδιάμεσες προβλέψεις. Η τελική απόφαση λαμβάνεται και πάλι πλειοψηφικά, με το αποτέλεσμα να αποδίδεται στο τελικό block FINAL. Για λόγους διαφάνειας, η μέθοδος συνθέτει ένα ενιαίο κείμενο που περιλαμβάνει όλα τα branches (διαχωρισμένα με ===== BRANCH =====) και την τελική επιλογή (===== SELECTOR =====), προσομοιώνοντας ένα δέντρο σκέψης εναρμονισμένο με τα πρότυπα του LangGraph, επιτρέποντας τη συγκριτική αξιολόγηση της ικανότητας κάθε μοντέλου να παράγει συνεκτικό συλλογισμό σε πολλαπλές εκδοχές.

Κοινός παρονομαστής όλων των pipelines είναι η χρήση τυποποιημένων δεικτών (FINAL:, TENTATIVE:), οι οποίοι διασφαλίζουν την αξιόπιστη εξαγωγή του αριθμητικού αποτελέσματος από το αρχείο evaluation.py, ακόμη και όταν το μοντέλο παράγει εκτενές κείμενο. Η συνάρτηση evaluate() εφαρμόζει την πλήρη ροή αξιολόγησης: καλεί την κατάλληλη μέθοδο εκτέλεσης, συγκρίνει το αποτέλεσμα με το πρότυπο (αγγλ. gold label) του GSM8K, και καταγράφει μετρικές ακρίβειας (αγγλ. Accuracy), καθυστέρησης (αγγλ. latency) και αποχής (αγγλ. abstention). Η μαζική εκτέλεση των πειραμάτων πραγματοποιείται μέσω του main_parallel.py, το οποίο αξιοποιεί την παράλληλη επεξεργασία (ProcessPoolExecutor) για την ταυτόχρονη εκτέλεση των μεθόδων Standard, CoT και ToT. Τα αποτελέσματα αποθηκεύονται σε αρχεία CSV για κάθε μέθοδο, συνοδευόμενα από πλήρη

μεταδεδομένα, ακατέργαστες απαντήσεις (αγγλ. raw outputs), εξαγόμενες τελικές απαντήσεις (αγγλ. extracted finals), πλήθος tokens, χρόνους εκτέλεσης, flags διαχείρισης σφαλμάτων και δείκτες επιτυχίας. Τα δεδομένα αυτά τροφοδοτούν στη συνέχεια τη δημιουργία συγκριτικών γραφημάτων μέσω των modules plots_and_tables.py και llm_metrics_charts.py. Συνοψίζοντας, η αρχιτεκτονική αυτή ακολουθεί με συνέπεια τη φιλοσοφία του LangChain, διασφαλίζει την επεκτασιμότητα, την αναπαραγωγιμότητα και την επιστημονική εγκυρότητα της σύγκρισης, τηρώντας τις αρχές της αρθρωτής (αγγλ. modular) ανάπτυξης και της σαφούς διάκρισης των επιπέδων ευθύνης.

4.6 Υλοποίηση ροών με LangGraph (CoT / ToT γραφήματα κόμβων - ακμών)

Η φιλοσοφία του LangGraph εδράζεται στην παραδοχή, ότι οι συλλογιστικές διαδικασίες ενός LLM δύνανται να μοντελοποιηθούν ως γράφος (αγγλ. graph) αποτελούμενος από κόμβους (αγγλ. nodes) και ακμές (αγγλ. edges). Στο πλαίσιο αυτό, κάθε κόμβος αντιστοιχεί σε ένα διακριτό βήμα σκέψης ή μια υποψήφια λύση, ενώ οι ακμές εκφράζουν τις μεταξύ τους συσχετίσεις, δηλαδή τη μετάβαση από ένα βήμα συλλογισμού (αγγλ. reasoning step) στο επόμενο. Στο παρόν σύστημα, μολονότι δεν γίνεται άμεση χρήση του πηγαίου κώδικα της βιβλιοθήκης LangGraph, η αρχιτεκτονική και η λογική της έχουν ενσωματωθεί δομικά στις μεθόδους CoT και ToT, όπως αυτές υλοποιούνται στο αρχείο pipeline.py. Η προσέγγιση αυτή αναβαθμίζει τις μεθόδους CoT και ToT, μετατρέποντάς τις από απλές ακολουθίες προτροπών σε ολοκληρωμένες ροές συλλογισμού, διατυπωμένες με όρους θεωρίας γραφημάτων.

Στην περίπτωση της CoT, κάθε εκτέλεση του μοντέλου αντιμετωπίζεται ως ένας γραμμικός κόμβος σκέψης, όπου το μοντέλο παράγει μια ενιαία ακολουθία συλλογισμού που καταλήγει σε ένα τελικό αποτέλεσμα. Στο pipeline.py, η συνάρτηση run_cot() δρομολογεί πολλαπλές διαδοχικές εκτελέσεις της ίδιας προτροπής, στο πλαίσιο της τεχνικής SC. Κάθε επιμέρους εκτέλεση λειτουργεί ως ανεξάρτητος κόμβος με το δικό του μοναδικό μονοπάτι συλλογισμού, οδηγώντας σε μια αυτόνομη τελική τιμή. Το σύνολο των εκτελέσεων αυτών συγκεντρώνεται σε μία δομή λίστας, η οποία αναπαριστά εννοιολογικά έναν γράφο με πολλαπλούς παράλληλους κόμβους χωρίς ενδιάμεσες διακλαδώσεις. Οι κόμβοι αυτοί συνδέονται μεταξύ τους όχι με σχέση αιτιότητας, αλλά μέσω της συμμετοχής τους στη συλλογική λήψη απόφασης. Η τελική επιλογή προκύπτει μέσω της διαδικασίας πλειοψηφίας (αγγλ. majority vote), η οποία λειτουργεί ως ένας υπερ-κόμβος (αγγλ. super-node) επιλογής. Ο κόμβος αυτός δέχεται ως είσοδο όλες τις τιμές από τις επιμέρους διαδρομές και εξάγει τη συνολική πρόβλεψη, δομή που αντικατοπτρίζει πιστά τη λογική ενός γράφου με παράλληλες ακμές που συγκλίνουν σε έναν κόμβο απόφασης, όπως περιγράφεται στον LangGraph για πολυσταδιακή αξιολόγηση απαντήσεων.

Αντιθέτως, η μέθοδος ToT υλοποιεί μια πολυπλοκότερη τοπολογία γράφου, οργανωμένη σε διακριτούς κόμβους-κλάδους (αγγλ. branches). Στο pipeline.py, η συνάρτηση run_tot() καλεί το μοντέλο b φορές, δημιουργώντας σε κάθε επανάληψη έναν νέο κλάδο που αντιστοιχεί σε ένα πιθανό μονοπάτι σκέψης. Σε αντίθεση με την CoT, εδώ η διαφοροποίηση δεν έγκειται μόνο στην πολλαπλότητα των κόμβων, αλλά και στον λειτουργικό τους ρόλο, όπου κάθε κόμβος εκφράζει μια διαφορετική στρατηγική επίλυσης και όχι απλώς μια παραλλαγή της διατύπωσης. Κατ' αυτόν τον τρόπο, η ToT δημιουργεί ένα σύνολο εναλλακτικών διαδρομών που επιτρέπουν στο μοντέλο να εξερευνήσει τον χώρο λύσεων με μεγαλύτερο πλούτο και δημιουργικότητα. Οι ακμές που συνδέουν τους κλάδους αυτούς με τον τελικό κόμβο απόφασης υλοποιούνται μέσω της εξαγωγής των τιμών TENTATIVE και της επακόλουθης πλειοψηφικής επιλογής. Η συνάρτηση οργανώνει όλα τα branches σε μια ενιαία συμβολοσειρά, όπου κάθε κλάδος εμφανίζεται ως διακριτό μπλοκ (διαχωρισμένο με ===== BRANCH =====), ενώ στο τέλος κατασκευάζεται το μπλοκ ===== SELECTOR ===== που λειτουργεί ως ο τελικός κόμβος επιλογής. Αυτή η δομή αντικατοπτρίζει πλήρως τη LangGraph λογική, όπου πολλοί κόμβοι συγκλίνουν σε έναν κεντρικό κόμβο που υπολογίζει το τελικό αποτέλεσμα. Το γεγονός ότι κάθε branch έχει τη δική

του ανεξάρτητη ToT προτροπή ενισχύει ακόμη περισσότερο την αναλογία με τα υπογράμματα σκέψης του LangGraph, όπου κάθε διαδρομή έχει τη δική της αυτονομία, πριν την τελική σύγκριση.

Η λειτουργία των παραπάνω ροών ενσωματώνει πλήρως τις έννοιες του κόμβου συλλογισμού, της ακμής μετάβασης και του κόμβου επιλογής, όπως αυτές ορίζονται στην τεκμηρίωση του LangGraph. Το CoT pipeline υλοποιεί έναν γραμμικό, παράλληλο γράφο ανεξάρτητων μονοπατιών, ενώ το ToT pipeline υλοποιεί ένα δενδροειδές γράφημα στρατηγικής εξερεύνησης. Η γραφηματική αυτή θεώρηση δεν λειτουργεί μόνο ως θεωρητικό μοντέλο, αλλά και ως πρακτικός μηχανισμός αξιολόγησης, καθώς επιτρέπει στο σύστημα να εξετάζει την ποικιλομορφία και τη σταθερότητα των μονοπατιών σκέψης. Μέσω της υλοποίησης αυτής στο pipeline.py, η σκέψη του μοντέλου δεν αντιμετωπίζεται ως ένα ενιαίο βήμα αλλά ως γράφημα πολλών πιθανών διαδρομών, προσφέροντας βαθύτερη κατανόηση της συμπεριφοράς των μοντέλων και αναδεικνύοντας τις ποιοτικές διαφορές μεταξύ των μεθόδων προτροπής.

4.7 Ενσωμάτωση αξιολόγησης μέσω LangSmith

Η αρχιτεκτονική της αξιολόγησης στο παρόν σύστημα έχει σχεδιαστεί ώστε να αντικατοπτρίζει τις θεμελιώδεις αρχές με τις οποίες το LangSmith οργανώνει και παρακολουθεί την ποιότητα των εκτελέσεων στα LLMs. Μολονότι η εργασία δεν κάνει άμεση χρήση του περιβάλλοντος νέφους της πλατφόρμας LangSmith, ο πηγαίος κώδικας ενσωματώνει πλήρως τη φιλοσοφία της: την καταγραφή πλήρους ίχνους (αγγλ. trace) για κάθε εκτέλεση, τον αυστηρό διαχωρισμό της παραγωγής απαντήσεων από την αξιολόγησή τους και την τυποποιημένη παραγωγή μεταδεδομένων για συγκριτική ανάλυση. Η πρώτη μορφή αξιολόγησης υλοποιείται στο αρχείο evaluation.py, όπου για κάθε μέθοδο προτροπής (Standard Prompting, CoT, ToT) καταγράφονται συστηματικά πληροφορίες που αντιστοιχούν στα ίχνη του LangSmith. Κάθε εκτέλεση επιστρέφει το πρωτογενές κείμενο συλλογισμού (αγγλ. raw reasoning), το τελικό αριθμητικό αποτέλεσμα και αναλυτικές πληροφορίες για την κατανάλωση tokens. Τα δεδομένα αυτά συνδυάζονται με τον χρόνο εκτέλεσης και την πρότυπη απάντηση (αγγλ. gold label) του συνόλου δεδομένων GSM8K, δημιουργώντας ένα ολοκληρωμένο σύνολο μεταδεδομένων. Με τον τρόπο αυτό, κάθε πρόβλεψη μετατρέπεται σε ένα πλήρες αρχείο αξιολόγησης (αγγλ. evaluation record), ισοδύναμο με αυτό που θα κατέγραφε το LangSmith σε ένα πραγματικό παραγωγικό pipeline.

Η πλέον κρίσιμη πτυχή της ενσωμάτωσης αφορά τη διαδικασία «LLM-as-a-Judge», η οποία υλοποιείται στο αρχείο judge_evaluation.py. Στο τμήμα αυτό, ένα δεύτερο, ανεξάρτητο LLM λειτουργεί ως αυτόνομος κριτής, αξιολογώντας την ποιοτική επάρκεια της απάντησης του pipeline. Ο κριτής λαμβάνει ως είσοδο τόσο τον πλήρη συλλογισμό όσο και το τελικό αποτέλεσμα (FINAL) και καλείται να βαθμολογήσει την απόκριση βάσει τεσσάρων κριτηρίων: Ορθότητα (αγγλ. Correctness), Λογική (αγγλ. Logic), Σαφήνεια (αγγλ. Clarity) και Συντομία (αγγλ. Conciseness). Η προσέγγιση αυτή αντιστοιχεί άμεσα στον μηχανισμό «Evaluation with LLM graders», που αποτελεί κεντρικό πυλώνα του LangSmith. Ο κώδικας, μάλιστα, δεν επιστρέφει απλώς αριθμητικές βαθμολογίες, αλλά και το σκεπτικό του κριτή, προσφέροντας πλήρη αναπαραγωγιμότητα και δυνατότητα ελέγχου των αποφάσεων του μοντέλου-κριτή. Τον συντονισμό της διαδικασίας αναλαμβάνει το judge_pipeline.py, το οποίο εφαρμόζει την ίδια ροή ελέγχου σε όλα τα αποτελέσματα που παράγονται από την κύρια προσομοίωση. Έτσι, επιτυγχάνεται ο καθαρός διαχωρισμός της εκτέλεσης των pipelines (Standard Prompting, CoT, ToT) από την ποιοτική αξιολόγηση, όπως ακριβώς προβλέπεται στις ροές αξιολόγησης του LangSmith, όπου η παραγωγή και ο έλεγχος αποτελούν δύο ανεξάρτητα αλλά αλληλένδετα στάδια.

Σε επίπεδο διαχείρισης δεδομένων, το αρχείο main_parallel.py διασφαλίζει την ακεραιότητα των αποτελεσμάτων αποθηκεύοντας κάθε εκτέλεση σε δομημένα αρχεία CSV, συνοδευόμενα από

αναλυτικά μεταδεδομένα: reasoning, final answers, tokens, δείκτες ακρίβειας, χρόνο απόκρισης (αγγλ. latency) και πιθανές ενδείξεις αποχής (αγγλ. abstention). Το σύστημα υιοθετεί τη λογική συστηματικής καταγραφής του LangSmith, ώστε τα αποτελέσματα να μπορούν να αναπαρασταθούν μετέπειτα σε γραφήματα και αναλυτικούς πίνακες, προσφέροντας πλήρη εικόνα της συμπεριφοράς των μοντέλων. Συνοψίζοντας, ο κώδικας ενσωματώνει τη φιλοσοφία του LangSmith σε τρία διακριτά επίπεδα:

1) καταγραφή λεπτομερούς ίχνους συλλογισμού από κάθε pipeline,
2) εφαρμογή αξιολόγησης μέσω δεύτερου LLM με τυποποιημένα κριτήρια ποιότητας και
3) παραγωγή πλούσιων μεταδεδομένων κατάλληλων για συγκριτική μελέτη και οπτικοποίηση.

Με την οργάνωση αυτή, το σύστημα δεν περιορίζεται στην απλή παραγωγή αριθμητικών αποτελεσμάτων, αλλά λειτουργεί ως ένα πλήρες ερευνητικό πλαίσιο, ικανό να προσφέρει βαθιά κατανόηση της ποιότητας, της συνέπειας και της συλλογιστικής συμπεριφοράς των μοντέλων.

4.8 Πειραματικές ρυθμίσεις και διαχείριση παραμέτρων εκτέλεσης

Η παραμετροποίηση των μοντέλων αποτελεί κρίσιμο στοιχείο της παρούσας ερευνητικής πλατφόρμας, καθώς καθορίζει τόσο τη συμπεριφορά των LLMs όσο και την αξιοπιστία των συγκριτικών αποτελεσμάτων. Στο αναπτυχθέν σύστημα, η διαχείριση των παραμέτρων είναι πλήρως αυτοματοποιημένη και βασίζεται στη χρήση μεταβλητών περιβάλλοντος, επιτρέποντας την εκτέλεση πολλαπλών μοντέλων και μεθόδων προτροπής χωρίς να απαιτούνται μεταβολές στον πυρήνα του κώδικα. Η προσέγγιση αυτή υλοποιεί την αρχή της «ανεξάρτητης από το μοντέλο αξιολόγησης» (αγγλ. model-agnostic evaluation), η οποία αποτελεί κεντρική σχεδιαστική κατεύθυνση τόσο του πλαισίου LangChain όσο και των σύγχρονων συστημάτων ανάπτυξης LLMs. Στο αρχείο pipeline.py, υλοποιείται μια ενιαία κλάση LLM που λειτουργεί ως καθολική διεπαφή, συνδέοντας το σύστημα με τέσσερις διαφορετικούς τύπους παρόχων: OpenAI, Google Gemini, DeepSeek, καθώς και τοπικά μοντέλα μέσω της πλατφόρμας Ollama. Η επιλογή του παρόχου πραγματοποιείται δυναμικά κατά την αρχικοποίηση της κλάσης, βάσει της διαθεσιμότητας κλειδιών API ή ρυθμίσεων τοπικής εκτέλεσης. Μέσω αυτής της αρχιτεκτονικής, η μέθοδος complete() εφαρμόζει ενιαία λογική κλήσης ανεξαρτήτως της υποκείμενης υποδομής (αγγλ. backend), διασφαλίζοντας τη σταθερότητα στην αλληλεπίδραση των ροών εργασίας με διαφορετικές αρχιτεκτονικές μοντέλων (Transformers, mixture-of-experts, reasoning-enhanced LLMs).

Οι κρίσιμες παράμετροι λειτουργίας των μοντέλων καθορίζονται μέσω μεταβλητών περιβάλλοντος που φορτώνονται στο αρχείο main_parallel.py. Η μεταβλητή MODEL_NAME/OLLAMA_MODEL/ /OPENAI_MODEL επιλέγει το συγκεκριμένο LLM που θα χρησιμοποιηθεί (π.χ. gpt-4o-mini, gemini-2.5-pro, deepseek-r1, llama3:8b), επιτρέποντας την εκτέλεση της ίδιας πειραματικής διαδικασίας με διαφορετικά μοντέλα για λόγους συγκρισιμότητας. Η παράμετρος NUM_PREDICT ορίζει το ανώτατο όριο παραγόμενων tokens κατά την ολοκλήρωση μιας προτροπής, στοιχείο που καθορίζει τη σταθερότητα του συλλογισμού, ιδιαίτερα στις μεθόδους CoT και ToT. Επιπλέον, χρησιμοποιούνται διακριτές ρυθμίσεις θερμοκρασίας (TEMPERATURE_STD, TEMPERATURE_COT, TEMPERATURE_TOT) για κάθε τύπο pipeline: χαμηλές τιμές στο Standard Prompting για determinism και υψηλότερες στο ToT για την ενίσχυση της διαφοροποίησης των κλάδων. Ιδιαίτερη σημασία έχει η παράμετρος SC_K (αγγλ. Self-Consistency multiplier), η οποία καθορίζει τον αριθμό των ανεξάρτητων ακολουθιών CoT που παράγονται για κάθε ερώτημα. Όταν η τιμή οριστεί στο 1, το σύστημα εκτελεί απλή CoT χωρίς μηχανισμό SC, ενώ υψηλότερες τιμές οδηγούν σε αυξημένη αξιοπιστία μέσω της μεθόδου πλειοψηφίας, με το ανάλογο κόστος σε tokens. Αντίστοιχα, η παράμετρος TOT_BRANCHES ορίζει τον αριθμό των κλάδων σκέψης που θα δημιουργήσει η μέθοδος ToT, επηρεάζοντας την ικανότητα εξερεύνησης του χώρου λύσεων.

Το σύνολο των παραμέτρων αυτών μεταβιβάζεται απευθείας στον περιέκτη (αγγλ. wrapper) του LLM και αξιοποιείται από τις μεθόδους run_standard(), run_cot() και run_tot() για τον καθορισμό της συμπεριφοράς του μοντέλου. Η παραμετροποίηση αυτή επηρεάζει άμεσα όχι μόνο την ποιότητα των απαντήσεων αλλά και την κατανάλωση πόρων, δεδομένο που καταγράφεται λεπτομερώς στο evaluation.py. Το αρχείο αυτό συλλέγει τα δεδομένα και τα ενσωματώνει ως μεταδεδομένα σε κάθε εγγραφή των αποτελεσμάτων (CSV), επιτρέποντας την ανάλυση της σχέσης κόστους - απόδοσης (αγγλ. trade-off) μέσω μετρικών όπως η ακρίβεια (αγγλ. Accuracy), ο χρόνος απόκρισης (αγγλ. latency) και ο συνολικός αριθμός tokens. Η διαδικασία ολοκληρώνεται στο main_parallel.py, όπου το σύστημα οργανώνει την παράλληλη εκτέλεση όλων των pipelines, δημιουργώντας ανεξάρτητες διεργασίες για κάθε συνδυασμό μοντέλου και μεθόδου. Συνολικά, ο μηχανισμός παραμετροποίησης εξασφαλίζει την πλήρη ανεξαρτησία των pipelines από τον πάροχο, παρέχει ευελιξία για γρήγορη προσαρμογή των πειραματικών συνθηκών, επιτρέπει την ομοιόμορφη διαχείριση παραμέτρων για κάθε μέθοδο προτροπής (Standard Prompting, CoT, ToT) και τη συγκριτική αξιολόγηση μεγάλης κλίμακας, ενώ παράλληλα προσφέρει τη δυνατότητα αναλυτικής καταγραφής της συμπεριφοράς κάθε μοντέλου.

4.9 Διαδικασία εκτέλεσης πειραμάτων (Standard, CoT, ToT)

Η διαδικασία εκτέλεσης των πειραμάτων έχει σχεδιαστεί και οργανωθεί με γνώμονα την εξασφάλιση πλήρους αυτοματοποίησης, αναπαραγωγιμότητας και του σαφούς λειτουργικού διαχωρισμού μεταξύ των ροών Standard Prompting, CoT και ToT. Η έναρξη της εκτέλεσης πραγματοποιείται από το αρχείο main_parallel.py, το οποίο λειτουργεί ως ο κεντρικός συντονιστής του συστήματος. Σε πρώτο στάδιο, φορτώνεται το σύνολο των προβλημάτων του GSM8K μέσω του load_data.py και ακολούθως δημιουργούνται τρεις ανεξάρτητες εργασίες, καθεμία εκ των οποίων αντιστοιχίζεται σε μία διαφορετική μέθοδο προτροπής. Για την επίτευξη βέλτιστης ταχύτητας και την απομόνωση των διεργασιών, οι εργασίες εκτελούνται παράλληλα, επιτρέποντας στο σύστημα να αξιοποιεί ταυτόχρονα πολλαπλούς πυρήνες επεξεργασίας. Κάθε επιμέρους εργασία δημιουργεί ένα νέο στιγμιότυπο (αγγλ. instance) του περιέκτη LLM και καλεί τη συνάρτηση evaluate() με τον κατάλληλο οδηγό εκτέλεσης (αγγλ. pipeline runner): run_standard(), run_cot() ή run_tot(). Η επιλογή της ανεξάρτητης κλήσης για κάθε μέθοδο διασφαλίζει την ακεραιότητα των αποτελεσμάτων, αποτρέποντας την επίδραση από προηγούμενες εκτελέσεις ή διαφορετικές παραμετροποιήσεις.

Εντός της συνάρτησης evaluate(), κάθε πρόβλημα υπόκειται σε σειριακή επεξεργασία: αρχικά δημιουργείται η κατάλληλη προτροπή, στη συνέχεια το LLM παράγει την απάντησή του και τέλος εξάγεται το τελικό αριθμητικό αποτέλεσμα προκειμένου να συγκριθεί με την ορθή λύση (gold label). Παράλληλα με τη διαδικασία αυτή, καταγράφονται συστηματικά ο χρόνος εκτέλεσης, η κατανάλωση tokens και οι ενδείξεις αποχής (αγγλ. abstention flags), ώστε να καταστεί δυνατή η αντικειμενική σύγκριση των τριών μεθόδων. Αξίζει να σημειωθεί η δομική διαφοροποίηση των ροών: η μέθοδος Standard παράγει μία μοναδική απόκριση ανά πρόβλημα, η ροή CoT δύναται να δημιουργεί πολλαπλές αποκρίσεις βάσει του αριθμού επαναλήψεων της SC, ενώ η ροή ToT δημιουργεί πολλαπλούς ανεξάρτητους κλάδους (αγγλ. branches). Παρά τις διαφορές αυτές στη δομή και το πλήθος των απαιτούμενων εκτελέσεων, και οι τρεις μέθοδοι ενσωματώνονται στην ίδια ενιαία διαδικασία αξιολόγησης, επιτρέποντας την άμεση και ισότιμη σύγκριση των αποτελεσμάτων.

Η πειραματική διαδικασία ολοκληρώνεται με την αποθήκευση των δεδομένων κάθε μεθόδου σε διακριτά αρχεία τύπου CSV, καθώς και σε αρχεία μεταδεδομένων συνολικής απόδοσης. Τα αρχεία αυτά διατηρούν τόσο το τελικό αποτέλεσμα όσο και τα ενδιάμεσα στοιχεία συλλογισμού, τα οποία είναι απαραίτητα για τη μεταγενέστερη οπτικοποίηση ή την ποιοτική αξιολόγηση μέσω του judge pipeline. Με τον τρόπο αυτό, η διαδικασία εκτέλεσης μετατρέπεται σε μια πλήρως αυτοματοποιημένη ροή

εργασίας, η οποία επιτρέπει τη σύγκριση των διαφορετικών μεθόδων συλλογισμού με συνέπεια, διαφάνεια και υψηλό επίπεδο ελέγχου.

4.10 Έλεγχος της συνέπειας ανά μέθοδο και μοντέλο (Consistency Check)

Η αξιολόγηση των LLMs δεν περιορίζεται αποκλειστικά στην εξέταση της ορθότητας των απαντήσεων, αλλά επεκτείνεται και στη μελέτη της σταθερότητας του παραγόμενου συλλογισμού. Προς την κατεύθυνση αυτή, στην παρούσα εργασία αναπτύχθηκε και ενσωματώθηκε μια εξειδικευμένη διαδικασία ελέγχου της συνέπειας (αγγλ. consistency check), η οποία υλοποιείται μέσω του αρχείου consistency_per_method.py. Κύριος στόχος της διαδικασίας είναι η απομόνωση της παραμέτρου της «βεβαιότητας» του μοντέλου από την ακρίβειά του, επιτρέποντας τη διερεύνηση του κατά πόσον οι διαφορετικές μέθοδοι προτροπής (Standard Prompting, CoT, ToT) επιτυγχάνουν τη μείωση της εγγενούς στοχαστικότητας των μοντέλων. Για τις ανάγκες του πειράματος, επιλέγεται ένα αντιπροσωπευτικό υποσύνολο 100 ερωτήσεων από το σύνολο δεδομένων αξιολόγησης του GSM8K. Ο αλγόριθμος υποβάλλει κάθε ερώτηση στο μοντέλο σε πολλαπλές διαδοχικές επαναλήψεις (αγγλ. runs), διατηρώντας απολύτως σταθερές τις παραμέτρους εκτέλεσης, προκειμένου να καταγραφεί η διακύμανση στις αποκρίσεις.

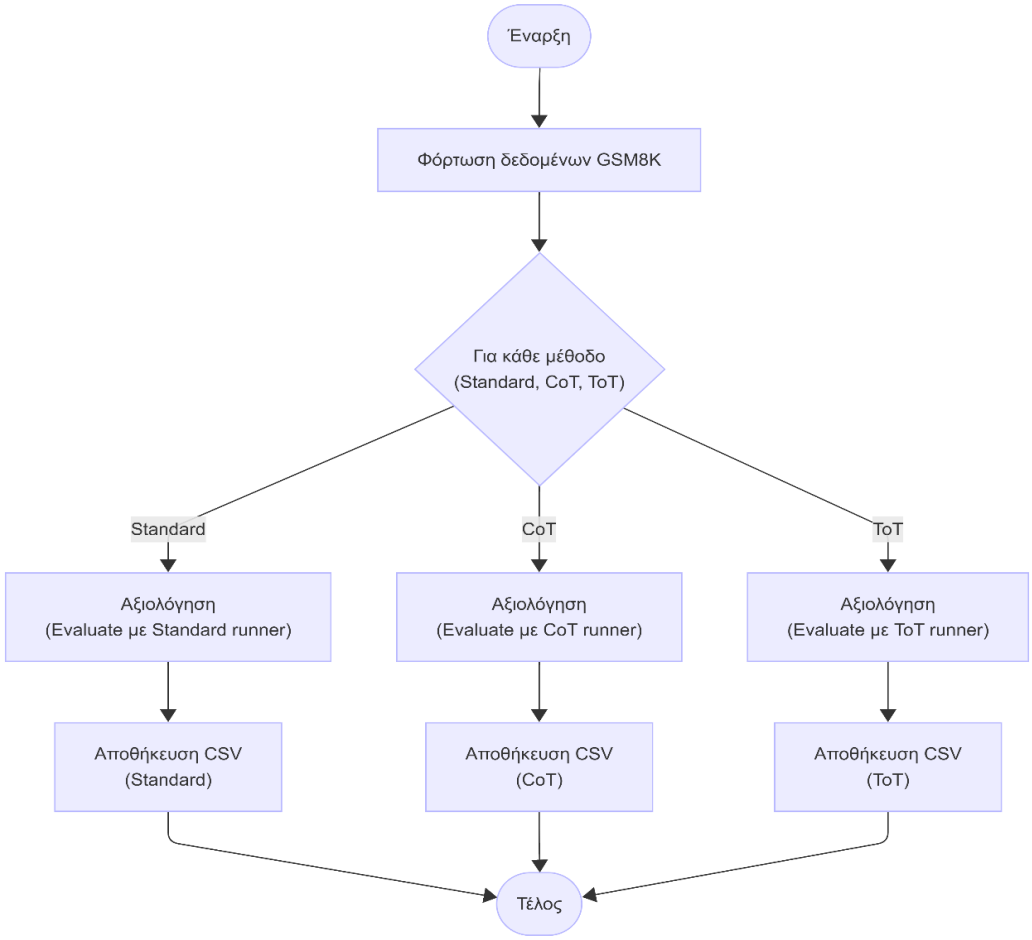
Η καταγραφή των αποτελεσμάτων πραγματοποιείται σε αναλυτικά αρχεία τύπου CSV, η δομή των οποίων έχει σχεδιαστεί ώστε να αποτυπώνει λεπτομερώς τη συμπεριφορά του μοντέλου σε κάθε επανάληψη. Για κάθε μοναδικό αναγνωριστικό ερώτησης, δημιουργούνται διακριτές εγγραφές που περιλαμβάνουν τον δείκτη της επανάληψης (run_index), την παραχθείσα πρόβλεψη (pred) και την ένδειξη ορθότητας (is_correct). Ένα κρίσιμο τεχνικό χαρακτηριστικό του αλγορίθμου είναι η εφαρμογή ελέγχου σημασιολογικής αριθμητικής συνέπειας (αγγλ. semantic numeric consistency). Αντί να περιορίζεται σε απλή σύγκριση συμβολοσειρών, ο κώδικας μετατρέπει τις αποκρίσεις σε αριθμητικές τιμές, διασφαλίζοντας ότι απαντήσεις με διαφορετική μορφοποίηση (π.χ. «6», «6.0», «6.00») αναγνωρίζονται ως ταυτόσημες. Βάσει των δεδομένων αυτών, υπολογίζονται δύο κρίσιμες μετρικές ανά ερώτηση: η συνέπεια (αγγλ. consistency), η οποία ορίζεται ως το ποσοστό των επαναλήψεων κατά τις οποίες το μοντέλο κατέληξε στην ίδια κυρίαρχη αριθμητική τιμή, και η ακρίβεια ανά ερώτηση (per_question_accuracy), που εκφράζει τον μέσο όρο των ορθών απαντήσεων στις δοκιμές αυτές.

Τα πρωτογενή αυτά δεδομένα συγκεντρώνονται σε ένα τελικό αρχείο αναφοράς, το οποίο αθροίζει τις επιδόσεις και υπολογίζει τη μέση συνέπεια και ακρίβεια για το σύνολο των ερωτήσεων. Η διαδικασία ολοκληρώνεται με τη χρήση του consistency_plots.py, το οποίο αξιοποιεί τα συγκεντρωτικά δεδομένα για την παραγωγή συγκριτικών ραβδογραμμάτων. Τα διαγράμματα αυτά απεικονίζουν παράλληλα την ακρίβεια και τη συνέπεια για κάθε μέθοδο προτροπής για το εκάστοτε μοντέλο, προσφέροντας μια εποπτική εικόνα της συσχέτισης μεταξύ της υπολογιστικής σταθερότητας και της μαθηματικής ορθότητας. Μέσω της μεθοδολογικής αυτής προσέγγισης, δεν αξιολογείται μόνο το τελικό αποτέλεσμα που παράγει το μοντέλο, αλλά και ο βαθμός στον οποίο η κάθε τεχνική προτροπής συμβάλλει στη δομική σταθερότητα της σκέψης του.

4.11 Ψευδοκώδικες και διαγράμματα ροής

Πίνακας 8 Η βασική ροή εκτέλεσης

Περιγραφή
Αρχή προγράμματος (main_parallel.py) Φόρτωσε δεδομένα GSM8K από load_data.py Καθόρισε τις μεθόδους: methods = [standard, cot, tot] Για κάθε μέθοδο m στις methods: Δημιούργησε εργασία (task) σε ξεχωριστή διεργασία: - Φόρτωσε/δημιούργησε LLM instance (pipeline.get_model) - Κάλεσε evaluation.evaluate() με: runner = pipeline.run_standard / run_cot / run_tot data = GSM8K subset - Πάρε αναλυτικά αποτελέσματα (per example) και συνοπτικές μετρικές Περίμενε να ολοκληρωθούν όλες οι διεργασίες Συγκέντρωσε τις μετρικές όλων των μεθόδων Αποθήκευσε αρχεία CSV για κάθε μέθοδο Τέλος προγράμματος

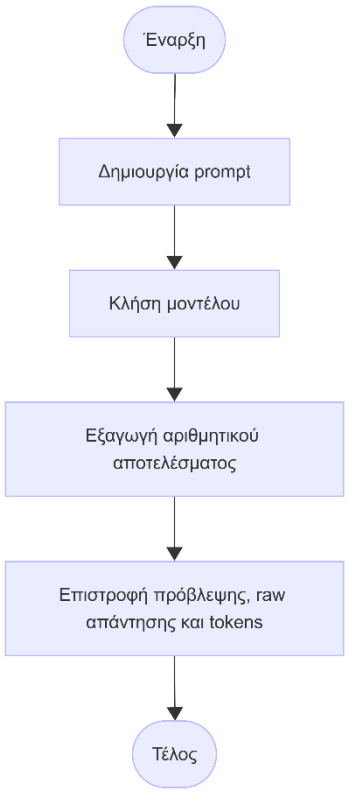


Εικόνα 1 Διάγραμμα ροής βασικής ροής εκτέλεσης

Πίνακας 9 Υλοποίηση της Standard Prompting

Περιγραφή

Συνάρτηση run_standard(problem, llm):
  prompt = φτιάξε_prompt_standard(problem)
 	
  (answer_text, usage) = llm.complete(prompt, temperature = TEMPERATURE_STD)
 	
  final_number = εξήγαγε_το_αποτέλεσμα_μετά_το("FINAL:", answer_text)
 	
  επιστροφή:
  prediction = final_number
  raw = answer_text
  tokens = usage

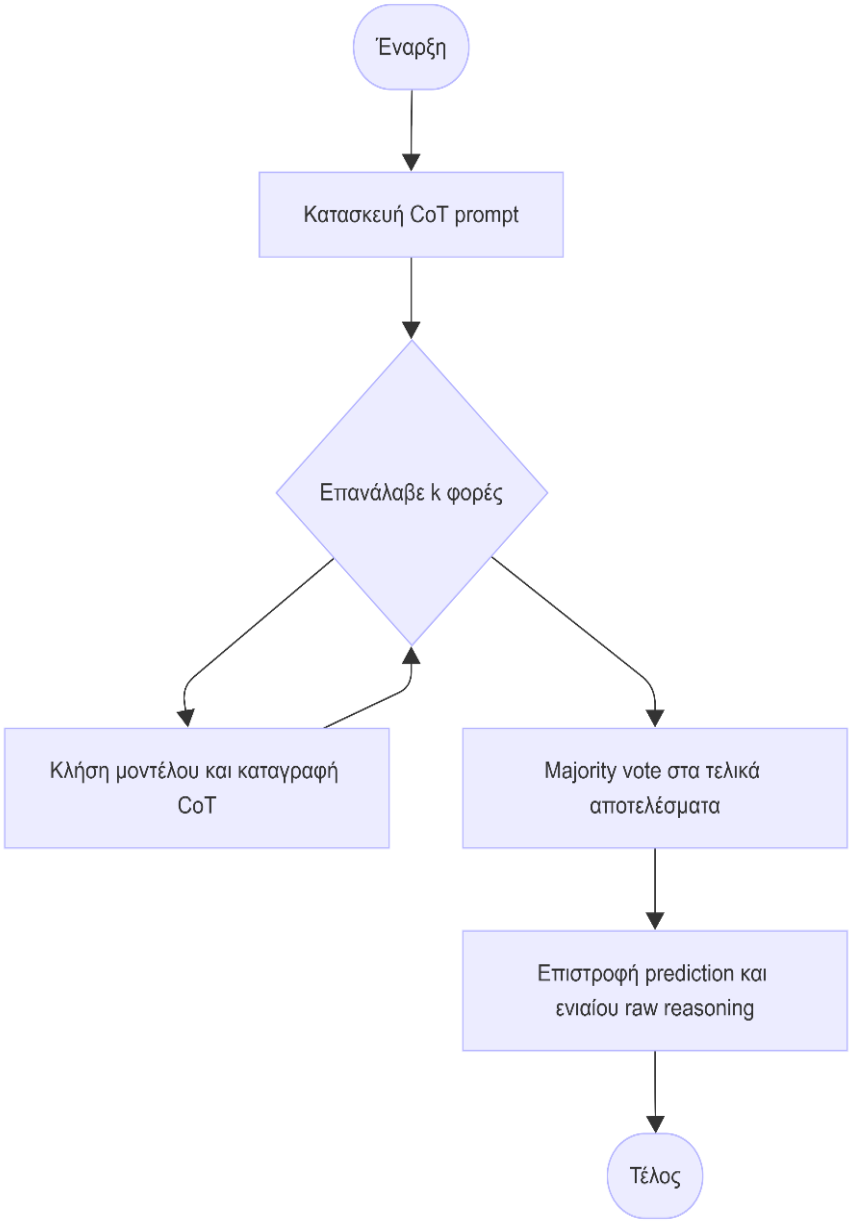


Εικόνα 2 Διάγραμμα ροής Standard Prompting εκτέλεσης

Πίνακας 10 Υλοποίηση της CoT με SC (k επαναλήψεις)

Περιγραφή

Συνάρτηση run_cot(problem, llm, k):
  prompt = φτιάξε_prompt_cot(problem)
 	
  λίστα predictions = []
  λίστα raw_cots = []
  συνολικά_tokens = 0
 	
  Για i από 1 έως k:
  (answer_text, usage) = llm.complete(prompt, temperature = TEMPERATURE_COT)
 	
  final_number = εξήγαγε_το_αποτέλεσμα_μετά_το("FINAL:", answer_text)
 	
  πρόσθεσε final_number στη λίστα predictions
  πρόσθεσε answer_text στη λίστα raw_cots
  συνολικά_tokens += usage.total
 	
  τελικό_result = τιμή_που_εμφανίζεται_πιο_συχνά(predictions)
 	
  raw_all = ένωση_όλων_των_raw_cots_με_διαχωριστικό("-----")
 	
  επιστροφή:
  prediction = τελικό_result
  raw = raw_all
  tokens = συνολικά_tokens

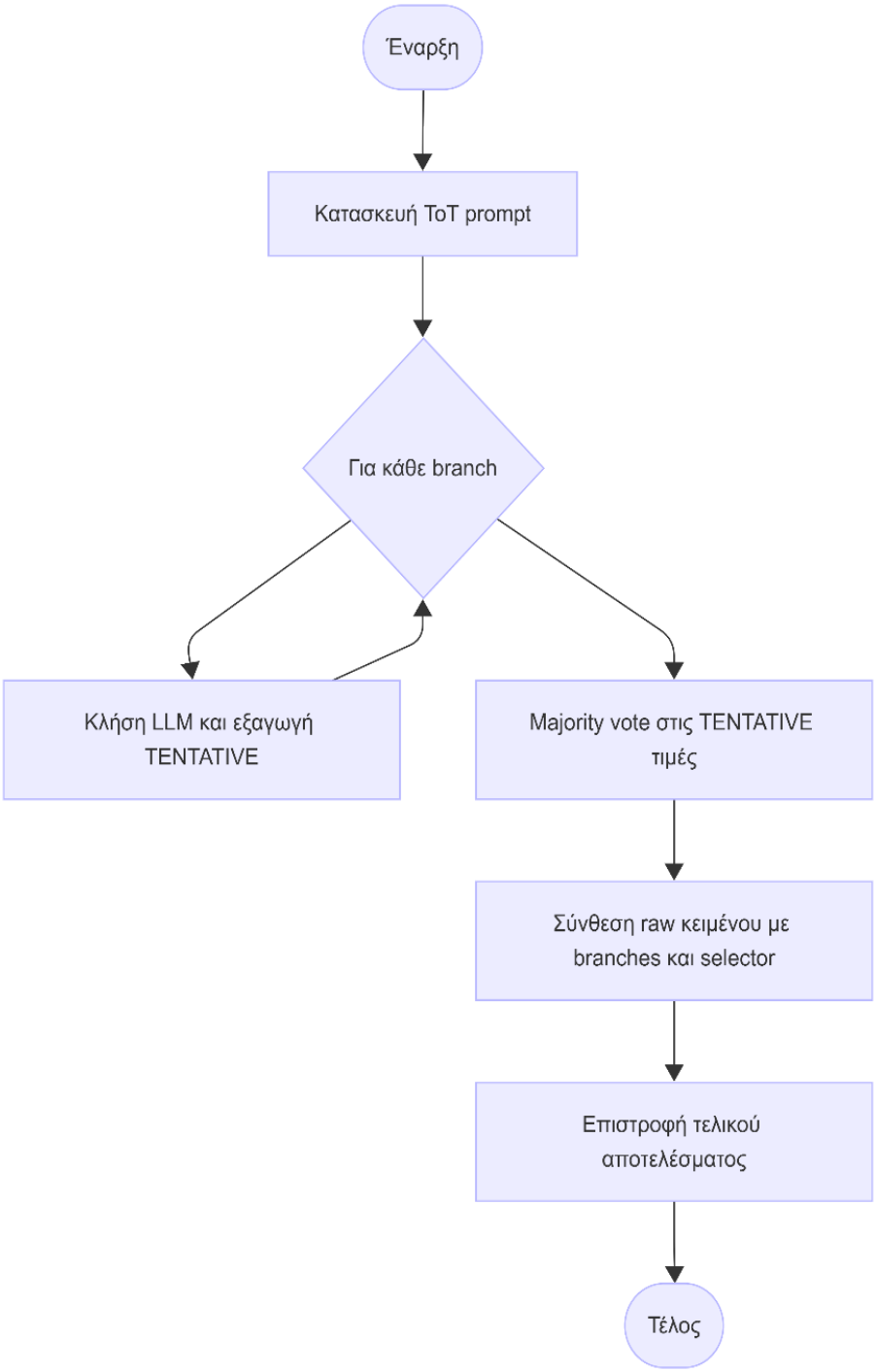


Εικόνα 3 Διάγραμμα ροής CoT εκτέλεσης

Πίνακας 11 Υλοποίηση της ToT με b branches

Περιγραφή

Συνάρτηση run_tot(problem, llm, b):
  prompt = φτιάξε_prompt_tot_branch(problem)
 	
  λίστα tentative_numbers = []
  λίστα branches_text = []
  συνολικά_tokens = 0
 	
  Για κάθε branch από 1 έως b:
  (answer_text, usage) = llm.complete(prompt, temperature = TEMPERATURE_TOT)
 	
  tentative = εξήγαγε_το_αποτέλεσμα_μετά_το("TENTATIVE:", answer_text)
 	
  πρόσθεσε tentative στη λίστα tentative_numbers
  πρόσθεσε answer_text στη λίστα branches_text
  συνολικά_tokens += usage.total
 	
  τελικό_result = τιμή_που_εμφανίζεται_πιο_συχνά(tentative_numbers)
 	
  raw_all = ένωση_όλων_των branches_text με διαχωριστικό:
  "===== BRANCH ====="
 	
  πρόσθεσε στο τέλος του raw_all block:
  "===== SELECTOR ====="
  "FINAL: " + τελικό_result
 	
  επιστροφή:
  prediction = τελικό_result
  raw = raw_all
  tokens = συνολικά_tokens



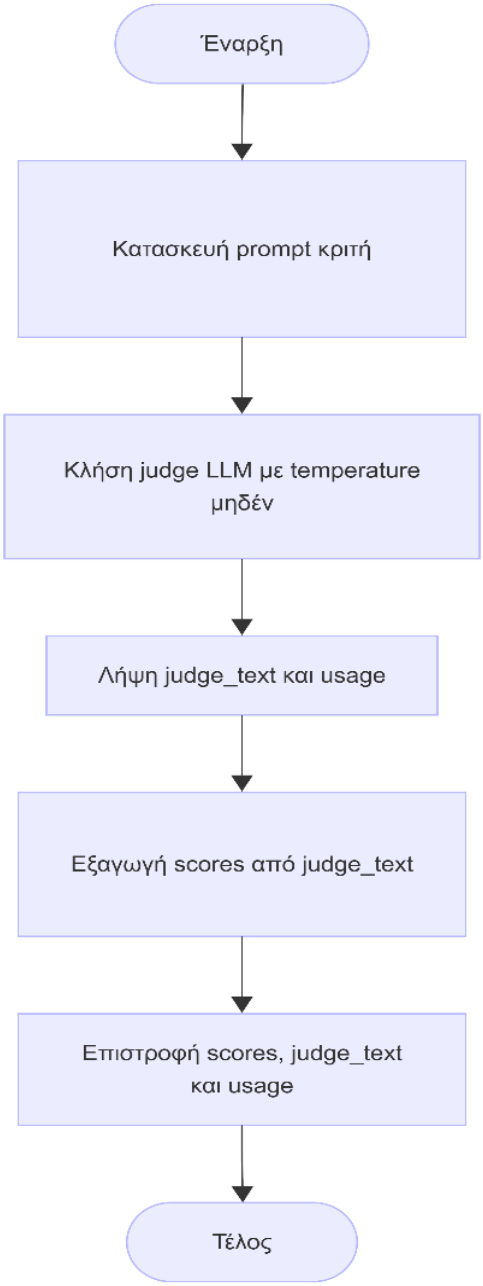
Εικόνα 4 Διάγραμμα ροής ToT εκτέλεσης

Πίνακας 12 Γενική διαδικασία αξιολόγησης για μια μεθόδου

Περιγραφή
Συνάρτηση evaluate(data, runner, llm): λίστα records = [] σωστά = 0 συνολικά = 0 Για κάθε problem στο data: ξεκίνα_χρονόμετρο() αποτέλεσμα_runner = runner(problem, llm) prediction = αποτέλεσμα_runner.prediction raw = αποτέλεσμα_runner.raw tokens = αποτέλεσμα_runner.tokens σταμάτα_χρονόμετρο() latency = χρόνος_εκτέλεσης gold = σωστή_απάντηση_από_GSM8K(problem) is_correct = (prediction == gold) αν is_correct: σωστά += 1 συνολικά += 1 πρόσθεσε στα records: (id, gold, prediction, is_correct, raw, tokens, latency) accuracy = σωστά / συνολικά επιστροφή: metrics = {accuracy, total=συνολικά, correct=σωστά} detailed_records = records

Πίνακας 13 Ροή αξιολόγησης με δεύτερο μοντέλο (Judge LLM)

Περιγραφή
Συνάρτηση judge_one(example, judge_llm): prompt_judge = φτιάξε_prompt_κριτή(problem_text, μοντέλο_που_έδωσε_τη_λύση, raw_reasoning, final_answer) (judge_text, usage) = judge_llm.complete(prompt_judge, temperature = 0) scores = εξήγαγε_scores_από(judge_text) # π.χ. correctness, logic, clarity, conciseness επιστροφή: scores, judge_text, usage
Και στη συνέχεια:
Συνάρτηση judge_all(results_csv, judge_llm): Για κάθε γραμμή στο results_csv: κάλεσε judge_one() με τα στοιχεία της γραμμής αποθήκευσε τις βαθμολογίες σε νέο αρχείο (judge_results.csv)
Με αυτό τον τρόπο, η διαδικασία αξιολόγησης αποτελεί ένα ξεχωριστό “διάγραμμα ροής” πάνω από τα ήδη υπάρχοντα pipelines.



