



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ  
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ  
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ  
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΜΑΘΗΣΗΣ

# Pitfalls of Scale: Investigating the Inverse Task of Redefinition in Large Language Models

DIPLOMA THESIS

by

**Eleni Stringli**

**Επιβλέπων:** Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2025





Εθνικό Μετσόβιο Πολυτεχνείο  
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών  
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών  
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

# Pitfalls of Scale: Investigating the Inverse Task of Redefinition in Large Language Models

## DIPLOMA THESIS

by

**Eleni Stringli**

**Επιβλέπων:** Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 2<sup>η</sup> Ιουλίου, 2025.

.....  
Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής Ε.Μ.Π.

.....  
Α.-Γ. Σταφυλοπάτης  
Ομότιμος Καθηγητής Ε.Μ.Π.

.....  
Γεώργιος Στάμου  
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2025

.....  
**ΕΛΕΝΗ ΣΤΡΙΓΓΛΗ**  
Διπλωματούχος Ηλεκτρολόγος Μηχανικός  
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © – All rights reserved Eleni Stringli, 2025.

Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.



# Περίληψη

Καθώς τα Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ) συνεχώς επεκτείνονται σε κλίμακα, παρουσιάζουν ολοένα και πιο εξελιγμένες συμπεριφορές, επιδεικνύοντας ακόμα και ικανότητες που παραπέμπουν σε λογική σκέψη. Ωστόσο, η γνησιότητα αυτών των ικανοτήτων αποτελεί συχνά αντικείμενο αμφισβήτησης, με πολλούς να υποστηρίζουν ότι δεν προέρχονται από πραγματική κατανόηση, αλλά απλή απομνημόνευση και αναγνώριση προτύπων. Για τη διερεύνηση αυτών των ελλείψεων έχουν κατασκευαστεί πειράματα που απαιτούν την υπέρβαση βαθιά ριζωμένων συσχετίσεων, αποκαλύπτοντας σημαντικές αδυναμίες στη συλλογιστική και τη προσαρμοστικότητα των ΜΓΜ, σε σύγκριση με τις ανθρώπινες νοητικές ικανότητες. Στα προβλήματα αντίστροφης κλιμάκωσης η απόδοση των μοντέλων παραδόξως χειροτερεύει όσο αυξάνεται το μέγεθός τους, με αποτέλεσμα τέτοιες αδυναμίες να έρχονται στην επιφάνεια, και μάλιστα εκθέτοντας τα μεγαλύτερα, πιο "εξελιγμένα" ΜΓΜ. Η παρούσα διπλωματική εργασία μελετά το πρόβλημα του επαναορισμού, στο οποίο τα ΜΓΜ καλούνται να υιοθετήσουν εναλλακτικούς ορισμούς για γνωστές επιστημονικές σταθερές και μονάδες μέτρησης, και στη συνέχεια να απαντήσουν σε ερωτήσεις σχετικά με αυτές τις νέες τιμές. Τα αποτελέσματά μας δείχνουν ότι τα μεγαλύτερα μοντέλα δυσκολεύονται περισσότερο να προσαρμοστούν στους νέους ορισμούς, προσκολλημένα στις τιμές που έχουν απομνημονεύσει κατά τη διαδικασία της προεκπαίδευσής τους, και, εκτός αυτού, προτιμούν να δίνουν απαντήσεις, ακόμα και εσφαλμένες, παρά να απέχουν. Επίσης, παρά το γεγονός ότι παράγοντες όπως η μορφοποίηση των απαντήσεων και οι τεχνικές προτροπής μπορούν να επηρεάσουν τη συμπεριφορά των μοντέλων, καμία μέθοδος δεν καταφέρνει να εξαλείψει πλήρως την τάση των μεγαλύτερων ΜΓΜ να προσκολλώνται στην εσωτερικευμένη γνώση τους.

**Λέξεις-κλειδιά** — Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ), μηχανική προτροπών, αντίστροφη κλιμάκωση, ικανότητες συλλογιστικής, προσαρμοστικότητα, απομνημόνευση, δυσκολία δειγμάτων, ερμηνευσιμότητα.