Εικόνα 5 Διάγραμμα ροής αξιολόγησης με δεύτερο μοντέλο (Judge LLM)

4.12 Συγκριτική επισκόπηση των μεθόδων Prompt Engineering

Ο Πίνακας 14 συνοψίζει τις τρεις ροές προτροπής που αξιολογούνται στην παρούσα εργασία (Standard, CoT με SC και ToT), όπως αυτές υλοποιούνται στο pipeline.py και αξιολογούνται στο evaluation.py. Η αποτύπωση αυτή αναδεικνύει τις διαφορές τους ως προς το πλήθος κλήσεων, τον τρόπο εξαγωγής/επιλογής τελικής απάντησης και τα αναμενόμενα αντισταθμίσματα σε κατανάλωση tokens και χρόνο απόκρισης, υπό κοινές και τυποποιημένες πειραματικές συνθήκες.

Πίνακας 14 Συγκριτική αποτύπωση των ροών προτροπής και αξιολόγησης (Standard / CoT-SC / ToT)

Διάσταση σύγκρισης	**Standard Prompting (run_standard)**	**CoT + Self-Consistency (run_cot, SC_K = k)**	**ToT (run_tot, TOT_BRANCHES = b)**
Στόχος/Ρόλος στη μελέτη	Baseline αναφοράς	Ενίσχυση συλλογισμού μέσω βημάτων + σταθεροποίηση μέσω επαναλήψεων	Εξερεύνηση εναλλακτικών λύσεων μέσω κλάδων (branches)
Μορφή προτροπής	Άμεση προτροπή, χωρίς απαίτηση ανάλυσης	Προτροπή τύπου “solve step by step” (αλυσίδα συλλογισμού)	Προτροπή δημιουργίας κλάδου λύσης ανά εκτέλεση
Δομή εξόδου που αναμένει το σύστημα	FINAL: <number>	FINAL: <number> ανά επανάληψη	TENTATIVE: <number> ανά branch και τελικό FINAL: από selector
Πλήθος κλήσεων στο LLM ανά πρόβλημα	1	k (όσες ορίζει η SC_K)	b (όσες ορίζει η TOT_BRANCHES)
Μηχανισμός τελικής επιλογής	Μοναδική πρόβλεψη από 1 εκτέλεση	Πλειοψηφία (majority vote) πάνω στις k τελικές τιμές FINAL:	Πλειοψηφία (majority vote) πάνω στις b τιμές TENTATIVE: και σύνθεση FINAL:
Συγκέντρωση reasoning (raw)	1 raw κείμενο	Συρραφή όλων των raw CoT με διαχωριστικό (----)	Συρραφή όλων των branch κειμένων με ===== BRANCH ===== + ===== SELECTOR =====
Κύριες παράμετροι εκτέλεσης	TEMPERATURE_STD, NUM_PREDICT	TEMPERATURE_COT, SC_K, NUM_PREDICT	TEMPERATURE_TOT, TOT_BRANCHES, NUM_PREDICT
Αναμενόμενες επιδράσεις σε tokens/κόστος	Χαμηλότερα tokens (μία κλήση)	Αυξημένα tokens ~ k φορές (πολλαπλές κλήσεις)	Αυξημένα tokens ~ b φορές (πολλαπλά branches)
Αναμενόμενη επίδραση σε latency	Μικρότερο latency	Μεγαλύτερο latency λόγω k επαναλήψεων	Μεγαλύτερο latency λόγω b κλάδων

Συμπεριφορά ως προς συνέπεια (Consistency)	Πιθανή αστάθεια σε επαναλήψεις (λόγω στοχαστικότητας)	Στόχος η αύξηση σταθερότητας μέσω πλειοψηφίας	Στόχος η αύξηση σταθερότητας/εξερεύνησης με branches και πλειοψηφία
Τυποποίηση αξιολόγησης (evaluation.py)	Εξαγωγή αριθμού μετά το FINAL: και σύγκριση με gold	Εξαγωγή αριθμών από όλα τα FINAL: και πλειοψηφία → σύγκριση με gold	Εξαγωγή TENTATIVE: από branches → πλειοψηφία → FINAL: → σύγκριση με gold
Μετρικές που καταγράφονται	Accuracy, tokens, latency, raw output	Accuracy, tokens (συνολικά), latency, όλα τα raw CoT	Accuracy, tokens (συνολικά), latency, raw branches + selector
Ποιοτική αξιολόγηση (LLM-as-a-Judge)	Προαιρετικά μέσω judge pipeline	Προαιρετικά μέσω judge pipeline	Προαιρετικά μέσω judge pipeline (χρήσιμο λόγω πολυπλοκότητας reasoning)

Κεφάλαιο 5ο: Αξιολόγηση και μετρικές απόδοσης

Η αξιολόγηση των LLMs συνιστά μια θεμελιώδη και κρίσιμη διαδικασία σε κάθε μελέτη που επιχειρεί τη συγκριτική ανάλυση διαφορετικών μεθόδων προτροπής, όπως η Standard Prompting, η CoT, η SC και η ToT. Καθώς τα μοντέλα αυτά επιδεικνύουν ολοένα και πιο σύνθετες συμπεριφορές συλλογισμού, η ανάγκη για την υιοθέτηση πολυδιάστατων μετρικών καθίσταται επιτακτική. Η σύγχρονη βιβλιογραφία αναγνωρίζει ευρέως ότι η μονοδιάστατη μέτρηση της ακρίβειας δεν επαρκεί για την πλήρη αποτίμηση της ικανότητας ενός μοντέλου να παράγει αποκρίσεις που χαρακτηρίζονται από αξιοπιστία, σταθερότητα και επαληθευσιμότητα [1]. Επιπροσθέτως, η σχετική βιβλιογραφία που εστιάζει στην αξιολόγηση της CoT υπογραμμίζει ότι ο συλλογισμός των μοντέλων, δεν πρέπει να αντιμετωπίζεται ως διαδικασία μη παρατηρήσιμη ή μη ερμηνεύσιμη. Αντιθέτως, απαιτείται η ανάλυση και ο έλεγχός του μέσω σαφών και τεκμηριωμένων μεθοδολογιών, οι οποίες αξιολογούν όχι μόνο την ορθότητα του τελικού αποτελέσματος αλλά και την εγκυρότητα της ακολουθούμενης διαδρομής συλλογισμού [5].

5.1 Κριτήρια αξιολόγησης

Στο πλαίσιο της διεθνούς ερευνητικής κοινότητας των LLMs, η αξιολόγηση θεμελιώνεται στη χρήση πολλαπλών μετρικών, οι οποίες εξετάζουν διακριτές και αλληλοσυμπληρούμενες πτυχές της συμπεριφοράς ενός μοντέλου. Όπως τεκμηριώνεται σε σύγχρονες έρευνες που πραγματεύονται τις τεχνικές προτροπών και τον μηχανισμό συλλογισμού, η υιοθέτηση μιας ολιστικής προσέγγισης κρίνεται απαραίτητη προκειμένου να διαχωριστεί η επιφανειακή απόδοση από την ουσιαστική ικανότητα γενίκευσης και λογικής σκέψης του μοντέλου [3]. Η προσέγγιση αυτή επιτρέπει την εξαγωγή ασφαλέστερων συμπερασμάτων σχετικά με την καταλληλότητα κάθε μεθόδου για συγκεκριμένα προβλήματα και εφαρμογές.

5.1.1 Ακρίβεια (Accuracy)

Η ακρίβεια (αγγλ. Accuracy) αποτελεί μία από τις θεμελιώδεις και πλέον διαδεδομένες μετρικές αξιολόγησης στην επιστήμη των υπολογιστών και ειδικότερα στη μελέτη των LLMs. Σε θεωρητικό επίπεδο, ορίζεται ως το ποσοστό των περιπτώσεων κατά τις οποίες το μοντέλο παράγει την ορθή τελική απάντηση εντός ενός συνόλου δεδομένων όπου η σωστή τιμή είναι προκαθορισμένη (αγγλ. ground truth). Στη βιβλιογραφία των LLMs, η χρήση της κυριαρχεί σε εργασίες με μονοσήμαντο αποτέλεσμα, όπως τα αριθμητικά προβλήματα, η ταξινόμηση κειμένου και οι ερωτοαπαντήσεις κλειστού τύπου. Ωστόσο, παρά την ευρεία εφαρμογή της, η θεωρητική βάση της μετρικής αποκαλύπτει εγγενείς περιορισμούς. Ο Naveed και συν. (2024) υπογραμμίζουν ότι η ακρίβεια περιορίζεται στην καταγραφή της τελικής απόδοσης, αγνοώντας τη διαδικασία μέσω της οποίας προέκυψε η απάντηση. Η διάκριση αυτή είναι ιδιαίτερα κρίσιμη στα σύγχρονα LLMs, όπου η συμπεριφορά του συλλογισμού διαδραματίζει καθοριστικό ρόλο στην αξιοπιστία. Είναι χαρακτηριστικό ότι δύο μοντέλα ενδέχεται να παρουσιάζουν ταυτόσημη ακρίβεια, αλλά να διαφέρουν ποιοτικά: το ένα να καταλήγει στη λύση μέσω λογικά συνεκτικών βημάτων, ενώ το άλλο να επιτυγχάνει το ίδιο αποτέλεσμα μέσω επιφανειακών στατιστικών συσχετίσεων ή ακόμη και τυχαία.

Η αδυναμία της ακρίβειας να αποτυπώσει τη λογική εγκυρότητα επισημαίνεται περαιτέρω στη βιβλιογραφία [5], καθώς η μετρική δεν ενσωματώνει στοιχεία αξιολόγησης για την ποιότητα του συλλογισμού, τη συνοχή των ενδιάμεσων βημάτων ή την ύπαρξη λανθασμένων συλλογιστικών μοτίβων

(αγγλ. reasoning fallacies) που οδηγούν συμπτωματικά σε ορθό αποτέλεσμα. Ως εκ τούτου, η ακρίβεια θεωρείται επαρκής δείκτης κυρίως όταν τα δεδομένα αξιολόγησης δεν απαιτούν ερμηνεύσιμη ή πολυβηματική σκέψη. Παράλληλα, ο Wei και συν. (2023) παρέχουν τη θεωρητική τεκμηρίωση για την αύξηση της ακρίβειας όταν εφαρμόζονται τεχνικές εξαγωγής συλλογισμού, όπως η CoT. Συγκεκριμένα, υποστηρίζουν ότι τα μοντέλα μεγάλης κλίμακας διαθέτουν εγγενείς και αναδυόμενες ικανότητες (αγγλ. Emergent Abilities), οι οποίες παραμένουν ανενεργές με την απλή προτροπή, αλλά εκδηλώνονται όταν το μοντέλο καθοδηγείται να διατυπώσει ρητά τα ενδιάμεσα βήματα σκέψης. Υπό το πρίσμα αυτό, η CoT λειτουργεί ως μηχανισμός που καθιστά τον συλλογισμό ρητό και ελεγχόμενο, μειώνοντας την εξάρτηση από επιφανειακές στρατηγικές και ενισχύοντας την πιθανότητα παραγωγής ορθών απαντήσεων μέσω λογικής αλληλουχίας.

Συμπερασματικά, η θεωρία αναδεικνύει τον χαρακτήρα της ακρίβειας ως μετρικής επιπέδου εξόδου (αγγλ. output-level metric), η οποία αδυνατεί να αξιολογήσει τη διαδρομή της σκέψης, να συλλάβει τον βαθμό αβεβαιότητας ή να διακρίνει την τυχαία επιτυχία. Γι' αυτόν τον λόγο, σύγχρονες προσεγγίσεις στη μηχανική προτροπών και την αξιολόγηση (S. Vatsal και H. Dubey (2024)) προτείνουν τον συνδυασμό της ακρίβειας με προηγμένες μετρικές που εξετάζουν τη συνέπεια, τη βαθμολογική αυτοπεποίθηση (αγγλ. calibration) και την ποιότητα του συλλογισμού, στοχεύοντας στη δημιουργία ενός πληρέστερου και πιο αξιόπιστου πλαισίου αξιολόγησης.

5.1.2 Αξιοπιστία και αυτο-συνέπεια (Reliability, Self-consistency)

Η αξιοπιστία (αγγλ. reliability) ενός LLM ορίζεται ως η ικανότητά του να παράγει σταθερές και επαναλαμβανόμενες προβλέψεις όταν καλείται να ανταποκριθεί στην ίδια είσοδο υπό αμετάβλητες συνθήκες. Σε θεωρητικό επίπεδο, η ιδιότητα αυτή αποτελεί θεμελιώδες χαρακτηριστικό κάθε συστήματος πρόβλεψης, καθώς συνδέεται άρρηκτα με την εσωτερική σταθερότητα των μηχανισμών λήψης αποφάσεων και την ποιότητα της μοντελοποιημένης γνώσης. Η Gu και συν. (2025) επισημαίνουν ότι η συνέπεια (αγγλ. consistency) δεν λειτουργεί απλώς συμπληρωματικά ως προς την ακρίβεια, αλλά συνιστά ένα αυτόνομο κριτήριο κριτικής αξιοπιστίας. Η παρατήρηση αυτή ισχύει ιδιαίτερα για μοντέλα που προορίζονται να λειτουργήσουν ως αξιολογητές ή συστήματα γνωστικής υποστήριξης, πεδία όπου η αναπαραγωγιμότητα των αποτελεσμάτων είναι επιτακτική ανάγκη. Η θεωρητική απαίτηση για τη μέτρηση της συνέπειας απορρέει από την εγγενή φύση των LLMs ως στοχαστικών συστημάτων. Η διαδικασία αποκωδικοποίησης της απάντησης (αγγλ. decoding) επηρεάζεται από παραμέτρους όπως η θερμοκρασία (αγγλ. temperature), με αποτέλεσμα η ίδια είσοδος να δύναται να παράγει διαφοροποιημένες απαντήσεις. Το φαινόμενο αυτό αναδεικνύει τον θόρυβο που ενυπάρχει στον μηχανισμό παραγωγής κειμένου και καθιστά απαραίτητη τη διάκριση ανάμεσα στην περιστασιακή επιτυχία και τη σταθερή γνώση.

Προς αντιμετώπιση του ζητήματος αυτού, έχει προταθεί η θεωρητική έννοια της SC, η οποία στοχεύει στον περιορισμό της στοχαστικότητας μέσω της παραγωγής πολλαπλών ανεξάρτητων απαντήσεων και της επιλογής της επικρατέστερης [2]. Η συγκεκριμένη μέθοδος εδράζεται στην αρχή ότι η πιθανότητα επιλογής της ορθής λογικής διαδρομής αυξάνεται όταν εξετάζονται πολλαπλές υποψήφιες ακολουθίες συλλογισμού, μειώνοντας κατ' αυτόν τον τρόπο το ενδεχόμενο τυχαίων αποκλίσεων και ενισχύοντας τη λογική σταθερότητα του συστήματος. Επιπροσθέτως, η συνέπεια δεν περιορίζεται αποκλειστικά στο τελικό αποτέλεσμα, αλλά επεκτείνεται και στην εσωτερική συνοχή των βημάτων συλλογισμού (αγγλ. Reasoning Paths) που οδηγούν σε αυτό [5]. Η απουσία συνέπειας στα ενδιάμεσα στάδια μπορεί να υποδηλώνει ανεπαρκή μάθηση, χαμηλή ποιότητα της μοντελοποιημένης γνώσης ή υπερβολική εξάρτηση από στατιστικές συσχετίσεις αντί ουσιαστικής κατανόησης. Σε αυτό το θεωρητικό πλαίσιο, η συνέπεια λειτουργεί ως δείκτης δομικής κατανόησης και όχι απλώς ως μέτρο επαναληψιμότητας των

εξόδων. Συνοψίζοντας, το θεωρητικό υπόβαθρο της αξιοπιστίας υπογραμμίζει ότι η αξιολόγηση ενός LLM δεν πρέπει να βασίζεται μόνο στην ορθότητα του τελικού αποτελέσματος, αλλά και στη σταθερότητα, την αναπαραγωγιμότητα και τη λογική συνοχή των απαντήσεών του, στοιχεία που αντικατοπτρίζουν την εσωτερική ορθότητα και την προβλεψιμότητα της συμπεριφοράς του.

5.1.3 Ρυθμός αποχής (Abstention Rate)

Ο ρυθμός αποχής (αγγλ. Abstention Rate) συνιστά μια θεωρητικά θεμελιώδη μετρική αξιολόγησης για συστήματα τεχνητής νοημοσύνης που εκτελούν καθήκοντα λογικής, λήψης αποφάσεων και απάντησης ερωτημάτων. Ορίζεται ως το ποσοστό των περιπτώσεων κατά τις οποίες ένα μοντέλο επιλέγει συνειδητά να μην απαντήσει σε ένα ερώτημα, ενεργοποιώντας έναν μηχανισμό αποφυγής σφάλματος όταν το επίπεδο εμπιστοσύνης του είναι χαμηλό ή όταν η εσωτερική του συλλογιστική διαδικασία δεν οδηγεί σε σταθερή και αξιόπιστη πρόβλεψη. Η θεωρητική βάση της μετρικής αυτής συνδέεται άρρηκτα με την ανάγκη αξιολόγησης της γνωσιακής αβεβαιότητας των LLMs, ήτοι της ικανότητάς τους να αναγνωρίζουν πότε δεν διαθέτουν επαρκή πληροφορία ή βεβαιότητα για την παραγωγή έγκυρης απάντησης. Σύμφωνα με την Aithal και συν. (2025), η αποχή συνδέεται άμεσα με το ζήτημα της λογικής αξιοπιστίας, καθώς μειώνει την πιθανότητα παραγωγής λανθασμένων απαντήσεων που συνοδεύονται από ατεκμηρίωτη αυτοπεποίθηση. Η μελέτη επισημαίνει ότι τα LLMs τείνουν συχνά να συμπληρώνουν απαντήσεις ακόμη και ελλείψει επαρκών συλλογιστικών ενδείξεων, οδηγώντας σε φαινόμενα παραισθήσεων (αγγλ. hallucination) ή μη ρεαλιστικής βεβαιότητας. Ως εκ τούτου, ο ρυθμός αποχής προσφέρει έναν μηχανισμό μέτρησης και ελέγχου αυτής της συμπεριφοράς, λειτουργώντας ως δείκτης εσωτερικής αναστολής και ενδογενούς αυτορρύθμισης.

Επιπροσθέτως, από το θεωρητικό πλαίσιο που αναλύεται από της Gu και συν. (2025), προκύπτει ότι η αποχή αποτελεί κρίσιμη διάσταση της συνολικής αξιοπιστίας ενός μοντέλου. Ένα σύστημα που αποφεύγει να απαντήσει σε συνθήκες υψηλής αβεβαιότητας δύναται να θεωρηθεί περισσότερο κριτικά αξιόπιστο, καθώς αναγνωρίζει τα γνωστικά του όρια και δεν παράγει αυθαίρετες ή ψευδώς βέβαιες απαντήσεις. Η αποχή, επομένως, δεν εκλαμβάνεται ως ένδειξη αδυναμίας, αλλά ως τεκμήριο ώριμης γνωσιακής συμπεριφοράς, ιδίως σε κρίσιμες εφαρμογές όπου η λανθασμένη απάντηση ενέχει μεγαλύτερο κόστος από τη μη απάντηση. Η θεωρητική σημασία της μετρικής επεκτείνεται επίσης στη μελέτη της βαθμονόμησης (αγγλ. calibration) των LLMs. Σύμφωνα με τον Naveed και συν. (2024), τα μοντέλα αυτά τείνουν να εμφανίζουν σφάλμα υπερβολικής αυτοπεποίθησης (overconfidence bias), παράγοντας απαντήσεις με υψηλή βεβαιότητα ακόμη και όταν η γνώση τους είναι περιορισμένη. Στο πλαίσιο αυτό, ο ρυθμός αποχής λειτουργεί ως αντισταθμιστικός μηχανισμός που επιτρέπει την ανίχνευση και τον περιορισμό περιπτώσεων όπου η αυτοπεποίθηση του μοντέλου δεν ευθυγραμμίζεται με την πραγματική του ικανότητα. Συμπερασματικά, το θεωρητικό υπόβαθρο του ρυθμού αποχής υπογραμμίζει ότι η δυνατότητα ενός μοντέλου να σιωπά σε καταστάσεις γνωσιακής αβεβαιότητας αποτελεί χαρακτηριστικό εξίσου σημαντικό με την ικανότητά του να απαντά ορθά, καθώς συνδέεται άμεσα με τη γνωσιακή ωριμότητα, τη βαθμονόμηση της εμπιστοσύνης και την αποφυγή επικίνδυνα παραπλανητικών αποτελεσμάτων.

5.1.4 Αναμενόμενο σφάλμα βαθμονόμησης (Expected Calibration Error, ECE)

Η έννοια της βαθμονόμησης (αγγλ. calibration) στα LLMs αναφέρεται στη σχέση ανάμεσα στην εκπεφρασμένη εμπιστοσύνη ενός μοντέλου και την πραγματική πιθανότητα ορθότητας των απαντήσεών του. Θεωρητικά, ένα άριστα βαθμονομημένο μοντέλο χαρακτηρίζεται από πλήρη αντιστοιχία: όταν δηλώνει εμπιστοσύνη 70%, οι απαντήσεις του οφείλουν να είναι ορθές στο 70% των περιπτώσεων. Ωστόσο, στη σύγχρονη βιβλιογραφία [1], αναδεικνύεται ως κρίσιμο ζήτημα το φαινόμενο

της υπερβολικής αυτοπεποίθησης (αγγλ. overconfidence), όπου τα μοντέλα τείνουν να αναθέτουν υψηλές πιθανότητες ορθότητας στις προβλέψεις τους, ανεξαρτήτως της πραγματικής εγκυρότητας της συλλογιστικής τους διαδικασίας. Η ασυμφωνία αυτή μεταξύ υποκειμενικής εμπιστοσύνης και αντικειμενικής ακρίβειας δημιουργεί σοβαρά προβλήματα αξιοπιστίας, ιδιαίτερα σε καθήκοντα που απαιτούν αξιόπιστη λήψη αποφάσεων ή κρίσιμη συλλογιστική.

Για την ποσοτικοποίηση αυτής της απόκλισης χρησιμοποιείται η μετρική του Αναμενόμενου Σφάλματος Βαθμονόμησης (αγγλ. Expected Calibration Error - ECE), η οποία μετρά τη μέση διαφορά ανάμεσα στο επίπεδο αυτοπεποίθησης του μοντέλου και στην πραγματική του επιτυχία. Η ECE υπολογίζεται μέσω της ομαδοποίησης των προβλέψεων σε διαστήματα εμπιστοσύνης (αγγλ. confidence bins) και της εκτίμησης της απόλυτης διαφοράς μεταξύ της μέσης εμπιστοσύνης και της πραγματικής ακρίβειας σε κάθε διάστημα. Μια υψηλή τιμή ECE υποδηλώνει κακή βαθμονόμηση και αδυναμία του μοντέλου να διαμορφώσει μια σταθερή και αξιόπιστη εσωτερική αναπαράσταση της αβεβαιότητάς του. Η αναγκαιότητα της βαθμονόμησης τονίζεται ιδιαίτερα από την Aithal και συν. (2025), οι οποίοι επισημαίνουν ότι τα μοντέλα που παράγουν εκτενείς αλυσίδες σκέψης (CoT) συχνά συνοδεύουν τις απαντήσεις τους με υψηλά επίπεδα βεβαιότητας, ακόμη και όταν σφάλλουν. Η τάση αυτή οδηγεί στην παραγωγή πειστικών αλλά λανθασμένων απαντήσεων, γεγονός που καθιστά την ECE έναν απαραίτητο δείκτη για τον εντοπισμό της ψευδαίσθησης γνώσης.

Επιπλέον, η ορθή βαθμονόμηση κρίνεται απαραίτητη όταν τα LLMs αξιοποιούνται ως κριτές (LLM-as-a-Judge) για την αξιολόγηση άλλων συστημάτων [12]. Ένα μοντέλο χωρίς αξιόπιστη βαθμονόμηση κινδυνεύει να υπερεκτιμήσει την ποιότητα μιας απάντησης ή να αποτύχει να αναγνωρίσει σημεία ασάφειας. Συνεπώς, η ECE δεν συνιστά απλώς μια τεχνική μετρική, αλλά ένα ουσιαστικό θεωρητικό εργαλείο για την κατανόηση της γνωσιακής πιστότητας του μοντέλου. Το θεωρητικό υπόβαθρο της ECE αναδεικνύει ότι η εμπιστοσύνη ενός LLM πρέπει να ερμηνεύεται με προσοχή: ένα μοντέλο μπορεί να είναι ακριβές αλλά κακώς βαθμονομημένο, ή βαθμονομημένο αλλά όχι ακριβές. Η ECE μετρά ακριβώς αυτή τη διάσταση, λειτουργώντας ως βασικός δείκτης ποιότητας της εσωτερικής αξιολόγησης αβεβαιότητας και καθιστώντας την αναπόσπαστο μέρος της συνολικής θεωρητικής αξιολόγησης.

5.1.5 Χρόνος απόκρισης ή καθυστέρηση (Latency)

Ο χρόνος απόκρισης ή καθυστέρηση (αγγλ. latency) συνιστά μία από τις θεμελιώδεις μετρικές στην αξιολόγηση της απόδοσης των LLMs, καθώς αποτυπώνει τη χρονική καθυστέρηση που μεσολαβεί μεταξύ της εισαγωγής της προτροπής και της παραγωγής της τελικής απάντησης. Σε θεωρητικό επίπεδο, η έννοια του χρόνου απόκρισης συναρτάται άμεσα με το υπολογιστικό κόστος, την πολυπλοκότητα της αρχιτεκτονικής και τον όγκο των παραγόμενων tokens, χαρακτηριστικά που καθορίζουν την πρακτική χρησιμότητα ενός LLM σε περιβάλλοντα υψηλών απαιτήσεων. Σύμφωνα με τον Naveed και συν. (2024), ο χρόνος απόκρισης αποτελεί κρίσιμο δείκτη της υπολογιστικής αποδοτικότητας, δεδομένου ότι τα μοντέλα με δισεκατομμύρια παραμέτρους απαιτούν εκτεταμένους πόρους για την παραγωγή κάθε token. Η αύξηση του μεγέθους των μοντέλων και του μήκους των ακολουθιών συνεπάγεται εκθετική αύξηση της καθυστέρησης, γεγονός που θέτει περιορισμούς στη χρήση τους σε εφαρμογές πραγματικού χρόνου.

Παράλληλα, η βιβλιογραφία σχετικά με τον πολυσταδιακό συλλογισμό, όπως περιγράφεται από την Aithal και συν. (2025), επισημαίνει ότι τεχνικές προτροπής όπως η CoT και η ToT επιβαρύνουν σημαντικά τον χρόνο απόκρισης λόγω της αναγκαιότητας παραγωγής εκτενών ενδιάμεσων βημάτων. Η μεν CoT απαιτεί από το μοντέλο να διατυπώσει πλήρεις λογικές αλυσίδες, οι οποίες πολλαπλασιάζουν τον αριθμό των tokens πριν από την τελική απάντηση η δε ToT εισάγει επιπρόσθετους εσωτερικούς

κλάδους εξερεύνησης (αγγλ. branches), δημιουργώντας πολλαπλά υποψήφια μονοπάτια συλλογισμού. Η αυξημένη αυτή πολυπλοκότητα μεγεθύνει τον απαιτούμενο χρόνο επεξεργασίας και τη συνολική καθυστέρηση του συστήματος. Επιπλέον, του Oshin και συν. (2025) υπογραμμίζουν ότι η καθυστέρηση καθίσταται κρίσιμη μετρική στον σχεδιασμό σύνθετων ροών εργασίας (αγγλ. pipelines). Όταν ένα σύστημα ενσωματώνει πολλαπλά στάδια, όπως επαναληπτικό συλλογισμό, εργαλεία ανάκτησης πληροφορίας ή δειγματοληψία αυτο-συνέπειας (αγγλ. SC sampling), κάθε στάδιο προσθέτει αθροιστικά στη συνολική καθυστέρηση.

Η θεωρητική πρόκληση έγκειται στην επίτευξη μιας βέλτιστης ισορροπίας μεταξύ της ποιότητας του συλλογισμού και της ταχύτητας απόκρισης, καθώς τα μοντέλα που παράγουν υψηλής ποιότητας λογικές αλυσίδες συχνά υστερούν σε ταχύτητα. Το ζήτημα αυτό συνδέεται και με την έννοια της αποδοτικής αποκωδικοποίησης (αγγλ. efficient decoding). Οι Vatsal και Dubey (2024) σημειώνουν ότι επιλογές όπως η υψηλή θερμοκρασία ή η δειγματοληψία nucleus (αγγλ. nucleus sampling) έχουν άμεσο αντίκτυπο στον χρόνο εκτέλεσης. Στο θεωρητικό πλαίσιο της αξιολόγησης, ο χρόνος απόκρισης δεν εξετάζεται ως μεμονωμένη μετρική, αλλά ως αναπόσπαστη συνιστώσα ενός «τριγώνου απόδοσης» που περιλαμβάνει την ποιότητα του συλλογισμού, το υπολογιστικό κόστος και την ταχύτητα παραγωγής. Η επίτευξη ανώτερου συλλογισμού μέσω CoT ή ToT οδηγεί θεωρητικά σε αύξηση του χρόνου απόκρισης, ενώ η βεβιασμένη μείωσή του συχνά συνεπάγεται έκπτωση στην τεκμηρίωση και τη λογική συνοχή. Συνεπώς, η θεωρητική ανάλυση αναδεικνύει τον χρόνο απόκρισης ως έναν από τους βασικούς περιοριστικούς παράγοντες των LLMs και ως κρίσιμο κριτήριο διαφοροποίησης στη συγκριτική μελέτη των μεθόδων προτροπής.

5.2 Πειραματικές ρυθμίσεις (Αριθμός Προτροπών, Temperature, Context Size)

Οι πειραματικές ρυθμίσεις συνιστούν κρίσιμο παράγοντα για την ορθή και αξιόπιστη αξιολόγηση των LLMs, ιδίως όταν επιχειρείται η σύγκριση διαφορετικών τεχνικών προτροπής, όπως η Standard Prompting, CoT, SC και ToT. Η σχετική βιβλιογραφία υπογραμμίζει ότι ακόμη και οριακές μεταβολές στις παραμέτρους εκτέλεσης δύνανται να επηρεάσουν δραστικά την απόδοση ενός μοντέλου, καθιστώντας επιτακτική την αυστηρή τήρηση και τεκμηρίωση των συνθηκών του πειράματος. Στο πλαίσιο αυτό, αναδεικνύονται τρεις κρίσιμες παράμετροι: ο αριθμός των προτροπών, η θερμοκρασία αποκωδικοποίησης (αγγλ. temperature) και το μέγεθος του context window (αγγλ. context size) [3] [2] [5].

Ο αριθμός των προτροπών καθορίζει το μέγεθος του δείγματος επί του οποίου υπολογίζονται οι μετρικές απόδοσης. Θεωρητικά, η αύξηση του πλήθους των προτροπών ενισχύει τη στατιστική ισχύ και την αξιοπιστία των αποτελεσμάτων. Ειδικότερα σε μελέτες αξιολόγησης συλλογισμού, τονίζεται ότι οι εργασίες υψηλής πολυπλοκότητας απαιτούν επαρκές δείγμα για την εξαγωγή ασφαλών συμπερασμάτων, ειδικά κατά την εφαρμογή τεχνικών όπως η SC ή η ToT, όπου κάθε προτροπή μπορεί να παράγει πολλαπλές ανεξάρτητες λογικές διαδρομές [5]. Η αύξηση του δείγματος συμβάλλει στον περιορισμό της τυχαιότητας που εισάγει η στοχαστική αποκωδικοποίηση (αγγλ. stochastic decoding), καθιστώντας τα αποτελέσματα πιο αντιπροσωπευτικά της πραγματικής συμπεριφοράς του μοντέλου [2]. Στην περίπτωση της μεθόδου SC, ο αριθμός των εκτελέσεων ανά προτροπή (K samples) συνιστά μια επιπρόσθετη εσωτερική παράμετρο που επιδρά καθοριστικά στις τελικές μετρικές αξιολόγησης.

Η θερμοκρασία αποκωδικοποίησης (αγγλ. temperature) αποτελεί σημαντική ρυθμιστική παράμετρο του μηχανισμού δειγματοληψίας (αγγλ. sampling), καθορίζοντας τον βαθμό στοχαστικότητας κατά την παραγωγή των tokens. Θεωρητικά, χαμηλές τιμές ($T \approx 0$) οδηγούν σε ντετερμινιστική συμπεριφορά, ενώ υψηλότερες τιμές επιτρέπουν μεγαλύτερη ποικιλομορφία στις απαντήσεις. Τα μοντέλα τείνουν να

επιδεικνύουν βελτιωμένη ικανότητα παραγωγής λογικών ακολουθιών υπό αυξημένη θερμοκρασία, καθώς η συνθήκη αυτή ενθαρρύνει τη διερεύνηση διαφορετικών μονοπατιών σκέψης [3]. Η ποικιλομορφία αυτή είναι απαραίτητη για τεχνικές όπως η SC, όπου η ακρίβεια βελτιώνεται μέσω της σύνθεσης πολλαπλών υποψήφιων αλυσίδων συλλογισμού, αλλά και για τη μέθοδο ToT, όπου η υψηλή στοχαστικότητα είναι αναγκαία για την εξερεύνηση διαφορετικών κλάδων (αγγλ. branches) και την παραγωγή εναλλακτικών λύσεων. Ωστόσο, όπως επισημαίνουν η Gu και συν. (2025), η υπερβολική στοχαστικότητα μπορεί να επιφέρει αστάθεια στη σύγκριση των μοντέλων, απαιτώντας προσεκτική ρύθμιση για τη διασφάλιση της επαναληψιμότητας.

Το μέγεθος χωρητικότητας προτροπής (αγγλ. context size) ορίζει τον μέγιστο αριθμό tokens που δύναται να επεξεργαστεί το μοντέλο κατά την παραγωγή μιας απάντησης, επιδρώντας άμεσα στην ικανότητά του για σύνθετο συλλογισμό. Ένα εκτεταμένο παράθυρο συμφραζομένων επιτρέπει την αποτελεσματική διαχείριση μακροσκελών αλυσίδων CoT, πολύπλοκων οδηγιών και εκτενών συλλογιστικών μονοπατιών [1]. Αντιθέτως, ο περιορισμός της χωρητικότητας οδηγεί σε απώλεια κρίσιμων πληροφοριών, διακοπή των λογικών συνδέσεων και σε φαινόμενα υπερχείλισης πλαισίου (αγγλ. context overflow). Η σημασία της παραμέτρου αυτής επιβεβαιώνεται από μελέτες που καταδεικνύουν ότι η ποιότητα του συλλογισμού υποβαθμίζεται όταν οι αλυσίδες CoT υπερβαίνουν το διαθέσιμο πλαίσιο, αναγκάζοντας το μοντέλο να ξεχάσει προηγούμενα βήματα της λογικής πορείας [5]. Η ανάγκη για ορθή διαχείριση του context window καθίσταται ακόμα πιο επιτακτική σε πολυσταδιακές ροές εργασίας, όπου τα ενδιάμεσα βήματα συλλογισμού καταναλώνουν σημαντικό μέρος της μνήμης εργασίας, οδηγώντας σε αυξημένο χρόνο απόκρισης και υποβάθμιση των μετρικών ποιότητας [20]. Συμπερασματικά, η τυποποίηση αυτών των τριών παραμέτρων: αριθμός προτροπών, θερμοκρασία και context size, συνθέτει τον πυρήνα της πειραματικής εγκυρότητας, καθώς επηρεάζουν άμεσα την ακρίβεια, τη συνέπεια, τη βαθμονόμηση, την ποιότητα και τον συνολικό χρόνο απόκρισης των υπό εξέταση συστημάτων.

5.3 Μετρική Self-Consistency

Η μετρική Self-Consistency (SC) αποτελεί μία από τις πλέον θεμελιώδεις θεωρητικές προσεγγίσεις για την αξιολόγηση της αξιοπιστίας και της λογικής σταθερότητας των LLMs, ιδίως όταν αυτά εφαρμόζουν διαδικασίες συλλογισμού όπως η μέθοδος CoT. Η αναγκαιότητα εισαγωγής της συγκεκριμένης μετρικής απορρέει από τον εγγενώς στοχαστικό χαρακτήρα των LLMs, ο οποίος οδηγεί στο φαινόμενο της μη ντετερμινιστικής συμπεριφοράς, όπου το ίδιο ερώτημα δύναται να παράγει διαφορετικές απαντήσεις σε διαδοχικές εκτελέσεις, ιδιαίτερα όταν χρησιμοποιούνται μέθοδοι αποκωδικοποίησης με θερμοκρασία μεγαλύτερη του μηδενός. Η στοχαστικότητα αυτή δεν επηρεάζει αποκλειστικά την τελική απάντηση αλλά διαχέεται και στα ενδιάμεσα βήματα συλλογισμού, δημιουργώντας πολλαπλές πιθανές λογικές διαδρομές [2]. Θεωρητικά, οι διαδρομές αυτές ενδέχεται να είναι εξίσου ορθές ή, αντίθετα, να αποκλίνουν σημαντικά από μια συνεκτική και έγκυρη ακολουθία σκέψης. Η SC προτείνεται ως μέθοδος άμβλυνσης αυτής της αστάθειας, βασιζόμενη στην υπόθεση ότι τα LLMs διαθέτουν τη δυνατότητα να παράγουν πολλαπλές, εν δυνάμει ορθές λογικές πορείες και ότι η στατιστικά συχνότερη απάντηση αποτελεί την πλέον αξιόπιστη εκτίμηση της πραγματικής γνώσης του μοντέλου.

Η θεωρητική τεκμηρίωση της SC ενισχύεται περαιτέρω από μελέτες αξιολόγησης συλλογισμού, όπου επισημαίνεται ότι η συνέπεια ενός μοντέλου δεν περιορίζεται στην απλή επαναληψιμότητα του τελικού αποτελέσματος, αλλά επεκτείνεται στην εσωτερική συνοχή και τη σταθερότητα των παραγόμενων μονοπατιών σκέψης [5]. Ένα μοντέλο που παράγει λογικά συνεπείς και σταθερές ακολουθίες θεωρείται ότι έχει αναπτύξει πιο αξιόπιστες γνωσιακές δομές, εν αντιθέσει με την ύπαρξη έντονης αστάθειας που υποδηλώνει ελλιπή μοντελοποίηση γνώσης ή υπερβολική εξάρτηση από επιφανειακά στατιστικά

μοτίβα. Επιπροσθέτως, η συνέπεια αναδεικνύεται ως θεμελιώδης δείκτης κριτικής αξιοπιστίας (αγγλ. critical reliability), ειδικά για συστήματα που καλούνται να λειτουργήσουν ως αξιολογητές της ποιότητας άλλων αποτελεσμάτων [12]. Στο πλαίσιο αυτό, η SC εντάσσεται ως μέθοδος που ενισχύει τη σταθερότητα του συλλογισμού και ελαχιστοποιεί τα σποραδικά σφάλματα που προκύπτουν από μεμονωμένες αποκωδικοποιήσεις. Το θεωρητικό υπόβαθρο της μεθόδου αντιμετωπίζει τη λογική αστάθεια των LLMs όχι ως πρόβλημα προς εξάλειψη, αλλά ως μετρήσιμο φαινόμενο που δύναται να βελτιωθεί μέσω της παραγωγής πολλαπλών ακολουθιών και της επιλογής της επικρατέστερης, συμβάλλοντας σε μια ολοκληρωμένη αξιολόγηση της ικανότητας του μοντέλου για σταθερή και συνεκτική λογική επεξεργασία.

Κεφάλαιο 6ο: Αποτελέσματα και συγκριτική ανάλυση

Το παρόν κεφάλαιο παραθέτει τα πειραματικά αποτελέσματα και τη συγκριτική ανάλυση των μεθόδων προτροπής και συλλογισμού που εφαρμόστηκαν στο σύνολο δεδομένων GSM8K, με στόχο την συστηματική αποτίμηση του τρόπου με τον οποίο διαφορετικές τεχνικές καθοδήγησης επηρεάζουν την ικανότητα των LLMs στην επίλυση προβλημάτων πολυσταδιακού μαθηματικού συλλογισμού. Η ανάλυση καλύπτει διαδοχικά τις μεθόδους Standard Prompting, CoT, SC-CoT και ToT, η οποία εξετάζεται σε δύο διακριτές ρυθμίσεις με 3 και 5 κλάδους (αγγλ. branches) αντίστοιχα. Σε κάθε περίπτωση, η διερεύνηση δεν περιορίζεται αποκλειστικά στην καθαρή επίδοση, δηλαδή στην ορθότητα της απάντησης, αλλά επεκτείνεται στη λειτουργική συμπεριφορά της εκάστοτε μεθόδου, εξετάζοντας παραμέτρους όπως το υπολογιστικό κόστος, ο χρόνος απόκρισης, η σταθερότητα και η ποιότητα του παραγόμενου συλλογισμού.

Σημαντική αρχή του κεφαλαίου είναι ότι τα πειραματικά αποτελέσματα δεν προκύπτουν από θεωρητικές προσεγγίσεις ούτε βασίζονται σε εξωτερικά σημεία αναφοράς (αγγλ. Benchmarks), αλλά εξάγονται αποκλειστικά από το πειραματικό σύστημα (αγγλ. pipeline) που αναπτύχθηκε στο πλαίσιο της παρούσας εργασίας και τα παραγόμενα αρχεία δεδομένων. Συγκεκριμένα, η ανάλυση βασίζεται στα πρωτογενή δεδομένα ανά εκτέλεση από τα αρχεία καταγραφής (gsm8k_ALL_methods_ALL_runs_rows), στις συγκεντρωτικές μετρικές ανά μέθοδο και ρύθμιση από τα αρχεία σύνοψης (summary_all_models_configs), καθώς και στα αποτελέσματα της ποιοτικής αξιολόγησης «LLM-as-a-Judge» (judge_summary_all_models_configs), όπου οι απαντήσεις βαθμολογούνται σε διαστάσεις όπως η ορθότητα (αγγλ. correctness), η λογική (αγγλ. logic), η σαφήνεια (αγγλ. clarity) και η συντομία (αγγλ. conciseness). Με τον τρόπο αυτό, κάθε αριθμητικό εύρος ή συμπέρασμα που κατατίθεται στις ενότητες που ακολουθούν (6.1 έως 6.7) είναι άμεσα επαληθεύσιμο και μπορεί να αναπαραχθεί από τα αρχεία των αποτελεσμάτων.

Η συγκριτική αποτίμηση οργανώνεται γύρω από ένα συνεκτικό σύνολο μετρικών, οι οποίες υπολογίζονται με ενιαίο τρόπο για όλες τις μεθόδους ώστε να διασφαλίζεται η ισότιμη σύγκριση. Η ακρίβεια (αγγλ. Accuracy) υπολογίζεται ως το ποσοστό των προβλημάτων όπου η τελική απάντηση ταυτίζεται με την ορθή λύση (αγγλ. ground truth) του GSM8K. Ο ρυθμός αποχής (αγγλ. Abstention Rate) καταγράφει το ποσοστό των περιπτώσεων όπου το σύστημα επιλέγει να μην απαντήσει, στοιχείο κρίσιμο για την αξιοπιστία σε συνθήκες υψηλής αβεβαιότητας, ο χρόνος απόκρισης (αγγλ. latency) εκφράζει τη χρονική διάρκεια επεξεργασίας ανά ερώτηση, ενώ η κατανάλωση tokens αποτυπώνει το υπολογιστικό και, κατ’ επέκταση, το οικονομικό κόστος. Παράλληλα, ο έλεγχος συνέπειας (αγγλ. Consistency Check) αξιοποιείται για τη μέτρηση της σταθερότητας των απαντήσεων σε επαναλαμβανόμενες εκτελέσεις, αποτυπώνοντας τη συχνότητα με την οποία το σύστημα καταλήγει στο ίδιο αποτέλεσμα υπό συνθήκες στοχαστικής αποκωδικοποίησης. Τέλος, οι δείκτες του LLM-as-a-Judge λειτουργούν ως συμπληρωματική, ποιοτική αξιολόγηση της λογικής συνοχής.

Η παρουσίαση των αποτελεσμάτων γίνεται, όπου κρίνεται απαραίτητο, με τη χρήση τυπικών στατιστικών μεγεθών (π.χ. median, mean, τεταρτημόρια) και εύρους τιμών (min–max), προκειμένου να αποτυπωθούν όχι μόνο οι μέσες επιδόσεις αλλά και οι ακραίες συμπεριφορές ή οι περιπτώσεις αστοχίας, οι οποίες είναι ιδιαίτερα σημαντικές για την πλήρη κατανόηση της απόδοσης σύνθετων μεθόδων όπως η ToT, όπου η απόδοση μπορεί να είναι εξαιρετική όταν η αναζήτηση ολοκληρώνεται ομαλά, αλλά να καταρρέει όταν οι πόροι δεν επαρκούν.

6.1 Απόδοση Standard Prompting

Η μέθοδος Standard Prompting λειτουργεί ως σημείο αναφοράς (αγγλ. Baseline) για την αποτίμηση της αποτελεσματικότητας των πλέον εξελιγμένων τεχνικών συλλογισμού. Τα αριθμητικά αποτελέσματα που προέκυψαν από την αξιολόγηση στο σύνολο δεδομένων GSM8K αποκαλύπτουν όχι μόνο το επίπεδο απόδοσης της μεθόδου, αλλά και τους εγγενείς δομικούς περιορισμούς της όταν καλείται να διαχειριστεί προβλήματα πολυσταδιακού μαθηματικού συλλογισμού.

Ακρίβεια (Accuracy: 29% - 94%, Μ.Ο. ≈ 64,1%, Προσαρμοσμένος Μ.Ο. ≈ 67%)

Η ακρίβεια της Standard Prompting παρουσιάζει σημαντική διακύμανση, κυμαινόμενη από 29% έως 94%, με τον συνολικό μέσο όρο να διαμορφώνεται στο 64,1%. Ωστόσο, εξαιρώντας τις διαμορφώσεις που παρουσιάζουν ακραία τεχνική αστάθεια, ο προσαρμοσμένος μέσος όρος ανέρχεται στο 67%, αποτυπώνοντας με μεγαλύτερη ακρίβεια την πραγματική δυναμική της μεθόδου στις λειτουργικές διαμορφώσεις.

Στα μικρότερης κλίμακας ή παλαιότερης γενιάς μοντέλα, η απόδοση παραμένει χαμηλή (περίπου 29%–31%), γεγονός που επιβεβαιώνει ότι η απουσία ρητής καθοδήγησης περιορίζει τη δυνατότητα διαχείρισης προβλημάτων πολυσταδιακού μαθηματικού συλλογισμού, όπως εκείνα του GSM8K, τα οποία απαιτούν σαφή ακολουθία 2 έως 8 λογικών βημάτων. Σε αυτές τις περιπτώσεις, το μοντέλο τείνει να συμπυκνώνει τη λύση σε μία ενιαία παραγωγή, με αποτέλεσμα η ύπαρξη λανθασμένου ενδιάμεσου συλλογισμού να μην εντοπίζεται ούτε να διορθώνεται.

Ωστόσο, στα ισχυρότερα και νεότερης γενιάς μοντέλα (κυρίως στις οικογένειες DeepSeek και Gemini), η ακρίβεια υπερβαίνει το 90%, καταδεικνύοντας ότι η Standard Prompting δεν είναι εγγενώς ανεπαρκής. Αντιθέτως, η αποτελεσματικότητά της εξαρτάται σε μεγάλο βαθμό από την εσωτερική συλλογιστική ικανότητα του ίδιου του μοντέλου. Αξίζει να σημειωθεί ότι ορισμένα μοντέλα κορυφαίας κατηγορίας (π.χ. GPT-4o & GPT-5.1) κυμάνθηκαν σε μεσαία επίπεδα (~54%) υπό τη συνθήκη Standard, υπογραμμίζοντας ότι η υπολογιστική ισχύς από μόνη της δεν εγγυάται πάντα την άμεση επίλυση χωρίς βηματική καθοδήγηση.

Συνεπώς, η χαμηλή ακρίβεια που παρατηρείται σε ορισμένες περιπτώσεις δεν αποτελεί καθολικό χαρακτηριστικό της μεθόδου, αλλά συνάρτηση της αρχιτεκτονικής και της υπολογιστικής ισχύος του εκάστοτε μοντέλου. Η Standard Prompting λειτουργεί ως ουδέτερη βάση αξιολόγησης, της οποίας η απόδοση αντανακλά τις εγγενείς δυνατότητες του υποκείμενου LLM.

Ρυθμός αποχής (Abstention Rate: 0 - 67%, Μ.Ο. ≈ 7,4%, Προσαρμοσμένος Μ.Ο. ≈ 1,9%)

Ο ρυθμός αποχής στη Standard Prompting παρουσιάζει έντονη ανομοιογένεια, με το εύρος τιμών να εκτείνεται από 0,00 έως το εξαιρετικά υψηλό 0,67. Ο συνολικός μέσος όρος διαμορφώνεται στο 7,4%, ωστόσο η τιμή αυτή επηρεάζεται καθοριστικά από συγκεκριμένες διαμορφώσεις των μοντέλων Gemini (π.χ. Gemini 2.5 Pro με 0,67 και Gemini 3 Pro Preview με 0,49). Εξαιρώντας αυτές τις ακραίες τιμές τεχνικής αστάθειας, ο προσαρμοσμένος μέσος όρος υποχωρεί στο 1,9%, αποτυπώνοντας τη συνήθη συμπεριφορά της μεθόδου.

Στη συντριπτική πλειονότητα των υπόλοιπων μοντέλων, ο ρυθμός αποχής παραμένει μηδενικός ή αμελητέος (κάτω από 0,02), γεγονός που επιβεβαιώνει ότι η Standard Prompting είναι μια άμεση μέθοδος που σπάνια οδηγεί σε άρνηση απόκρισης. Η δραματική αύξηση της αποχής σε συγκεκριμένες εκδόσεις Gemini υποδηλώνει δομική ευπάθεια των εν λόγω διαμορφώσεων στη διαχείριση της συγκεκριμένης προτροπής, όπου η έλλειψη βηματικής καθοδήγησης φαίνεται να προκαλεί συστηματική αδυναμία παραγωγής έγκυρης εξόδου, αλλοιώνοντας τη στατιστική εικόνα της ακρίβειας.

Χρόνος απόκρισης (Latency: 450 ms - 13.600 ms, M.O. = 3.046 ms)

Ο χρόνος απόκρισης της Standard Prompting παρουσιάζει διακύμανση από περίπου 450 χιλιοστά του δευτερολέπτου έως και 13.600 χιλιοστά, ανάλογα με το υποκείμενο μοντέλο, με τη μέση τιμή να διαμορφώνεται στα 3.046 ms.

Συγκεκριμένα, τα ελαφρύτερα ή βελτιστοποιημένα μοντέλα επιτυγχάνουν χρόνους κάτω του ενός δευτερολέπτου, καθιστώντας τη μέθοδο ιδανική για εφαρμογές που ο χρόνος εκτέλεσης αποτελεί σημαντικό παράγοντα. Αντιθέτως, οι υψηλές τιμές αφορούν κυρίως μοντέλα που ενσωματώνουν εγγενείς μηχανισμούς συλλογισμού (reasoning models) ή διαθέτουν πολύ μεγάλο αριθμό παραμέτρων, όπου η καθυστέρηση οφείλεται στην πολυπλοκότητα της εσωτερικής επεξεργασίας και όχι στη δομή της προτροπής.

Η Standard Prompting, λόγω της απουσίας διαδικασιών πολυβηματικού συλλογισμού, δειγματοληψίας ή διακλαδώσεων, διατηρεί το χαμηλότερο υπολογιστικό φορτίο σε σύγκριση με τεχνικές όπως CoT ή ToT. Ωστόσο, το πλεονέκτημα αυτό είναι πρωτίστως τεχνικής φύσεως. Η μειωμένη υπολογιστική επιβάρυνση δεν συνεπάγεται κατ' ανάγκη ποιοτική ανωτερότητα, καθώς ο τελικός χρόνος απόκρισης καθορίζεται σε μεγάλο βαθμό από την αρχιτεκτονική και την ταχύτητα επεξεργασίας του ίδιου του μοντέλου, παρά από τη δομική απλότητα της προτροπής.

Συνεπώς, η Standard Prompting μπορεί να χαρακτηριστεί υπολογιστικά αποδοτική, χωρίς όμως ο χρόνος απόκρισης να αποτελεί από μόνος του δείκτη γνωστικής επάρκειας.

Κατανάλωση Tokens (140 - 750 tokens, M.O. ≈ 240 tokens)

Η συνολική κατανάλωση tokens στη Standard Prompting κυμαίνεται από 140 έως 750 tokens ανά κλήση, με τον μέσο όρο να διαμορφώνεται στα 240 tokens. Η σημαντική αυτή διακύμανση εξαρτάται άμεσα από το υποκείμενο μοντέλο: τα ελαφρύτερα μοντέλα (π.χ. GPT-3.5 και ορισμένες εκδόσεις Flash) κινούνται στο κάτω άκρο του εύρους (140–150 tokens), ενώ ισχυρότερα ή πιο αναλυτικά μοντέλα μπορεί να φθάνουν συνολικά τα 400–750 tokens.

Η διαφοροποίηση αυτή δεν οφείλεται στη μέθοδο prompting, αλλά στον τρόπο με τον οποίο κάθε μοντέλο δομεί και αναπτύσσει την απάντησή του. Η Standard Prompting δεν επιβάλλει ρητή αλυσίδα συλλογισμού ούτε μηχανισμό πολλαπλής εξερεύνησης, γεγονός που περιορίζει τη συστηματική αύξηση των tokens μέσω εξωτερικευμένων βημάτων σκέψης. Ωστόσο, σε ισχυρότερα μοντέλα παρατηρείται αυξημένη παραγωγή tokens ακόμη και χωρίς ρητή καθοδήγηση, υποδηλώνοντας ότι μέρος της συλλογιστικής διεργασίας ενσωματώνεται εσωτερικά στην απάντηση.

Συνεπώς, η κατανάλωση πόρων στη Standard Prompting δεν μπορεί να χαρακτηριστεί καθολικά «χαμηλή», αλλά μάλλον ελεγχόμενη και εξαρτώμενη από την αρχιτεκτονική του μοντέλου. Σε σύγκριση με τεχνικές όπως η CoT ή η ToT, η μέθοδος παραμένει γενικά πιο αποδοτική, χωρίς όμως η οικονομία tokens να συνεπάγεται περιορισμό γνωστικής ικανότητας, ιδίως στα πλέον ισχυρά LLMs.

Έλεγχος συνέπειας (Consistency: 96% - 100%, M.O. ≈ 98%)

Σημειώνεται ότι η συνέπεια υπολογίστηκε σε ένα υποσύνολο 100 ερωτήσεων, όπου κάθε ερώτηση υποβλήθηκε σε τρεις διαδοχικές εκτελέσεις, προκειμένου να ελεγχθεί η σταθερότητα των αποκρίσεων υπό ίδιες συνθήκες.

Η ποσοτική ανάλυση καταγράφει εξαιρετικά υψηλές τιμές, με το εύρος να κυμαίνεται σταθερά μεταξύ 96% και 100% και τον συνολικό μέσο όρο να αγγίζει το 98% για το σύνολο των εξεταζόμενων μοντέλων.

Σε αντίθεση με την έντονη μεταβλητότητα που παρατηρήθηκε στην ακρίβεια και τον χρόνο απόκρισης στο πλήρες σύνολο δεδομένων, η μετρική της συνέπειας στο ελεγχόμενο υποσύνολο εμφανίζει εντυπωσιακή ομοιογένεια. Το εύρημα αυτό επιβεβαιώνει ότι η μέθοδος Standard Prompting, ελλείψει μηχανισμών στοχαστικής διακλάδωσης, λειτουργεί με σχεδόν ντετερμινιστικό τρόπο, συγκλίνοντας στην ίδια ακριβώς έξοδο σε διαδοχικές προσπάθειες.

Στατιστικά, είναι αξιοσημείωτο ότι η υψηλή αυτή μέση τιμή διατηρείται αμετάβλητη ακόμη και σε μοντέλα με χαμηλά ποσοστά ακρίβειας. Χαρακτηριστικότερο παράδειγμα αποτελεί το μοντέλο GPT-4o-mini, το οποίο, παρότι σημειώνει τη χαμηλότερη ακρίβεια (~34%), διατηρεί μία από τις υψηλότερες τιμές συνέπειας (98%). Αυτό αποδεικνύει ότι, στο πλαίσιο της Standard Prompting, η επαναληψιμότητα της απάντησης είναι μέγεθος ανεξάρτητο από την ορθότητά της, καθώς το σύστημα αναπαράγει με την ίδια σταθερότητα και αυτοπεποίθηση τόσο τις επιτυχείς όσο και τις λανθασμένες προβλέψεις.

LLM-as-a-Judge Scores (2.57 - 8.97, M.O. ≈ 6.40)

Η ποιοτική αξιολόγηση μέσω της μεθόδου «LLM-as-a-Judge» για τη Standard Prompting καταγράφει συνολικό μέσο όρο 6,40 (σε κλίμακα 0-10), με το εύρος τιμών να εκτείνεται από 2,57 έως 8,97. Δεδομένου ότι στη συγκεκριμένη μέθοδο η απόκριση των μοντέλων περιορίζεται αποκλειστικά στην παροχή του τελικού αριθμητικού αποτελέσματος, η βαθμολογία του κριτή συνδέεται άρρηκτα με την ορθότητα της απάντησης.

Σε αυτό το πλαίσιο, οι υψηλές βαθμολογίες μοντέλων όπως το Gemini 2.5 Flash (8,97) και το DeepSeek Chat (8,90) αντικατοπτρίζουν την υψηλή τους ευστοχία. Αντιθέτως, οι χαμηλότερες βαθμολογίες σε μοντέλα όπως το GPT-4o (6,97) και το GPT-5.1 (6,74) οφείλονται αποκλειστικά στις συχνότερες λανθασμένες απαντήσεις, καθώς ένα εσφαλμένο αποτέλεσμα χωρίς συλλογιστική τεκμηρίωση οδηγεί σε χαμηλή βαθμολόγηση της λογικής και της σαφήνειας.

Τέλος, οι ιδιαίτερα χαμηλές τιμές σε ορισμένες εκδόσεις (π.χ. Gemini 3.0 Pro Preview στο 2,57) ενσωματώνονται στον τελικό υπολογισμό, αντανακλώντας τη χαμηλή ακρίβεια που προκύπτει από τα υψηλά ποσοστά αποχής (abstention). Στις περιπτώσεις αυτές, η συστηματική άρνηση παραγωγής απάντησης οδηγεί σε αισθητή μείωση της ποιότητας, επηρεάζοντας τη συνολική στατιστική εικόνα της μεθόδου.

Συνολική ερμηνεία των αποτελεσμάτων

Συνοψίζοντας τα ευρήματα της πειραματικής διαδικασίας, καθίσταται σαφές ότι η μέθοδος Standard Prompting δεν εμφανίζει ενιαία συμπεριφορά, αλλά η απόδοσή της διαφοροποιείται σημαντικά ανάλογα με το υποκείμενο μοντέλο. Σε μοντέλα περιορισμένης συλλογιστικής ισχύος ή σε πειραματικές εκδόσεις με υψηλά ποσοστά αποχής (abstention), η μέθοδος παρουσιάζει χαμηλή ακρίβεια, γεγονός που υποδηλώνει τεχνική αδυναμία ή εγγενή δυσκολία στη διαχείριση πολυσταδιακού μαθηματικού συλλογισμού χωρίς καθοδήγηση. Αντιθέτως, σε συγκεκριμένα μοντέλα νεότερης γενιάς, όπως το Gemini 2.5 Flash και το DeepSeek Chat, η Standard Prompting επιτυγχάνει ιδιαίτερα υψηλή ακρίβεια και ποιότητα απαντήσεων, υπερβαίνοντας ακόμη και ισχυρότερα μοντέλα.

Ένα από τα πιο αξιοσημείωτα ευρήματα είναι ότι η υψηλή συνέπεια που παρατηρείται στην πλειονότητα των διαμορφώσεων δεν συνεπάγεται κατ' ανάγκη και υψηλή ορθότητα. Η περίπτωση του GPT-4o-mini αναδεικνύει το παράδοξο μιας «σταθερά λανθασμένης» συμπεριφοράς, όπου το μοντέλο αποκρίνεται με μέγιστη σταθερότητα (98%) αλλά χαμηλή ακρίβεια (~34%). Η συνολική εικόνα καταδεικνύει ότι η Standard Prompting λειτουργεί κυρίως ως βάση αξιολόγησης, της οποίας η αποτελεσματικότητα δεν εξαρτάται από τη μέθοδο, αλλά αντανακλά την εγγενή συλλογιστική ικανότητα του εκάστοτε μοντέλου.

Συνεπώς, η μέθοδος δεν μπορεί να χαρακτηριστεί καθολικά ως ανεπαρκής για το απαιτητικό περιβάλλον του GSM8K. Η επίδοσή της είναι άμεσα συνυφασμένη με την αρχιτεκτονική του υποκείμενου LLM. Το γεγονός ότι ορισμένα μοντέλα αριστεύουν στη Standard συνθήκη, ενώ άλλα υστερούν σημαντικά, υπογραμμίζει ότι η ποιότητα της απάντησης δεν είναι προϊόν της προτροπής, αλλά της ικανότητας του μοντέλου να ανακαλεί και να εφαρμόζει εσωτερικά τις απαραίτητες λογικές διαδρομές.

6.2 Απόδοση Chain-of-Thought (CoT)

6.2.1 CoT Standard

Η μέθοδος CoT-Standard (K=1) εφαρμόστηκε στο σύνολο δεδομένων GSM8K με πρωταρχικό στόχο τη διερεύνηση της επίδρασης που ασκεί η ρητή, βηματική συλλογιστική στη δυνατότητα επίλυσης πολυσταδιακών μαθηματικών προβλημάτων. Σε πλήρη αντιδιαστολή με τη μέθοδο Standard Prompting, η προσέγγιση CoT ενθαρρύνει το μοντέλο να εξωτερικεύσει τη διαδικασία σκέψης του, παράγοντας μια συνεκτική ακολουθία ενδιάμεσων λογικών και αριθμητικών βημάτων πριν καταλήξει στην τελική απάντηση. Η διαδικασία αυτή μετατρέπει τη διαδικασία επίλυσης του προβλήματος σε μια παρατηρήσιμη ροή συλλογισμού.

Ακρίβεια (Accuracy: 0 - 98%, Μ.Ο. = 77%, Προσαρμοσμένος Μ.Ο. ≈ 87,9%)

Η ακρίβεια της μεθόδου CoT Standard (K=1) κυμάνθηκε στο εύρος 0,00 - 0,98, με συνολικό μέσο όρο 77%. Ωστόσο, εξαιρώντας τις διαμορφώσεις που παρουσιάζουν κατάρρευση απόδοσης (0,00 και 0,02), ο προσαρμοσμένος μέσος όρος διαμορφώνεται στο 87,9% αποτυπώνοντας με μεγαλύτερη ακρίβεια την πραγματική δυναμική της τεχνικής στις λειτουργικές διαμορφώσεις.

Σε ισχυρά μοντέλα, όπως τα GPT-5.1 (0,98), DeepSeek Chat (0,97) και Gemini 2.0 Flash (0,97), η επίδοση είναι ιδιαίτερα υψηλή, καταδεικνύοντας ότι η ρητή αλυσίδα σκέψης ενισχύει σημαντικά τη διαχείριση πολυσταδιακού συλλογισμού. Αντιθέτως, οι εξαιρετικά χαμηλές τιμές δεν αποτελούν χαρακτηριστικό της μεθόδου καθαυτής, αλλά αποτέλεσμα ασυμβατότητας συγκεκριμένων διαμορφώσεων με την τεχνική CoT.

Ρυθμός αποχής (Abstention Rate: 0% - 100%, Μ.Ο. ≈ 10,5%, Προσαρμοσμένος Μ.Ο. ≈ 4,5%)

Ο ρυθμός αποχής της μεθόδου CoT Standard (K=1) κυμαίνεται μεταξύ 0,00 και 1,00, με τον συνολικό μέσο όρο να διαμορφώνεται στο 10,5%. Στη συντριπτική πλειονότητα των διαμορφώσεων υψηλής απόδοσης, το ποσοστό αποχής παραμένει μηδενικό. Ωστόσο, παρατηρούνται συγκεκριμένες περιπτώσεις πλήρους αποχής (1,00), οι οποίες επιβαρύνουν σημαντικά τη στατιστική εικόνα του συνολικού μέσου όρου.

Εξαιρώντας την τιμή της πλήρους αποχής, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στο 4,53%, γεγονός που αποτυπώνει πιο ρεαλιστικά τη συνήθη συμπεριφορά της μεθόδου. Συνεπώς, η CoT Standard διατηρεί γενικά χαμηλό επίπεδο αποχής στις λειτουργικές διαμορφώσεις, ενώ οι ακραίες τιμές συνδέονται με την αστάθεια συγκεκριμένων αρχιτεκτονικών και συνοδεύονται από δραματική πτώση της ακρίβειας.

Χρόνος απόκρισης (Latency: 1.100 - 18.500 ms, Μ.Ο. ≈ 5.600 ms, Προσαρμοσμένος Μ.Ο. ≈ 4.740 ms)

Ο χρόνος απόκρισης της μεθόδου CoT Standard (K=1) κυμαίνεται μεταξύ 1.100 ms και 18.500 ms, με συνολικό μέσο όρο 5.600 ms. Εξαιρώντας τη μέγιστη τιμή για τον υπολογισμό του προσαρμοσμένου

μέσου όρου, ο μέσος χρόνος διαμορφώνεται στα 4.740 ms, αποτυπώνοντας πιο ρεαλιστικά τη συνήθη συμπεριφορά της μεθόδου.

Η αύξηση σε σχέση με τη Standard Prompting είναι εμφανής και αποτελεί άμεση συνέπεια της παραγωγής μεγαλύτερου αριθμού tokens, καθώς και της πρόσθετης υπολογιστικής επιβάρυνσης που συνεπάγεται η διατύπωση ενδιάμεσων βημάτων συλλογισμού. Ιδιαίτερα στα μοντέλα συλλογισμού (reasoning models), η χρονική επιβάρυνση είναι εντονότερη, αντανακλώντας την αυξημένη πολυπλοκότητα της συλλογιστικής διαδικασίας. Παρά τη βελτίωση της ακρίβειας σε αρκετές διαμορφώσεις, το κόστος σε χρόνο εκτέλεσης παραμένει σαφώς αυξημένο.

Συνεπώς, η CoT Standard ενισχύει συχνά την επίδοση εις βάρος της αποδοτικότητας, αναδεικνύοντας τον κλασικό συμβιβασμό μεταξύ ποιότητας συλλογισμού και υπολογιστικού κόστους.

Κατανάλωση Tokens (0 - 1.070 tokens, Μ.Ο. ≈ 434 tokens, Προσαρμοσμένος Μ.Ο. ≈ 463 tokens)

Η κατανάλωση tokens της μεθόδου CoT Standard (K=1) κυμαίνεται μεταξύ 0 tokens (ελάχιστη τιμή) και 1.070 tokens (μέγιστη τιμή), με συνολικό μέσο όρο 434 tokens ανά κλήση. Εξαιρώντας τη μηδενική τιμή που αντιστοιχεί σε αποτυχία παραγωγής λύσης, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στα 463 tokens, αποτυπώνοντας πιο ρεαλιστικά τη συνήθη συμπεριφορά της μεθόδου.

Η αύξηση σε σχέση με τη Standard Prompting είναι εμφανής και οφείλεται στην παραγωγή ρητών ενδιάμεσων βημάτων συλλογισμού. Σε ισχυρά μοντέλα υψηλής επίδοσης, η κατανάλωση κυμαίνεται συνήθως μεταξύ 205 και 680 tokens, ενώ σε ορισμένες περιπτώσεις (π.χ. Gemini 2.5 Pro) φθάνει τα 1.070 tokens. Αντίθετα, σε διαμορφώσεις όπου παρατηρείται αποτυχία παραγωγής λύσης (π.χ. Gemini 3.0 Pro Preview), η κατανάλωση μπορεί να είναι μηδενική.

Το αυξημένο υπολογιστικό κόστος αντανακλά την επένδυση πόρων για τη ρητή διατύπωση της συλλογιστικής διαδικασίας. Ωστόσο, η αύξηση των tokens δεν εγγυάται καθολικά υψηλότερη ακρίβεια, καθώς η αποτελεσματικότητα της CoT εξαρτάται από τη συμβατότητά της με το υποκείμενο μοντέλο.

Έλεγχος συνέπειας (Consistency: 96% - 99,7%, Μ.Ο. ≈ 98,29%)

Η αξιολόγηση της συνέπειας για τη μέθοδο CoT Standard πραγματοποιήθηκε με χρήση παραθύρου πλαισίου (context window) 2048 tokens, προκειμένου να διασφαλιστεί επαρκής χώρος για την ανάπτυξη της συλλογιστικής αλυσίδας. Η ανάλυση βασίστηκε σε τρεις διαδοχικές εκτελέσεις ανά ερώτηση στο υποσύνολο των 100 ερωτήσεων, καταγράφοντας εντυπωσιακά υψηλά επίπεδα σταθερότητας.

Το εύρος τιμών κυμαίνεται από 96% έως 99,7%, με τον συνολικό μέσο όρο να διαμορφώνεται στο 98,29%. Την κορυφαία επίδοση σημείωσε το Gemini 2.5 Flash με το σχεδόν απόλυτο 99,7%, επιδεικνύοντας ελάχιστη απόκλιση στις διαδοχικές εκτελέσεις, ενώ ακολούθησε το GPT-5.1 με 99%. Στον αντίποδα, το GPT-4o-mini κατέγραψε τη χαμηλότερη τιμή (96%), η οποία ωστόσο παραμένει σε εξαιρετικά υψηλά επίπεδα.

Είναι αξιοσημείωτο ότι η υψηλή συνέπεια συνοδεύεται και από υψηλή ακρίβεια σε όλα τα μοντέλα. Ακόμη και το μοντέλο με τη χαμηλότερη συνέπεια (GPT-4o-mini) επέτυχε ακρίβεια 92%, ενώ τα μοντέλα με τη μέγιστη σταθερότητα (DeepSeek Chat, Gemini 2.5 Flash, GPT-5.1) άγγιξαν το 98% σε ακρίβεια. Το εύρημα αυτό επιβεβαιώνει ότι η CoT Standard, υπό συνθήκες επαρκούς context, λειτουργεί ως μια εξαιρετικά στιβαρή μέθοδος, όπου η εξωτερίκευση της σκέψης δεν εισάγει θόρυβο, αλλά παγιώνει τη συλλογιστική πορεία προς το ορθό αποτέλεσμα.

LLM-as-a-Judge Scores (0,08 - 9,23, Μ.Ο. ≈ 7,24, Προσαρμοσμένος Μ.Ο. ≈ 7,72)

Οι βαθμολογίες της ποιοτικής αξιολόγησης LLM-as-a-Judge για την CoT Standard (K=1) κυμάνθηκαν μεταξύ 0,08 (ελάχιστη τιμή) και 9,23 (μέγιστη τιμή), με συνολικό μέσο όρο 7,24 (σε κλίμακα 0-10). Εξαιρώντας την τιμή που τείνει στο μηδέν (0,08) λόγω τεχνικής αστοχίας, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στο 7,72, αποτυπώνοντας πιο ρεαλιστικά την ποιότητα των λειτουργικών διαμορφώσεων.

Σε ισχυρές αρχιτεκτονικές υψηλής ακρίβειας, οι βαθμολογίες προσεγγίζουν τις ανώτατες τιμές της κλίμακας (άνω του 8,8), γεγονός που υποδηλώνει υψηλή λογική συνοχή, σαφή διάρθρωση της συλλογιστικής πορείας και επαληθεύσιμη αιτιολόγηση της λύσης. Η εξωτερικευμένη αλυσίδα σκέψης αξιολογείται θετικά από τον κριτή, καθώς καθιστά τη διαδικασία επίλυσης διαφανή και επιτρέπει την αναλυτική αποτίμηση των ενδιάμεσων βημάτων. Ωστόσο, σε διαμορφώσεις χαμηλής απόδοσης, οι βαθμολογίες καταρρέουν, αντανακλώντας αστοχίες στη λογική δομή ή πλήρη αποτυχία παραγωγής έγκυρης λύσης.

Αξιοσημείωτο είναι ότι, ακόμη και σε περιπτώσεις όπου το τελικό αριθμητικό αποτέλεσμα είναι λανθασμένο, η CoT λαμβάνει συχνά υψηλότερη ποιοτική αξιολόγηση σε σύγκριση με τη Standard Prompting, εφόσον η συλλογιστική πορεία εμφανίζει εσωτερική συνέπεια και σαφήνεια. Το εύρημα αυτό αναδεικνύει ότι η CoT δεν βελτιώνει μόνο την ακρίβεια, αλλά και τη δομική ποιότητα της παραγόμενης γνώσης όταν η τεχνική εφαρμόζεται σε συμβατά μοντέλα.

Συνολική ερμηνεία των αποτελεσμάτων CoT Standard

Συνοψίζοντας, η μέθοδος CoT Standard (K=1) παρουσιάζει σημαντική ενίσχυση της συλλογιστικής ικανότητας σε σχέση με τη Standard Prompting. Η ακρίβεια κυμαίνεται στο εύρος 0,00 - 0,98, με συνολικό μέσο όρο 77%. Εξαιρώντας τις διαμορφώσεις που εμφανίζουν κατάρρευση απόδοσης, ο προσαρμοσμένος μέσος όρος ανέρχεται σε 87,9%, γεγονός που καταδεικνύει ότι στις λειτουργικές περιπτώσεις η ρητή αλυσίδα σκέψης συμβάλλει ουσιαστικά στη βελτίωση της επίδοσης.

Ωστόσο, η βελτίωση αυτή δεν είναι καθολική. Σε ορισμένες διαμορφώσεις παρατηρείται πλήρης ή σχεδόν πλήρης αποτυχία, γεγονός που υποδηλώνει ότι η αποτελεσματικότητα της CoT εξαρτάται απόλυτα από τη συμβατότητα της τεχνικής με την αρχιτεκτονική του εκάστοτε μοντέλου.

Η μέθοδος συνοδεύεται από αυξημένο υπολογιστικό κόστος, με μέσο χρόνο απόκρισης 5.600 ms (Προσαρμοσμένος Μ.Ο. 4.740 ms) και μέση κατανάλωση 434 tokens (Προσαρμοσμένος Μ.Ο. 463 tokens). Το εύρημα αυτό επιβεβαιώνει τον κλασικό συμβιβασμό μεταξύ ποιότητας συλλογισμού και αποδοτικότητας (trade-off).

Όσον αφορά τη σταθερότητα, ο έλεγχος συνέπειας κατέδειξε εξαιρετικά υψηλή επίδοση (98,29%), διαψεύδοντας την ανησυχία ότι η αυξημένη παραγωγή κειμένου οδηγεί απαραίτητα σε στοχαστική αστάθεια. Αντιθέτως, αποδείχθηκε ότι η εξωτερίκευση της σκέψης σε ισχυρά μοντέλα παγιώνει τη συλλογιστική πορεία, προσφέροντας υψηλή επαναληψιμότητα.

Συνεπώς, η CoT Standard λειτουργεί ως ισχυρός μηχανισμός ενίσχυσης της συλλογιστικής διαδικασίας σε συμβατά μοντέλα, χωρίς όμως να εγγυάται καθολική βελτίωση. Αποτελεί ένα κρίσιμο ενδιάμεσο στάδιο μετάβασης από την απλή απάντηση προς πιο εξελιγμένες τεχνικές πολυσταδιακού συλλογισμού.

6.2.2 Self-Consistent CoT

Η μέθοδος SC-CoT (αγγλ. Self-Consistent Chain-of-Thought) συνιστά τη μεθοδολογική επέκταση της βασικής προσέγγισης CoT, στοχεύοντας στην περαιτέρω θωράκιση της αξιοπιστίας και της σταθερότητας των παραγόμενων απαντήσεων. Η θεμελιώδης διαφορά της από την CoT Standard κατά

την οποία παράγεται μία και μοναδική αλυσίδα συλλογισμού, έγκειται στη δημιουργία πολλαπλών ανεξάρτητων αλυσίδων σκέψης για το ίδιο πρόβλημα, με την τελική απάντηση να προκύπτει μέσω της διαδικασίας πλειοψηφικής συμφωνίας (αγγλ. majority voting).

Ακρίβεια (Accuracy: 50% - 97%, Μ.Ο. ≈ 84,6%, Προσαρμοσμένος Μ.Ο. ≈ 90,25%)

Η ακρίβεια της μεθόδου Self-Consistent CoT (K=3) κυμάνθηκε στο εύρος 50% - 97%, με τον συνολικό μέσο όρο να διαμορφώνεται στο 84,6%. Εξαιρώντας τις χαμηλές ακραίες τιμές που οφείλονται σε λειτουργικές αστοχίες συγκεκριμένων διαμορφώσεων για τον υπολογισμό του προσαρμοσμένου μέσου όρου, η μέση ακρίβεια ανέρχεται στο 90,25%, γεγονός που αποτυπώνει τη σταθερά υψηλή απόδοση των περισσότερων διαμορφώσεων.

Σε ισχυρά μοντέλα, όπως τα GPT-5.1 (97%), Gemini 2.0 Flash (97%) και GPT-4o (96%), η απόδοση παραμένει ιδιαίτερα υψηλή, επιβεβαιώνοντας ότι η πλειοψηφική επιλογή μεταξύ πολλαπλών αλυσίδων συλλογισμού λειτουργεί αποτελεσματικά ως μηχανισμός σταθεροποίησης της τελικής πρόβλεψης. Η βελτίωση αυτή δεν απορρέει από αλλαγή στη μορφή της προτροπής, αλλά από τη στατιστική αξιοποίηση πολλαπλών συλλογιστικών διαδρομών, η οποία περιορίζει την επίδραση μεμονωμένων λανθασμένων αλυσίδων και ενισχύει τις λύσεις που επαναλαμβάνονται συχνότερα.

Ωστόσο, η αποτελεσματικότητα της SC-CoT δεν είναι καθολική. Σε ορισμένες περιπτώσεις (π.χ. Gemini 2.5 Flash με 50%), παρατηρείται σημαντική πτώση απόδοσης, γεγονός που υποδηλώνει ότι η πολλαπλή δειγματοληψία δεν εγγυάται αυτόματα βελτίωση αν οι επιμέρους αλυσίδες είναι συστηματικά λανθασμένες.

Ρυθμός αποχής (Abstention Rate: 0% - 17%, Προσαρμοσμένος Μ.Ο. ≈ 3,25%)

Ο ρυθμός αποχής της Self-Consistent CoT (K=3) κυμάνθηκε μεταξύ 0,00 και 0,17. Εξαιρώντας τις ακραίες τιμές για τον υπολογισμό του προσαρμοσμένου μέσου όρου, ο μέσος ρυθμός αποχής διαμορφώνεται στο 3,25%.

Η συνολική εικόνα υποδηλώνει ότι η SC-CoT διατηρεί χαμηλή τάση αποχής στις περισσότερες διαμορφώσεις. Οι αυξημένες τιμές εμφανίζονται σε περιπτώσεις όπου οι παραγόμενες αλυσίδες συλλογισμού αποκλίνουν σημαντικά ή δεν σχηματίζεται σαφής πλειοψηφία, οπότε η αποχή λειτουργεί ως μηχανισμός αποφυγής αβέβαιων προβλέψεων. Συνεπώς, η SC-CoT δεν παρουσιάζει συστηματική αύξηση αποχής, αλλά περιορισμένες και μοντελοεξαρτώμενες αποκλίσεις.

Χρόνος απόκρισης (Latency: 4.000 - 56.000 ms, Μ.Ο. ≈ 17.417 ms, Προσαρμοσμένος Μ.Ο. ≈ 9.700 ms)

Ο χρόνος απόκρισης της Self-Consistent CoT (K=3) κυμαίνεται μεταξύ 4.000 ms και 56.000 ms. Η ελάχιστη τιμή καταγράφεται σε ελαφρύτερα μοντέλα, ενώ η μέγιστη τιμή εμφανίζεται σε reasoning διαμορφώσεις υψηλής υπολογιστικής επιβάρυνσης.

Ο συνολικός μέσος χρόνος απόκρισης διαμορφώνεται στα 17.417 ms. Ωστόσο, εξαιρώντας την ακραία τιμή των 56.000 ms, ο προσαρμοσμένος μέσος όρος υποχωρεί στα 9.700 ms, αποτυπώνοντας πιο ρεαλιστικά τη συνήθη συμπεριφορά της μεθόδου.

Η σημαντική αυτή αύξηση σε σχέση με την CoT (K=1) και τη Standard Prompting αποτελεί άμεση συνέπεια της δημιουργίας πολλαπλών ανεξάρτητων αλυσίδων συλλογισμού και της μεταγενέστερης διαδικασίας σύγκρισης. Η SC-CoT επενδύει πολλαπλάσιους υπολογιστικούς πόρους προκειμένου να μειώσει τον στοχαστικό θόρυβο. Ως εκ τούτου, η μέθοδος κρίνεται σαφώς λιγότερο κατάλληλη για

εφαρμογές που απαιτούν άμεση απόκριση (low-latency), καθώς η βελτίωση της σταθερότητας επιτυγχάνεται εις βάρος της ταχύτητας.

Κατανάλωση Tokens (630 - 1.620 tokens, Μ.Ο. ≈ 1.068 tokens, Προσαρμοσμένος Μ.Ο. ≈ 958 tokens)

Η κατανάλωση tokens της Self-Consistent CoT (K=3) κυμαίνεται μεταξύ 630 και 1.620 tokens ανά κλήση. Ο συνολικός μέσος όρος κατανάλωσης διαμορφώνεται στα 1.068 tokens, ενώ ο προσαρμοσμένος μέσος όρος (χωρίς τη μέγιστη τιμή) ανέρχεται στα 958 tokens, αποτυπώνοντας πιο ρεαλιστικά τη συνήθη συμπεριφορά της μεθόδου.

Η αυξημένη κατανάλωση αντανακλά την παραγωγή πολλαπλών αλυσίδων σκέψης. Το κόστος αυτό αποτελεί αναμενόμενη συνέπεια της στατιστικής ενίσχυσης της αξιοπιστίας, επιβεβαιώνοντας τον σαφή συμβιβασμό (trade-off) μεταξύ υπολογιστικής επιβάρυνσης και σταθερότητας της τελικής πρόβλεψης.

LLM-as-a-Judge Scores (4,22 – 9,23, Μ.Ο. ≈ 8,40, Προσαρμοσμένος Μ.Ο. ≈ 9,00)

Οι ποιοτικές αξιολογήσεις LLM-as-a-Judge για τη SC-CoT (K=3) κυμάνθηκαν μεταξύ 4,22 και 9,23 (σε κλίμακα 0–10), με συνολικό μέσο όρο περίπου 8,40. Εξαιρώντας τις χαμηλές ακραίες τιμές, ο προσαρμοσμένος μέσος όρος ανέρχεται περίπου στο 9,00.

Το αποτέλεσμα αυτό υποδηλώνει ότι, στις λειτουργικές διαμορφώσεις, η SC-CoT παράγει απαντήσεις εξαιρετικά υψηλής λογικής συνοχής και σαφούς τεκμηρίωσης. Η χρήση πολλαπλών συλλογιστικών διαδρομών και η επιλογή μέσω πλειοψηφίας φαίνεται να ενισχύουν σημαντικά τη δομική ποιότητα της παραγόμενης γνώσης, ελαχιστοποιώντας τις αστοχίες.

Συνολική ερμηνεία των αποτελεσμάτων SC-CoT

Η μέθοδος SC-CoT επιτυγχάνει υψηλότερη ακρίβεια σε σχέση με την απλή CoT, συνοδευόμενη από εξαιρετική συνέπεια απαντήσεων και βέλτιστη βαθμονόμηση. Ωστόσο, τα πλεονεκτήματα αυτά συνεπάγονται σημαντική επιβάρυνση στον χρόνο απόκρισης και το υπολογιστικό κόστος, ενώ απαιτείται προσεκτική παραμετροποίηση του πλήθους των δειγμάτων (k) για τη βελτιστοποίηση της απόδοσης.

Συνοψίζοντας, η SC-CoT αποτελεί ένα ισχυρό βήμα προς τον αξιόπιστο συλλογισμό, με σαφές ωστόσο ισοζύγιο κόστους-οφέλους, προετοιμάζοντας το έδαφος για τις πλέον δομημένες προσεγγίσεις όπως η Tree-of-Thought (ToT).