# Abstract

As Large Language Models (LLMs) continue to grow in scale, they exhibit increasingly sophisticated behaviors, including abilities that resemble logical reasoning. However, the authenticity of such advances remains a subject of debate, with many arguing that they are largely a byproduct of memorization and advanced statistical pattern recognition rather than genuine understanding. To shed light on these limitations, researchers have developed experimental conditions that challenge LLMs to override entrenched associations, highlighting gaps in reasoning and adaptability when compared to human cognition. Inverse scaling tasks are designed to uncover such weaknesses by revealing a paradoxical decline in performance as model scale increases, thereby exposing critical blind spots in scale-driven improvements. In this thesis, we explore the redefinition task, which challenges LLMs to adopt nonstandard definitions for familiar scientific constants and units of measurement and then respond based on these altered values. We evaluate state-of-the-art models from multiple LLM families and demonstrate that larger LLMs not only perform worse at following redefinitions, anchoring more strongly to their memorized knowledge, but also demonstrate increased confidence in generating false responses rather than choosing to abstain. In addition, although factors such as response formatting and prompting techniques can influence these behaviors, no strategy fully counteracts the tendency of larger models to revert to pretraining priors.

**Keywords** — Large Language Models (LLMs), prompt engineering, inverse scaling, reasoning capabilities, adaptability, memorization, hardness of samples, interpretability.



# Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά όλους αυτούς που στάθηκαν δίπλα μου, όχι μόνο κατά τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας, αλλά και καθ' όλη τη διάρκεια των σπουδών μου στη σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών. Ο καθένας με τον δικό του τρόπο συνέβαλε καθοριστικά στην ολοκλήρωση αυτής της μακράς και απαιτητικής, αλλά βαθιά σημαντικής για μένα, πορείας.

Ευχαριστώ θερμά, καταρχάς, τον επιβλέποντά μου, κ. Αθανάσιο Βουλόδημο, για την εμπιστοσύνη που μου έδειξε και τη σταθερή υποστήριξή του. Ιδιαίτερες ευχαριστίες, επίσης, οφείλω στη Μαρία Λυμπεραίου και τον Γιώργο Φιλανδριανό, των οποίων η βοήθεια ήταν καθοριστική σε όλα τα στάδια της εργασίας. Είμαι πολύ ευγνώμων για την εξαιρετική συνεργασία μας και ειλικρινά θαυμάζω τις γνώσεις και τις ικανότητές τους.

Μεγάλη είναι, φυσικά, και η ευγνωμοσύνη μου προς την οικογένειά μου, η οποία με στήριξε σε αυτήν την προσπάθεια, όπως άλλωστε με στηρίζει και σε κάθε μου βήμα. Θα ήθελα να αναφέρω την ξαδέρφη και ταυτόχρονα συμφοιτήτριά μου, Πηνελόπη, καθώς – για να επαναλάβω τα λόγια της – "είμαι περήφανη για τον εαυτό μας". Τέλος, ευχαριστώ ολόψυχα τους φίλους μου για την πολύτιμη συντροφιά τους όλα αυτά τα χρόνια, χωρίς την οποία αδυνατώ να φανταστώ πώς θα ήταν αυτή η εμπειρία.