6.3 Απόδοση Tree-of-Thought (ToT)

6.3.1 Tree-of-Thought (ToT) – tot_b3

Η μέθοδος ToT με ρύθμιση tot_b3 (branching factor b=3) αξιολογήθηκε στο σύνολο δεδομένων GSM8K ως μια εξερευνητική στρατηγική επίλυσης, η οποία διαφοροποιείται δομικά από τη γραμμική φύση της CoT. Αντί να παράγει μία μοναδική αλυσίδα συλλογισμού, το μοντέλο δημιουργεί πολλαπλές υποψήφιες πορείες σκέψης ανά βήμα, τις αξιολογεί και επιλέγει δυναμικά την καλύτερη προς επέκταση. Στόχος αυτής της διαδικασίας είναι η ελαχιστοποίηση των σφαλμάτων που προκύπτουν από την υιοθέτηση ενός λανθασμένου συλλογιστικού μονοπατιού, με τίμημα ωστόσο το σημαντικά αυξημένο υπολογιστικό κόστος.

Ακρίβεια (Accuracy: 0% - 97%, Μ.Ο. ≈ 72,5%, Προσαρμοσμένος Μ.Ο. ≈ 86,8%, Διάμεσος = 90%, Q3 = 96%)

Η ακρίβεια της μεθόδου tot_b3 κυμαίνεται μεταξύ 0% και 97%, με τον συνολικό μέσο όρο να διαμορφώνεται στο 72,5%. Ωστόσο, εξαιρώντας τις διαμορφώσεις που παρουσιάζουν κατάρρευση απόδοσης (τιμές < 2% λόγω τεχνικών αστοχιών), ο προσαρμοσμένος μέσος όρος ανέρχεται στο 86,8%, αποτυπώνοντας την υψηλή δυναμική της μεθόδου στις λειτουργικές περιπτώσεις.

Η διάμεσος τιμή (90%) και το άνω τεταρτημόριο (96%) υποδεικνύουν ότι τουλάχιστον το 50% των διαμορφώσεων επιτυγχάνει ακρίβεια ≥ 90%, ενώ το ανώτερο 25% υπερβαίνει το 96%. Το εύρημα αυτό καταδεικνύει ότι η ToT μπορεί να προσφέρει ιδιαίτερα υψηλή επίδοση σε συμβατές και ισχυρές αρχιτεκτονικές, ξεπερνώντας συχνά τις γραμμικές μεθόδους.

Ρυθμός αποχής (Abstention Rate: 0% - 100%, Μ.Ο. ≈ 8,4%, Προσαρμοσμένος Μ.Ο. ≈ 3,0%, Διάμεσος = 0%, Q3 = 1%)

Ο ρυθμός αποχής της μεθόδου tot_b3 κυμαίνεται μεταξύ 0,00 και 1,00, με συνολικό μέσο όρο 8,4%. Εξαιρώντας την τιμή της πλήρους αποχής (1,00) που σημειώθηκε λόγω ασυμβατότητας, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στο 3,0%.

Η διάμεσος (0%) και το άνω τεταρτημόριο (1%) καταδεικνύουν ότι στο 75% των εκτελέσεων η αποχή είναι πρακτικά μηδενική. Ωστόσο, η ύπαρξη πλήρους αποχής σε μεμονωμένη διαμόρφωση υποδηλώνει ότι η μέθοδος, αν και σταθερή στη συνήθη λειτουργία της, παρουσιάζει ευθραυστότητα σε περιπτώσεις υπέρβασης πόρων ή τεχνικής αστοχίας, οδηγώντας σε δραματική πτώση της ακρίβειας.

Χρόνος απόκρισης (Latency: 3.500 - 62.000 ms, Μ.Ο. ≈ 14.183 ms, Προσαρμοσμένος Μ.Ο. ≈ 11.370 ms, Διάμεσος = 6.300 ms, Q3 = 14.500 ms)

Ο χρόνος απόκρισης της μεθόδου tot_b3 κυμαίνεται μεταξύ 3.500 ms και 62.000 ms. Εξαιρώντας την ακραία μέγιστη τιμή, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στα 11.370 ms, αποτυπώνοντας πιο ρεαλιστικά τη συνήθη χρονική συμπεριφορά.

Η διάμεσος (6.300 ms) υποδηλώνει ότι στο 50% των εκτελέσεων ο χρόνος απόκρισης παραμένει κάτω από ~6,3 δευτερόλεπτα, ενώ το άνω τεταρτημόριο (14.500 ms) δείχνει ότι το 25% των περιπτώσεων απαιτεί σημαντικά περισσότερο χρόνο. Η μεγάλη αυτή διακύμανση καταδεικνύει την αυξημένη ευαισθησία της ToT στην πολυπλοκότητα της δενδρικής αναζήτησης, καθιστώντας τη λιγότερο κατάλληλη για εφαρμογές πραγματικού χρόνου.

Κατανάλωση Tokens (0 - 3.300 tokens, Μ.Ο. ≈ 1.161 tokens, Προσαρμοσμένος Μ.Ο. ≈ 1.099 tokens, Διάμεσος = 770 tokens, Q3 = 1.780 tokens)

Η κατανάλωση tokens για την tot_b3 κυμαίνεται από 0 έως 3.300 tokens ανά κλήση, με συνολικό μέσο όρο 1.161 tokens. Εξαιρώντας τις ακραίες τιμές (0 και 3.300), ο προσαρμοσμένος μέσος όρος διαμορφώνεται στα 1.099 tokens.

Η τιμή αυτή είναι υπερδιπλάσια της απλής CoT, αντικατοπτρίζοντας το κόστος της εξερεύνησης πολλαπλών υποθέσεων (branches). Η διάμεσος τιμή (770 tokens) βρίσκεται κοντά στα επίπεδα των απαιτητικών CoT, ωστόσο το άνω τεταρτημόριο (1.780 tokens) αποκαλύπτει ότι σε πολλές περιπτώσεις το κόστος υπερδιπλασιάζεται, επιβεβαιώνοντας το υψηλό υπολογιστικό τίμημα της εξερευνητικής στρατηγικής.

Έλεγχος συνέπειας (Consistency: 96,3% - 100%, Μ.Ο. ≈ 98,6%)

Η αξιολόγηση της συνέπειας για τη μέθοδο ToT κατέγραψε εξαιρετικά υψηλή σταθερότητα, με τις τιμές να κυμαίνονται από 96,3% έως 100% και τον συνολικό μέσο όρο να αγγίζει το 98,6%. Μοντέλα όπως το GPT-5.1 και το DeepSeek Chat σημείωσαν την απόλυτη τιμή (100%), αποδεικνύοντας ότι η

δομημένη διαδικασία της δενδρικής αναζήτησης, παρά την πολυπλοκότητά της, οδηγεί σε συγκλίνουσες και επαναλήψιμες λύσεις.

Ιδιαίτερη σημασία έχει η αντιπαραβολή της συνέπειας με την ακρίβεια στο συγκεκριμένο υποσύνολο ελέγχου. Παρατηρείται ότι η υψηλή συνέπεια συνοδεύεται από εξίσου υψηλή ακρίβεια, η οποία κυμαίνεται στο εύρος 92,6% - 99% (με μέσο όρο 95,86%). Το γεγονός αυτό επιβεβαιώνει ότι η σταθερότητα της ToT δεν είναι πλασματική, αλλά αποτέλεσμα μιας στιβαρής συλλογιστικής πορείας που οδηγεί συστηματικά στο ορθό αποτέλεσμα. Σε αντίθεση με μεθόδους που μπορεί να είναι "σταθερά λανθασμένες", η ToT αποδεικνύεται "σταθερά ορθή" στις λειτουργικές της εκτελέσεις, ενισχύοντας την εμπιστοσύνη στην αξιοπιστία της τελικής απόφασης.

LLM-as-a-Judge Scores (0,02 – 9,19, Μ.Ο. ≈ 8,15, Προσαρμοσμένος Μ.Ο. ≈ 8,94)

Οι ποιοτικές αξιολογήσεις μέσω του μηχανισμού LLM-as-a-Judge για τη μέθοδο tot_b3 κυμαίνονται μεταξύ 0,02 και 9,19 (σε κλίμακα 0-10), με συνολικό μέσο όρο 8,15. Εξαιρώντας τις ελάχιστες τιμές (κάτω του 0,70) που οφείλεται σε περιορισμένη ανταπόκριση του μοντέλου, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στο 8,9.

Το αποτέλεσμα αυτό υποδηλώνει ότι, στις λειτουργικές διαμορφώσεις, η ToT παράγει απαντήσεις κορυφαίας λογικής συνοχής και αναλυτικής τεκμηρίωσης. Οι ιδιαίτερα χαμηλές τιμές περιορίζονται σε περιπτώσεις λειτουργικής αστοχίας ή ολικής αποχής, όπου η διαδικασία δενδρικής αναζήτησης δεν ολοκληρώνεται επιτυχώς. Συνεπώς, η συστηματική εξερεύνηση ενδιάμεσων κόμβων οδηγεί σε λύσεις που προσεγγίζουν το άριστα, εφόσον η εκτέλεση ολοκληρωθεί επιτυχώς.

Συνολικό συμπέρασμα για ToT (tot_b3)

Συνοψίζοντας, η μέθοδος ToT (tot_b3) αναδεικνύεται ως μία από τις ισχυρότερες στρατηγικές συλλογισμού για το GSM8K, καθώς στις λειτουργικές διαμορφώσεις επιτυγχάνει πολύ υψηλή επίδοση (Ακρίβεια με διάμεσο 90% και Q3 = 96%, με προσαρμοσμένο Μ.Ο. = 86,8%). Η εικόνα αυτή δείχνει ότι, όταν η δενδρική αναζήτηση ολοκληρώνεται ομαλά, η ToT μπορεί να ξεπεράσει τη Standard Prompting και να προσεγγίσει την κορυφαία απόδοση των πιο επιτυχημένων CoT/SC-CoT ρυθμίσεων, προσφέροντας ταυτόχρονα εξαιρετική ποιοτική τεκμηρίωση (LLM-as-a-Judge Μ.Ο. = 8,15 και Προσαρμοσμένος Μ.Ο. = 8,94).

Ωστόσο, η ToT δεν εμφανίζει πλήρως σταθερή συμπεριφορά. Οι πολύ χαμηλές τιμές ακρίβειας (0.00–0.02) και οι περιπτώσεις πλήρους αποχής (Abstention έως 1.00) υποδηλώνουν ότι η μέθοδος είναι πιο «εύθραυστη» από τη Standard Prompting και τη βασική CoT, καθώς μπορεί να περάσει από κανονική λειτουργία υψηλής απόδοσης σε κατάσταση αστοχίας όταν οι πόροι/ρυθμίσεις δεν επαρκούν. Παράλληλα, το κόστος είναι σημαντικά αυξημένο: ο χρόνος απόκρισης παρουσιάζει μεγάλο εύρος (3.500 - 62.000 ms, με διάμεσο 6.300 ms και Q3 = 14.500 ms), γεγονός που καθιστά τη μέθοδο λιγότερο κατάλληλη για εφαρμογές πραγματικού χρόνου, ενώ επίσης ιδιαίτερα αυξημένη παρουσιάζεται και η κατανάλωση σε tokens (Προσαρμοσμένος Μ.Ο. ≈ 1.099 tokens).

Συνεπώς, σε σχέση με τη Standard Prompting (Baseline), η ToT (tot_b3) προσφέρει σαφώς ανώτερη ποιότητα και υψηλότερη ακρίβεια στις περισσότερες επιτυχημένες εκτελέσεις, αλλά με αντάλλαγμα αισθητά μεγαλύτερο υπολογιστικό κόστος και αυξημένη ευαισθησία σε ασυμβατότητες/περιορισμούς πόρων. Είναι ιδανική όταν προτεραιότητα είναι η μέγιστη ορθότητα και υπάρχει επαρκές χρονικό/υπολογιστικό περιθώριο, ενώ απαιτεί προσεκτική παραμετροποίηση για να αποφευχθούν σπάνιες αλλά κρίσιμες καταρρεύσεις απόδοσης.

6.3.2 Tree-of-Thought (ToT) – tot_b5

Η ρύθμιση ToT με παράγοντα διακλάδωσης b=5 (tot_b5) συνιστά την πλέον εκτεταμένη εκδοχή της μεθοδολογίας που εξετάστηκε, επεκτείνοντας τη βασική ιδέα της δενδρικής αναζήτησης μέσω της αύξησης του αριθμού των υποψήφιων συλλογιστικών διαδρομών σε κάθε βήμα. Σε σύγκριση με την εκδοχή tot_b3, η προσέγγιση αυτή επιδιώκει μια βαθύτερη και ευρύτερη εξερεύνηση του χώρου λύσεων, με στόχο τη μεγιστοποίηση της πιθανότητας εντοπισμού της βέλτιστης πορείας συλλογισμού. Ωστόσο, η αύξηση του εύρους αναζήτησης συνεπάγεται εκθετική αύξηση του υπολογιστικού κόστους, γεγονός που αποτυπώνεται ξεκάθαρα στα πειραματικά αποτελέσματα.

Ακρίβεια (Accuracy: 22% – 97%, Μ.Ο. ≈ 76,9%, Προσαρμοσμένος Μ.Ο. ≈ 86,0%, Διάμεσος = 89%, Q3 = 97%)

Η ανάλυση της ακρίβειας για τη μέθοδο tot (b=5) καταδεικνύει έντονη διαφοροποίηση μεταξύ των μοντέλων, με τιμές που κυμαίνονται από 22% έως 97%. Ο συνολικός μέσος όρος (76,9%) επηρεάζεται αισθητά από τη χαμηλότερη επίδοση του GPT-3.5 Turbo, ενώ ο προσαρμοσμένος μέσος όρος (86%), που προκύπτει από την εξαίρεση των ακραίων τιμών, αποτυπώνει πιο ρεαλιστικά την τυπική συμπεριφορά της μεθόδου. Η διάμεσος τιμή (89%) και το άνω τεταρτημόριο (97%) επιβεβαιώνουν ότι το μεγαλύτερο μέρος των διαμορφώσεων κινείται σε κορυφαία επίπεδα ακρίβειας, υποδηλώνοντας ότι η αύξηση του παράγοντα διακλάδωσης μπορεί να ενισχύσει την επίδοση σε ισχυρές αρχιτεκτονικές, αν και δεν εξαλείφει τις περιπτώσεις αστοχίας σε μοντέλα με περιορισμένες ικανότητες συλλογισμού.

Ρυθμός αποχής (Abstention Rate: 0% – 16%, Μ.Ο. ≈ 4,0%, Προσαρμοσμένος Μ.Ο. ≈ 2,0%, Διάμεσος = 0%, Q3 = 6%)

Ο ρυθμός αποχής της μεθόδου tot_b5 κυμαίνεται μεταξύ 0 και 16%, με συνολικό μέσο όρο 4%. Εξαιρώντας την ακραία τιμή, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στο 2%, αποτυπώνοντας πιο ρεαλιστικά τη συνήθη συμπεριφορά της μεθόδου. Η διάμεσος δείχνει ότι στο 50% των διαμορφώσεων δεν παρατηρείται αποχή, ενώ το άνω τεταρτημόριο Q3 (6%) υποδηλώνει ότι στο 75% των περιπτώσεων η αποχή παραμένει εξαιρετικά χαμηλή (≤6%).

Σε σύγκριση με την tot_b3, η αύξηση του παράγοντα διακλάδωσης δεν οδηγεί σε δραματική αύξηση της αποχής, αλλά ενδέχεται να ενισχύσει την πιθανότητα αστάθειας σε πιο απαιτητικές διαμορφώσεις, καθώς η διαδικασία σύγκλισης καθίσταται δυσκολότερη.

Χρόνος απόκρισης (Latency: 6.800 - 96.000 ms, Μ.Ο. ≈ 24.557 ms, Προσαρμοσμένος Μ.Ο. ≈ 12.650 ms, Διάμεσος = 11.800 ms, Q3 = 18.750 ms)

Ο χρόνος απόκρισης της μεθόδου tot_b5 παρουσιάζει τεράστιο εύρος, από 6.800 ms έως 96.000 ms, με συνολικό μέσο όρο 24.557 ms. Εξαιρώντας την ακραία μέγιστη τιμή, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στα 12.650 ms. Η διάμεσος (11.800 ms) δείχνει ότι στο 50% των εκτελέσεων ο χρόνος υπερβαίνει τα 11,8 δευτερόλεπτα, ενώ το άνω τεταρτημόριο (18.750 ms) υποδηλώνει ότι το 25% των περιπτώσεων ξεπερνά τα 18,75 δευτερόλεπτα. Η έντονη αυτή χρονική επιβάρυνση οφείλεται στην εκθετική αύξηση των κόμβων που απαιτούν αξιολόγηση, καθιστώντας τη διαμόρφωση tot_b5 ακατάλληλη για εφαρμογές πραγματικού χρόνου.

Κατανάλωση Tokens (340 - 2.900 tokens, Μ.Ο. ≈ 1.563 tokens, Προσαρμοσμένος Μ.Ο. ≈ 1.340 tokens, Διάμεσος = 1.250 tokens, Q3 = 2.000 tokens)

Η κατανάλωση tokens της μεθόδου tot_b5 κυμαίνεται μεταξύ 340 tokens (ελάχιστη τιμή) και 2.900 tokens (μέγιστη τιμή), με συνολικό μέσο όρο 1.563 tokens ανά κλήση. Εξαιρώντας την ακραία μέγιστη τιμή (2900 tokens), ο προσαρμοσμένος μέσος όρος διαμορφώνεται στα 1.340 tokens, αποτυπώνοντας

πιο ρεαλιστικά τη συνήθη κατανάλωση πόρων. Η διάμεσος (1250 tokens) υποδηλώνει ότι στο 50% των εκτελέσεων η κατανάλωση υπερβαίνει τα 1250 tokens, ενώ το άνω τεταρτημόριο Q3 δείχνει ότι το 25% των περιπτώσεων απαιτεί περισσότερα από 2000 tokens.

Η κατανάλωση αυτή είναι πολλαπλάσια σε σχέση με τη Standard Prompting και αισθητά αυξημένη σε σχέση με την CoT (K=1), γεγονός που αντανακλά τον πολλαπλασιασμό των παραγόμενων κλαδιών και τη συνεχή αξιολόγηση ενδιάμεσων κόμβων. Η αύξηση του παράγοντα διακλάδωσης σε b=5 συνεπάγεται σημαντική διόγκωση του χώρου αναζήτησης και, κατ' επέκταση, του αριθμού παραγόμενων tokens. Συνεπώς, η ρύθμιση tot_b5 αποτελεί ιδιαίτερα δαπανηρή επιλογή από πλευράς υπολογιστικών πόρων.

LLM-as-a-Judge Scores (4,38 – 9,19, Μ.Ο. ≈ 8,15, Προσαρμοσμένος Μ.Ο. ≈ 8,69)

Οι ποιοτικές αξιολογήσεις για τη μέθοδο tot (b=5) κυμαίνονται μεταξύ 4,38 και 9,19, με μέσο όρο 8,15. Εξαιρώντας την ελάχιστη τιμή (4,38) που προέκυψε από τη λειτουργική αστοχία του GPT-3.5 Turbo, ο προσαρμοσμένος μέσος όρος διαμορφώνεται στο 8,69. Το αποτέλεσμα επιβεβαιώνει ότι, στις λειτουργικές της εκτελέσεις, η tot_b5 παράγει απαντήσεις εξαιρετικής λογικής συνοχής και δομικής πληρότητας.

Η εκτεταμένη εξερεύνηση πολλαπλών κλάδων συλλογισμού και η διαδοχική αξιολόγηση ενδιάμεσων κόμβων φαίνεται να ενισχύουν περαιτέρω τη δομική ποιότητα της παραγόμενης γνώσης σε σχέση με το tot_b3, οδηγώντας σε απαντήσεις που προσεγγίζουν ποιοτικά το άριστα όταν η διαδικασία ολοκληρώνεται ομαλά.

Συνολικό συμπέρασμα για ToT (tot_b5)

Συμπερασματικά, η μέθοδος ToT με παράγοντα διακλάδωσης b=5 αναδεικνύεται ως η πλέον ισχυρή αλλά και η πλέον απαιτητική υπολογιστικά προσέγγιση μεταξύ όλων των εξεταζόμενων τεχνικών. Η ακρίβεια φθάνει σε κορυφαία επίπεδα (+97%), ενώ οι διάμεσες επιδόσεις και οι ποιοτικές βαθμολογίες LLM-as-a-Judge επιβεβαιώνουν την υψηλή αξιοπιστία του συλλογισμού. Ωστόσο, η εντατική εξερεύνηση συνεπάγεται δραματική αύξηση του χρόνου απόκρισης και της κατανάλωσης tokens, με έντονη διακύμανση και περιπτώσεις υπέρβασης υπολογιστικών ορίων. Κατ' ουσίαν, η tot_b5 μεγιστοποιεί την ποιότητα, αλλά με βαρύ κόστος σε αποδοτικότητα. Ενδείκνυται για σενάρια όπου η ορθότητα αποτελεί απόλυτη προτεραιότητα, ενώ για γενική χρήση η ρύθμιση tot_b3 προσφέρει σαφώς πιο ισορροπημένο ισοζύγιο απόδοσης-κόστους.

6.4 Οπτικοποίηση αποτελεσμάτων

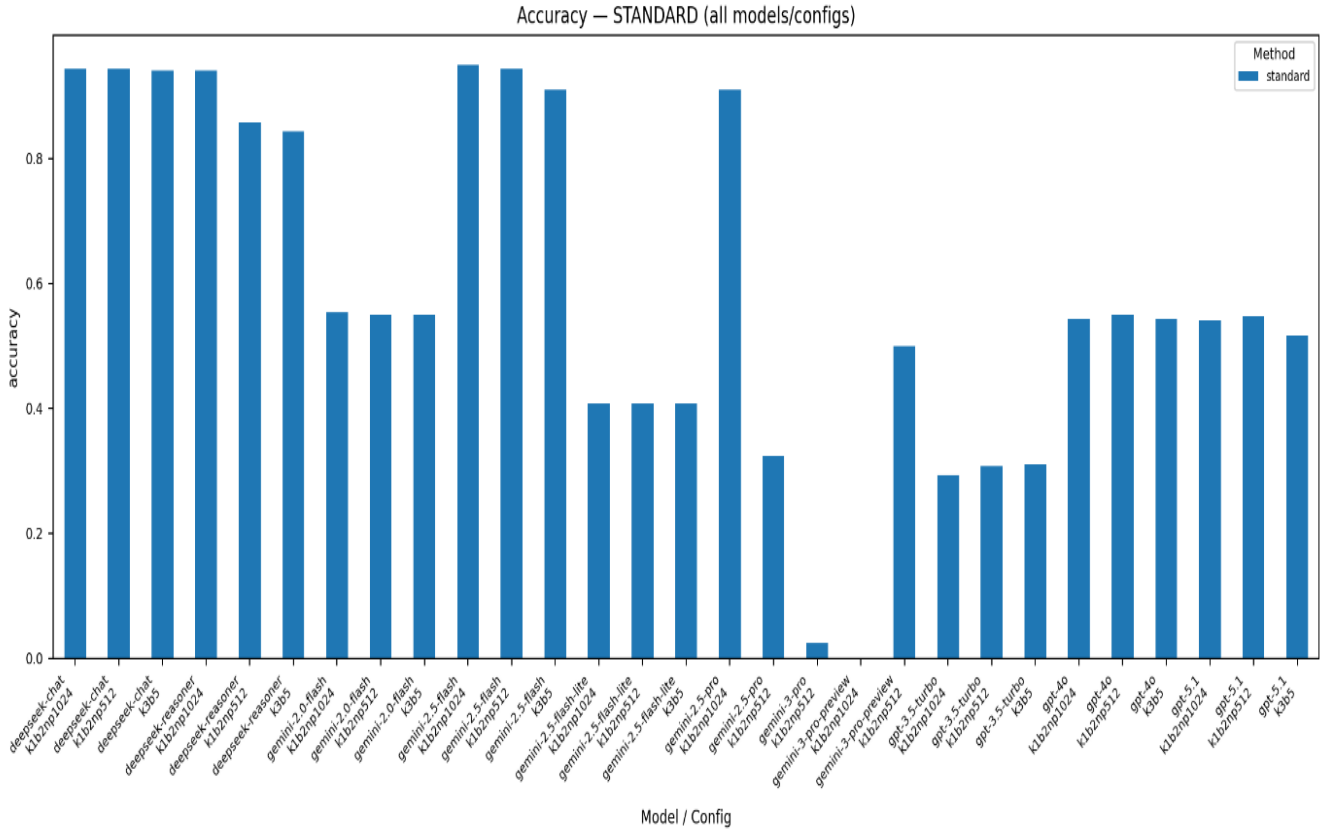
Στην παρούσα ενότητα παρουσιάζεται η οπτικοποίηση και η αναλυτική ερμηνεία των πειραματικών αποτελεσμάτων για τις τρεις εξεταζόμενες μεθόδους (Standard Prompting, CoT, ToT). Η ανάλυση βασίζεται σε βασικούς δείκτες απόδοσης, όπως η ακρίβεια (αγγλ. Accuracy), ο ρυθμός αποχής (αγγλ. Abstention Rate), ο χρόνος απόκρισης (αγγλ. latency), η κατανάλωση πόρων (αγγλ. tokens) και η συνέπεια (αγγλ. consistency), όπως προέκυψαν από τον μηχανισμό σύνοψης των εκτελέσεων ανά συνδυασμό μοντέλου και διαμόρφωσης.

6.4.1 Οπτικοποίηση αποτελεσμάτων μεθόδου Standard Prompting

Η μέθοδος Standard Prompting αξιολογήθηκε ως το θεμελιώδες σημείο αναφοράς (αγγλ. Baseline) της παρούσας μελέτης, λειτουργώντας χωρίς ρητή καθοδήγηση συλλογισμού ή παραγωγή ενδιάμεσων βημάτων. Τα αποτελέσματα που παρουσιάζονται στα αντίστοιχα γραφήματα προκύπτουν από τον μηχανισμό σύνοψης του ανά μέθοδο και ανά συνδυασμό μοντέλου/διαμόρφωσης.

Ακρίβεια (Accuracy)

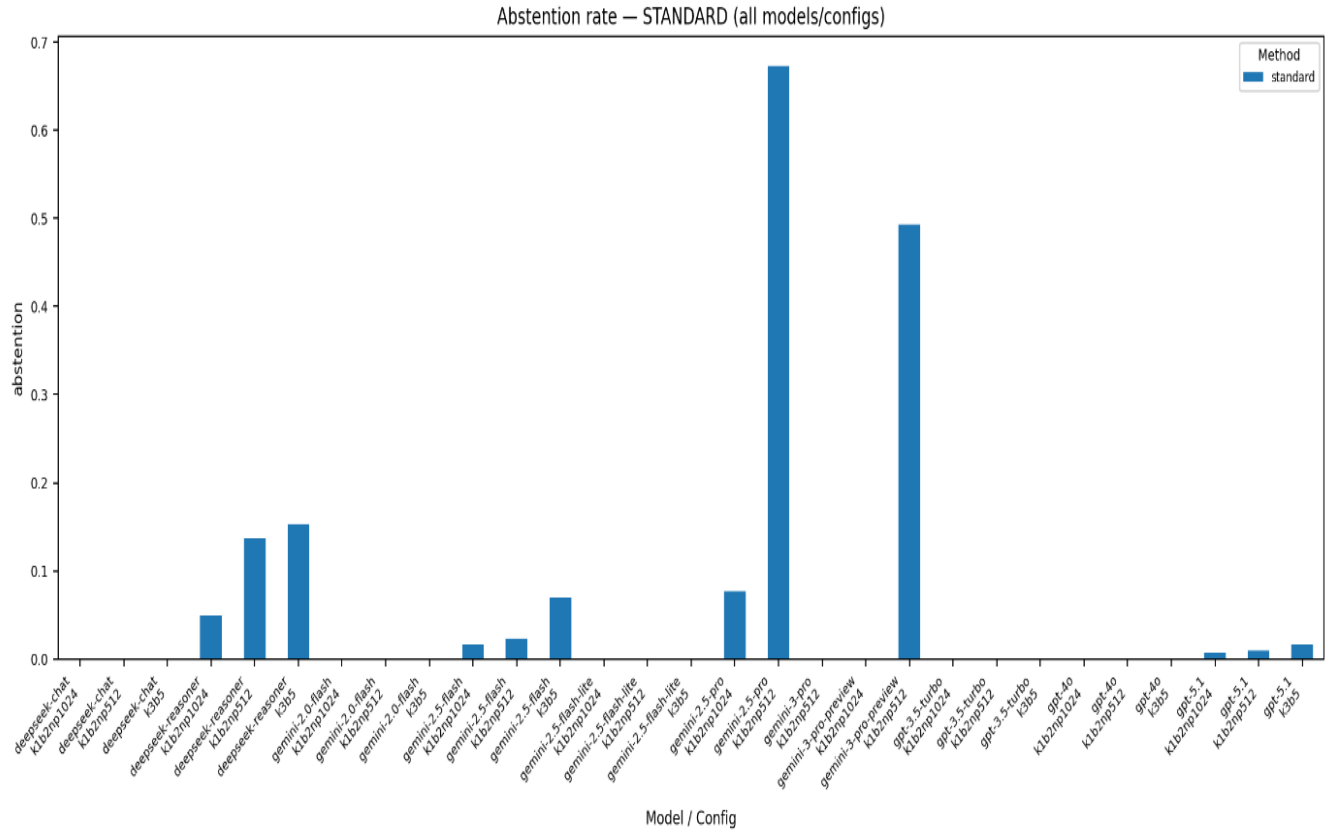
Η ανάλυση της ακρίβειας για τη μέθοδο Standard Prompting καταδεικνύει ένα ιδιαίτερα ευρύ φάσμα τιμών, το οποίο εκτείνεται από εξαιρετικά χαμηλά επίπεδα έως και ποσοστά που προσεγγίζουν το 85–94% σε συγκεκριμένα μοντέλα (π.χ. deepseek-chat, gemini-2.5-flash). Ωστόσο, η υψηλή αυτή επίδοση δεν εμφανίζεται με συστηματικό τρόπο σε όλες τις διαμορφώσεις (αγγλ. configurations). Η παρατηρούμενη μεταβλητότητα υποδηλώνει ότι η επιτυχία της μεθόδου εξαρτάται σε μεγάλο βαθμό από τη μορφολογία του προβλήματος και την ικανότητα του μοντέλου για επιφανειακή αντιστοίχιση προτύπων (αγγλ. pattern matching), παρά από μια σταθερή ικανότητα πολυσταδιακού συλλογισμού. Σε συνθετότερα προβλήματα, η απόδοση τείνει να καταρρέει, εύρημα που επιβεβαιώνει ότι η Standard Prompting, ενώ αποτελεί χρήσιμο Baseline, δεν συνιστά αξιόπιστη στρατηγική για την επίλυση απαιτητικών μαθηματικών προβλημάτων.



Εικόνα 6 Ακρίβεια (Accuracy) Standard Prompting

Ρυθμός αποχής (Abstention Rate)

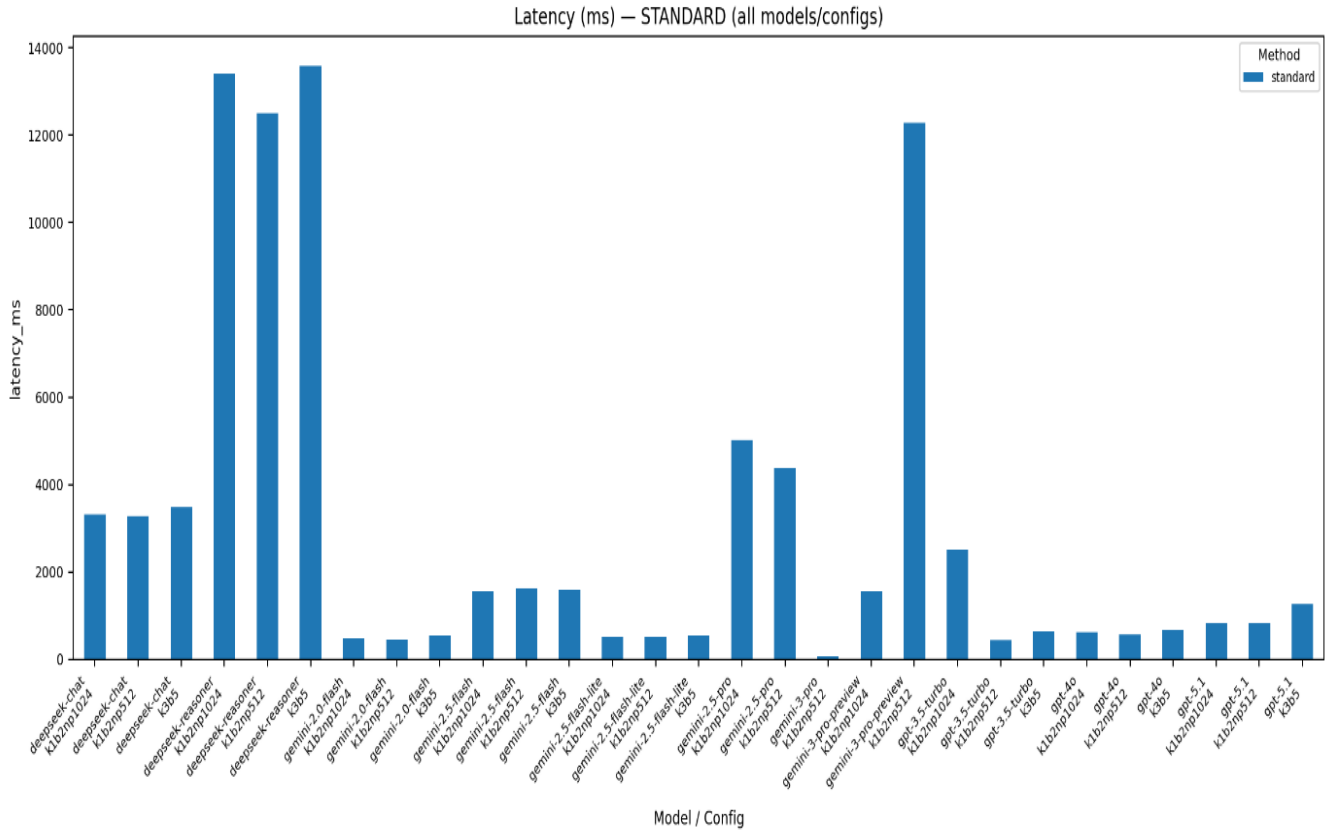
Ο ρυθμός αποχής παρουσιάζει έντονη ανομοιογένεια μεταξύ των διαφορετικών μοντέλων. Ορισμένες διαμορφώσεις εμφανίζουν σχεδόν μηδενική αποχή, γεγονός που υποδηλώνει ότι το μοντέλο εξαναγκάζεται να παράγει απάντηση ανεξαρτήτως της πολυπλοκότητας του προβλήματος. Αντιθέτως, σε άλλες περιπτώσεις (π.χ. gemini-2.5-pro σε συγκεκριμένα configurations), τα ποσοστά αποχής υπερβαίνουν το 60%. Η συμπεριφορά αυτή ερμηνεύεται ως απόρροια της απουσίας ρητού μηχανισμού συλλογισμού. Σε προβλήματα που απαιτούν αλληλουχία 2 - 8 πράξεων, το μοντέλο συχνά αδυνατεί να διατηρήσει συνεπή εσωτερική κατάσταση, οδηγούμενο είτε σε αποχή είτε σε αποσπασματική παραγωγή. Συνεπώς, η χαμηλή αποχή στη μέθοδο Standard δεν αποτελεί κατ' ανάγκη ένδειξη ικανότητας, αλλά συχνά δείγμα "αναγκαστικής" και πιθανώς εσφαλμένης απόκρισης.



Εικόνα 7 Ρυθμός Αποχής (Abstention Rate) Standard Prompting

Χρόνος απόκρισης (Latency)

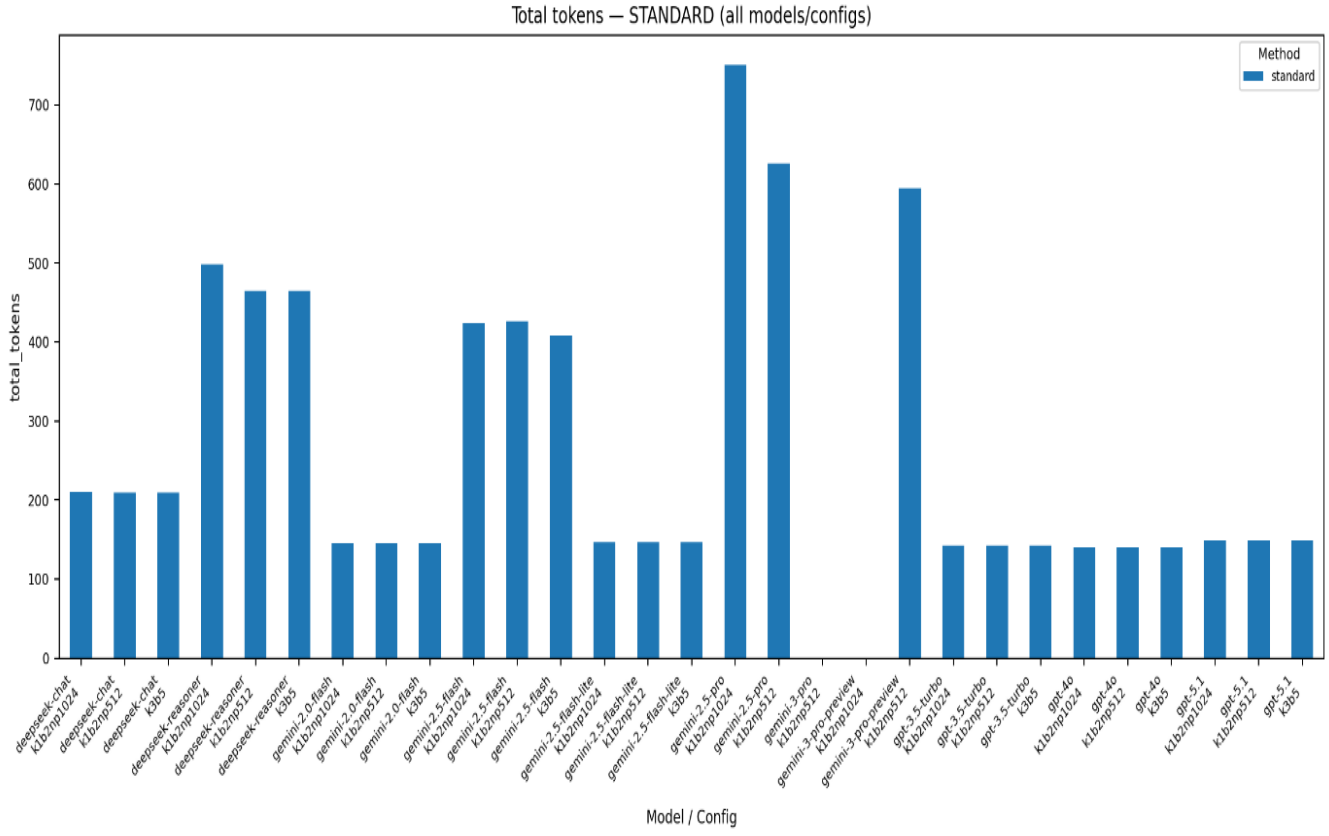
Τα δεδομένα της μετρικής latency αποκαλύπτουν σημαντικές αποκλίσεις, με τιμές που κυμαίνονται από μερικές εκατοντάδες milliseconds έως και άνω των 13 δευτερολέπτων (π.χ. deepseek-reasoner). Αξιοσημείωτο είναι ότι ο αυξημένος χρόνος απόκρισης δεν συσχετίζεται γραμμικά με υψηλότερη ακρίβεια. Το γεγονός αυτό υποδηλώνει ότι, στο πλαίσιο της Standard Prompting, ο επιπλέον υπολογιστικός χρόνος δεν αξιοποιείται παραγωγικά για δομημένη σκέψη. Αντίθετα, φαίνεται να καταναλώνεται σε μη ελεγχόμενη παραγωγή κειμένου, χωρίς την ύπαρξη μηχανισμών ελέγχου προόδου ή αναθεώρησης των ενδιάμεσων συμπερασμάτων.



Εικόνα 8 Χρόνος Απόκρισης (Latency) Standard Prompting

Συνολική κατανάλωση Tokens (Total Tokens)

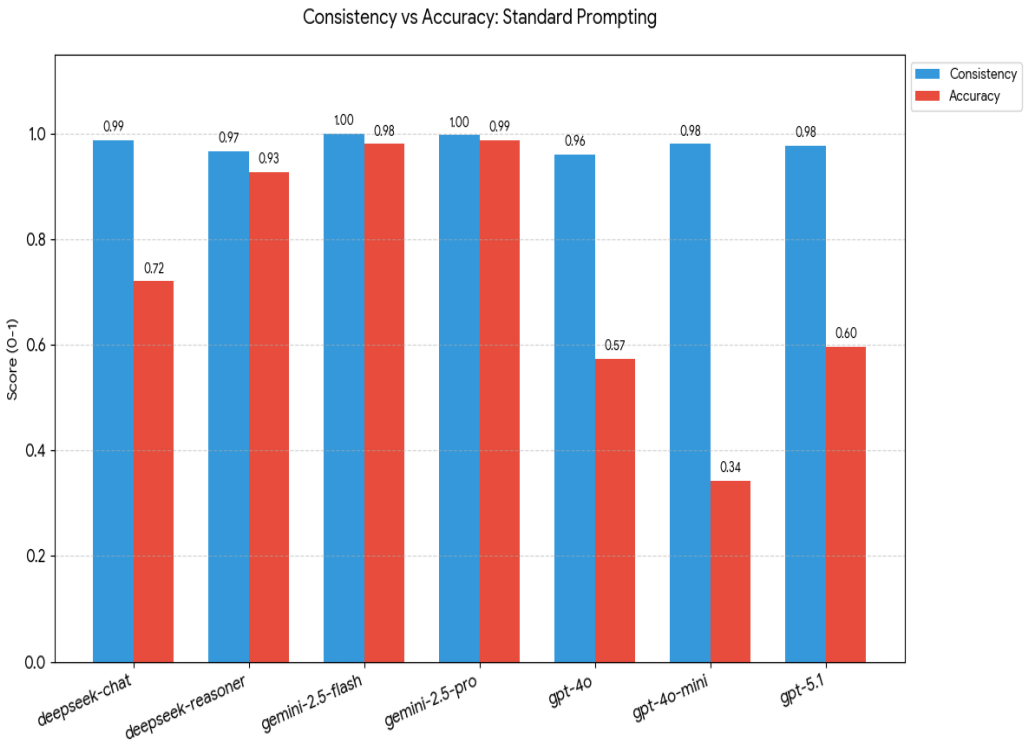
Η μέθοδος Standard Prompting παραμένει στη γενική περίπτωση, η πλέον οικονομική ως προς την κατανάλωση tokens. Παρ' όλα αυτά, παρατηρούνται περιπτώσεις μοντέλων με αυξημένη κατανάλωση (π.χ. gemini-3-pro-preview) χωρίς την αντίστοιχη βελτίωση στην ακρίβεια. Το εύρημα αυτό υπογραμμίζει ότι η απλή αύξηση της παραγόμενης πληροφορίας δεν συνεπάγεται ποιοτικότερη λύση όταν απουσιάζει ο ρητός έλεγχος της συλλογιστικής διαδικασίας, καθιστώντας τον λόγο κόστους-απόδοσης (αγγλ. tokens-to-accuracy) χαμηλό για τα σύνθετα προβλήματα.



Εικόνα 9 Συνολική Κατανάλωση Tokens (Total Tokens) Standard Prompting

Έλεγχος συνέπειας (Consistency Check)

Η εξέταση της μετρικής της συνέπειας για τη μέθοδο Standard Prompting αναδεικνύει ένα φαινομενικά παράδοξο εύρημα. Η μέθοδος επιτυγχάνει καθολικά υψηλά ποσοστά συνέπειας (>95%) μεταξύ των διαδοχικών εκτελέσεων για την πλειονότητα των μοντέλων. Ωστόσο, η υψηλή αυτή σταθερότητα βρίσκεται σε πλήρη αντιδιαστολή με τη χαμηλή ακρίβεια σε ορισμένους συνδυασμούς μοντέλου/μεθόδου (π.χ gpt-4o, gpt-4o-mini). Το γεγονός αυτό υποδηλώνει ότι η απουσία καθοδηγούμενου συλλογισμού οδηγεί τα μοντέλα σε μια δογματική συμπεριφορά, όπου αντί να εξερευνούν εναλλακτικές λύσεις, τείνουν να αναπαράγουν στερεοτυπικά και με απόλυτη βεβαιότητα το ίδιο σφάλμα. Συνεπώς, στην περίπτωση της Standard Prompting, η υψηλή συνέπεια δεν αποτελεί δείκτη αξιοπιστίας, αλλά συχνά ένδειξη συστηματικής αποτυχίας.



Εικόνα 10 Συνέπεια (Consistency) Vs Ακρίβεια (Accuracy) Standard Prompting

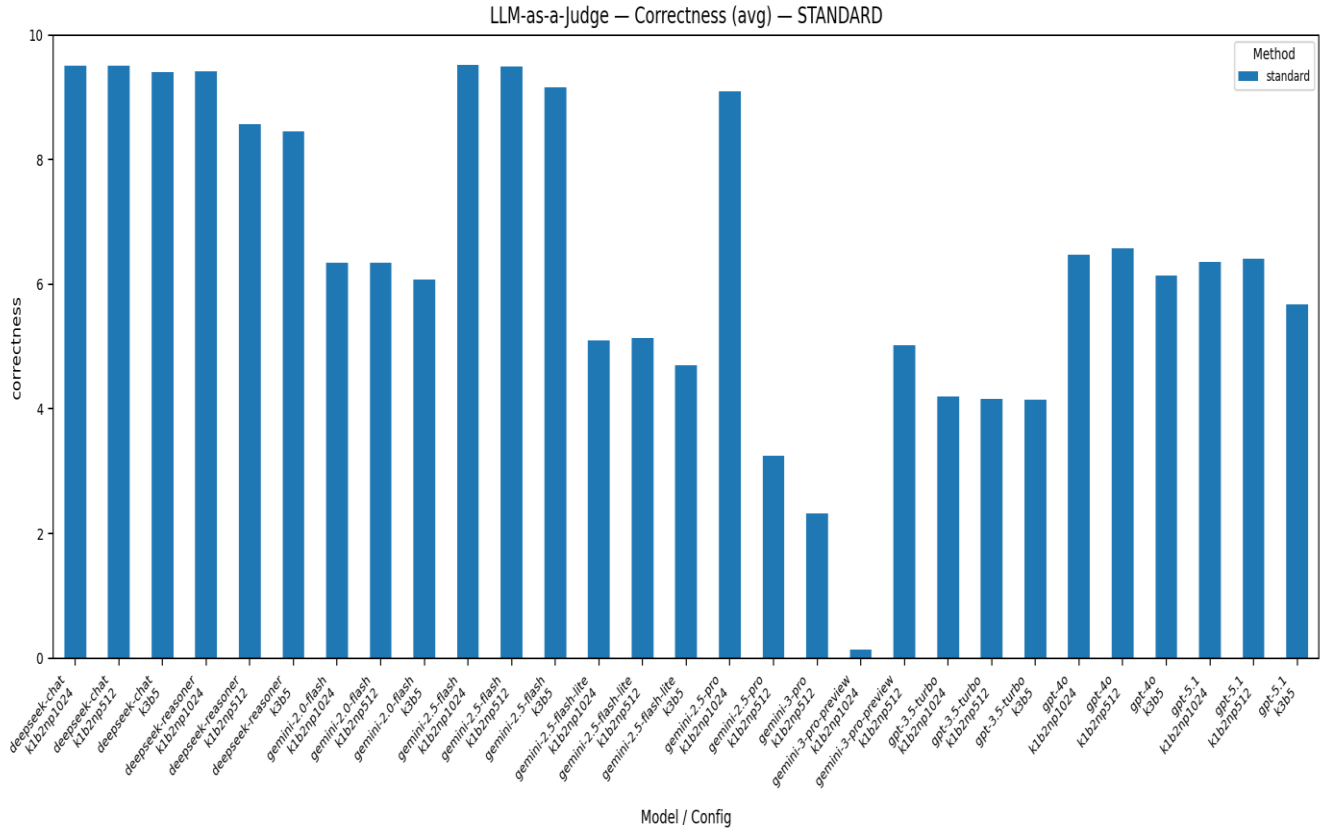
Πίνακας 15 Συγκεντρωτικά Αποτελέσματα Standard Prompting

Model	**Config**	**Accuracy**	**Abstention Rate**	**Latency (ms)**	**Total Tokens**
deepseek-chat	k12bnp1024	**0.93**	0.00	3 300	210
deepseek-chat	k12bnp512	**0.93**	0.00	3 250	210
deepseek-chat	k3b5	**0.93**	0.00	3 450	210
deepseek-reasoner	k12bnp1024	0.92	0.05	**13 400**	500
deepseek-reasoner	k12bnp512	0.86	0.14	**12 500**	465
deepseek-reasoner	k3b5	0.85	0.15	**13 600**	465
gemini-2.0-flash	k12bnp1024	0.55	0.00	450	145
gemini-2.0-flash	k12bnp512	0.55	0.00	480	145
gemini-2.0-flash	k3b5	0.55	0.00	520	145

gemini-2.5-flash	k12bnp1024	**0.94**	0.02	1 550	425
gemini-2.5-flash	k12bnp512	0.93	0.03	1 650	425
gemini-2.5-flash	k3b5	0.90	0.07	1 600	405
gemini-2.5-lite	k12bnp1024	0.41	0.00	500	145
gemini-2.5-lite	k12bnp512	0.41	0.00	520	145
gemini-2.5-lite	k3b5	0.41	0.00	550	145
gemini-2.5-pro	k12bnp1024	0.90	0.08	5 000	**750**
gemini-2.5-pro	k12bnp512	0.32	**0.67**	4 400	**625**
gemini-3-pro-preview	k12bnp1024	0.50	**0.49**	**12 300**	**595**
gpt-3.5-turbo	k12bnp1024	0.29	0.00	2 500	140
gpt-3.5-turbo	k12bnp512	0.31	0.00	450	140
gpt-3.5-turbo	k3b5	0.31	0.00	650	140
gpt-4o	k12bnp1024	0.54	0.00	650	140
gpt-4o	k12bnp512	0.55	0.00	600	140
gpt-4o	k3b5	0.54	0.00	580	140
gpt-5.1	k12bnp1024	0.54	0.01	720	150
gpt-5.1	k12bnp512	0.55	0.01	820	150
gpt-5.1	k3b5	0.52	0.02	1 250	150

Μέση βαθμολογία Correctness (LLM-as-a-Judge) για τη Standard Prompting ανά μοντέλο και διαμόρφωση

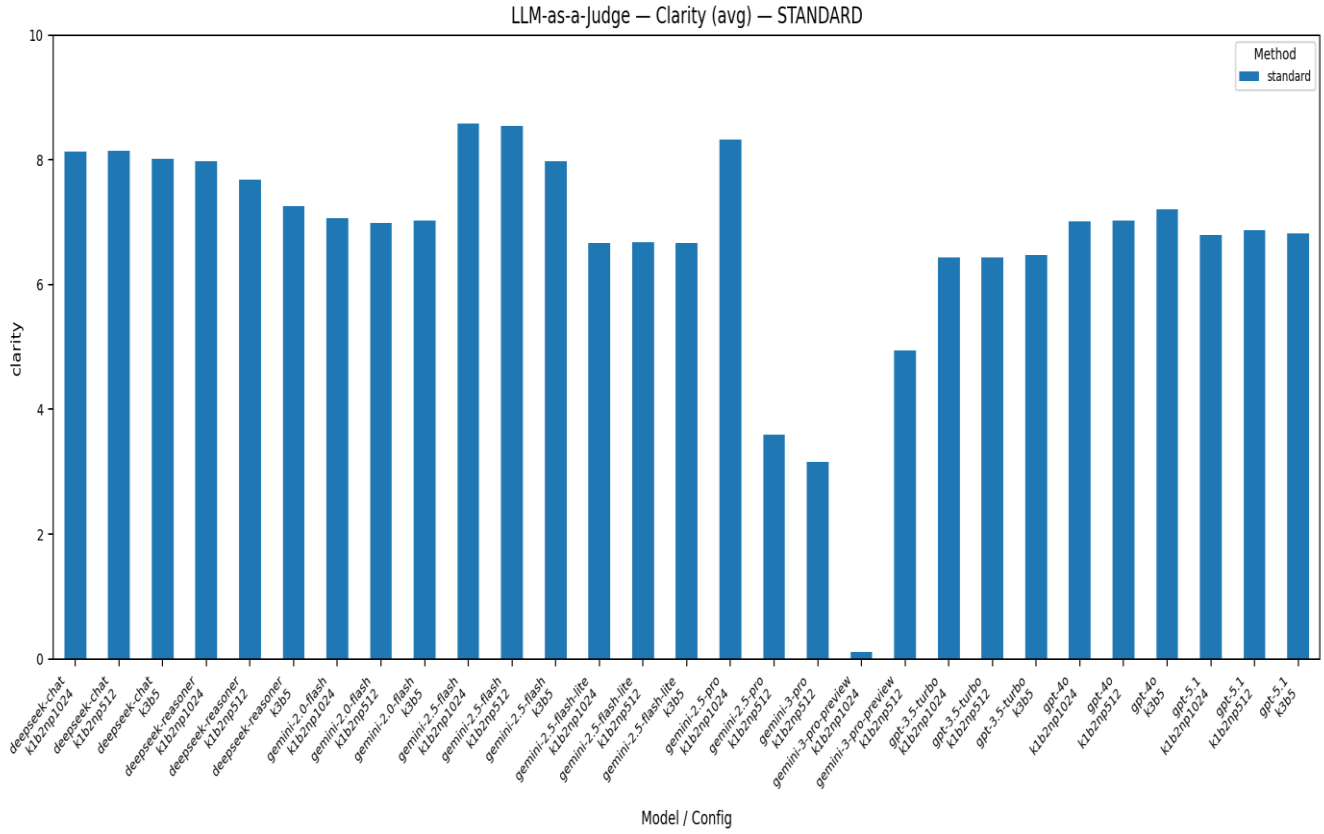
Η ποιότητα ως προς την ορθότητα παρουσιάζει έντονη εξάρτηση από την αρχιτεκτονική του μοντέλου: μοντέλα υψηλής απόδοσης (π.χ. DeepSeek, Gemini 2.5 Flash) επιτυγχάνουν πολύ υψηλές βαθμολογίες (≈9–9.5), ενώ άλλα ισχυρά μοντέλα (π.χ. GPT-4o) καθώς και ασθενέστερες διαμορφώσεις εμφανίζουν μετριοπαθή έως χαμηλή επίδοση. Η Standard Prompting λειτουργεί επομένως ως καθρέφτης της εγγενούς συλλογιστικής ικανότητας του μοντέλου, χωρίς μηχανισμό διόρθωσης ή σταθεροποίησης σφαλμάτων.



Εικόνα 11 Μέση βαθμολογία Correctness (LLM-as-a-Judge) για τη Standard Prompting ανά μοντέλο και διαμόρφωση

Μέση βαθμολογία Clarity (LLM-as-a-Judge) για τη Standard Prompting ανά μοντέλο και διαμόρφωση.

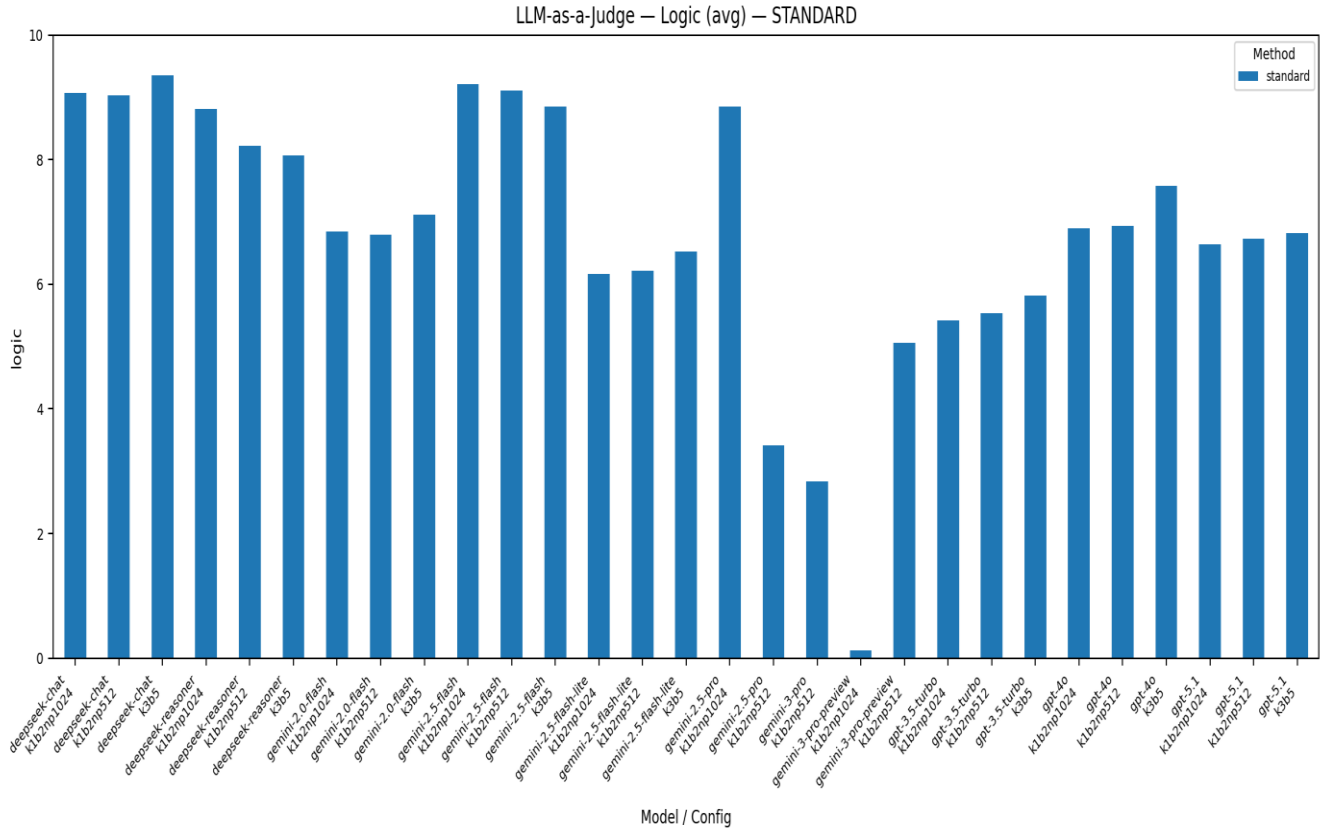
Η σαφήνεια των απαντήσεων παραμένει σε γενικά ικανοποιητικά επίπεδα στα ισχυρότερα μοντέλα (≈8–8.5), ενώ σε ασθενέστερες ή μη σταθερές διαμορφώσεις παρατηρείται αισθητή πτώση. Η Standard Prompting δεν επιβάλλει δομημένη ανάπτυξη σκέψης, με αποτέλεσμα η καθαρότητα της διατύπωσης να εξαρτάται άμεσα από την εσωτερική γλωσσική και συλλογιστική ικανότητα του εκάστοτε μοντέλου.



Εικόνα 12 Μέση βαθμολογία Clarity (LLM-as-a-Judge) για τη Standard Prompting ανά μοντέλο και διαμόρφωση.

Μέση βαθμολογία Logic (LLM-as-a-Judge) για τη Standard Prompting ανά μοντέλο και διαμόρφωση.

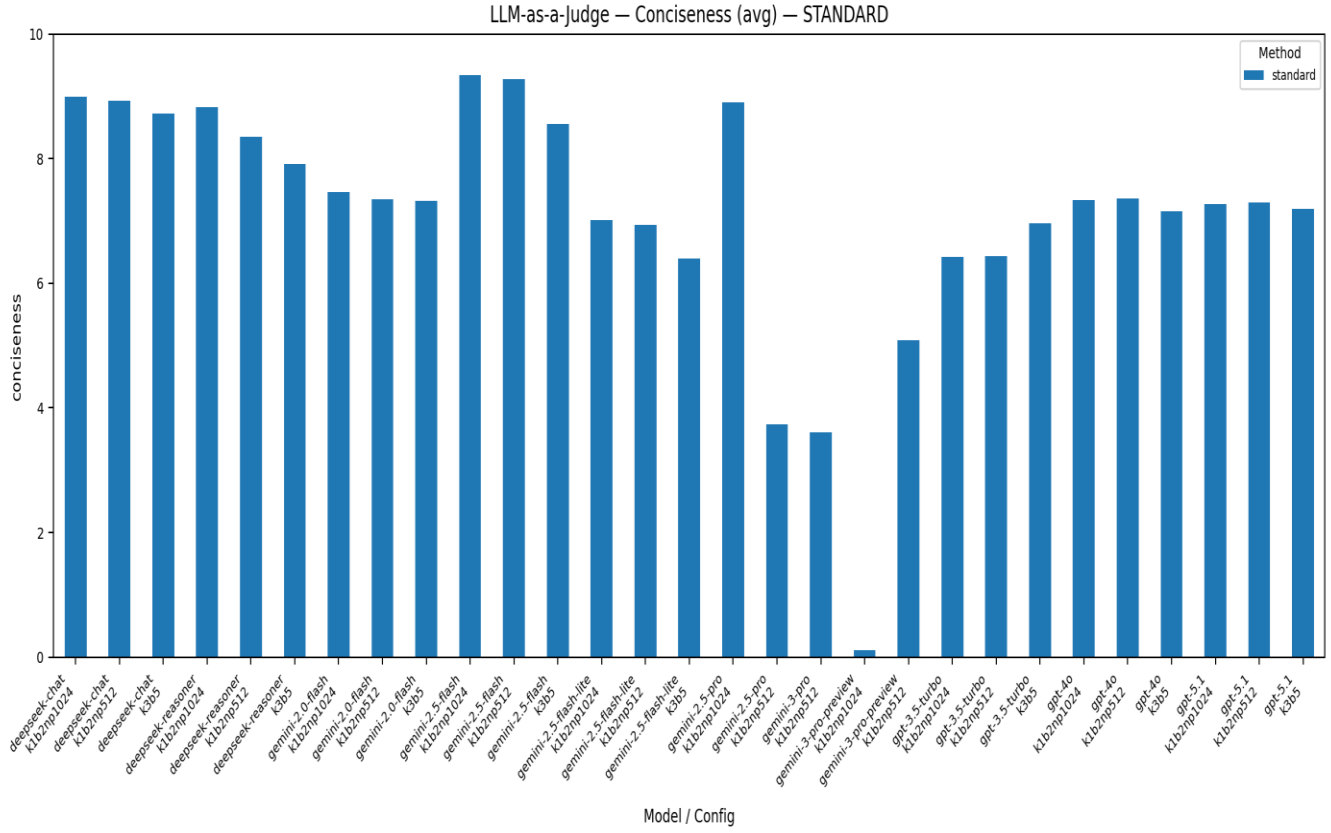
Η λογική συνοχή των απαντήσεων εμφανίζει έντονη εξάρτηση από το υποκείμενο μοντέλο. Τα ισχυρότερα LLMs επιτυγχάνουν υψηλές τιμές (≈8.5–9.3), υποδηλώνοντας ικανοποιητική εσωτερική δομή συλλογισμού ακόμη και χωρίς ρητή αλυσίδα σκέψης. Αντίθετα, σε ασθενέστερες ή ασταθείς διαμορφώσεις παρατηρείται σημαντική πτώση της λογικής συνέπειας, γεγονός που αναδεικνύει ότι η Standard Prompting δεν ενισχύει εγγενώς τη δομημένη συλλογιστική, αλλά αντανακλά κυρίως τις εγγενείς δυνατότητες του μοντέλου.



Εικόνα 13 Μέση βαθμολογία Logic (LLM-as-a-Judge) για τη Standard Prompting ανά μοντέλο και διαμόρφωση.

Μέση βαθμολογία Conciseness (LLM-as-a-Judge) για τη Standard Prompting ανά μοντέλο και διαμόρφωση.

Η συντομία των απαντήσεων παρουσιάζει μέτρια έως υψηλή επίδοση στα ισχυρότερα μοντέλα (≈7–9), υποδηλώνοντας ικανότητα παραγωγής περιεκτικών απαντήσεων χωρίς ρητή εξωτερικευμένη συλλογιστική. Ωστόσο, σε ορισμένες διαμορφώσεις παρατηρείται αισθητή πτώση, γεγονός που καταδεικνύει ότι η περιεκτικότητα εξαρτάται άμεσα από την εσωτερική αρχιτεκτονική του μοντέλου και δεν αποτελεί εγγενές χαρακτηριστικό της Standard Prompting.



Εικόνα 14 Μέση βαθμολογία Conciseness (LLM-as-a-Judge) για τη Standard Prompting ανά μοντέλο και διαμόρφωση.

Πίνακας 16 LLM-as-a-Judge για τη Standard Prompting

Model	**Config**	**Correct.**	**Clarity**	**Logic**	**Concise**	**MO**	**Παρατήρηση**
deepseek-chat	k1b2np1024	9.50	8.12	9.06	8.99	**8.92**	Πολύ καλή απόδοση
deepseek-chat	k1b2np512	9.50	8.14	9.03	8.93	**8.90**	Πολύ καλή απόδοση
deepseek-chat	k3b5	9.41	8.01	9.35	8.72	**8.87**	Πολύ καλή απόδοση
deepseek-reasoner	k1b2np1024	9.42	7.98	8.81	8.82	**8.76**	Πολύ καλή απόδοση
deepseek-reasoner	k1b2np512	8.56	7.67	8.22	8.34	**8.20**	Πολύ καλή απόδοση

deepseek-reasoner	k3b5	8.45	7.25	8.06	7.91	**7.92**	Πολύ καλή απόδοση
gemini-2.0-flash	k1b2np1024	6.34	7.06	6.84	7.46	**6.93**	Μέτρια απόδοση
gemini-2.0-flash	k1b2np512	6.34	6.98	6.80	7.35	**6.87**	Μέτρια απόδοση
gemini-2.0-flash	k3b5	6.08	7.02	7.12	7.31	**6.88**	Μέτρια απόδοση
gemini-2.5-flash	k1b2np1024	9.52	8.58	9.20	9.34	**9.16**	Εξαιρετική απόδοση
gemini-2.5-flash	k1b2np512	9.49	8.53	9.11	9.27	**9.10**	Εξαιρετική απόδοση
gemini-2.5-flash	k3b5	9.15	7.97	8.85	8.55	**8.63**	Πολύ καλή απόδοση
gemini-2.5-flash-lite	k1b2np1024	5.10	6.66	6.17	7.01	**6.24**	Μέτρια απόδοση
gemini-2.5-flash-lite	k1b2np512	5.13	6.68	6.21	6.94	**6.24**	Μέτρια απόδοση
gemini-2.5-flash-lite	k3b5	4.70	6.66	6.53	6.39	**6.07**	Μέτρια απόδοση
gemini-2.5-pro	k1b2np1024	9.09	8.32	8.84	8.90	**8.79**	Πολύ καλή απόδοση
gemini-2.5-pro	k1b2np512	3.25	3.59	3.41	3.73	**3.50**	Χαμηλή απόδοση
gemini-3-pro	k1b2np512	2.32	3.15	2.83	3.60	**2.98**	Πολύ χαμηλή απόδοση
gemini-3-pro-preview	k1b2np1024	0.13	0.10	0.12	0.10	**0.12**	Πολύ χαμηλή απόδοση
gemini-3-pro-preview	k1b2np512	5.02	4.94	5.06	5.09	**5.03**	Μέτρια απόδοση
gpt-3.5-turbo	k1b2np1024	4.20	6.44	5.41	6.41	**5.62**	Μέτρια απόδοση
gpt-3.5-turbo	k1b2np512	4.16	6.44	5.53	6.44	**5.64**	Μέτρια απόδοση
gpt-3.5-turbo	k3b5	4.14	6.47	5.81	6.95	**5.85**	Μέτρια απόδοση
gpt-4o	k1b2np1024	6.47	7.00	6.89	7.33	**6.93**	Μέτρια απόδοση
gpt-4o	k1b2np512	6.57	7.02	6.94	7.36	**6.97**	Μέτρια απόδοση
gpt-4o	k3b5	6.13	7.20	7.58	7.15	**7.02**	Πολύ καλή απόδοση
gpt-5.1	k1b2np1024	6.36	6.79	6.64	7.26	**6.77**	Μέτρια απόδοση
gpt-5.1	k1b2np512	6.40	6.87	6.72	7.29	**6.82**	Μέτρια απόδοση

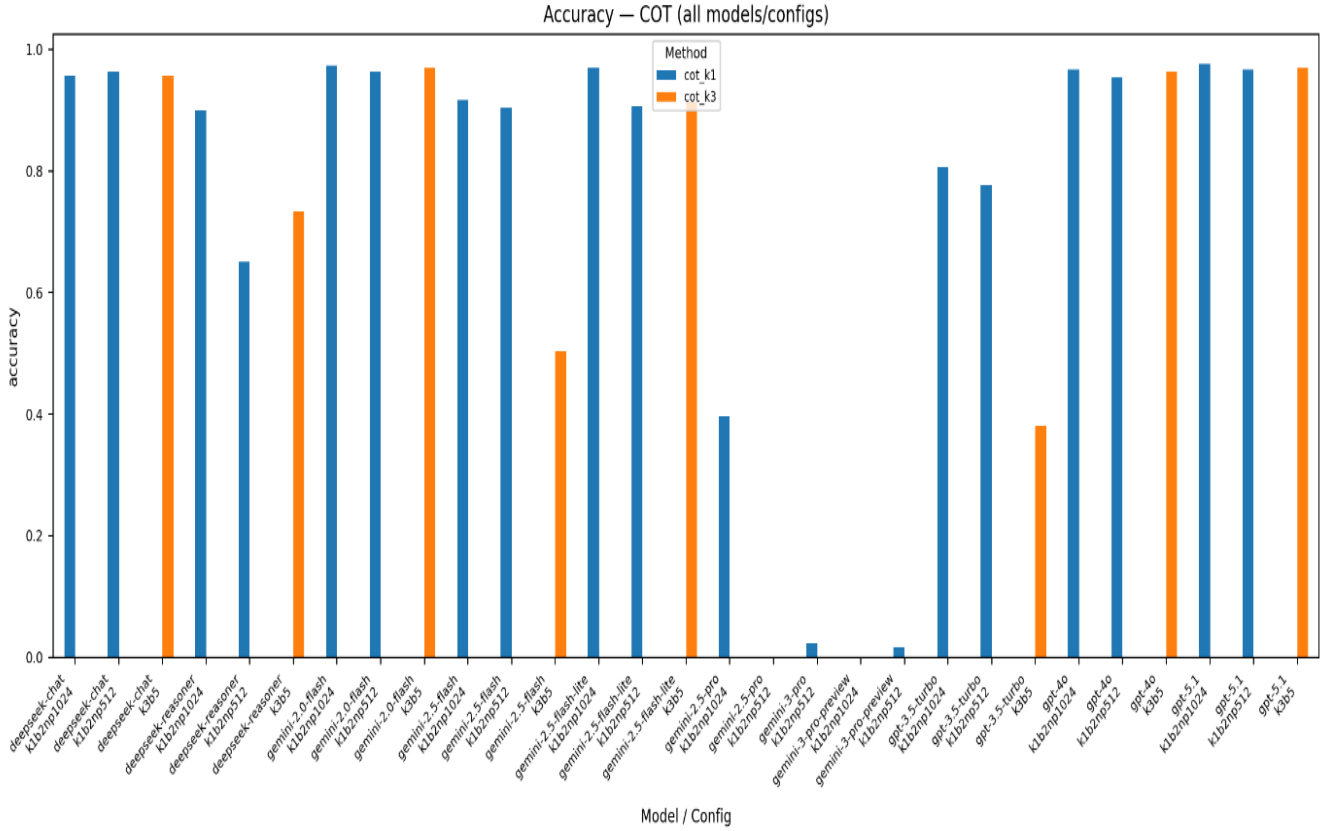
gpt-5.1	k3b5	5.67	6.82	6.82	7.18	**6.63**	Μέτρια απόδοση

6.4.2 Οπτικοποίηση αποτελεσμάτων μεθόδου CoT

Η μέθοδος CoT αξιολογήθηκε μέσω δύο διακριτών ρυθμίσεων, cot_k1 και cot_k3, οι οποίες διαφοροποιούνται ως προς το πλήθος των παραγόμενων συλλογιστικών αλυσίδων. Σε πλήρη αντιδιαστολή με τη μέθοδο Standard Prompting, η προσέγγιση CoT επιβάλλει τη ρητή εξωτερίκευση της συλλογιστικής διαδικασίας, επιτρέποντας στο μοντέλο να αποδομήσει και να επιλύσει σταδιακά τα σύνθετα, πολυσταδιακά προβλήματα του GSM8K.

Ακρίβεια (Accuracy)

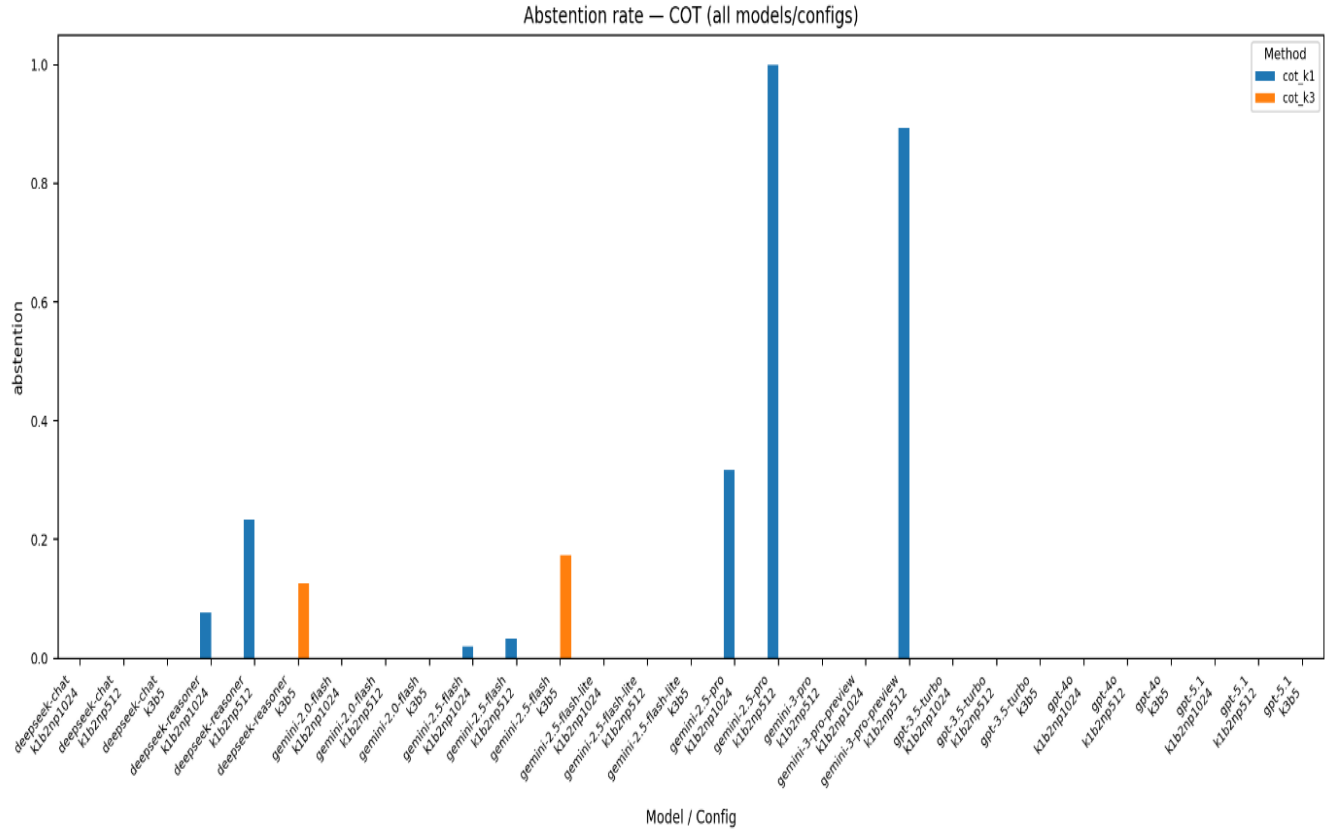
Τα πειραματικά δεδομένα καταδεικνύουν σαφή και ουσιαστική βελτίωση της ακρίβειας έναντι της Standard Prompting. Τα κορυφαία μοντέλα επιτυγχάνουν ποσοστά που προσεγγίζουν ή και αγγίζουν το 97–98%, τόσο στη ρύθμιση cot_k1 όσο και στην cot_k3. Η επίδοση αυτή επιβεβαιώνει ότι η δομημένη ακολουθία σκέψης λειτουργεί ως καταλύτης για την ορθή επίλυση, σταθεροποιώντας τη στοχαστική συμπεριφορά των μοντέλων. Η πλειονότητα των εξεταζόμενων συστημάτων διατηρεί ακρίβεια άνω του 80%, ωστόσο εντοπίζονται μεμονωμένες περιπτώσεις (π.χ. gemini-2.5-pro σε συγκεκριμένα configurations) με εξαιρετικά χαμηλή επίδοση (<5%), φαινόμενο που συνδέεται άμεσα με υψηλά ποσοστά αποχής ή αδυναμία παραγωγής έγκυρης συλλογιστικής αλυσίδας.



Εικόνα 15 Ακρίβεια (Accuracy) CoT

Ρυθμός αποχής (Abstention Rate)

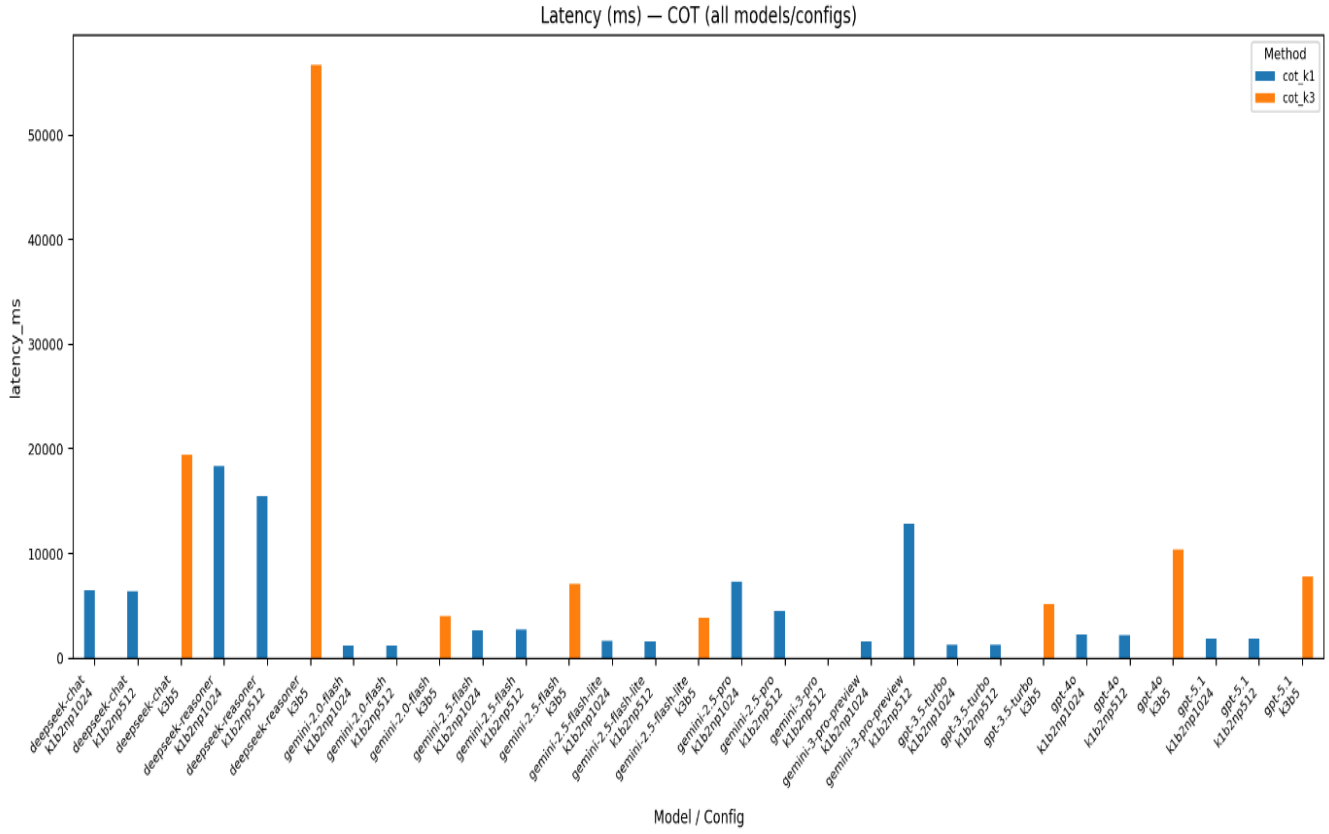
Η εφαρμογή της CoT οδηγεί σε αισθητή μείωση του ρυθμού αποχής για την πλειονότητα των μοντέλων, συγκριτικά με την Standard Prompting. Εντούτοις, καταγράφονται ακραίες αποκλίσεις που υπογραμμίζουν την εξάρτηση της μεθόδου από τις δυνατότητες και τη συμβατότητα του εκάστοτε μοντέλου. Χαρακτηριστικά, παρατηρείται περίπτωση ολικής αποχής (100%) σε διαμόρφωση cot_k1 και np_512, καθώς και υψηλά ποσοστά της τάξης του 90% και 30% σε άλλες ρυθμίσεις. Τα ευρήματα αυτά υποδηλώνουν ότι η απαίτηση για ρητό, βηματικό συλλογισμό μπορεί να λειτουργήσει αποτρεπτικά για μοντέλα που δεν διαθέτουν ικανότητα διαχείρισης πολυπλοκότητας ή παρουσιάζουν ασυμβατότητα σε συγκεκριμένες παραμετροποιήσεις, οδηγώντας τα σε αδυναμία παραγωγής απάντησης αντί για βελτίωση.



Εικόνα 16 Ρυθμός Αποχής (Abstention Rate) CoT

Χρόνος απόκρισης (Latency)

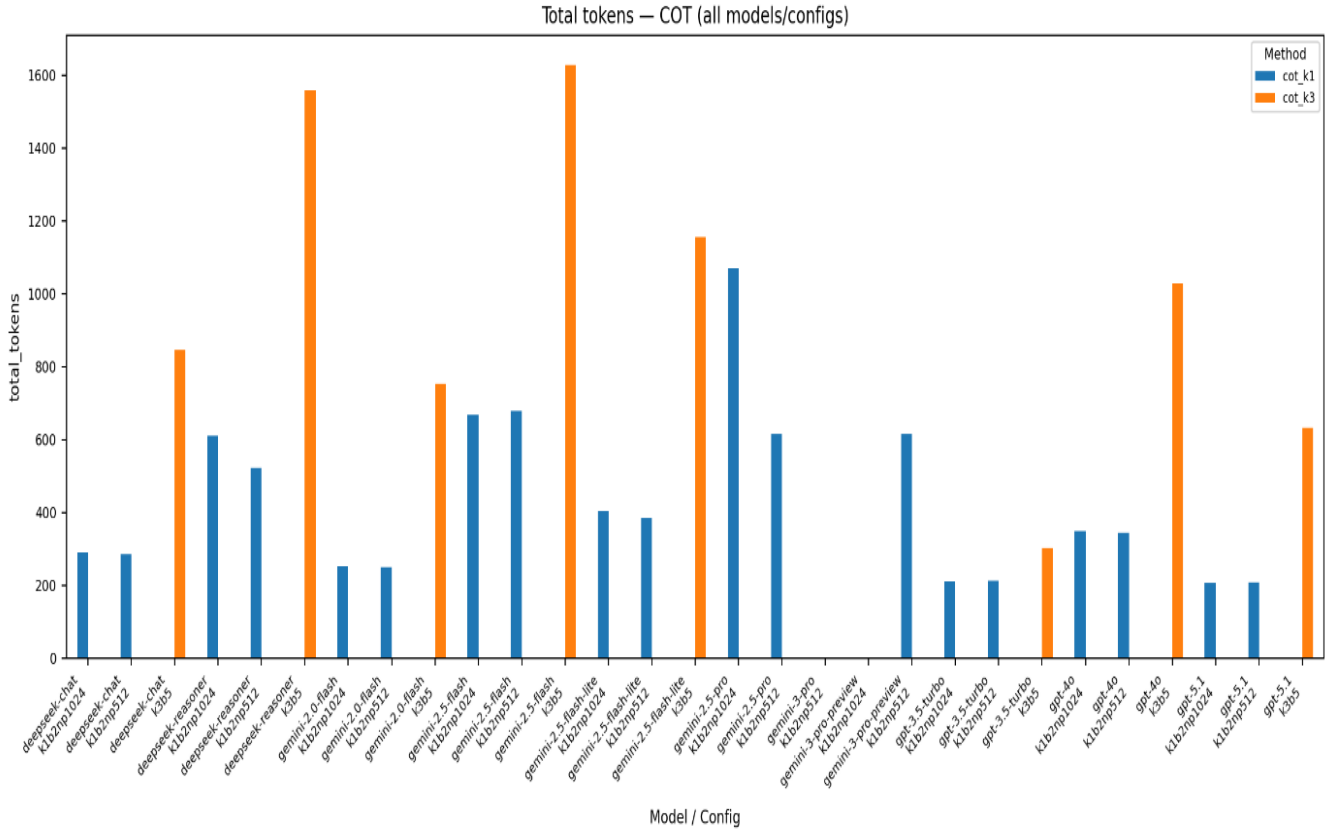
Η ρητή εξωτερίκευση της συλλογιστικής διαδικασίας επιφέρει αναπόφευκτα σημαντική επιβάρυνση στον χρόνο απόκρισης. Η διαφορά μεταξύ των ρυθμίσεων cot_k1 και cot_k3 είναι εμφανής, με την τελευταία να καταγράφει τιμές που υπερβαίνουν τα 55.000 ms σε ακραίες περιπτώσεις (deepseek-reasoner), ακολουθούμενες από μετρήσεις της τάξης των 15.000–19.000 ms. Παρόλο που η αυξημένη χρονική επένδυση συχνά μεταφράζεται σε υψηλότερη ακρίβεια, αναδεικνύεται ένα σαφές ισοζύγιο (αγγλ. trade-off) μεταξύ ποιότητας αποτελέσματος και χρονικής αποδοτικότητας, καθιστώντας την cot_k3 λιγότερο ελκυστική για εφαρμογές πραγματικού χρόνου.



Εικόνα 17 Χρόνος Απόκρισης (Latency) CoT

Συνολική κατανάλωση Tokens (Total Tokens)

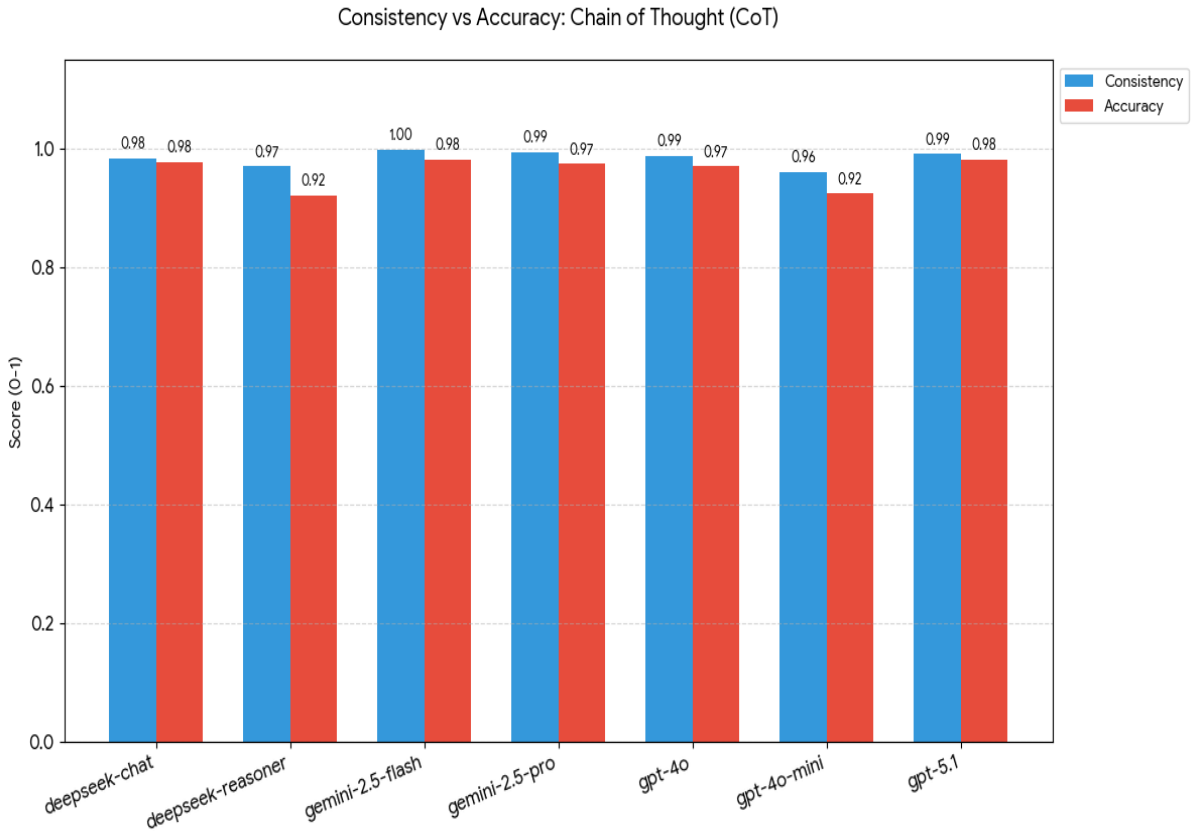
Η κατανάλωση πόρων παρουσιάζει σημαντική αύξηση σε σχέση με τη Standard Prompting, ως άμεση συνέπεια της αναλυτικής καταγραφής των βημάτων σκέψης. Οι υψηλότερες τιμές εντοπίζονται κυρίως στη ρύθμιση cot_k3, όπου σε πολλές περιπτώσεις ξεπερνούν τα 1.000 tokens, με τις μέγιστες τιμές να καταγράφονται στα μοντέλα gemini-2.5-flash (1.620 tokens) και deepseek-reasoner (1.550 tokens). Το εύρημα αυτό είναι αναμενόμενο και συμβατό με τη φύση της μεθόδου. Ωστόσο, η αυξημένη κατανάλωση περιορίζει την αποδοτικότητα της μεθόδου σε περιβάλλοντα με αυστηρούς περιορισμούς πόρων, παρά τη βελτιωμένη ακρίβεια που προσφέρει.