Ελένη Στριγγλή, Ιούλιος 2025



# Contents

<b>Contents</b>	<b>11</b>
<b>List of Figures</b>	<b>14</b>
<b>List of Tables</b>	<b>16</b>
<b>1 Εκτεταμένη Περίληψη στα Ελληνικά</b>	<b>21</b>
1.1 Εισαγωγή	22
1.2 Θεωρητικό Υπόβαθρο	22
1.2.1 Εισαγωγή στα Μεγάλα Γλωσσικά Μοντέλα	22
1.2.2 Μηχανική Προτροπών	23
1.2.3 Αξιολόγηση με MFM	23
1.2.4 Συλλογιστική στα MFM	24
1.2.5 Προβλήματα αντίστροφης κλιμάκωσης	24
1.3 Μέθοδος	25
1.3.1 Σύνολα Δεδομένων	25
1.3.1.1 Επαναορισμός Φυσικών Σταθερών	25
1.3.1.2 Επαναορισμός Μονάδων Μέτρησης	26
1.3.1.3 Μορφοποίηση των Ερωτήσεων	26
1.3.1.4 Υλοποίηση	27
1.3.2 Μετρικές και Αξιολόγηση	27
1.3.3 Σχεδίαση Προτροπών	27
1.3.4 Επιλογή MFM	28
1.3.5 Πειραματική Υλοποίηση	28
1.4 Πειραματικά Αποτελέσματα	29
1.4.1 Επαναορισμός Επιστημονικών Σταθερών	29
1.4.1.1 Προσκόλληση στις Πραγματικές Τιμές	29
1.4.1.2 Αντίστροφη Κλιμάκωση	30
1.4.1.3 Μορφοποίηση Απαντήσεων	31
1.4.1.4 Τύπος Επαναορισμού	32
1.4.1.5 Λειτουργία Εκτεταμένης Σκέψης	32
1.4.1.6 Επίδραση Προτροπών	33
1.4.1.7 Άρνηση Απόκρισης	34
1.4.2 Επαναορισμός Μονάδων Μέτρησης	35
1.4.2.1 Προσκόλληση στις Πραγματικές Τιμές	35
1.4.2.2 Αντίστροφη Κλιμάκωση	36
1.4.2.3 Μορφοποίηση Απαντήσεων	37
1.4.2.4 Επίδραση Προτροπών	37
1.4.2.5 Άρνηση Απόκρισης	38
1.5 Συμπεράσματα	38
<b>2 Introduction</b>	<b>41</b>

<b>3</b>	<b>Background</b>	<b>43</b>
3.1	Introduction to Large Language Models	43
3.1.1	Framework and Objective	43
3.1.2	Parameters and Scaling Laws	44
3.1.3	Capabilities and Emergent Behaviors	45
3.2	Prompt Engineering	46
3.2.1	Prompts and Prompt Engineering	46
3.2.2	Prompt Templates	46
3.2.3	Prompting Techniques	46
3.2.3.1	Zero-Shot Prompting	46
3.2.3.2	One-Shot Prompting	47
3.2.3.3	Few-Shot Prompting	47
3.2.3.4	Chain-of-Thought Prompting	47
3.3	Evaluation with LLMs	48
3.4	Related Work	50
3.4.1	True Reasoning in Large Language Models	50
3.4.1.1	Emerging but Limited: Reasoning in LLMs	50
3.4.1.2	Evidence of Fragile Reasoning	50
3.4.1.3	Memorization behind LLM Reasoning	50
3.4.2	Inverse scaling problems	51
<b>4</b>	<b>Methodology</b>	<b>55</b>
4.1	Datasets	55
4.1.1	Redefinition of Scientific Constants Dataset	56
4.1.1.1	Selection of Constants	56
4.1.1.2	Redefinition Scenarios	56
4.1.1.3	Question Levels	57
4.1.2	Redefinition of Units of Measure Dataset	58
4.1.2.1	Selection of Units	58
4.1.2.2	Redefinition Scenarios	58
4.1.2.3	Question Levels	59
4.1.3	Question Formats	59
4.1.4	Dataset Format and Implementation	61
4.2	Metrics and Evaluation	61
4.3	Prompting Details	62
4.3.1	No Redefinition Prompts	62
4.3.2	Redefinition Prompts	65
4.3.3	Evaluation Prompts	68
4.4	LLM selection	71
4.5	Experimental setup	72
<b>5</b>	<b>Experimental Results</b>	<b>73</b>
5.1	Results on constants redefinition	73
5.1.1	Anchoring to default values	73
5.1.2	Inverse trends	74
5.1.3	Response format	77
5.1.4	Assignment vs Swapping	78
5.1.5	Extended Thinking Blocks	80
5.1.6	The influence of prompting	81
5.1.7	Completely Wrong Responses Analysis	82
5.1.7.1	Refusal to Respond	82
5.1.7.2	Case studies on Refusal and Overconfidence	84
5.1.7.3	Refusal-Adjusted Anchoring: A Refined Perspective	87
5.2	Results on units redefinition	88
5.2.1	Anchoring to default values	88
5.2.2	Inverse Trends	89



5.2.3	Response Format . . . . .	91
5.2.4	The influence of prompting . . . . .	92
5.2.5	Completely Wrong Responses Analysis . . . . .	93
5.2.5.1	Refusal to respond . . . . .	93
5.2.5.2	Case studies on Refusal and Overconfidence . . . . .	94
<b>6</b>	<b>Conclusion</b>	<b>97</b>
6.1	Conclusion . . . . .	97
6.2	Reasoning vs. Robustness Trade-Off . . . . .	98
<b>7</b>	<b>Bibliography</b>	<b>99</b>



# List of Figures

1.4.1 Ποσοστά προσκόλλησης για μοντέλα διαφορετικού μεγέθους στην οικογένεια Llama (μορφή πολλαπλών επιλογών). . . . .	30
1.4.2 Ποσοστά προσκόλλησης για μοντέλα διαφορετικού μεγέθους στην οικογένεια Mistral (μορφή πολλαπλών επιλογών). . . . .	31
1.4.3 Αποτελέσματα για τα μοντέλα των οικογενειών Mistral και Llama στις ερωτήσεις του τρίτου επιπέδου δυσκολίας με προτροπές χωρίς παραδείγματα. Η σειρά των ράβδων ανά τύπο/επίπεδο επαναορισμού αντιστοιχεί σε αύξουσα σειρά μεγέθους μοντέλου. . . . .	31
1.4.4 Ανάλυση απαντήσεων για το Llama 70B στο πρώτο επίπεδο ερωτήσεων και όλες τις στρατηγικές προτροπών. Σε κάθε τύπο/επίπεδο επαναορισμού, οι ράβδοι αντιστοιχούν με τη σειρά σε: χωρίς πασαδείγματα, με αλυσίδες σκέψης, με παραδείγματα. . . . .	32
1.4.5 Ανάλυση απαντήσεων για το Claude 3.5 Sonnet στο τρίτο επίπεδο ερωτήσεων και όλες τις στρατηγικές προτροπών. Σε κάθε τύπο/επίπεδο επαναορισμού, οι ράβδοι αντιστοιχούν με τη σειρά σε: χωρίς πασαδείγματα, με αλυσίδες σκέψης, με παραδείγματα. . . . .	32
1.4.6 Ανάλυση απαντήσεων Claude 3.7 Sonnet χωρίς και με Thinking. . . . .	33
1.4.7 Σύγκριση των ποσοστών προσκόλλησης για τις ερωτήσεις $Q_3$ και το επίπεδο επαναορισμών $R_{s2}$ για όλα τα MGM. . . . .	33
1.4.8 Σύγκριση των απαντήσεων των Mistral7B και Mistral Large (123B) σε ερωτήσεις πολλαπλών επιλογών για επαναορισμούς μονάδων μέτρησης. . . . .	36
1.4.9 Αποτελέσματα για τα μοντέλα των οικογενειών Mistral και Llama στις ερωτήσεις του τρίτου επιπέδου δυσκολίας με προτροπές χωρίς παραδείγματα. Η σειρά των ράβδων ανά τύπο/επίπεδο επαναορισμού αντιστοιχεί σε αύξουσα σειρά μεγέθους μοντέλου. . . . .	37
3.1.1 Empirical scaling laws showing power-law relationships between model performance and compute, dataset size, and parameter size. [51]. . . . .	44
3.1.2 Emergent abilities of LLMs: Performance of large language models on eight benchmark tasks, showing a sudden jump in capability once the model surpasses a certain parameter scale threshold. [139]. . . . .	45
3.2.1 Prompt template example for the task of tweet sentiment analysis [105]. . . . .	46
3.2.2 Comparison between the Zero-Shot, One-Shot and Few-Shot prompting techniques with examples on the English-to-French translation task [8]. . . . .	47
3.2.3 Application of Chain-of-Thought prompting for arithmetic reasoning [14]. . . . .	48
3.3.1 Overview of the LLMs-as-judges systems [64]. . . . .	49
4.0.1 Redefined reasoning pathways. . . . .	55
5.1.1 Number of anchored responses for models of varying sizes in the Llama family (MC response format). . . . .	75
5.1.2 Number of anchored responses for models of varying sizes in the Mistral family (MC response format). . . . .	76
5.1.3 Comparison of Mistral 7B and Mistral Large responses on the MC response format. . . . .	76
5.1.4 Comparison of Mistral 7B and Mistral Large responses on the MC response format. . . . .	77