Εικόνα 18 Συνολική Κατανάλωση Tokens (Total Tokens) CoT Prompting

Έλεγχος συνέπειας (Consistency Check)

Η ανάλυση της συνέπειας για τη μέθοδο CoT αποκαλύπτει μια εξαιρετικά θετική εικόνα, διαφοροποιώντας τη ριζικά από τη Standard Prompting. Όπως καταδεικνύεται στο συγκριτικό διάγραμμα, η μέθοδος επιτυγχάνει υψηλότατα επίπεδα συνέπειας, τα οποία για την πλειονότητα των μοντέλων υπερβαίνουν το 98%. Σημαντική είναι η ισχυρή θετική συσχέτιση μεταξύ συνέπειας και ακρίβειας, όπου η σταθερότητα των απαντήσεων δεν είναι προϊόν επαναλαμβανόμενου λάθους, αλλά αποτέλεσμα μιας στιβαρής συλλογιστικής διαδικασίας. Η ρητή καταγραφή των ενδιάμεσων βημάτων λειτουργεί προς όφελος της αξιοπιστίας, μειώνοντας δραστικά την πιθανότητα τυχαίας απόκλισης και οδηγώντας το μοντέλο να συγκλίνει με αξιοσημείωτη σταθερότητα στην ορθή λύση, μετατρέποντας την αβεβαιότητα σε επαληθεύσιμη λογική βεβαιότητα.



Εικόνα 19 Συνέπεια (Consistency) Vs Ακρίβεια (Accuracy) CoT

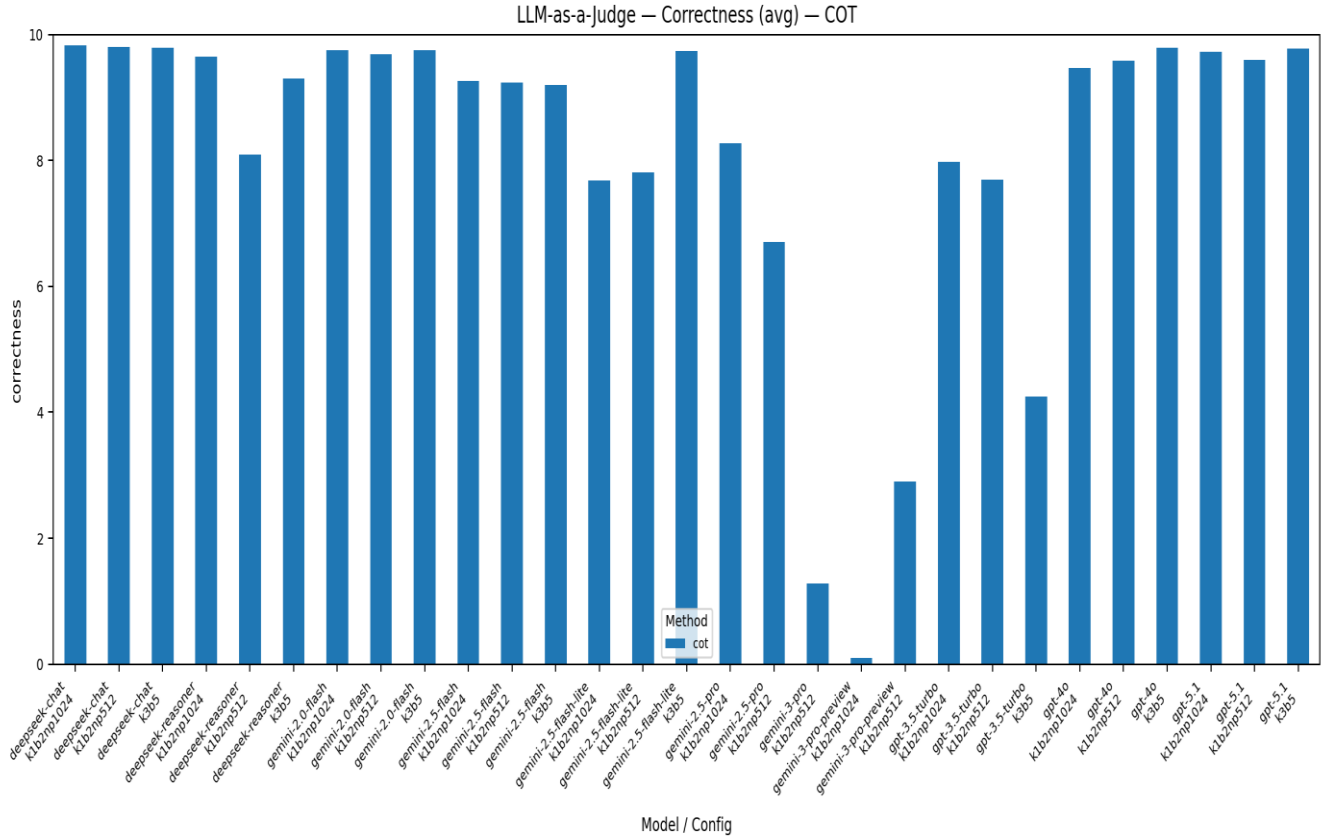
Πίνακας 17 Συγκεντρωτικά Αποτελέσματα CoT

Model	Config	Method	Accuracy	Abstention	Latency (ms)	Total Tokens
deepseek-chat	k12bnp1024	cot_k1	**0.96**	0.00	6 500	290
deepseek-chat	k12bnp512	cot_k1	**0.97**	0.00	6 300	280
deepseek-chat	k3b5	cot_k3	**0.95**	0.00	19 500	840
deepseek-reasoner	k12bnp1024	cot_k1	0.90	0.08	18 500	610
deepseek-reasoner	k12bnp512	cot_k1	0.65	0.23	15 500	520
deepseek-reasoner	k3b5	cot_k3	0.73	0.13	**56 000**	**1 550**
gemini-2.0-flash	k12bnp1024	cot_k1	**0.97**	0.00	1 100	250
gemini-2.0-flash	k12bnp512	cot_k1	**0.96**	0.00	1 200	250

gemini-2.0-flash	k3b5	cot_k3	**0.97**	0.00	4 000	750
gemini-2.5-flash	k12bnp1024	cot_k1	0.92	0.02	2 800	670
gemini-2.5-flash	k12bnp512	cot_k1	0.90	0.03	2 900	680
gemini-2.5-flash	k3b5	cot_k3	0.50	0.17	7 000	**1 620**
gemini-2.5-pro	k12bnp1024	cot_k1	0.40	0.32	7 300	**1 070**
gemini-2.5-pro	k12bnp512	cot_k1	0.00	**1.00**	4 500	610
gemini-3-pro-preview	k12bnp1024	cot_k1	0.02	0.00	1 600	0
gpt-3.5-turbo	k12bnp1024	cot_k1	0.81	0.00	12 800	610
gpt-4o	k12bnp1024	cot_k1	**0.96**	0.00	2 400	350
gpt-4o	k12bnp512	cot_k1	0.95	0.00	2 300	340
gpt-4o	k3b5	cot_k3	**0.96**	0.00	10 200	**1 020**
gpt-5.1	k12bnp1024	cot_k1	**0.98**	0.00	1 900	205
gpt-5.1	k12bnp512	cot_k1	0.97	0.00	2 000	210
gpt-5.1	k3b5	cot_k3	**0.97**	0.00	7 800	630

Μέση βαθμολογία Correctness (LLM-as-a-Judge) για τη μέθοδο CoT ανά μοντέλο και διαμόρφωση.

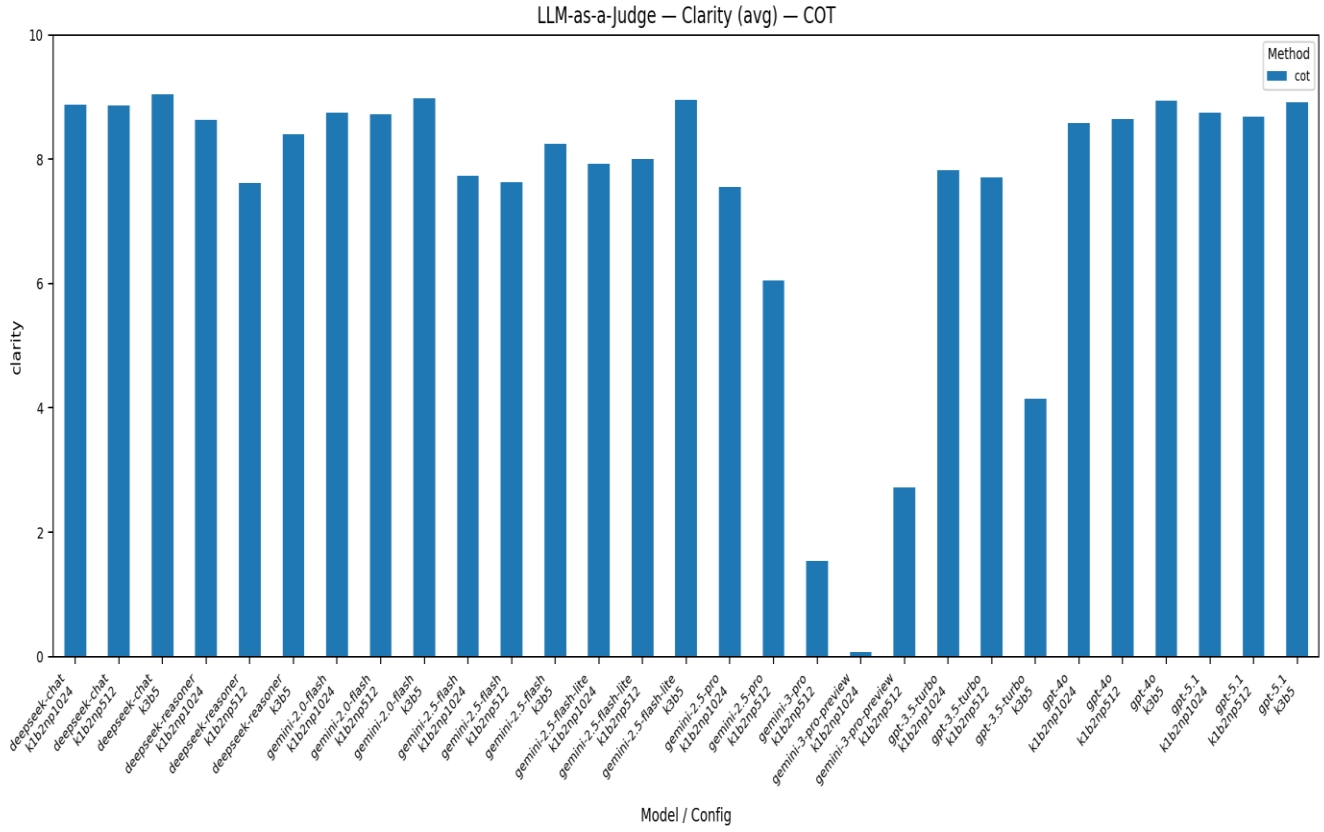
Η CoT εμφανίζει συνολικά υψηλότερες τιμές ορθότητας σε σύγκριση με τη Standard Prompting, με τις περισσότερες ισχυρές αρχιτεκτονικές να προσεγγίζουν βαθμολογίες κοντά στο άριστα (≈9–9.8). Οι χαμηλές τιμές περιορίζονται σε συγκεκριμένες διαμορφώσεις reasoning ή περιπτώσεις αστοχίας, γεγονός που υποδηλώνει ότι η ρητή αλυσίδα συλλογισμού ενισχύει σημαντικά την ορθότητα όταν εφαρμόζεται σε συμβατά μοντέλα, χωρίς όμως να εξαλείφει πλήρως φαινόμενα κατάρρευσης απόδοσης.



Εικόνα 20 Μέση βαθμολογία Correctness (LLM-as-a-Judge) για τη μέθοδο CoT ανά μοντέλο και διαμόρφωση.

Μέση βαθμολογία Clarity (LLM-as-a-Judge) για τη μέθοδο CoT ανά μοντέλο και διαμόρφωση.

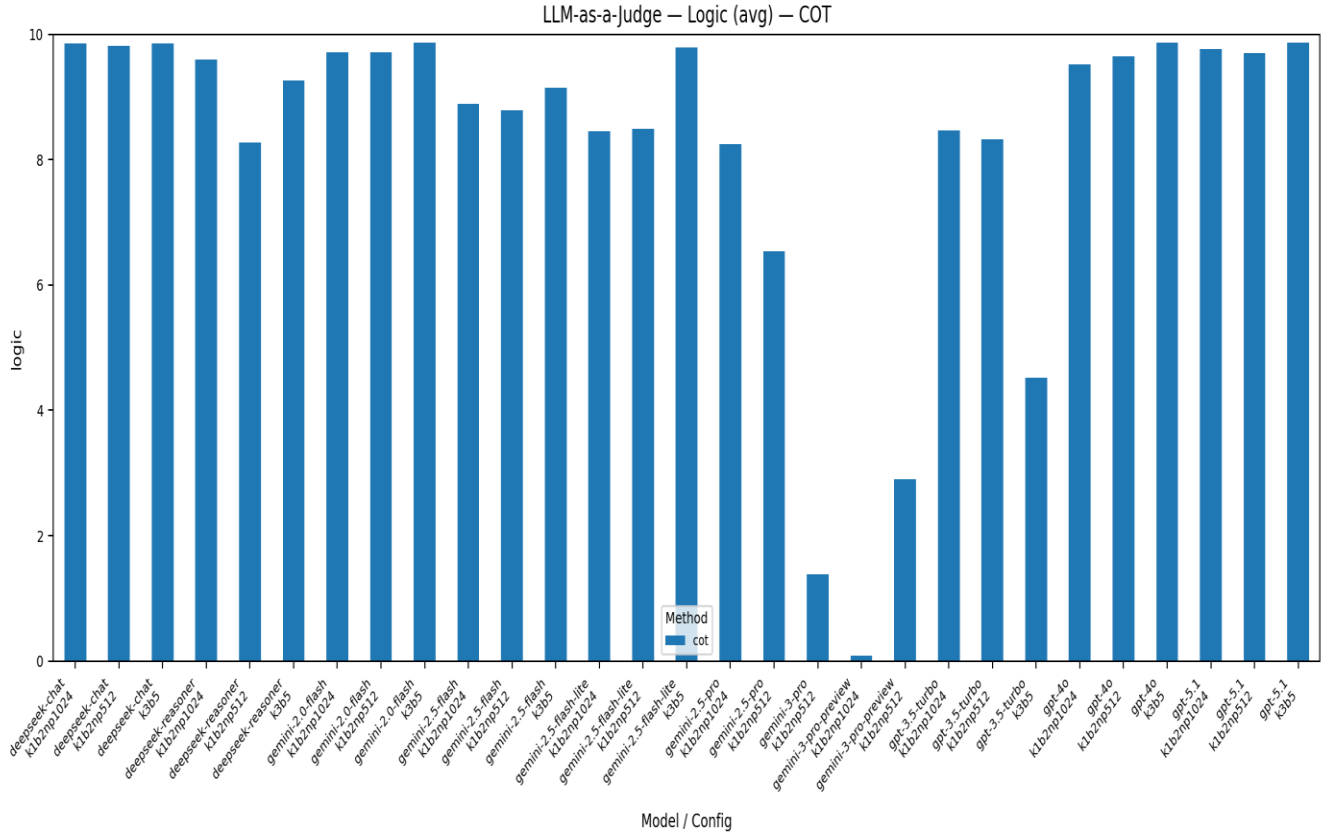
Η CoT παρουσιάζει γενικά υψηλά επίπεδα σαφήνειας, με τις περισσότερες ισχυρές αρχιτεκτονικές να κινούνται στο εύρος 8–9. Η ρητή εξωτερικευμένη αλυσίδα σκέψης συμβάλλει στη βελτίωση της δομής και της αναγνωσιμότητας της λύσης. Ωστόσο, σε ορισμένες διαμορφώσεις reasoning ή σε περιπτώσεις λειτουργικής αστοχίας παρατηρούνται σημαντικές πτώσεις, γεγονός που υποδηλώνει ότι η σαφήνεια εξαρτάται όχι μόνο από τη μέθοδο αλλά και από τη συμβατότητα της τεχνικής με το εκάστοτε μοντέλο.



Εικόνα 21 Μέση βαθμολογία Clarity (LLM-as-a-Judge) για τη μέθοδο CoT ανά μοντέλο και διαμόρφωση.

Μέση βαθμολογία Logic (LLM-as-a-Judge) για τη μέθοδο CoT ανά μοντέλο και διαμόρφωση.

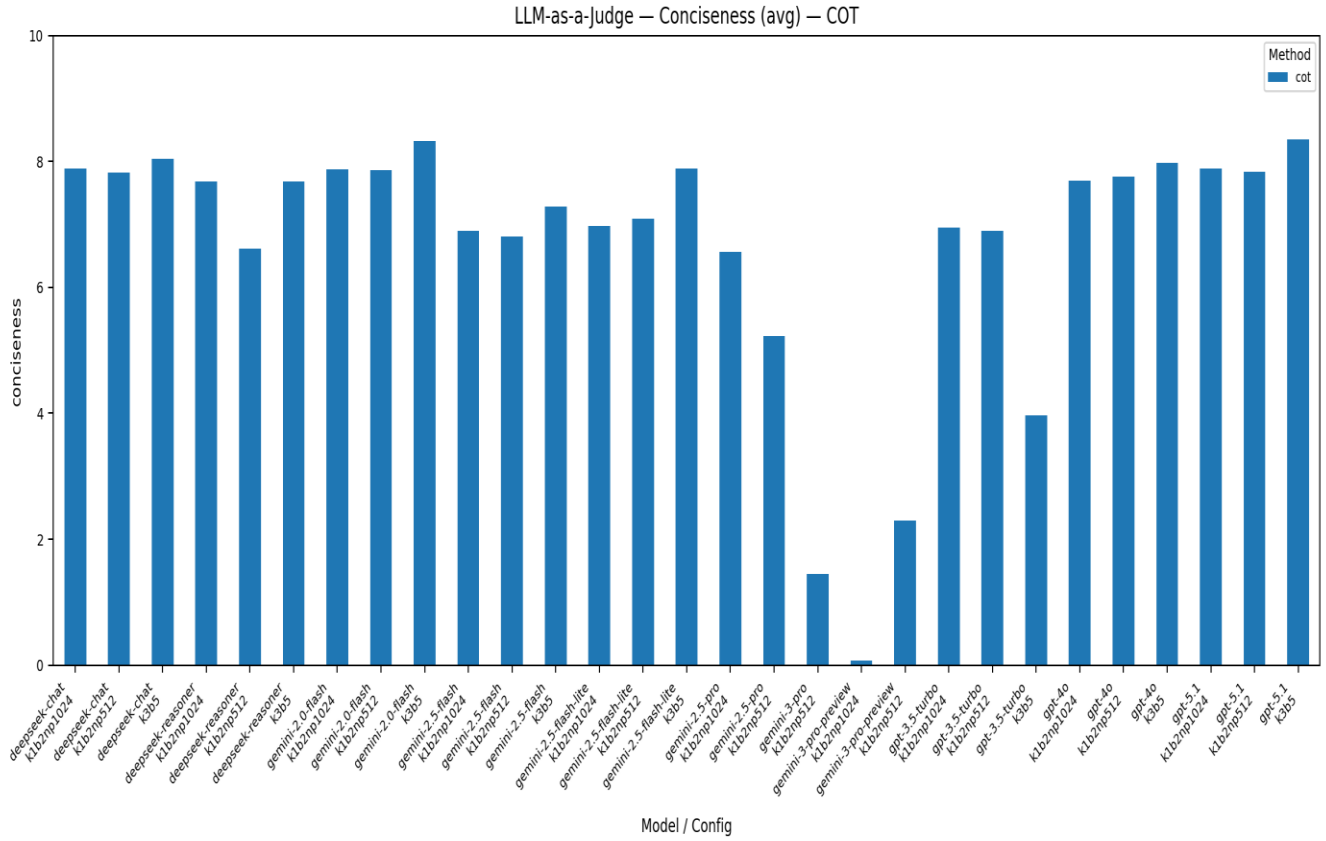
Η CoT εμφανίζει ιδιαίτερα υψηλές επιδόσεις στον άξονα της λογικής συνοχής, με τα ισχυρότερα μοντέλα να προσεγγίζουν τιμές 9-10. Η ρητή διατύπωση ενδιάμεσων βημάτων ενισχύει τη δομική συνέπεια και τη διαφάνεια του συλλογισμού. Ωστόσο, σε ορισμένες reasoning διαμορφώσεις ή περιπτώσεις λειτουργικής αστοχίας παρατηρούνται έντονες αποκλίσεις, γεγονός που επιβεβαιώνει ότι η αποτελεσματικότητα της CoT εξαρτάται ουσιωδώς από τη συμβατότητα της τεχνικής με την αρχιτεκτονική του εκάστοτε μοντέλου.



Εικόνα 22 Μέση βαθμολογία Logic (LLM-as-a-Judge) για τη μέθοδο CoT ανά μοντέλο και διαμόρφωση.

Μέση βαθμολογία Conciseness (LLM-as-a-Judge) για τη μέθοδο CoT ανά μοντέλο και διαμόρφωση.

Η CoT παρουσιάζει ικανοποιητική αλλά όχι κορυφαία επίδοση στον άξονα της συντομίας, καθώς η ρητή διατύπωση ενδιάμεσων βημάτων οδηγεί εγγενώς σε εκτενέστερες απαντήσεις. Τα ισχυρότερα μοντέλα διατηρούν τιμές κοντά στο 8, ενώ σε ορισμένες reasoning διαμορφώσεις παρατηρείται αισθητή πτώση, γεγονός που αντανακλά την αύξηση του όγκου κειμένου και τη μειωμένη περιεκτικότητα σε περιπτώσεις αστάθειας. Η εικόνα αυτή επιβεβαιώνει τον συμβιβασμό μεταξύ διαφάνειας συλλογισμού και συντομίας.



Εικόνα 23 Μέση βαθμολογία Conciseness (LLM-as-a-Judge) για τη μέθοδο CoT ανά μοντέλο και διαμόρφωση.

Πίνακας 18 LLM-as-a-Judge για τη μέθοδο CoT

Model	**Config**	**Correct.**	**Clarity**	**Logic**	**Concise**	**MO**	**Παρατήρηση**
deepseek-chat	k1b2np1024	9.82	8.87	9.85	7.88	**9.11**	Εξαιρετική απόδοση
deepseek-chat	k1b2np512	9.80	8.86	9.81	7.82	**9.07**	Εξαιρετική απόδοση
deepseek-chat	k3b5	9.79	9.04	9.85	8.04	**9.18**	Εξαιρετική απόδοση
deepseek-reasoner	k1b2np1024	9.65	8.63	9.59	7.68	**8.89**	Πολύ καλή απόδοση
deepseek-reasoner	k1b2np512	8.09	7.62	8.28	6.61	**7.65**	Πολύ καλή απόδοση

deepseek-reasoner	k3b5	9.29	8.39	9.26	7.68	**8.66**	Πολύ καλή απόδοση
gemini-2.0-flash	k1b2np1024	9.75	8.75	9.71	7.88	**9.02**	Εξαιρετική απόδοση
gemini-2.0-flash	k1b2np512	9.69	8.71	9.71	7.85	**8.99**	Πολύ καλή απόδοση
gemini-2.0-flash	k3b5	9.75	8.97	9.86	8.32	**9.23**	Εξαιρετική απόδοση
gemini-2.5-flash	k1b2np1024	9.26	7.72	8.89	6.89	**8.19**	Πολύ καλή απόδοση
gemini-2.5-flash	k1b2np512	9.23	7.63	8.79	6.81	**8.12**	Πολύ καλή απόδοση
gemini-2.5-flash	k3b5	9.20	8.24	9.14	7.28	**8.47**	Πολύ καλή απόδοση
gemini-2.5-flash-lite	k1b2np1024	7.69	7.92	8.45	6.97	**7.76**	Πολύ καλή απόδοση
gemini-2.5-flash-lite	k1b2np512	7.81	8.00	8.49	7.08	**7.85**	Πολύ καλή απόδοση
gemini-2.5-flash-lite	k3b5	9.74	8.95	9.79	7.89	**9.09**	Εξαιρετική απόδοση
gemini-2.5-pro	k1b2np1024	8.27	7.56	8.25	6.56	**7.66**	Πολύ καλή απόδοση
gemini-2.5-pro	k1b2np512	6.71	6.05	6.54	5.22	**6.13**	Μέτρια απόδοση
gemini-3-pro	k1b2np512	1.28	1.53	1.37	1.45	**1.41**	Πολύ χαμηλή απόδοση
gemini-3-pro-preview	k1b2np1024	0.10	0.07	0.08	0.07	**0.08**	Πολύ χαμηλή απόδοση
gemini-3-pro-preview	k1b2np512	2.90	2.72	2.90	2.29	**2.70**	Πολύ χαμηλή απόδοση
gpt-3.5-turbo	k1b2np1024	7.98	7.82	8.46	6.95	**7.80**	Πολύ καλή απόδοση
gpt-3.5-turbo	k1b2np512	7.69	7.71	8.32	6.89	**7.65**	Πολύ καλή απόδοση
gpt-3.5-turbo	k3b5	4.25	4.15	4.51	3.96	**4.22**	Χαμηλή απόδοση
gpt-4o	k1b2np1024	9.46	8.57	9.52	7.69	**8.81**	Πολύ καλή απόδοση
gpt-4o	k1b2np512	9.59	8.65	9.65	7.76	**8.91**	Πολύ καλή απόδοση
gpt-4o	k3b5	9.79	8.94	9.87	7.97	**9.14**	Εξαιρετική απόδοση
gpt-5.1	k1b2np1024	9.72	8.74	9.76	7.88	**9.03**	Εξαιρετική απόδοση
gpt-5.1	k1b2np512	9.60	8.68	9.69	7.83	**8.95**	Πολύ καλή απόδοση

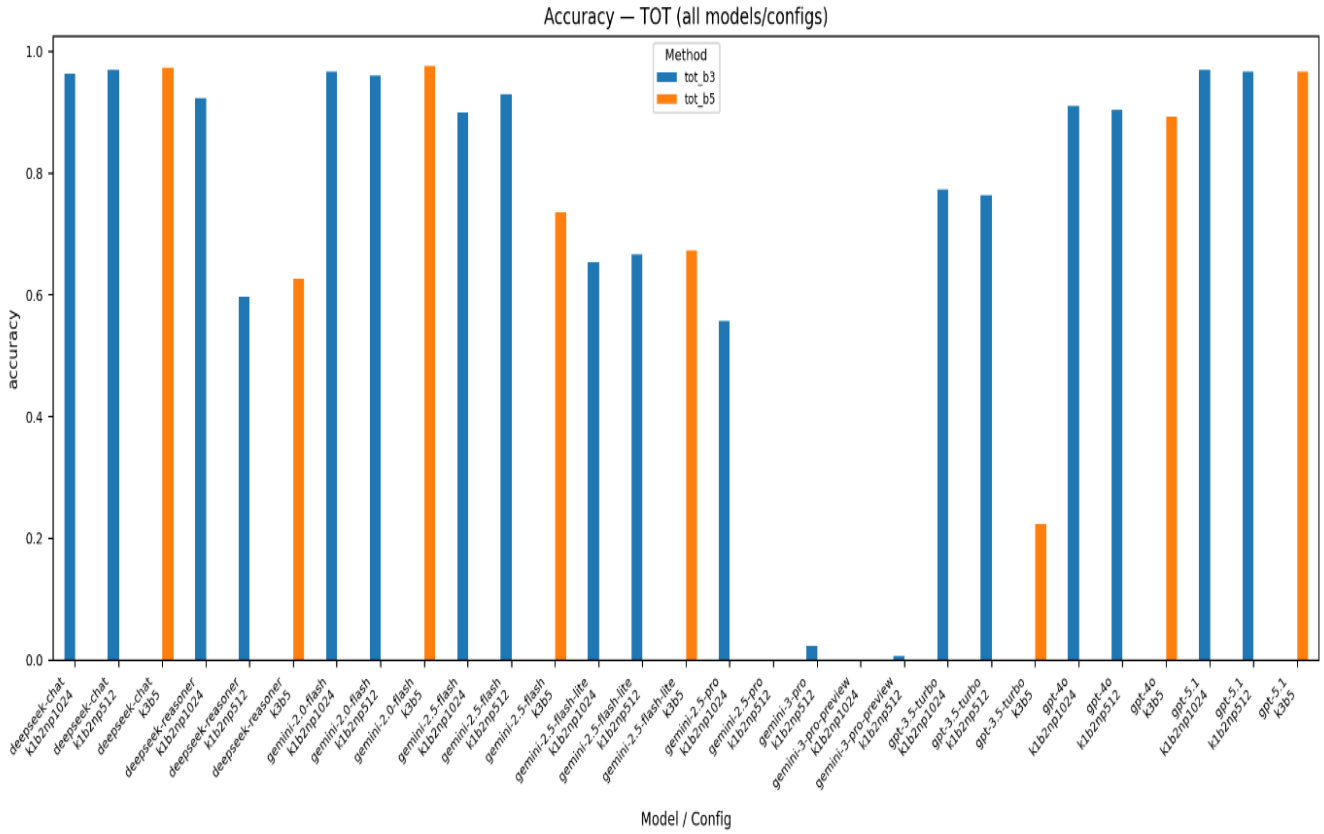
gpt-5.1	k3b5	9.77	8.91	9.86	8.35	**9.22**	Εξαιρετική απόδοση

6.4.3 Οπτικοποίηση αποτελεσμάτων μεθόδου ToT

Η μέθοδος ToT συνιστά την πλέον εξελιγμένη προσέγγιση συλλογισμού, επεκτείνοντας τη λειτουργία των LLMs από τη γραμμική ακολουθία σε μια δομημένη διαδικασία δενδρικής αναζήτησης. Το μοντέλο διερευνά παράλληλα πολλαπλά εναλλακτικά μονοπάτια λύσης, αξιολογώντας και επιλέγοντας το βέλτιστο μέσω εσωτερικών μηχανισμών κριτικής. Η αξιολόγηση πραγματοποιήθηκε με δύο ρυθμίσεις πολυπλοκότητας, tot_b3 (branching factor 3) και tot_b5 (branching factor 5), επιτρέποντας τη διερεύνηση τόσο της απόδοσης όσο και του κόστους κλιμάκωσης της μεθόδου.

Ακρίβεια (Accuracy)

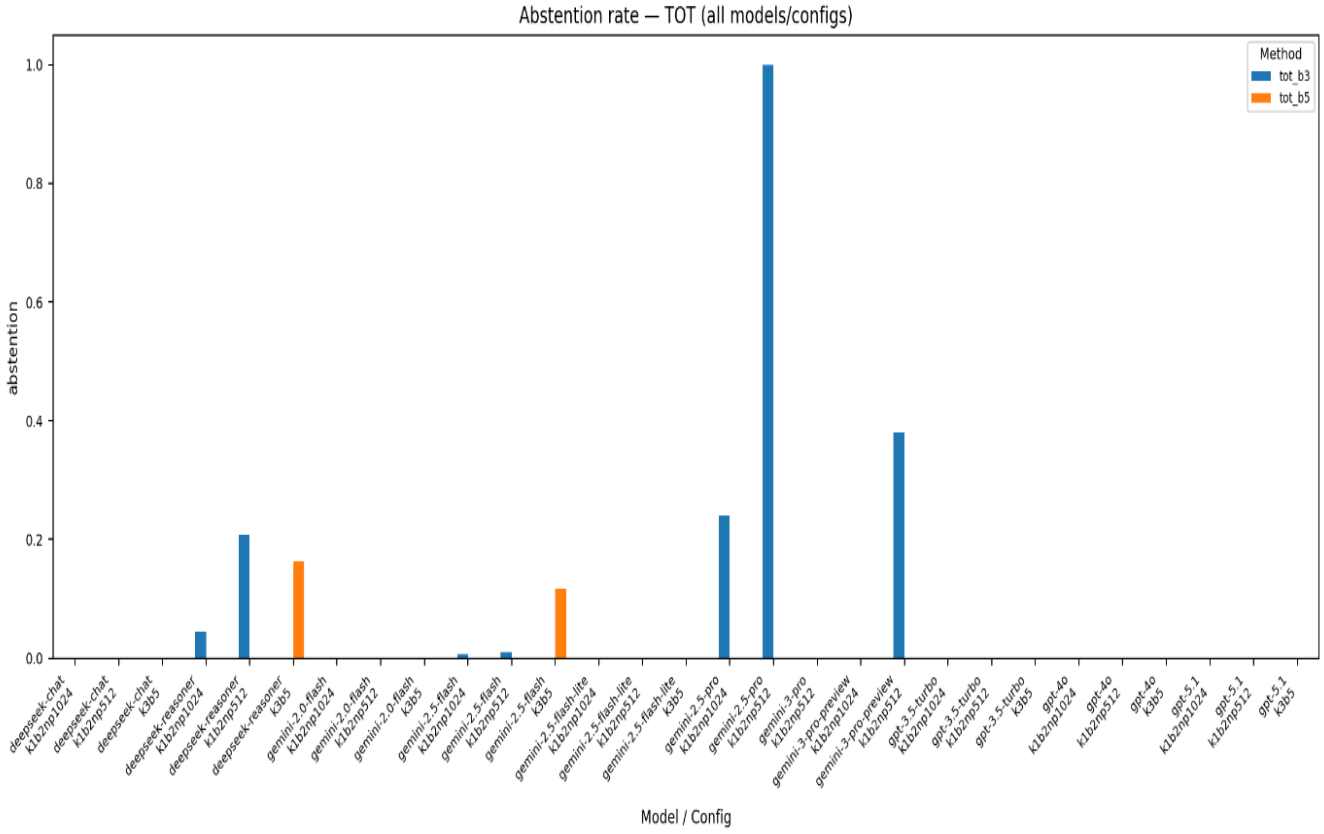
Η ακρίβεια αναδεικνύεται ως το αδιαμφισβήτητο πλεονέκτημα της μεθόδου ToT. Τα πειραματικά δεδομένα καταγράφουν εξαιρετικά υψηλές επιδόσεις, οι οποίες σε πολλές περιπτώσεις αγγίζουν ή υπερβαίνουν το 95%. Κορυφαία παραδείγματα αποτελούν τα μοντέλα gpt-5.1, deepseek-chat και gemini-2.0-flash (tot_b3/b5: ≈ 0.97). Τα αποτελέσματα αυτά επιβεβαιώνουν ότι η συστηματική εξερεύνηση εναλλακτικών σεναρίων ενισχύει δραστικά την ορθότητα της τελικής απόκρισης. Η σύγκριση μεταξύ tot_b3 και tot_b5 υποδεικνύει ότι η αύξηση του παράγοντα διακλάδωσης δύναται να προσφέρει οριακές αλλά μετρήσιμες βελτιώσεις, κυρίως σε μοντέλα υψηλής δυναμικότητας.



Εικόνα 24 Ακρίβεια (Accuracy) ToT

Ρυθμός αποχής (Abstention Rate)

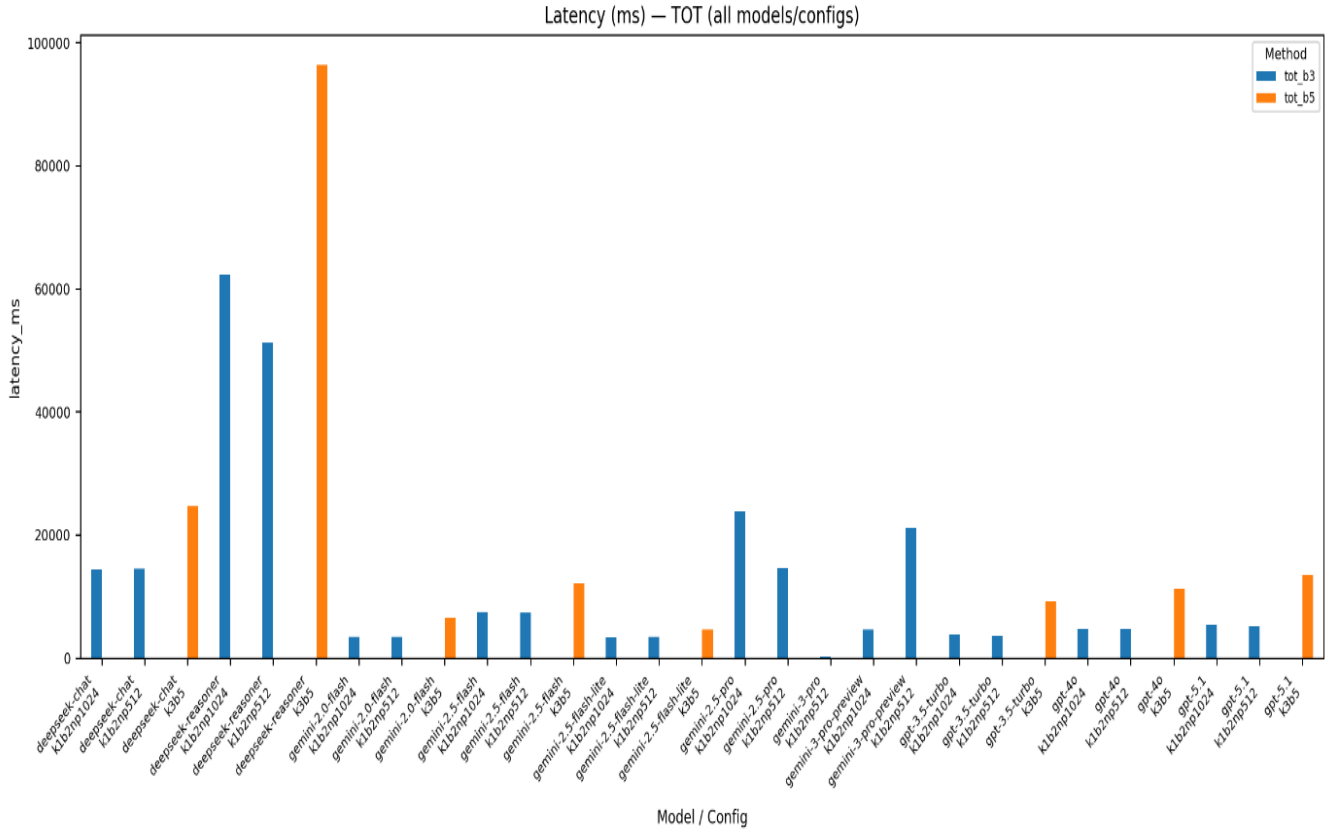
Η μέθοδος ToT διατηρεί γενικά χαμηλά ποσοστά αποχής για την πλειονότητα των μοντέλων, ένδειξη ότι η διερεύνηση πολλαπλών κλάδων μειώνει την αβεβαιότητα λήψης απόφασης. Ωστόσο, εντοπίζονται σοβαρές εξαιρέσεις που υποδηλώνουν αδυναμία σύγκλισης ή ασυμβατότητα σε συγκεκριμένες συνθήκες. Το πλέον χαρακτηριστικό παράδειγμα είναι το μοντέλο gemini-2.5-pro στη ρύθμιση (K=1, B=2, NP=512), όπου παρατηρείται πλήρης αποτυχία με ρυθμό αποχής 1.00 (100%). Υψηλά ποσοστά καταγράφονται επίσης στο gemini-3-pro-preview και το deepseek-reasoner. Τα ευρήματα αυτά καταδεικνύουν ότι, παρά τη θεωρητική ισχύ της, η αυξημένη πολυπλοκότητα της δενδρικής αναζήτησης μπορεί να οδηγήσει σε λειτουργικό αδιέξοδο όταν οι πόροι ή η ικανότητα του μοντέλου δεν επαρκούν.



Εικόνα 25 Ρυθμός Αποχής (Abstention Rate) ToT

Χρόνος απόκρισης (Latency)

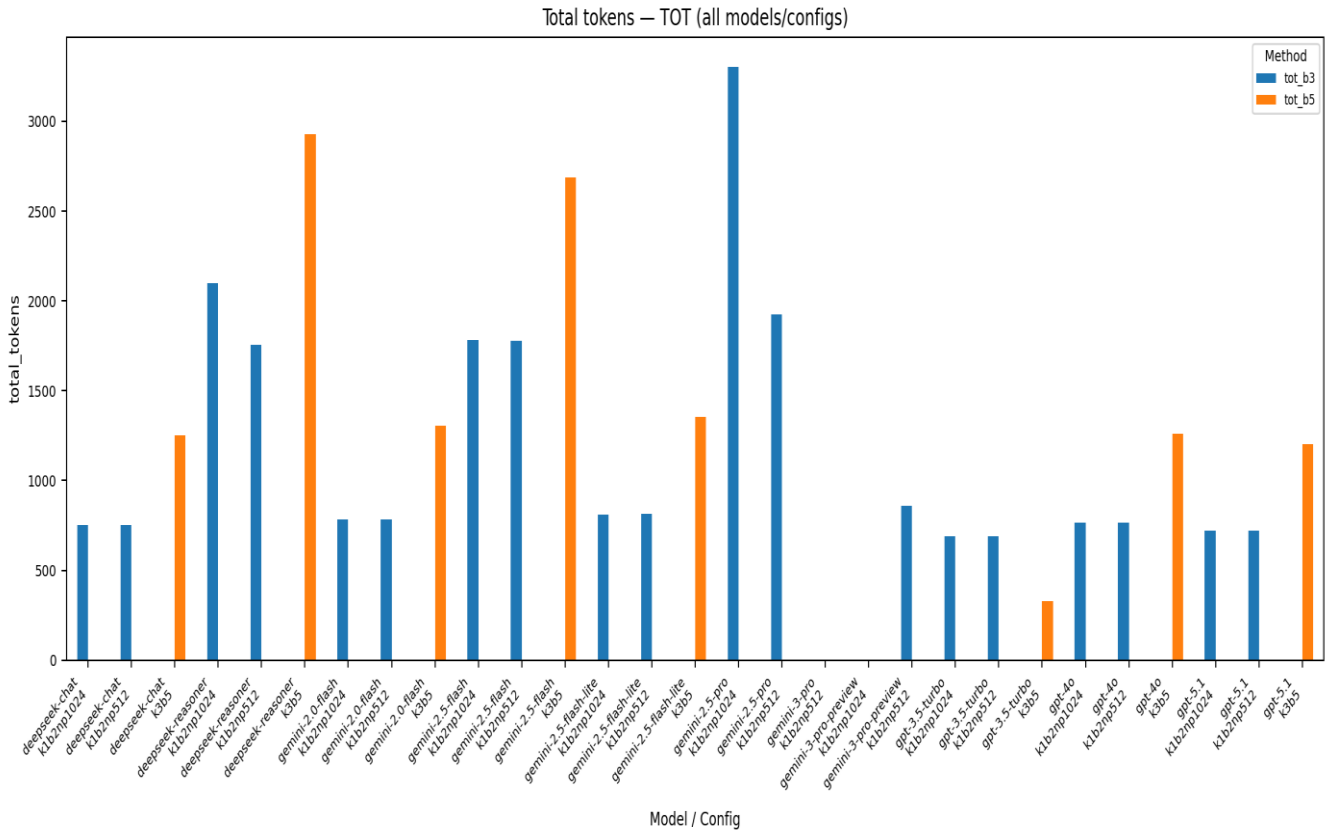
Η ποιοτική υπεροχή της ToT συνοδεύεται από εκθετική αύξηση του χρόνου απόκρισης. Οι απαιτήσεις για πολλαπλές παραγωγές κειμένου και επαναλαμβανόμενες αξιολογήσεις εκτοξεύουν τις τιμές, ιδιαίτερα στη ρύθμιση tot_b5. Το μοντέλο deepseek-reasoner καταγράφει τον υψηλότερο χρόνο με ~96.000 ms (σχεδόν 1.5 λεπτό ανά ερώτηση), ενώ ακόμη και στη ρύθμιση tot_b3 αγγίζει τα 62.000 ms. Αντίστοιχα υψηλές τιμές παρατηρούνται και στα deepseek-chat, gemini-2.5-pro και gemini-3-pro-preview. Είναι σαφές ότι η μέθοδος ToT καθίσταται απαγορευτική για εφαρμογές πραγματικού χρόνου, αποτελώντας λύση αποκλειστικά για σενάρια όπου η ακρίβεια ιεραρχείται ως η απόλυτη προτεραιότητα.