5.1.5 Results for the different Mistral and Llama models on $Q_3$ questions using ZS prompting. The order of the bars per redefinition type/level corresponds to increasing model size. The color coding is the same as in Figure 5.1.3. . . . .	78
5.1.6 Response breakdown for Llama 70B on $Q_1$ -level questions across all prompting strategies. Within each redefinition type/level, the bars are ordered as follows: Zero-Shot, Chain-of-Thought, and Few-Shot. . . . .	79
5.1.7 Response breakdown for Claude 3.5 Sonnet on $Q_3$ -level questions across all prompting strategies. Within each redefinition type/level, the bars are ordered as follows: Zero-Shot, Chain-of-Thought, and Few-Shot. . . . .	79
5.1.8 Claude 3.7 Sonnet results without and with Thinking. . . . .	80
5.1.9 Comparison of the anchored response rate for $Q_3$ questions in the $R_{s2}$ redefinition level for all LLMs. . . . .	81
5.1.10 Completely wrong responses breakdown for Llama8B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses. . . . .	84
5.1.11 Completely wrong responses breakdown for Llama70B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses . . . . .	85
5.1.12 Completely wrong responses breakdown for Mixtral8x7. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses . . . . .	86
5.1.13 Completely wrong responses breakdown for Claude v2. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses . . . . .	87
5.2.1 Comparison of Mistral7B and Mistral Large (123B) responses on the MC response format for units of measure redefinitions. . . . .	90
5.2.2 Results for the different Mistral and Llama models on $Q_3$ questions using ZS prompting. The order of the bars per redefinition type/level corresponds to increasing model size. . . . .	91
5.2.3 Completely wrong responses breakdown for Mistral 7B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses . . . . .	94
5.2.4 Completely wrong responses breakdown for Mixtral 8x7B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses . . . . .	95
5.2.5 Completely wrong responses breakdown for Mistral Large. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses . . . . .	96