Εικόνα 26 Latency (Χρόνος Απόκρισης) ToT

Συνολική κατανάλωση Tokens (Total Tokens)

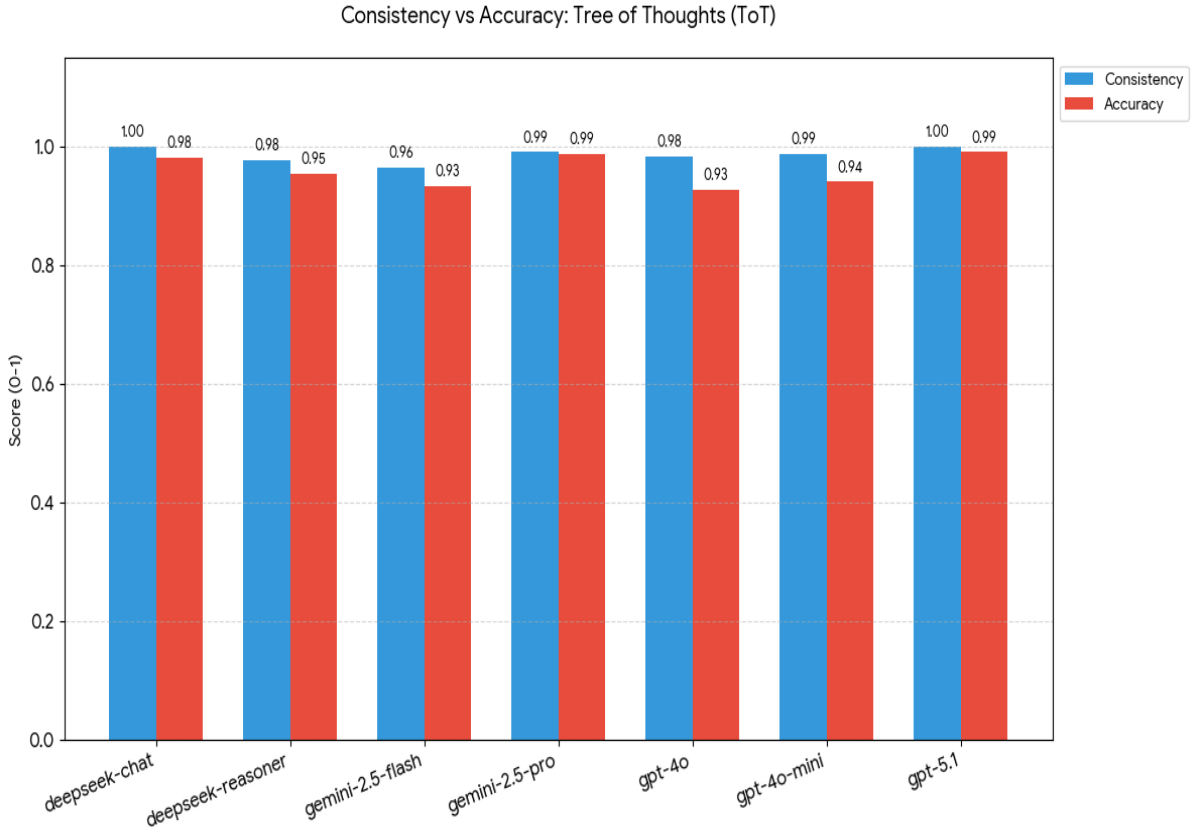
Η κατανάλωση tokens ακολουθεί την αυξητική τάση του χρόνου απόκρισης (αγγλ. latency), αντανακλώντας το πλήθος των εξεταζόμενων κλάδων και την έκταση των ενδιάμεσων συλλογισμών. Οι μέγιστες τιμές καταγράφονται στο gemini-2.5-pro (tot_b3: ~3.300 tokens), gemini-2.5-flash (tot_b5: ~2.700 tokens) και στο deepseek-reasoner (tot_b5: ~2.900 tokens). Η δραματική αυτή αύξηση επιβεβαιώνει ότι η ToT αποτελεί την πλέον κοστοβόρα υπολογιστικά μέθοδο, απαιτώντας σημαντικούς πόρους για κάθε μεμονωμένο ερώτημα.



Εικόνα 27 Συνολική Κατανάλωση Tokens (Total Tokens) ToT Prompting

Έλεγχος συνέπειας (Consistency Check)

Η ανάλυση της συνέπειας για τη μέθοδο ToT επιβεβαιώνει την ποιοτική υπεροχή της, καθώς καταγράφονται εξαιρετικά υψηλά ποσοστά που υπερβαίνουν το 96% για το σύνολο των εξεταζόμενων μοντέλων. Το εύρημα αυτό αποκτά βαρύνουσα σημασία όταν συνεξετάζεται με την ακρίβεια. Σε πλήρη αντιδιαστολή με τη μέθοδο Standard Prompting, όπου η σταθερότητα υποδήλωνε συχνά εμμονή στο σφάλμα, στην ToT η υψηλή συνέπεια συμβαδίζει απόλυτα με την ορθότητα. Η δομημένη διαδικασία της δενδρικής αναζήτησης λειτουργεί ως ισχυρός μηχανισμός σύγκλισης, καθώς η ενεργή απόρριψη των ασθενών κλάδων οδηγεί το σύστημα να καταλήγει συστηματικά στη βέλτιστη λύση, προσφέροντας έτσι επαληθεύσιμη σταθερότητα και ελαχιστοποιώντας τον στοχαστικό θόρυβο στις διαδοχικές εκτελέσεις.



Εικόνα 28 Συνέπεια (Consistency) Vs Ακρίβεια (Accuracy) ToT

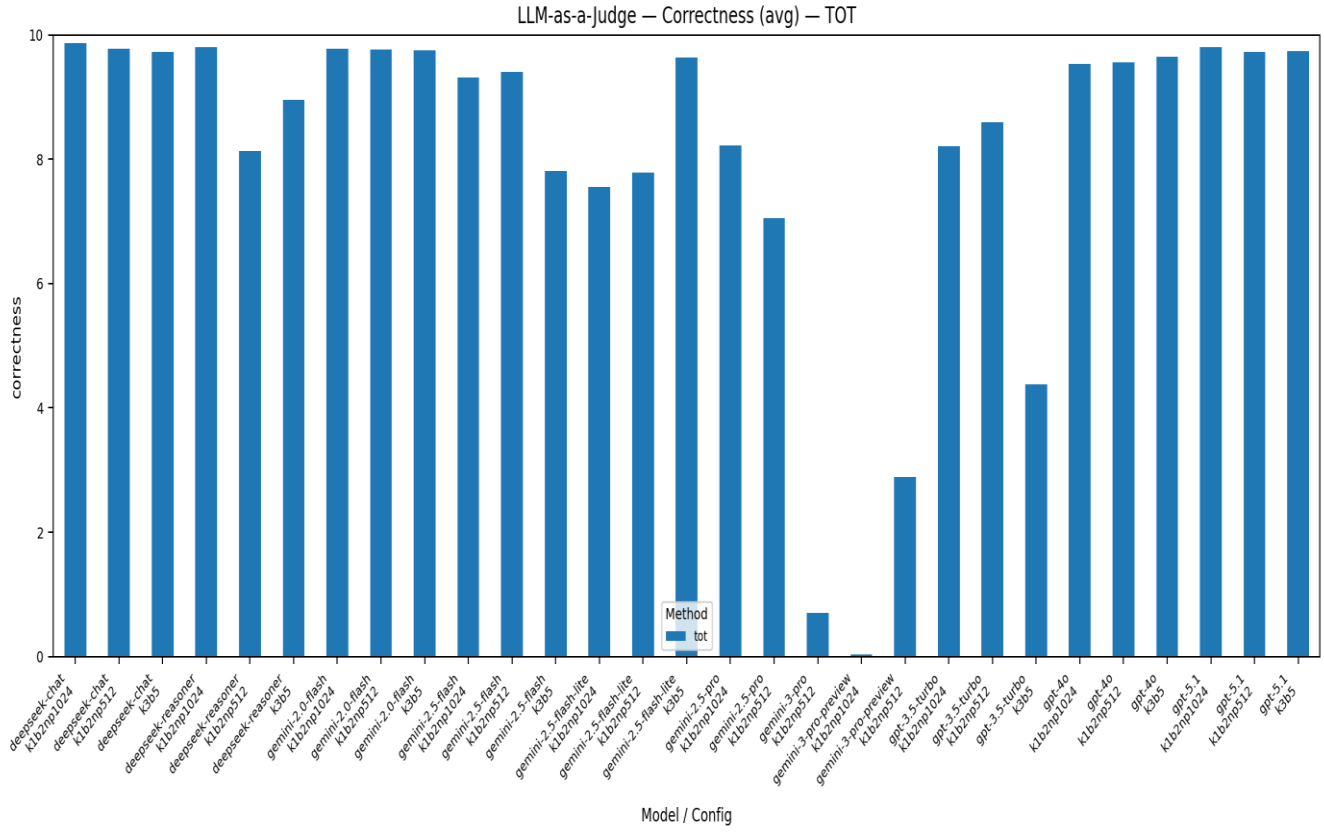
Πίνακας 19 Συγκεντρωτικά Αποτελέσματα ToT

Model	**Config**	**Method**	**Accuracy**	**Abstention**	**Latency (ms)**	**Total Tokens**
deepseek-chat	k1b2np1024	tot_b3	**0.96**	0.00	14,000	750
deepseek-chat	k1b2np512	tot_b3	**0.97**	0.00	14,200	750
deepseek-chat	k3b5	tot_b5	**0.97**	0.00	**24,500**	1,250
deepseek-reasoner	k1b2np1024	tot_b3	0.92	0.05	**62,000**	2,100
deepseek-reasoner	k1b2np512	tot_b3	0.60	**0.21**	**51,000**	1,750
deepseek-reasoner	k3b5	tot_b5	0.63	**0.16**	**96,000**	**2,900**
gemini-2.0-flash	k1b2np1024	tot_b3	**0.96**	0.00	3,500	780
gemini-2.0-flash	k1b2np512	tot_b3	0.95	0.00	3,600	780
gemini-2.0-flash	k3b5	tot_b5	**0.97**	0.00	6,800	1,300

gemini-2.5-flash	k1b2np1024	tot_b3	0.90	0.01	7,200	1,780
gemini-2.5-flash	k1b2np512	tot_b3	0.93	0.01	7,100	1,780
gemini-2.5-flash	k3b5	tot_b5	0.73	**0.12**	11,800	**2,700**
gemini-2.5-pro	k1b2np1024	tot_b3	0.56	**0.24**	**24,000**	**3,300**
gemini-2.5-pro	k1b2np512	tot_b3	0.00	**1.00**	14,500	1,920
gemini-3-pro-preview	k1b2np1024	tot_b3	0.02	0.00	4,500	0
gemini-3-pro-preview	k1b2np512	tot_b3	0.01	0.00	**21,000**	860
gpt-3.5-turbo	k1b2np1024	tot_b3	0.77	0.00	4,100	690
gpt-3.5-turbo	k1b2np512	tot_b3	0.76	0.00	4,000	690
gpt-3.5-turbo	k3b5	tot_b5	0.22	0.00	8,800	340
gpt-4o	k1b2np1024	tot_b3	0.91	0.00	4,800	760
gpt-4o	k1b2np512	tot_b3	0.90	0.00	4,900	760
gpt-4o	k3b5	tot_b5	0.89	0.00	11,000	1,250
gpt-5.1	k1b2np1024	tot_b3	**0.97**	0.00	5,500	720
gpt-5.1	k1b2np512	tot_b3	**0.96**	0.00	5,400	720
gpt-5.1	k3b5	tot_b5	**0.97**	0.00	13,000	1,200

Μέση βαθμολογία Correctness (LLM-as-a-Judge) για τη μέθοδο ToT ανά μοντέλο και διαμόρφωση.

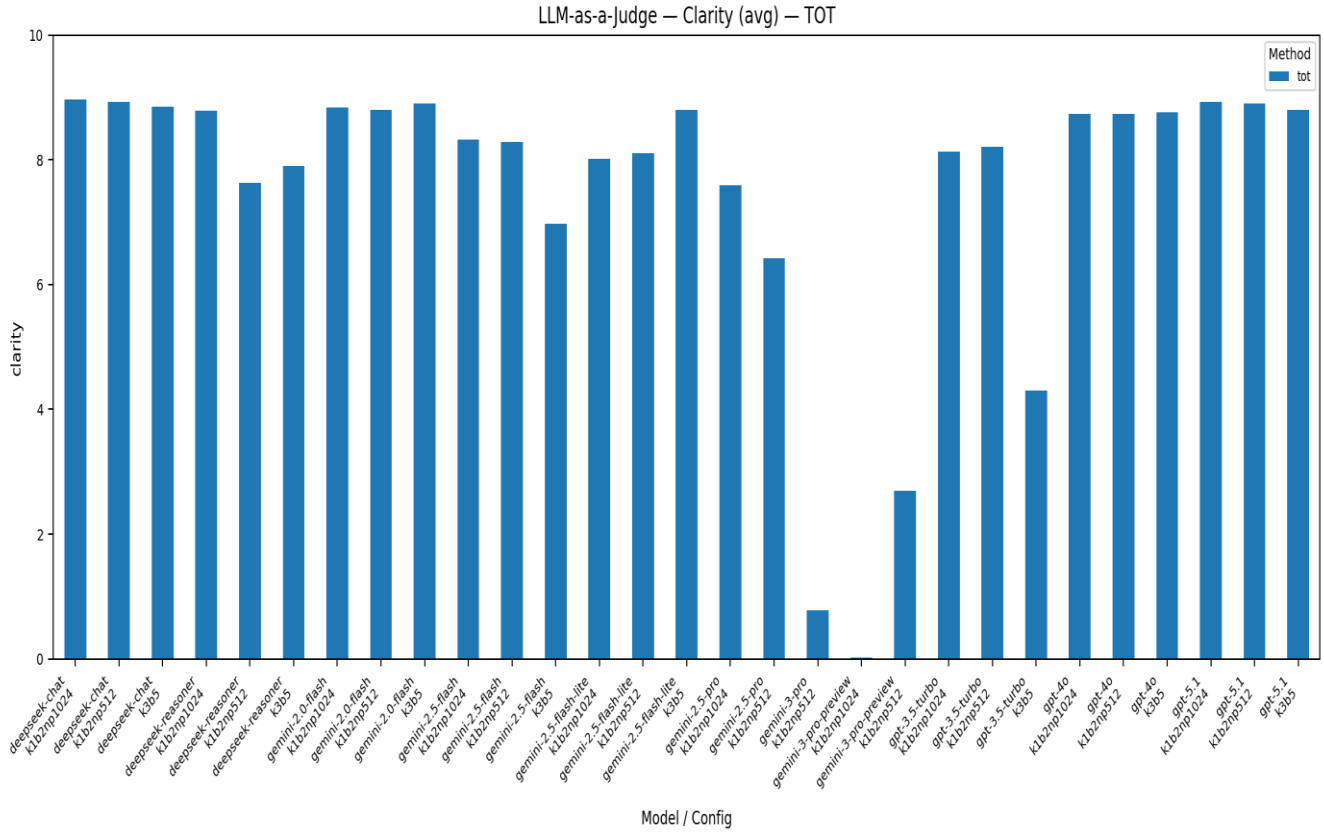
Η ToT εμφανίζει τις υψηλότερες συνολικά επιδόσεις στον άξονα της ορθότητας, με τα ισχυρότερα μοντέλα να προσεγγίζουν σταθερά το άριστα. Η συστηματική διερεύνηση πολλαπλών κλάδων συλλογισμού ενισχύει την πιθανότητα εύρεσης της βέλτιστης λύσης, οδηγώντας σε ιδιαίτερα υψηλές βαθμολογίες όταν η διαδικασία ολοκληρώνεται επιτυχώς. Ωστόσο, σε ορισμένες διαμορφώσεις παρατηρούνται έντονες πτώσεις, οι οποίες σχετίζονται με λειτουργικές αστοχίες ή αδυναμία σύγκλισης, επιβεβαιώνοντας ότι η μέγιστη ισχύς της μεθόδου συνοδεύεται από αυξημένη εκτελεστική ευαισθησία.



Εικόνα 29 Μέση βαθμολογία Correctness (LLM-as-a-Judge) για τη μέθοδο ToT ανά μοντέλο και διαμόρφωση.

Μέση βαθμολογία Clarity (LLM-as-a-Judge) για τη μέθοδο Tree-of-Thought (ToT) ανά μοντέλο και διαμόρφωση.

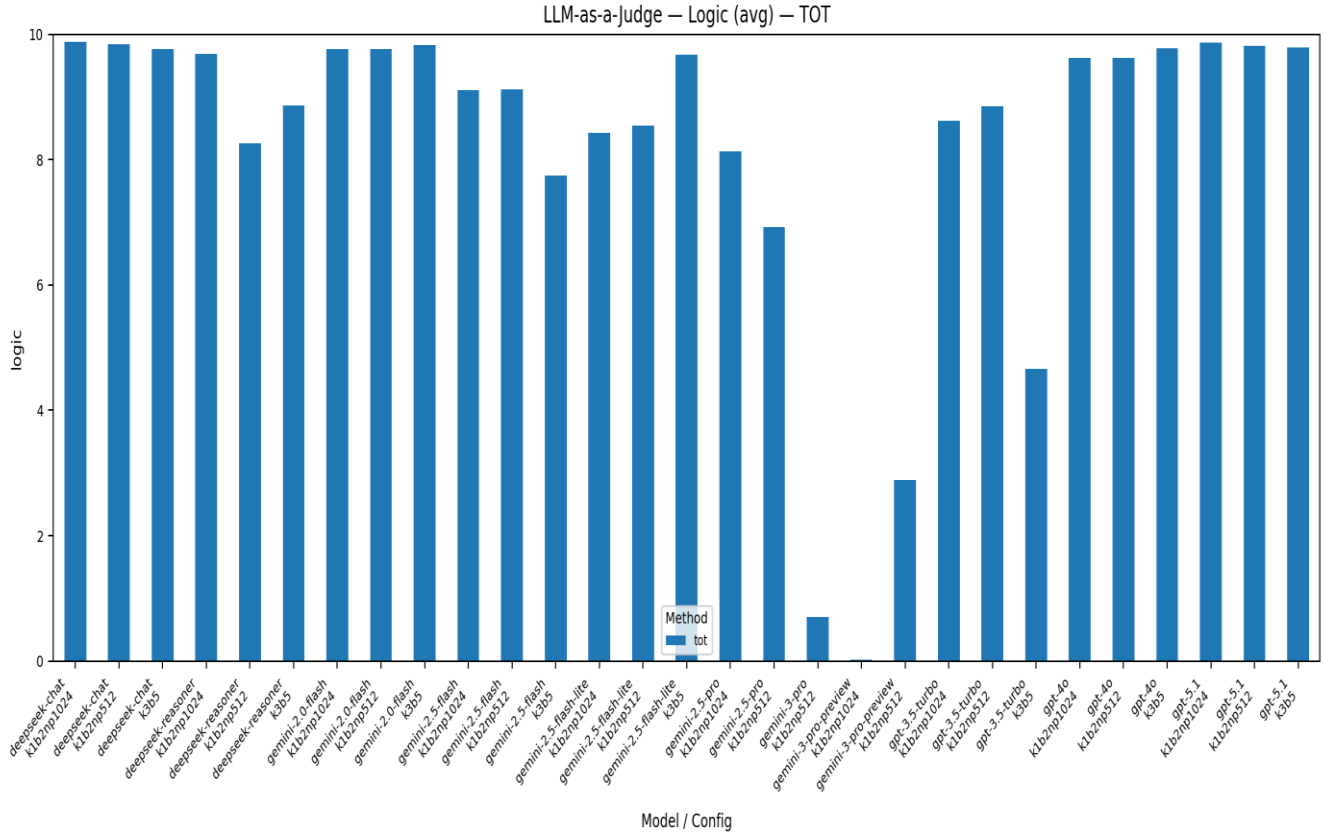
Η ToT διατηρεί υψηλά επίπεδα σαφήνειας στις περισσότερες διαμορφώσεις, με τις ισχυρότερες υλοποιήσεις να κινούνται κοντά στο 8,5–9/10. Η συστηματική ανάπτυξη πολλαπλών συλλογιστικών κλάδων συμβάλλει στη δομημένη και αναλυτική παρουσίαση της λύσης, ενισχύοντας την κατανόηση του τελικού αποτελέσματος. Ωστόσο, σε ορισμένες περιπτώσεις παρατηρούνται απότομες πτώσεις στη βαθμολογία, οι οποίες σχετίζονται με αστοχίες σύγκλισης ή υπέρμετρα εκτεταμένες απαντήσεις που μειώνουν τη συνολική αναγνωσιμότητα.



Εικόνα 30 Μέση βαθμολογία Clarity (LLM-as-a-Judge) για τη μέθοδο ToT ανά μοντέλο και διαμόρφωση.

Μέση βαθμολογία Logic (LLM-as-a-Judge) για τη μέθοδο ToT ανά μοντέλο και διαμόρφωση.

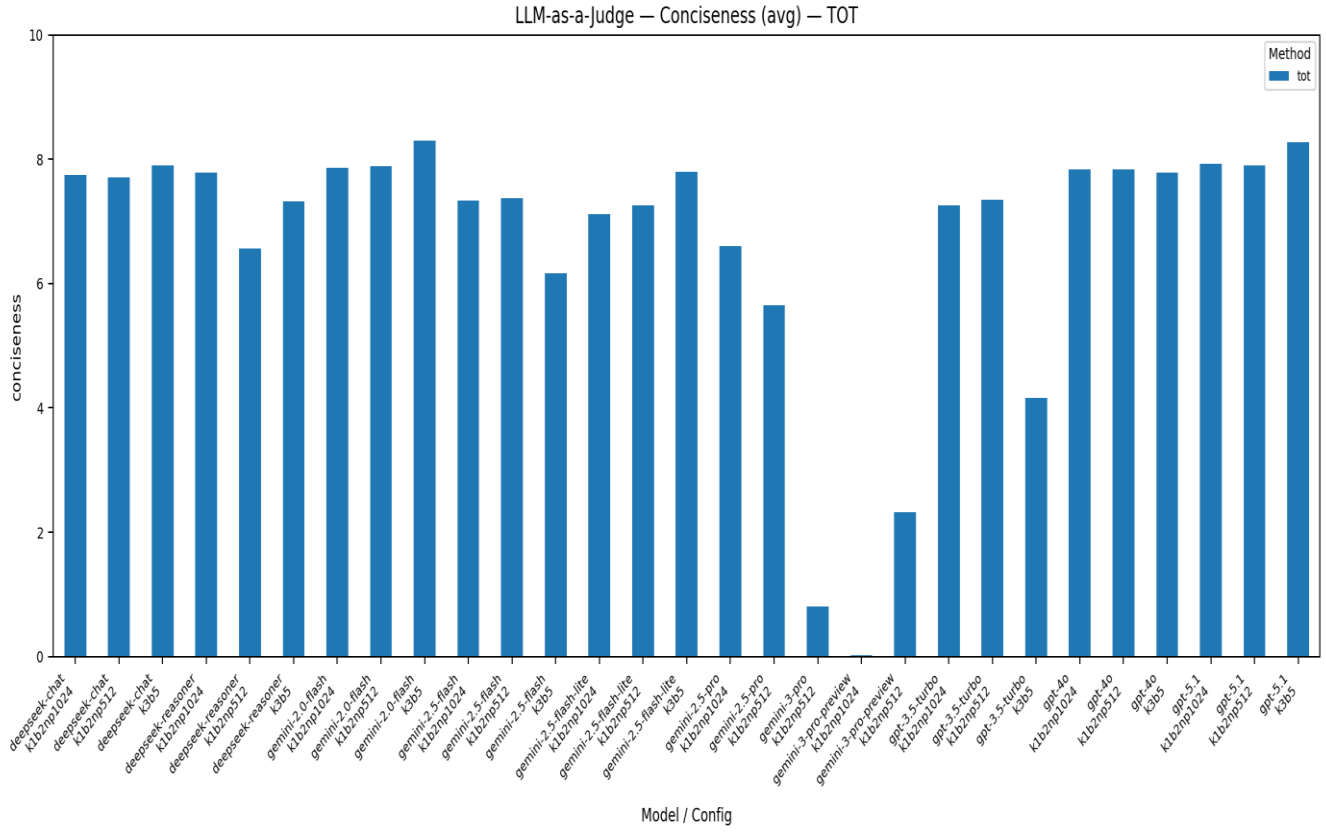
Η ToT επιτυγχάνει εξαιρετικά υψηλές βαθμολογίες λογικής συνοχής στις περισσότερες διαμορφώσεις, με πολλές τιμές να προσεγγίζουν το 9,5–10/10. Η δενδρική εξερεύνηση πολλαπλών κλάδων συλλογισμού και η αξιολόγηση ενδιάμεσων κόμβων ενισχύουν σημαντικά τη δομική ορθότητα των παραγόμενων απαντήσεων. Οι χαμηλές τιμές περιορίζονται σε μεμονωμένες περιπτώσεις αστοχίας ή ατελούς σύγκλισης, επιβεβαιώνοντας ότι, όταν η διαδικασία ολοκληρώνεται κανονικά, η ToT παράγει λύσεις υψηλής λογικής πληρότητας και συνέπειας.



Εικόνα 31 Μέση βαθμολογία Logic (LLM-as-a-Judge) για τη μέθοδο ToT ανά μοντέλο και διαμόρφωση.

Μέση βαθμολογία Conciseness (LLM-as-a-Judge) για τη μέθοδο Tree-of-Thought (ToT) ανά μοντέλο και διαμόρφωση.

Η ToT εμφανίζει ικανοποιητικές αλλά σαφώς χαμηλότερες επιδόσεις στον άξονα της συντομίας σε σχέση με τους δείκτες ορθότητας και λογικής. Οι περισσότερες διαμορφώσεις κυμαίνονται σε μεσαίες–υψηλές τιμές (περίπου 7–8/10), ωστόσο παρατηρείται μεγαλύτερη διακύμανση και ορισμένες έντονες αποκλίσεις σε περιπτώσεις αστοχίας. Η τάση αυτή επιβεβαιώνει ότι η εκτενής δενδρική εξερεύνηση οδηγεί σε πιο αναλυτικές και συχνά μακροσκελείς απαντήσεις, θυσιάζοντας τη συντομία προς όφελος της πληρότητας και της συλλογιστικής τεκμηρίωσης.



Εικόνα 32 Μέση βαθμολογία Conciseness (LLM-as-a-Judge) για τη μέθοδο ToT ανά μοντέλο και διαμόρφωση.

Πίνακας 20 LLM-as-a-Judge για τη μέθοδο ToT

Model	**Config**	**Correct.**	**Clarity**	**Logic**	**Concise**	**MO**	**Παρατήρηση**
deepseek-chat	k1b2np1024	9.86	8.96	9.87	7.74	**9.11**	Εξαιρετική απόδοση
deepseek-chat	k1b2np512	9.77	8.92	9.84	7.70	**9.06**	Εξαιρετική απόδοση
deepseek-chat	k3b5	9.72	8.85	9.76	7.90	**9.06**	Εξαιρετική απόδοση
deepseek-reasoner	k1b2np1024	9.80	8.78	9.69	7.78	**9.01**	Εξαιρετική απόδοση
deepseek-reasoner	k1b2np512	8.13	7.62	8.26	6.55	**7.64**	Πολύ καλή απόδοση

deepseek-reasoner	k3b5	8.95	7.89	8.86	7.32	**8.25**	Πολύ καλή απόδοση
gemini-2.0-flash	k1b2np1024	9.77	8.84	9.76	7.86	**9.06**	Εξαιρετική απόδοση
gemini-2.0-flash	k1b2np512	9.76	8.80	9.76	7.88	**9.05**	Εξαιρετική απόδοση
gemini-2.0-flash	k3b5	9.75	8.89	9.82	8.29	**9.19**	Εξαιρετική απόδοση
gemini-2.5-flash	k1b2np1024	9.31	8.32	9.11	7.34	**8.52**	Πολύ καλή απόδοση
gemini-2.5-flash	k1b2np512	9.40	8.28	9.12	7.37	**8.54**	Πολύ καλή απόδοση
gemini-2.5-flash	k3b5	7.81	6.97	7.75	6.16	**7.17**	Πολύ καλή απόδοση
gemini-2.5-flash-lite	k1b2np1024	7.55	8.02	8.42	7.11	**7.77**	Πολύ καλή απόδοση
gemini-2.5-flash-lite	k1b2np512	7.78	8.10	8.54	7.26	**7.92**	Πολύ καλή απόδοση
gemini-2.5-flash-lite	k3b5	9.63	8.79	9.67	7.80	**8.98**	Πολύ καλή απόδοση
gemini-2.5-pro	k1b2np1024	8.22	7.58	8.13	6.60	**7.63**	Πολύ καλή απόδοση
gemini-2.5-pro	k1b2np512	7.05	6.42	6.92	5.65	**6.51**	Μέτρια απόδοση
gemini-3-pro	k1b2np512	0.70	0.78	0.70	0.80	**0.75**	Πολύ χαμηλή απόδοση
gemini-3-pro-preview	k1b2np1024	0.03	0.02	0.02	0.02	**0.02**	Πολύ χαμηλή απόδοση
gemini-3-pro-preview	k1b2np512	2.88	2.69	2.88	2.32	**2.69**	Πολύ χαμηλή απόδοση
gpt-3.5-turbo	k1b2np1024	8.20	8.13	8.62	7.25	**8.05**	Πολύ καλή απόδοση
gpt-3.5-turbo	k1b2np512	8.59	8.21	8.85	7.35	**8.25**	Πολύ καλή απόδοση
gpt-3.5-turbo	k3b5	4.37	4.30	4.66	4.16	**4.37**	Χαμηλή απόδοση
gpt-4o	k1b2np1024	9.53	8.73	9.61	7.83	**8.93**	Πολύ καλή απόδοση
gpt-4o	k1b2np512	9.56	8.74	9.62	7.83	**8.94**	Πολύ καλή απόδοση
gpt-4o	k3b5	9.64	8.76	9.77	7.78	**8.99**	Πολύ καλή απόδοση
gpt-5.1	k1b2np1024	9.80	8.93	9.86	7.92	**9.13**	Εξαιρετική απόδοση
gpt-5.1	k1b2np512	9.73	8.89	9.81	7.89	**9.08**	Εξαιρετική απόδοση

gpt-5.1	k3b5	9.74	8.80	9.78	8.27	**9.15**	Εξαιρετική απόδοση

6.5 Ανάλυση μέγιστης ακρίβειας ανά συνδυασμό μοντέλου-μεθόδου (GSM8K)

Στον Πίνακα 21 παρουσιάζονται οι κορυφαίες επιδόσεις ακρίβειας ανά συνδυασμό μοντέλου και μεθόδου στο GSM8K.

Πίνακας 21 Μέγιστη Ακρίβεια ανά Συνδυασμό Μοντέλου-Μεθόδου (GSM8K)

Μοντέλο	**Μέθοδος**	**Ακρίβεια (%)**
gemini-2.0-flash	ToT (b=5)	**97.67**
deepseek-chat	ToT (b=5)	97.33
gpt-5.1	CoT (k=1)	97.17
gpt-5.1	CoT (k=3)	97.00
gemini-2.0-flash	CoT (k=3)	97.00

6.6 Ανάλυση ακρίβειας, χρόνου και αξιοπιστίας

Οι τρεις αυτές παράμετροι αποτελούν βασικά κριτήρια αξιολόγησης συστημάτων συλλογισμού βασισμένων σε LLMs, καθώς αποτυπώνουν τόσο την ποιοτική όσο και τη λειτουργική τους συμπεριφορά.

Ανάλυση ακρίβειας

Η ανάλυση της ακρίβειας καταδεικνύει μια σαφή ιεράρχηση των μεθόδων, η οποία αντικατοπτρίζει το βάθος της συλλογιστικής διαδικασίας που η καθεμία επιβάλλει. Η μέθοδος Standard Prompting καταγράφει τις χαμηλότερες και πιο αστάθμητες επιδόσεις, με έναν Προσαρμοσμένο Μέσο Όρο της τάξης του 67% και τεράστιο εύρος διακύμανσης (29% – 94%). Η μέθοδος αυτή εξαρτάται απόλυτα από την «εγγενή ευφυΐα» του μοντέλου, αποτυγχάνοντας συστηματικά σε προβλήματα που απαιτούν ενδιάμεσα βήματα. Είναι χαρακτηριστικό ότι ακόμη και κορυφαία μοντέλα (π.χ. GPT-4o, GPT-5.1) κυμάνθηκαν σε μεσαία επίπεδα (~54%) υπό τη συνθήκη Standard, υπογραμμίζοντας ότι η υπολογιστική ισχύς από μόνη της δεν εγγυάται την άμεση επίλυση χωρίς βηματική καθοδήγηση. Εξαίρεση αποτελούν συγκεκριμένα μοντέλα νεότερης γενιάς (π.χ. DeepSeek Chat, Gemini 2.5 Flash), που κατόρθωσαν να επιτύχουν υψηλή ακρίβεια (>90%).

Η μετάβαση στη ρητή βηματική συλλογιστική (CoT Standard) επιφέρει την πιο ουσιαστική βελτίωση, εκτοξεύοντας τον Προσαρμοσμένο Μέσο Όρο στο 87,9% (εύρος 0% – 98%). Η μέθοδος αυτή σταθεροποιεί την απόδοση ακόμη και σε μεσαίας δυναμικότητας μοντέλα, περιορίζοντας τα λογικά άλματα. Η εφαρμογή της Αυτο-Συνέπειας (SC CoT, k=3) ενισχύει περαιτέρω την ακρίβεια, φτάνοντας το 90,25% ως Προσαρμοσμένο Μέσο Όρο και αγγίζοντας το +97% σε ισχυρά μοντέλα, λειτουργώντας ως φίλτρο για την εξάλειψη τυχαίων λαθών. Ωστόσο, σε ασθενέστερα μοντέλα, η πολυπλοκότητα διαχείρισης πολλαπλών μονοπατιών μπορεί να οδηγήσει σε μείωση της ακρίβειας.

Τέλος, η μέθοδος ToT επιτυγχάνει σταθερά υψηλές επιδόσεις, με Προσαρμοσμένο Μέσο Όρο 86,8% για την tot_b3 και 86,0% για την tot_b5, προσφέροντας την ανώτατη δυνατή διασφάλιση ορθότητας στις λειτουργικές διαμορφώσεις. Παρόλα αυτά, η υπεροχή της έναντι της SC CoT δεν είναι πάντα

ποσοτική σε απόλυτα νούμερα ακρίβειας, αλλά ποιοτική, καθώς συνοδεύεται από αυστηρότερη τεκμηρίωση και εξάλειψη λαθών σε βέλτιστες εκτελέσεις.

Ανάλυση χρόνου απόκρισης

Η εξέταση του χρόνου απόκρισης αναδεικνύει το αναπόφευκτο τίμημα της ακρίβειας. Η Standard Prompting παραμένει η ταχύτερη μέθοδος με μέσο χρόνο περίπου 3.046 ms, προσφέροντας άμεση απόκριση, ιδανική για εφαρμογές που ο χρόνος εκτέλεσης είναι κρίσιμος.

Η CoT εισάγει μια λογική καθυστέρηση με Προσαρμοσμένο Μέσο Όρο περίπου 4.740 ms, η οποία είναι απολύτως βιώσιμη για τις περισσότερες εφαρμογές. Η SC CoT, λόγω της απαίτησης για παραγωγή και αξιολόγηση πολλαπλών δειγμάτων, αυξάνει σημαντικά τον χρόνο, με τον Προσαρμοσμένο Μέσο Όρο να φτάνει τα 9.700 ms, καθιστώντας την λιγότερο κατάλληλη για εφαρμογές άμεσης απόκρισης (low-latency).

Η ToT καθίσταται η πλέον χρονοβόρα, με τον Προσαρμοσμένο Μέσο Όρο να κυμαίνεται από 11.370 ms (tot_b3) έως 12.650 ms (tot_b5), και ακραίες τιμές που σε reasoning μοντέλα υπερβαίνουν το 1.5 λεπτό, καθιστώντας την απαγορευτική για εφαρμογές πραγματικού χρόνου, περιορίζοντας τη χρήση της σε σενάρια κρίσιμης σημασίας ή offline επεξεργασίας, όπου η ορθότητα του αποτελέσματος ιεραρχείται υψηλότερα από την ταχύτητα.

Ανάλυση αξιοπιστίας

Η έννοια της αξιοπιστίας δεν ορίζεται μονοσήμαντα ως η ικανότητα επανάληψης μιας απάντησης, αλλά ως μια σύνθετη παράμετρος που προκύπτει από τη συνδυαστική εξέταση τριών δεικτών:

- της σταθερότητας στις διαδοχικές εκτελέσεις (αγγλ. Consistency Check),
- της ικανότητας του συστήματος να αναγνωρίζει την αδυναμία απάντησης (αγγλ. Abstention Rate)
- και της ποιοτικής αξιολόγησης της λογικής δομής (αγγλ. LLM-as-a-Judge scores).

Στην περίπτωση της Standard Prompting, η αξιοπιστία κρίνεται χαμηλή και συχνά πλασματική. Παρά την υψηλή τυπική συνέπεια (>95%), ο συνδυασμός χαμηλών Judge Scores (6,40) και σχεδόν μηδενικής αποχής (1,9%) υποδηλώνει ότι το μοντέλο τείνει να αναπαράγει λανθασμένες απαντήσεις με υψηλή αυτοπεποίθηση, χωρίς να αναγνωρίζει την αδυναμία επίλυσης.

Αντιθέτως, οι μέθοδοι CoT και SC-CoT προσφέρουν αναβαθμισμένη αξιοπιστία μέσω της λογικής τεκμηρίωσης. Η SC-CoT επιτυγχάνει την κορυφαία ποιοτική αξιολόγηση με Judge Score 9,00, το οποίο αποτελεί την υψηλότερη τιμή μεταξύ όλων των μεθόδων. Η επίδοση αυτή οφείλεται στη στατιστική σύγκλιση πολλαπλών μονοπατιών σκέψης, η οποία εξαλείφει τον στοχαστικό θόρυβο και διασφαλίζει ότι η τελική απάντηση υποστηρίζεται από πλειοψηφική συναίνεση.

Η μέθοδος ToT (tot_b3), παρότι ακολουθεί με ελαφρώς χαμηλότερο Judge Score (8,94), εισάγει μια κρίσιμη παράμετρο αξιοπιστίας: τη δομική δικλείδα ασφαλείας. Μέσω της ενεργής αξιολόγησης των ενδιάμεσων κόμβων και των μη-μηδενικών ρυθμών αποχής, η ToT επιδεικνύει μια μορφή «γνωστικής μετριοπάθειας», προτιμώντας τη μη-απόκριση έναντι της παραγωγής εσφαλμένου αποτελέσματος. Συνεπώς, ενώ η SC-CoT αναδεικνύεται ως η πλέον αξιόπιστη μέθοδος ως προς την ποιοτική σταθερότητα της απάντησης, η ToT προσφέρει την υψηλότερη επαληθεύσιμη αξιοπιστία σε επίπεδο διαδικασίας, αποτρέποντας αποτελεσματικότερα τις λογικές παραισθήσεις.

Συνολική σύνθεση και ισοζύγια απόδοσης (Trade-offs)

Η διεξοδική συγκριτική ανάλυση των πειραματικών δεδομένων οδηγεί στο συμπέρασμα ότι δεν υφίσταται μια καθολικά βέλτιστη μεθοδολογία για την επίλυση μαθηματικών προβλημάτων μέσω

LLMs. Αντιθέτως, κάθε προσέγγιση καταγράφει σαφή πλεονεκτήματα και μειονεκτήματα λαμβάνοντας υπόψη τους παράγοντες «Ακρίβεια - Χρόνος - Αξιοπιστία», δημιουργώντας ισοζύγια που πρέπει να σταθμίζονται ανάλογα με την εφαρμογή. Συγκεκριμένα, το λειτουργικό προφίλ κάθε μεθόδου διαμορφώνεται ως εξής:

- **Standard Prompting:** Τοποθετείται στο άκρο της μέγιστης ταχύτητας και του ελάχιστου κόστους (Χρόνος απόκρισης ~3s, Πρ. Μ.Ο Ακρίβειας ~67%). Ωστόσο, η χαμηλή ποιοτική αξιολόγηση (Judge Score 6,40) και η αδυναμία διαχείρισης σύνθετων προβλημάτων σε μοντέλα χωρίς εγγενές reasoning, την καθιστούν κατάλληλη μόνο για απλούστερες εργασίες ή ως βάση αναφοράς για μοντέλα με υψηλή εσωτερική συλλογιστική ικανότητα.
- **CoT Standard:** Συνιστά τη βέλτιστη τομή μεταξύ απόδοσης και πόρων. Προσφέρει την πλέον ισορροπημένη σχέση ακρίβειας (Πρ. Μ.Ο: 87,9%) και χρόνου απόκρισης (~4,7s), καθιστώντας την την ενδεδειγμένη επιλογή για γενικές εφαρμογές που απαιτούν αξιόπιστο συλλογισμό (Judge Score 7,72) χωρίς υπερβολική υπολογιστική καθυστέρηση.
- **SC CoT:** Ενισχύει καθοριστικά την αξιοπιστία και τη σταθερότητα του συστήματος μέσω της πλειοψηφικής επιλογής. Επιτυγχάνει την υψηλότερη ποιοτική βαθμολογία (Judge Score 9,00) με μέτριο κόστος χρόνου (~9,7s). Η χρήση της κρίνεται απολύτως δικαιολογημένη όταν ζητούμενο είναι η ελαχιστοποίηση της στοχαστικής μεταβλητότητας και η μεγιστοποίηση της εμπιστοσύνης στο αποτέλεσμα.
- **ToT (b3):** Επιτυγχάνει εξαιρετικά υψηλή ακρίβεια (Πρ. Μ.Ο: 86,8%, με Q3 στο 96%) και δομική τεκμηρίωση (Judge Score 8,94), θυσιάζοντας ωστόσο την ταχύτητα (Πρ. Μ.Ο.: ~11,4s). Το κόστος της θεωρείται αυξημένο αλλά αποδεκτό για εφαρμογές που απαιτούν βαθύ συλλογισμό, επαληθεύσιμα βήματα και δεν υπόκεινται σε αυστηρούς χρονικούς περιορισμούς.
- **ToT (b5):** Αποτελεί το όριο της μέγιστης ποιότητας συλλογισμού και της δομικής πληρότητας (Judge Score 8,69), προσφέροντας οριακές βελτιώσεις στην ακρίβεια (Πρ. Μ.Ο: 86%, με Q3 = 97%) σε ισχυρές αρχιτεκτονικές. Ωστόσο, η διεύρυνση του παράγοντα διακλάδωσης συνεπάγεται δυσανάλογη αύξηση του κόστους, με τον χρόνο απόκρισης να εκτοξεύεται (Πρ. Μ.Ο: ~12,7s) και την κατανάλωση πόρων να αυξάνεται σημαντικά (~1.340 tokens). Λόγω της αυξημένης υπολογιστικής πολυπλοκότητας και της ευθραυστότητας σε λιγότερο ικανά μοντέλα, η χρήση της περιορίζεται αποκλειστικά σε εξειδικευμένα σενάρια υψηλού ρίσκου, όπου η ορθότητα είναι αδιαπραγμάτευτη και οι πόροι απεριόριστοι.

Συμπερασματικά, η ανάλυση αυτή επιβεβαιώνει τον κεντρικό στόχο της παρούσας εργασίας, ότι δεν υπάρχει μία μοναδική βέλτιστη στρατηγική συλλογισμού για κάθε περίσταση. Η επιλογή της κατάλληλης μεθόδου είναι μια δυναμική διαδικασία, η οποία εξαρτάται άμεσα από τους περιορισμούς της κάθε εφαρμογής. Ως εκ τούτου, απαιτείται πάντοτε η εξισορρόπηση ανάμεσα στην ανάγκη για υψηλή ακρίβεια, στους διαθέσιμους υπολογιστικούς πόρους και τον χρόνο εκτέλεσης.

6.7 Ανάλυση αιτιών αποχής και αστοχιών (Failure Analysis)

Πέραν της ποσοτικής αποτίμησης της ακρίβειας, του χρόνου εκτέλεσης και της αξιοπιστίας, κρίσιμη συνιστώσα για την αξιολόγηση της ικανότητας των LLMs αποτελεί και η διερεύνηση των περιπτώσεων όπου τα μοντέλα αδυνατούν να παράγουν έγκυρη απάντηση. Η ανάλυση των αρχείων καταγραφής (logs) ανέδειξε ότι ο δείκτης Abstention Rate δεν αποτελεί απλώς ένδειξη αβεβαιότητας, αλλά αποκαλύπτει δομικούς και λειτουργικούς περιορισμούς των εξεταζόμενων συστημάτων. Συγκεκριμένα, τα ευρήματα κατηγοριοποιούνται σε τρεις διακριτές δέσμες αιτιών:

6.7.1 Καθολική αποτυχία (Abstention Rate = 1.00)

Η πλέον αξιοσημείωτη παρατήρηση αφορά την περίπτωση καθολικής αποτυχίας του μοντέλου Gemini 2.5 Pro στη διαμόρφωση με περιορισμένο παράθυρο κειμένου (Context Size = 512 tokens), όπου τόσο στη μέθοδο ToT όσο και στην CoT κατέγραψε ποσοστό αποχής 100%.

Η εξέταση των πρωτογενών αποκρίσεων (αγγλ. raw_responses) κατέδειξε ότι η αιτία δεν ήταν γνωσιακή ανεπάρκεια, αλλά τεχνική ασφυξία λόγω του ορίου των tokens. Οι μέθοδοι CoT και ToT απαιτούν εκτενή παραγωγή ενδιάμεσων συλλογισμών, οι οποίοι συχνά υπερβαίνουν τα 512 tokens. Όταν το όριο τίθεται στα 512 tokens, η διαδικασία παραγωγής διακόπτεται βίαια πριν την ολοκλήρωση του συλλογισμού (π.χ. η απόκριση σταματά στη φράση "Here is the step-by..." ή "FINAL: N/A"). Ως αποτέλεσμα, οι μηχανισμοί εξαγωγής απάντησης αδυνατούν να εντοπίσουν τελικό αποτέλεσμα, οδηγώντας το σύστημα σε κατάσταση μόνιμης αποχής.

Το συμπέρασμα ότι η αστοχία οφείλεται αποκλειστικά στον περιορισμό πόρων επιβεβαιώνεται και από τα συγκριτικά αποτελέσματα του ελέγχου συνέπειας (αγγλ. consistency check). Στις δοκιμές αυτές, όπου το παράθυρο πλαισίου αυξήθηκε στα 2048 tokens, το ίδιο ακριβώς μοντέλο (gemini-2.5-pro) επέτυχε εξαιρετικά υψηλή απόδοση, εκμηδενίζοντας πλήρως το ποσοστό αποχής (Abstention Rate ≈ 0.00). Η αντιπαραβολή αυτή αποδεικνύει στην πράξη ότι οι προηγμένες τεχνικές συλλογισμού είναι λειτουργικά βιώσιμες μόνο όταν συνοδεύονται από επαρκείς πόρους παραγωγής.

6.7.2 Περιορισμοί πόρων και API

Μια δεύτερη κατηγορία αφορά περιπτώσεις με εξαιρετικά υψηλό ρυθμό αποχής (≈89%), όπως παρατηρήθηκε στο μοντέλο Gemini 3 Pro Preview. Σε αντίθεση με την προηγούμενη περίπτωση, εδώ η αποτυχία δεν οφείλεται σε εσωτερικό περιορισμό του μοντέλου, αλλά σε εξωτερικούς περιορισμούς της υποδομής διεπαφής (αγγλ. API limits). Η εφαρμογή της μεθόδου ToT, η οποία πολλαπλασιάζει τις κλήσεις προς το μοντέλο (λόγω των παραγόντων διακλάδωσης b=3, b=5), οδήγησε σε ταχεία εξάντληση των ορίων χρήσης (Rate Limiting), επιστρέφοντας σφάλματα τύπου "429 - Resource Exhausted". Το γεγονός αυτό υπογραμμίζει το υψηλό "κρυφό κόστος" των μεθόδων δενδρικής αναζήτησης: η αύξηση της ακρίβειας μέσω ToT απαιτεί εκθετικά διαθέσιμους πόρους, καθιστώντας τη μέθοδο ευάλωτη σε περιβάλλοντα με αυστηρούς περιορισμούς χρήσης (αγγλ. quotas).

6.7.3 Γλωσσική ασάφεια και λογικά αδιέξοδα

Τέλος, πέρα από τους τεχνικούς περιορισμούς, εντοπίστηκε ένα σύνολο προβλημάτων με ασυνήθιστη ή στρυφνή διατύπωση, που οδήγησαν σε αποχή λόγω σημασιολογικής πολυπλοκότητας ακόμη και όταν οι πόροι ήταν επαρκείς. Χαρακτηριστικό παράδειγμα αποτελεί το πρόβλημα test-209:

"Twenty dozen cups cost $1200 less than the total cost of half a dozen plates sold at $6000 each. Calculate the total cost of buying each cup."

Το συγκεκριμένο πρόβλημα απέτυχε συστηματικά τόσο σε μοντέλα της οικογένειας Gemini όσο και της DeepSeek. Στην περίπτωση αυτή, εντοπίστηκε μια γνωσιακή σύγκρουση. Η λέξη "each" επιβάλλει συντακτικά τον πολλαπλασιασμό, οδηγώντας σε αποτέλεσμα που αντιβαίνει στην κοινή λογική ($36.000 για 6 πιάτα). Οι μηχανισμοί αναζήτησης της ToT, παράγοντας διαφορετικούς κλάδους για τη συντακτική και την πραγματολογική ερμηνεία, οδηγήθηκαν σε ασυμφωνία και αδυναμία σύγκλισης, καταλήγοντας τελικά σε κενή απάντηση ή πλήρη αποχή. Αυτό καταδεικνύει ότι τα LLMs παραμένουν ευάλωτα σε διατυπώσεις που απαιτούν στάθμιση μεταξύ αυστηρής σύνταξης και πλαισιωμένης λογικής.

6.8 Ποσοτικά και ποιοτικά συμπεράσματα

Η παρούσα ενότητα επιχειρεί τη σύνθεση των ερευνητικών ευρημάτων σε δύο συμπληρωματικές διαστάσεις: (α) τα ποσοτικά συμπεράσματα, τα οποία απορρέουν από την ανάλυση των μετρικών επίδοσης και κόστους, και (β) τα ποιοτικά συμπεράσματα, τα οποία αφορούν τη λειτουργική

συμπεριφορά των μεθόδων, τα χαρακτηριστικά του παραγόμενου συλλογισμού και τη χρησιμότητά τους σε ρεαλιστικά σενάρια εφαρμογής.

Ποσοτικά συμπεράσματα

Η ανάλυση των αριθμητικών δεδομένων οδηγεί στη διατύπωση τεσσάρων βασικών παρατηρήσεων που διέπουν τη σχέση μεταξύ απόδοσης και πόρων:

1. **Θετική συσχέτιση ακρίβειας και υπολογιστικού κόστους:** Η βελτίωση της ακρίβειας δεν είναι ανέξοδη, αλλά συνδέεται άμεσα με την αύξηση του υπολογιστικού φόρτου. Οι τεχνικές που ενσωματώνουν εκτενή παραγωγή συλλογισμού (CoT), πολλαπλά δείγματα (SC) ή δενδρική αναζήτηση (ToT), εμφανίζουν σταθερά ανώτερη επίδοση, η οποία όμως συνοδεύεται από αναπόφευκτη επιβάρυνση σε χρόνο εκτέλεσης (αγγλ. latency) και κατανάλωση tokens.
2. **Η συνέπεια ως διακριτή μετρική αξιολόγησης:** Η συνέπεια (αγγλ. consistency) αποτελεί διακριτό στόχο από την ακρίβεια (αγγλ. Accuracy). Τα πειραματικά αποτελέσματα κατέδειξαν ότι δύο μέθοδοι δύνανται να παρουσιάζουν παρόμοια μέση ακρίβεια, αλλά να διαφέρουν δραματικά ως προς τη σταθερότητα αναπαραγωγής του αποτελέσματος. Συνεπώς, η αξιολόγηση των συστημάτων συλλογισμού οφείλει να είναι πολυδιάστατη, συνεξετάζοντας την ορθότητα με την επαναληψιμότητα.
3. **Στατιστική διασπορά στις μεθόδους ToT**: Οι μέθοδοι ToT παρουσιάζουν σημαντική διασπορά στις μετρικές τους. Η εμφάνιση ακραίων τιμών (αγγλ. outliers), όπως υπερβολικά υψηλό χρόνο απόκρισης (αγγλ. latency) ή περιπτώσεις αποχής, υποδηλώνει ότι η επίδοση των ToT δεν περιγράφεται επαρκώς από έναν απλό μέσο όρο. Η χρήση στατιστικών μεγεθών όπως η διάμεσος, τα τεταρτημόρια και το εύρος τιμών καθίσταται απαραίτητη για την αποτύπωση τόσο της τυπικής λειτουργίας όσο και των περιπτώσεων αστοχίας.
4. **Μη γραμμική κλιμάκωση κόστους:** Το κόστος κλιμακώνεται εκθετικά με την πολυπλοκότητα του μηχανισμού. Η μετάβαση από τη μοναδική απάντηση (Standard Prompting) στην αλυσίδα σκέψης (CoT), στα πολλαπλά δείγματα (SC-CoT) και τέλος στο δέντρο συλλογισμού (ToT) οδηγεί σε προοδευτική και συχνά δυσανάλογη αύξηση της κατανάλωσης πόρων σε σχέση με το κέρδος ακρίβειας.

Ποιοτικά συμπεράσματα

Πέραν των ποσοτικών δεικτών, η ανάλυση ανέδειξε σημαντικά ποιοτικά χαρακτηριστικά που αφορούν τη φύση του συλλογισμού:

1. **Μετασχηματισμός από Μηχανή Απάντησης σε Μηχανή Επίλυσης:** Η ρητή συλλογιστική μεταβάλλει τη φύση της λειτουργίας του μοντέλου. Οι τεχνικές CoT, SC-CoT και ToT δεν αυξάνουν απλώς την πιθανότητα σωστής απάντησης, αλλά καθιστούν τη διαδικασία διαφανή, ελέγξιμη και αξιολογήσιμη, επιτρέποντας στον χρήστη να επαληθεύσει τη λογική πορεία και όχι μόνο το τελικό αποτέλεσμα.
2. **Η Ανάγκη Μηχανισμών Εποπτείας:** Τα σφάλματα των LLMs στα μαθηματικά συχνά δεν οφείλονται σε έλλειψη γνώσης, αλλά σε έλλειψη επιτήρησης της διαδικασίας. Μέθοδοι όπως η SC-CoT και η ToT λειτουργούν ως μηχανισμοί εποπτείας που μειώνουν την επίδραση μεμονωμένων λανθασμένων συλλογισμών, είτε μέσω της πλειοψηφίας είτε μέσω της ενεργής αξιολόγησης και απόρριψης κλάδων.
3. **Η Διττή Φύση της Αποχής:** Η αποχή (αγγλ. abstention) λειτουργεί ως δείκτης υπεύθυνης συμπεριφοράς, ωστόσο απαιτεί προσεκτική ερμηνεία. Όταν προκύπτει από διαχείριση αβεβαιότητας, αποτελεί ένδειξη αξιοπιστίας. Όταν όμως είναι αποτέλεσμα λειτουργικού περιορισμού (π.χ. time-out αναζήτησης), αποτελεί ένδειξη ευθραυστότητας του συστήματος.
4. **Ευαισθησία της Μεθόδου ToT:** Παρόλο που η ToT παράγει λύσεις υψηλής λογικής συνοχής, παρουσιάζει αυξημένη ευαισθησία σε εξωτερικούς περιορισμούς. Λόγω της εξάρτησής της από πολλαπλές, αλυσιδωτές κλήσεις στο μοντέλο, είναι επιρρεπής σε διακοπές λόγω ορίων παραγωγής (αγγλ. context limits) ή χρονικών περιορισμών, γεγονός που μειώνει τη στιβαρότητά της σε μη ελεγχόμενα περιβάλλοντα.

5. **Πλαίσιο Επιλογής Βέλτιστης Μεθόδου:** Από τη σύνθεση των ευρημάτων προκύπτει ότι δεν υπάρχει μία μοναδική βέλτιστη λύση, αλλά ένα πλαίσιο λήψης απόφασης βάσει στόχων. Συγκεκριμένα, όταν προτεραιότητα είναι η ταχύτητα και το κόστος, απαιτείται απλούστερος μηχανισμός (Standard Prompting ή CoT Standard). Όταν προτεραιότητα είναι η αξιοπιστία και η επαναληψιμότητα, απαιτούνται τεχνικές σταθεροποίησης (SC-CoT), ενώ τέλος όταν προτεραιότητα είναι η μέγιστη ποιότητα συλλογισμού, απαιτείται εξερεύνηση λύσεων, με αναμενόμενο υψηλό κόστος (ToT).

Συνοψίζοντας, τα αποτελέσματα τεκμηριώνουν ότι η μετάβαση από την απλή προτροπή προς μηχανισμούς ρητής συλλογιστικής και αναζήτησης επιφέρει σημαντικά οφέλη ως προς την επίδοση και την ποιότητα του παραγόμενου αποτελέσματος, εισάγοντας ταυτόχρονα ουσιώδεις περιορισμούς σε κόστος και πολυπλοκότητα. Ως εκ τούτου, η κύρια συμβολή της παρούσας μελέτης δεν έγκειται απλά στην κατάταξη των τεχνικών προτροπής με βάση την απόδοσή τους, αλλά στην ανάδειξη των κρίσιμων συμβιβασμών (αγγλ. trade-offs) που απαιτούνται για την επιλογή της βέλτιστης μεθόδου σε ρεαλιστικές εφαρμογές.

Κεφάλαιο 7ο: Συζήτηση και ερμηνεία

Στο παρόν κεφάλαιο επιχειρείται η εις βάθος ερμηνεία των πειραματικών αποτελεσμάτων που παρουσιάστηκαν στο Κεφάλαιο 6, υπό το πρίσμα της σύγχρονης βιβλιογραφίας. Στόχος είναι η αποτίμηση της συμβολής των μεθόδων προτροπών στην ικανότητα μαθηματικού συλλογισμού των Μεγάλων Γλωσσικών Μοντέλων (LLMs), η ανάλυση των συγκριτικών τους πλεονεκτημάτων και η διερεύνηση των επιπτώσεων που έχουν αυτές οι τεχνικές στην ανάπτυξη εφαρμογών τεχνητής νοημοσύνης.

7.1 Τι δείχνουν τα αποτελέσματα για τις μεθόδους προτροπών

Τα αποτελέσματα της παρούσας διατριβής επιβεβαιώνουν την υπόθεση ότι η επίδοση ενός Μεγάλου Γλωσσικού Μοντέλου (LLM) σε σύνθετα προβλήματα, όπως αυτά του συνόλου δεδομένων GSM8K, δεν εξαρτάται αποκλειστικά από την αρχιτεκτονική ή το μέγεθός του. Αντιθέτως, επηρεάζεται σε καθοριστικό βαθμό από τη μέθοδο διατύπωσης και υποβολής του ερωτήματος (αγγλ. prompting strategy).

Ειδικότερα, η μετάβαση από τη Standard Prompting στη μέθοδο Chain-of-Thought (CoT) και, στη συνέχεια, στη Tree-of-Thought (ToT) καταδεικνύει ότι τα LLMs διαθέτουν λανθάνουσες ικανότητες συλλογισμού, οι οποίες ενεργοποιούνται μόνο μέσω κατάλληλης καθοδήγησης. Η Standard Prompting, η οποία βασίζεται κυρίως στην άμεση ανάκληση προτύπων (αγγλ. pattern matching), αποδείχθηκε ανεπαρκής για προβλήματα πολυσταδιακού συλλογισμού, οδηγώντας συχνά σε παραισθήσεις ή σε λογικά άλματα χωρίς ενδιάμεση τεκμηρίωση.

Αντιθέτως, η ρητή εξωτερίκευση της συλλογιστικής διαδικασίας μέσω της μεθόδου CoT, όπως περιγράφεται από τον Wei και συν. (2023), επιτρέπει στο μοντέλο να διασπά το πρόβλημα σε επιμέρους, διαχειρίσιμα βήματα. Με τον τρόπο αυτό, μια μη γραμμική και ενδεχομένως αδιαφανής διαδικασία σκέψης μετατρέπεται σε διαδοχική ακολουθία αιτιακών συνδέσεων, ενισχύοντας τόσο την ακρίβεια όσο και τη διαφάνεια της λύσης.

Επιπλέον, η εφαρμογή της μεθόδου ToT (Yao και συν., 2023) στα πειραματικά αποτελέσματα υποδηλώνει ότι ακόμη και η γραμμική βηματική σκέψη της CoT έχει εγγενή όρια. Η δυνατότητα του μοντέλου να εξερευνά εναλλακτικά μονοπάτια συλλογισμού, να επιστρέφει σε προηγούμενα βήματα και να αυτο-αξιολογεί ενδιάμεσες λύσεις προσεγγίζει περισσότερο τον ανθρώπινο τρόπο σκέψης. Κατ' αυτόν τον τρόπο, επιτυγχάνεται μέγιστη ακρίβεια σε περιπτώσεις όπου μια άμεση ή γραμμική απόκριση αποτυγχάνει, με αντίστοιχη όμως αύξηση της υπολογιστικής επιβάρυνσης.

7.2 Πλεονεκτήματα και περιορισμοί κάθε μεθόδου

Η συγκριτική ανάλυση ανέδειξε ότι κάθε μεθοδολογία συνοδεύεται από σαφή πλεονεκτήματα και εγγενείς περιορισμούς, διαμορφώνοντας ένα φάσμα επιλογών ανάλογα με τις απαιτήσεις.

7.2.1 Standard Prompting

Πλεονεκτήματα:
Αποτελεί τη πλέον αποδοτική λύση ως προς την ταχύτητα και το υπολογιστικό κόστος. Η ελάχιστη κατανάλωση tokens και ο χαμηλότερος μέσος χρόνος απόκρισης την καθιστούν κατάλληλη για

εφαρμογές πραγματικού χρόνου και για απλές εργασίες ανάκτησης ή παραγωγής πληροφορίας, όπου η υπολογιστική αποδοτικότητα υπερισχύει της μέγιστης ακρίβειας.

Περιορισμοί:
Παρουσιάζει τη χαμηλότερη μέση ακρίβεια σε προβλήματα πολυσταδιακού συλλογισμού. Αν και τα αποτελέσματα έδειξαν υψηλή συνέπεια (consistency), αυτή συχνά αντανακλά σταθερή αναπαραγωγή του ίδιου σφάλματος. Το φαινόμενο των «εμμονικών σφαλμάτων» (αγγλ. stubborn errors) παρατηρείται όταν το μοντέλο απαντά με υψηλή αυτοπεποίθηση αλλά λανθασμένα, χωρίς μηχανισμό αυτοδιόρθωσης ή αναθεώρησης της συλλογιστικής του πορείας. Ως εκ τούτου, η υψηλή επαναληψιμότητα δεν συνεπάγεται και υψηλή αξιοπιστία σε σύνθετα προβλήματα λογικής.

7.2.2 CoT και SC-CoT

Πλεονεκτήματα:
Η μέθοδος Chain-of-Thought (CoT) προσφέρει την πλέον ισορροπημένη προσέγγιση (αγγλ. trade-off) μεταξύ ακρίβειας και υπολογιστικών πόρων. Η ρητή παραγωγή ενδιάμεσων βημάτων καθιστά τη συλλογιστική διαδικασία διαφανή (interpretable) και επιτρέπει τον έλεγχο της λογικής συνέχειας των επιμέρους βημάτων. Στα πειραματικά αποτελέσματα, η CoT παρουσίασε σημαντική αύξηση ακρίβειας σε σχέση με τη Standard Prompting, με σχετικά περιορισμένη αύξηση σε χρόνο απόκρισης και κατανάλωση tokens, γεγονός που την καθιστά πρακτικά βιώσιμη λύση για απαιτητικά αλλά όχι ακραία σενάρια.
Η επέκταση Self-Consistency (SC-CoT), μέσω της παραγωγής πολλαπλών ανεξάρτητων αλυσίδων συλλογισμού και επιλογής της πλειοψηφικής απάντησης, ενίσχυσε περαιτέρω τη σταθερότητα και τη μέση ακρίβεια. Η προσέγγιση αυτή μείωσε τη στοχαστική μεταβλητότητα των αποκρίσεων και παρουσίασε την υψηλότερη ποιοτική αξιολόγηση στον δείκτη LLM-as-a-Judge, επιβεβαιώνοντας την αυξημένη αξιοπιστία των παραγόμενων λύσεων.

Περιορισμοί:
Η CoT αυξάνει τον χρόνο απόκρισης (latency) και την κατανάλωση tokens λόγω της παραγωγής εκτενέστερης συλλογιστικής αλυσίδας. Επιπλέον, είναι επιρρεπής στο φαινόμενο της διάδοσης σφάλματος (error propagation), όπου ένα λανθασμένο ενδιάμεσο βήμα οδηγεί αναπόφευκτα σε εσφαλμένο τελικό αποτέλεσμα.
Στην περίπτωση της SC-CoT, το υπολογιστικό κόστος αυξάνεται αναλογικά με τον αριθμό των δειγματοληψιών (K), καθώς απαιτείται πολλαπλή εκτέλεση της διαδικασίας συλλογισμού. Παρότι η πλειοψηφική σύγκλιση μειώνει τη στοχαστική αστάθεια, δεν εξαλείφει πλήρως τα συστηματικά σφάλματα, ιδιαίτερα όταν όλα τα παραγόμενα μονοπάτια συγκλίνουν προς την ίδια λανθασμένη λογική υπόθεση.

7.2.3 ToT

Πλεονεκτήματα:
Η μέθοδος Tree-of-Thought (ToT) προσφέρει την πλέον δομημένη μορφή συλλογιστικής διερεύνησης, επιτρέποντας την παράλληλη εξερεύνηση εναλλακτικών κλάδων σκέψης. Η δυνατότητα αναδρομής (backtracking) και αξιολόγησης ενδιάμεσων καταστάσεων μειώνει την πιθανότητα εγκλωβισμού σε πρώιμες λανθασμένες υποθέσεις, οι οποίες στη γραμμική CoT οδηγούν σε αναπόφευκτο τελικό σφάλμα. Στα πειραματικά αποτελέσματα, η ToT κατέγραψε τη μέγιστη επιμέρους ακρίβεια της μελέτης (97.67%) σε κατάλληλες παραμετροποιήσεις, καθώς και ιδιαίτερα υψηλά επίπεδα συνέπειας (consistency). Σε λειτουργικές διαμορφώσεις, η υψηλή συνέπεια συνοδεύτηκε από υψηλή ακρίβεια, γεγονός που υποδηλώνει σταθερό και επαληθεύσιμο συλλογισμό. Η μέθοδος παρήγαγε επίσης λύσεις με αυξημένη λογική συνοχή και δυνατότητα εσωτερικής αξιολόγησης.

Περιορισμοί:
Το υπολογιστικό κόστος αυξάνεται σημαντικά με την αύξηση του παράγοντα διακλάδωσης και του βάθους αναζήτησης, καθώς η μέθοδος απαιτεί πολλαπλές κλήσεις στο μοντέλο για την παραγωγή και αξιολόγηση κάθε υποψήφιου κλάδου. Η επιβάρυνση σε χρόνο απόκρισης και κατανάλωση tokens είναι

αισθητά υψηλότερη σε σχέση με CoT και SC-CoT, ενώ παρατηρήθηκαν και ακραίες αποκλίσεις (outliers) σε περιβάλλοντα με περιορισμούς πόρων.Επιπλέον, η πολυπλοκότητα υλοποίησης είναι αυξημένη, καθώς απαιτείται μηχανισμός διαχείρισης δέντρου αναζήτησης και στρατηγική αξιολόγησης κόμβων. Σε συνθήκες περιορισμένου context ή περιορισμών API, η μέθοδος μπορεί να καταστεί λιγότερο αποδοτική ή να οδηγήσει σε αποτυχίες εκτέλεσης.

7.3 Επιρροή του μεγέθους του μοντέλου και της διαμόρφωσης προτροπών

Τα ευρήματα της διατριβής συμφωνούν με τους νόμους κλιμάκωσης (αγγλ. scaling laws) και τη θεωρία των αναδυόμενων ικανοτήτων (Emergent Abilities), όπως περιγράφονται στη σχετική βιβλιογραφία.

Μέγεθος Μοντέλου: Παρατηρήθηκε ότι οι προηγμένες τεχνικές προτροπών (CoT, ToT) αποδίδουν τα μέγιστα σε μοντέλα μεγάλης κλίμακας. Τα μικρότερα μοντέλα, αν και ωφελούνται, συχνά αδυνατούν να διατηρήσουν τη μακροσκελή λογική συνοχή που απαιτεί η ToT, οδηγώντας σε υψηλότερα ποσοστά αποχής ή παραισθήσεων. Αυτό υποδηλώνει ότι η ικανότητα μετα-συλλογισμού (αγγλ. meta-reasoning) που απαιτεί η αξιολόγηση κλάδων στην ToT είναι χαρακτηριστικό που αναδύεται μετά από ένα ορισμένο κατώφλι παραμέτρων.

Διαμόρφωση Προτροπών (k, b): Η αύξηση των παραμέτρων k (αριθμός αλυσίδων στην SC) και b (branching factor στην ToT) βελτιώνει την ακρίβεια, αλλά με φθίνουσες αποδόσεις. Για παράδειγμα, η μετάβαση από το tot_b3 στο tot_b5 προσέφερε οριακή βελτίωση στην ακρίβεια, αλλά διπλασίασε σχεδόν το κόστος. Αυτό υποδεικνύει ότι υπάρχει ένα σημείο κορεσμού, πέραν του οποίου η περαιτέρω αναζήτηση δεν προσθέτει αξία, αλλά μόνο θόρυβο και καθυστέρηση.

7.4 Επιπτώσεις στην ανάπτυξη εφαρμογών βασισμένων σε LLMs

Η ανάλυση των αποτελεσμάτων έχει άμεσες πρακτικές προεκτάσεις για τη μηχανική λογισμικού και την ανάπτυξη εφαρμογών AI:

1. **Διαχείριση Κόστους-Απόδοσης:** Οι προγραμματιστές καλούνται να επιλέξουν στρατηγική προτροπών βάσει της κρισιμότητας της εφαρμογής. Για chatbots εξυπηρέτησης πελατών, η Standard Prompting ή η απλή CoT είναι επαρκή. Για εφαρμογές ιατρικής διάγνωσης, χρηματοοικονομικής ανάλυσης ή νομικής συμβουλευτικής, η χρήση SC-CoT ή ToT κρίνεται επιβεβλημένη παρά το κόστος, καθώς η ακρίβεια είναι αδιαπραγμάτευτη.
2. **Διαχείριση χρόνου:** Η υψηλή καθυστέρηση των μεθόδων ToT επιβάλλει ασύγχρονη αρχιτεκτονική στις εφαρμογές. Δεν είναι δυνατή η χρήση τους σε σενάρια αλληλεπίδρασης πραγματικού χρόνου (αγγλ. real-time) χωρίς σημαντική υποβάθμιση της εμπειρίας χρήστη.
3. **Υβριδικά Συστήματα:** Τα αποτελέσματα προκρίνουν την ανάπτυξη υβριδικών συστημάτων, όπου ένα "ελαφρύ" μοντέλο/μέθοδος (Standard) αξιολογεί τη δυσκολία του ερωτήματος και δρομολογεί τα σύνθετα προβλήματα σε πιο σύνθετες διαδικασίες συλλογισμού (ToT), βελτιστοποιώντας τη συνολική κατανάλωση πόρων.

7.5 Προτάσεις για μελλοντική έρευνα

Η παρούσα διπλωματική εργασία ανέδειξε τη σημαντική επίδραση των προηγμένων τεχνικών προτροπών στην ακρίβεια και την αξιοπιστία των LLMs. Ωστόσο, τα ευρήματα αυτά ανοίγουν νέους ορίζοντες για περαιτέρω διερεύνηση, τόσο σε επίπεδο βελτιστοποίησης των υπαρχουσών μεθόδων όσο και σε επίπεδο σύζευξης με νέες αρχιτεκτονικές.

Στην παρούσα μελέτη, οι παράμετροι των μεθόδων (π.χ. ο αριθμός των αλυσίδων k στην SC-CoT ή ο παράγοντας διακλάδωσης b στην ToT) ορίστηκαν στατικά. Μια σημαντική κατεύθυνση για μελλοντική έρευνα αφορά την ανάπτυξη δυναμικών αλγορίθμων (αγγλ. adaptive algorithms), οι οποίοι θα προσαρμόζουν αυτές τις παραμέτρους σε πραγματικό χρόνο ανάλογα με την εκτιμώμενη δυσκολία του

εκάστοτε ερωτήματος. Η χρήση πλαισίων όπως το DSPy (Declarative Self-improving Language Programs) θα μπορούσε να αυτοματοποιήσει τη διαδικασία βελτιστοποίησης των προτροπών, επιτρέποντας στο σύστημα να μαθαίνει τη βέλτιστη δομή συλλογισμού για κάθε τύπο προβλήματος, μεταβαίνοντας αυτόματα από CoT σε ToT όταν ανιχνεύεται χαμηλή αυτοπεποίθηση.

Επιπλέον, το υψηλό υπολογιστικό κόστος της μεθόδου ToT αποτελεί τροχοπέδη για την ευρεία υιοθέτησή της. Μια υποσχόμενη ερευνητική οδός είναι η εφαρμογή τεχνικών Απόσταξης Γνώσης (αγγλ. Knowledge Distillation). Συγκεκριμένα, προτείνεται η χρήση των υψηλής ποιότητας συλλογιστικών μονοπατιών που παράγει ένα μεγάλο μοντέλο (αγγλ. Teacher model) μέσω ToT, για την εκπαίδευση μικρότερων και ταχύτερων μοντέλων (αγγλ. Student models). Στόχος είναι τα μικρότερα μοντέλα να εσωτερικεύσουν τη διαδικασία δενδρικής αναζήτησης και να παράγουν απευθείας τη σωστή απάντηση ή μια βέλτιστη αλυσίδα σκέψης, χωρίς την ανάγκη εκτέλεσης πολλαπλών κλήσεων κατά τη φάση της παραγωγής, μειώνοντας έτσι δραστικά τον χρόνο εκτέλεσης.

Τα αποτελέσματα έδειξαν ότι ακόμη και με την ToT, τα μοντέλα ενδέχεται να σφάλλουν σε αριθμητικούς υπολογισμούς. Η μελλοντική έρευνα οφείλει να εξετάσει τη σύζευξη των μεθόδων προτροπών με εξωτερικά εργαλεία ελέγχου. Συγκεκριμένα, προτείνεται η ενσωμάτωση διερμηνέων κώδικα (π.χ. Python Code Interpreter) εντός των κόμβων του δέντρου ToT. Αντί το μοντέλο να εκτελεί τους υπολογισμούς, θα μπορούσε να παράγει κώδικα που εκτελείται ντετερμινιστικά, με την ToT να διαχειρίζεται τη λογική ροή και τον διερμηνέα να εγγυάται την αριθμητική ακρίβεια. Η προσέγγιση αυτή Program-of-Thought αναμένεται να εξαλείψει πλήρως τα λάθη υπολογισμού που παρατηρήθηκαν στο GSM8K.

Τέλος, με την έλευση των νεότερων και πιο ικανών μοντέλων (π.χ. GPT-5.2 και Gemini 3 Pro), η έρευνα μπορεί να επεκταθεί πέραν του κειμένου. Μια ενδιαφέρουσα κατεύθυνση είναι η εφαρμογή των μεθόδων CoT και ToT σε προβλήματα που συνδυάζουν εικόνα και κείμενο (π.χ. προβλήματα γεωμετρίας με διαγράμματα). Η διερεύνηση του κατά πόσο η οπτική πληροφορία μπορεί να λειτουργήσει ως ενδιάμεσο βήμα συλλογισμού ή ως μηχανισμός επαλήθευσης των παραγόμενων λύσεων αποτελεί ένα εξαιρετικά ενδιαφέρον ερευνητικό πεδίο.

Κεφάλαιο 8ο: Συμπεράσματα

Το παρόν κεφάλαιο συνοψίζει και συνθέτει τα ευρήματα της διατριβής, αποτιμώντας τη συμβολή της έρευνας στο πεδίο της μηχανικής προτροπών (Prompt Engineering). Παράλληλα, αναδεικνύονται τα κομβικά συμπεράσματα της συγκριτικής ανάλυσης, οι περιορισμοί της μελέτης και οι προοπτικές για μελλοντική έρευνα.

8.1 Συνοπτική αποτίμηση της ερευνητικής προσέγγισης

Στόχος της εργασίας ήταν η συστηματική διερεύνηση της επίδρασης διαφορετικών τεχνικών προτροπής στη συλλογιστική ικανότητα των Μεγάλων Γλωσσικών Μοντέλων (LLMs), εστιάζοντας σε προβλήματα πολυβηματικού συλλογισμού. Σε αντιδιαστολή με μεθόδους που απαιτούν κοστοβόρα επανεκπαίδευση, η παρούσα μελέτη επικεντρώθηκε στην αξιοποίηση των ενσωματωμένων και αναδυόμενων ικανοτήτων των μοντέλων μέσω βελτιστοποιημένων προτροπών.

Η μεθοδολογία θεμελιώθηκε σε ένα αυστηρό πειραματικό πλαίσιο, το οποίο εξασφάλισε τη δίκαιη σύγκριση μεταξύ της Τυπικής Προτροπής (Standard Prompting), της Αλυσίδας Συλλογισμού (Chain-of-Thought - CoT) και της Δενδρικής Συλλογιστικής (Tree-of-Thought - ToT). Ιδιαίτερη βαρύτητα δόθηκε όχι μόνο στην ποσοτική μέτρηση της ακρίβειας, αλλά και στην ποιοτική ανάλυση της συλλογιστικής διαδρομής.

8.2 Κύρια ερευνητικά ευρήματα

Η πειραματική αξιολόγηση κατέδειξε ότι η αρχιτεκτονική της προτροπής αποτελεί καθοριστικό ρυθμιστή της απόδοσης. Η Standard Prompting, ενώ επαρκεί για απλές εργασίες ανάκλησης, εμφάνισε δομικές αδυναμίες σε προβλήματα που απαιτούν ανάλυση, ενδιάμεσα βήματα και εσωτερική συνέπεια στη συλλογιστική πορεία.

Αντίθετα, η μέθοδος CoT αποδείχθηκε καταλύτης για την επίλυση σύνθετων προβλημάτων, καθώς η ρητή παραγωγή ενδιάμεσων βημάτων οδήγησε σε θεαματική αύξηση της ακρίβειας, της διαφάνειας και της ερμηνευσιμότητας. Ακόμη πιο εξελιγμένη συμπεριφορά επέδειξε η μέθοδος ToT, η οποία μέσω της μη γραμμικής εξερεύνησης εναλλακτικών σεναρίων, επέτρεψε στα μοντέλα να επανεξετάζουν ενδιάμεσα συμπεράσματα και να αποφεύγουν πρόωρες ή λανθασμένες αποφάσεις. Αν και υπολογιστικά απαιτητικότερη, η ToT αναδείχθηκε ως η βέλτιστη στρατηγική για προβλήματα πολύ υψηλής πολυπλοκότητας.

8.3 Συμβολή της εργασίας στη βιβλιογραφία

Η παρούσα διατριβή συνεισφέρει στη σύγχρονη βιβλιογραφία σε τρία επίπεδα. Πρώτον, προσφέρει μια δομημένη συγκριτική ανάλυση που γεφυρώνει τη θεωρία με την εμπειρική επαλήθευση. Δεύτερον, αναδεικνύει τη σημασία της συλλογιστικής διαδρομής και όχι μόνο της τελικής απάντησης, ως κρίσιμου κριτηρίου αξιολόγησης των LLMs. Τρίτον, τεκμηριώνει ότι η μηχανική προτροπών αποτελεί μια βιώσιμη, χαμηλού κόστους εναλλακτική λύση έναντι της επανεκπαίδευσης, ξεκλειδώνοντας λανθάνουσες ικανότητες των LLMs χωρίς μεταβολή των παραμέτρων τους.

8.4 Περιορισμοί της μελέτης

Παρά τα ενθαρρυντικά αποτελέσματα, η παρούσα εργασία υπόκειται σε ορισμένους περιορισμούς. Συγκεκριμένα, η αξιολόγηση επικεντρώθηκε σε συγκεκριμένο σύνολο δεδομένων (GSM8K), γεγονός που περιορίζει τη γενίκευση των συμπερασμάτων σε όλα τα πιθανά πεδία εφαρμογής. Επιπλέον, το αυξημένο υπολογιστικό κόστος και η καθυστέρηση απόκρισης των μεθόδων ToT ενδέχεται να περιορίζουν την πρακτική τους εφαρμογή σε συστήματα πραγματικού χρόνου.

Τέλος, η αξιολόγηση της ορθότητας και της ποιότητας της συλλογιστικής διαδικασίας παραμένει σε κάποιο βαθμό υποκειμενική, παρά τη χρήση μετρικών. Το ζήτημα αυτό αναδεικνύει τη γενικότερη πρόκληση της αξιόπιστης αξιολόγησης σύνθετων γνωστικών διεργασιών στα σύγχρονα συστήματα τεχνητής νοημοσύνης.

8.5 Τελικό συμπέρασμα

Συνοψίζοντας, η παρούσα εργασία αποδεικνύει ότι η ευφυΐα ενός Γλωσσικού Μοντέλου δεν είναι στατική ιδιότητα, αλλά δυναμικό μέγεθος που εξαρτάται άμεσα από την ποιότητα της καθοδήγησης που λαμβάνει. Η μετάβαση από το απλό ερώτημα στη δομημένη συλλογιστική (CoT/ToT) μετατρέπει τα LLMs από παθητικές μηχανές πρόβλεψης λέξεων σε ενεργούς επιλυτές προβλημάτων.

Το μείζον συμπέρασμα που προκύπτει είναι η μετατόπιση του ενδιαφέροντος από το «αποτέλεσμα» στη «διαδικασία». Η ικανότητα ενός συστήματος να αιτιολογεί, να αναθεωρεί και να επαληθεύει τα βήματά του, καθίσταται ο πλέον κρίσιμος παράγοντας για την ανάπτυξη αξιόπιστων και ερμηνεύσιμων εφαρμογών Τεχνητής Νοημοσύνης. Υπό αυτό το πρίσμα, η μηχανική προτροπών αναδεικνύεται όχι ως απλή τεχνική δεξιότητα, αλλά ως θεμελιώδης επιστημονικός κλάδος για την πλήρη αξιοποίηση των δυνατοτήτων των σύγχρονων υπολογιστικών συστημάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] H. Naveed *et al.*, “A Comprehensive Overview of Large Language Models,” 2024. [Online]. Available: https://arxiv.org/abs/2307.06435

[2] S. Vatsal and H. Dubey, “A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks,” 2024. [Online]. Available: https://arxiv.org/abs/2407.12994

[3] J. Wei *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[4] S. Yao *et al.*, “Tree of Thoughts: Deliberate Problem Solving with Large Language Models,” 2023. [Online]. Available: https://arxiv.org/abs/2305.10601

[5] S. G. Aithal, A. B. Rao, S. Singh, and others, “Chain-of-Thought Reasoning Evaluation Framework for Question Answering System,” in *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, 2025, pp. 725–730.

[6] Z. Wang, Z. Chu, T. V. Doan, S. Ni, M. Yang, and W. Zhang, “History, Development, and Principles of Large Language Models-An Introductory Survey,” 2024. [Online]. Available: https://arxiv.org/abs/2402.06853

[7] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[8] J. Yu, M. Luo, H. Zhou, and Z. Lan, “Unleashing the Second Brain: Enhancing Large Language Models through Chain of Thought with Human Feedback,” in *2023 16th International Symposium on Computational Intelligence and Design (ISCID)*, 2023, pp. 90–95. doi: 10.1109/ISCID59865.2023.00029.

[9] B. Wang *et al.*, “Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters,” Dec. 2022. doi: 10.48550/arXiv.2212.10001.

[10] L. Dai and K. Qin, “Fact-CoT: Explicit Fact-Enhanced Chain-of-Thought Prompting for Commonsense Reasoning,” in *2023 International Conference on Computer Science and Automation Technology (CSAT)*, 2023, pp. 335–338. doi: 10.1109/CSAT61646.2023.00092.

[11] A. Gao, “Prompt Engineering for Large Language Models: A brief guide with examples for non-technical readers,” *Available at SSRN 4504303*, 2023.

[12] J. Gu *et al.*, “A Survey on LLM-as-a-Judge,” 2025. [Online]. Available: https://arxiv.org/abs/2411.15594

[13] “Prompting Guide 101,” Oct. 2024, *Google*.

[14] J. Phoenix and M. Taylor, *Prompt engineering for generative AI*. “ O’Reilly Media, Inc.,” 2024.

[15] X. Wang *et al.*, “Self-Consistency Improves Chain of Thought Reasoning in Language Models,” 2023. [Online]. Available: https://arxiv.org/abs/2203.11171

[16] K. Liu, S. Zhang, Y. Wu, H. Zhang, X. Wu, and L. He, “Chain-of-Thought Prompting Enhanced by GPT-4 for Improving Reasoning Capabilities in Large Language Models,” in *2024 6th International*

Academic Exchange Conference on Science and Technology Innovation (IAECST), 2024, pp. 144–148. doi: 10.1109/IAECST64597.2024.11117945.

[17] J. Long, "Large Language Model Guided Tree-of-Thought," 2023. [Online]. Available: https://arxiv.org/abs/2305.08291

[18] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," 2021. [Online]. Available: https://arxiv.org/abs/2107.13586

[19] K. Cobbe *et al.*, "Training Verifiers to Solve Math Word Problems," 2021. [Online]. Available: https://arxiv.org/abs/2110.14168

[20] M. Oshin and N. Campos, *Learning LangChain*. " O'Reilly Media, Inc.," 2025.

[21] L. Reynolds and K. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," 2021. [Online]. Available: https://arxiv.org/abs/2102.07350

[22] V. Mavroudis, "LangChain v0.3," Dec. 2025. doi: 10.31219/osf.io/4gprt_v1.

[23] O. Topsakal and T. C. Akinci, "Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast," *International Conference on Applied Engineering and Natural Sciences*, vol. 1, pp. 1050–1056, Dec. 2023, doi: 10.59287/icaens.1127.

[24] S. Joshi, "Review of Prompt Engineering Techniques in Finance: An Evaluation of Chain-of-Thought, Tree-of-Thought, and Graph- of-Thought Approaches," *International Journal of Innovative Research in Computer Science & Technology*, vol. 13, pp. 2347–5552, Dec. 2025, doi: 10.55524/ijircst.2025.13.4.6.

[25] L. Nguyen and Y. Xu, "Reasoning for Translation: Comparative Analysis of Chain-of-Thought and Tree-of-Thought Prompting for LLM Translation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, J. Zhao, M. Wang, and Z. Liu, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 259–275. doi: 10.18653/v1/2025.acl-srw.17.

[26] L. Boonstra, "Prompt engineering," 2024, *Google*.

[27] A. Obuchowski, "Prompt engineering techniques."

[28] H. Zhao, H. Yilahun, and A. Hamdulla, "Pipeline Chain-of-Thought: A Prompt Method for Large Language Model Relation Extraction," in *2023 International Conference on Asian Language Processing (IALP)*, 2023, pp. 31–36. doi: 10.1109/IALP61005.2023.10337264.

[29] G. Chochlakis, N. M. Pandiyan, K. Lerman, and S. Narayanan, "Larger Language Models Don't Care How You Think: Why Chain-of-Thought Prompting Fails in Subjective Tasks," 2024. [Online]. Available: https://arxiv.org/abs/2409.06173