# List of Tables

1.1	Επίπεδα δυσκολίας επαναορισμών σταθερών (ανάθεση και αντικατάσταση).	25
1.2	Επαναορισμοί των σχέσεων μεταξύ μονάδων μέτρησης.	26
1.3	Ποσοστά προσκόλλησης όλων των MGM με προτροπές χωρίς παραδείγματα για τις πιο δύσκολες περιπτώσεις επαναορισμών σταθερών. Το υψηλότερο ποσοστό για κάθε οικογένεια μοντέλων επισημαίνεται με <b>έντονη γραφή</b> .	29
1.4	Μέση τιμή συσχέτισης μεταξύ επίδοσης στην εργασία χωρίς επαναορισμό και ποσοστών προσκόλλησης για τη στρατηγική χωρίς παραδείγματα. Τα κελιά με <b>ροζ</b> χρώμα υποδηλώνουν <b>υψηλή θετική συσχέτιση</b> ( $> 0.3$ ), ενώ αυτά με <b>πράσινο</b> χρώμα <b>υψηλή αρνητική συσχέτιση</b> ( $< -0.3$ ).	29
1.5	Ποσοστό των ορθών απαντήσεων χωρίς επαναορισμό (NR) και ποσοστό προσκόλλησης για ερωτήσεις ανοιχτού τύπου επαναορισμού μονάδων μέτρησης (χωρίς παραδείγματα). Τα χρωματισμένα κελιά υποδεικνύουν ποσοστά προσκόλλησης που αυξάνονται με το μέγεθος των MGM.	30
1.6	Μέσα ποσοστά άρνησης για όλα τα MGM (μικρότερες τιμές σε <b>έντονη γραφή</b> και μεγαλύτερες τιμές <u>υπογραμμισμένες</u> ). Δεν συμπεριλαμβάνονται τα μοντέλα που σημείωσαν μηδενικά ποσοστά σε όλες τις περιπτώσεις.	34
1.7	Ποσοστά προσκόλλησης όλων των MGM με προτροπές χωρίς παραδείγματα για τις πιο δύσκολες περιπτώσεις επαναορισμών μονάδων μέτρησης. Το υψηλότερο ποσοστό για κάθε οικογένεια μοντέλων επισημαίνεται με <b>έντονη γραφή</b> .	35
1.8	Ποσοστό των ορθών απαντήσεων χωρίς επαναορισμό (NR) και ποσοστό προσκόλλησης για ερωτήσεις ανοιχτού τύπου επαναορισμού μονάδων μέτρησης (χωρίς παραδείγματα).	36
1.9	Μέση τιμή συσχέτισης μεταξύ επίδοσης στην εργασία χωρίς επαναορισμό και ποσοστών προσκόλλησης για τη στρατηγική χωρίς παραδείγματα. Τα κελιά με <b>ροζ</b> χρώμα υποδηλώνουν <b>υψηλή θετική συσχέτιση</b> ( $> 0.3$ ), ενώ αυτά με <b>πράσινο</b> χρώμα <b>υψηλή αρνητική συσχέτιση</b> ( $< -0.3$ ).	37
1.10	Μέση τιμή συσχέτισης μεταξύ επίδοσης στην εργασία χωρίς επαναορισμό και ποσοστών προσκόλλησης για τη στρατηγική με παραδείγματα. Τα κελιά με <b>ροζ</b> χρώμα υποδηλώνουν <b>υψηλή θετική συσχέτιση</b> ( $> 0.3$ ), ενώ αυτά με <b>πράσινο</b> χρώμα <b>υψηλή αρνητική συσχέτιση</b> ( $< -0.3$ ).	38
1.11	Μέση τιμή συσχέτισης μεταξύ επίδοσης στην εργασία χωρίς επαναορισμό και ποσοστών προσκόλλησης για τη στρατηγική με αλυσίδες σχέψης. Τα κελιά με <b>ροζ</b> χρώμα υποδηλώνουν <b>υψηλή θετική συσχέτιση</b> ( $> 0.3$ ), ενώ αυτά με <b>πράσινο</b> χρώμα <b>υψηλή αρνητική συσχέτιση</b> ( $< -0.3$ ).	38
4.1	Varying levels of difficulty for constant redefinitions (assignments and swaps).	56
4.2	Questions of three difficulty levels ( $Q_1$ , $Q_2$ , $Q_3$ ) for units of measure.	57
4.3	Redefinitions of unit scaling between base and derived units.	59
4.4	Questions of three difficulty levels ( $Q_1$ , $Q_2$ , $Q_3$ ) for units of measure.	60
4.5	NR prompts for FF format across ZS, CoT and FS prompting strategies.	63
4.6	NR prompts for MC format across ZS, CoT and FS prompting strategies.	64
4.7	R prompts for FF format across ZS, CoT and FS prompting strategies.	66
4.8	R prompts for MC format across ZS, CoT and FS prompting strategies.	67

5.1	Anchoring response rate for all LLMs tested using ZS prompting for the most difficult cases in <i>assignment</i> ( $R_a3$ ) and <i>swapping</i> ( $R_s2$ ) redefinitions. The highest anchoring rate for each LLM family is marked in <b>bold</b> .	73
5.2	Correlation between average NR correct response rate with anchored response rate for each redefinition and question level in ZS setup. Cells in <b>pink</b> indicate a <b>high positive correlation</b> ( $> 0.3$ ), while cells in <b>green</b> indicate a <b>high negative correlation</b> ( $< -0.3$ ).	74
5.3	Correct response rate without redefinition (NR) versus post-redefinition anchoring rate in the free-form (FF) format, for LLMs with known sizes using ZS prompting. Colored cells indicate elevated anchoring with LLM scale.	75
5.4	Correlation between model performance before redefinition with the percentage of anchored answers for each type of constant redefinition and question level in FS setup. Cells highlighted in <b>pink</b> indicate a <b>high positive correlation</b> ( $> 0.3$ ), while cells in <b>green</b> indicate a <b>high negative correlation</b> ( $< -0.3$ ).	82
5.5	Correlation between model performance before redefinition with the percentage of anchored answers for each type of constant redefinition and question level in CoT setup. Cells highlighted in <b>pink</b> indicate a <b>high positive correlation</b> ( $> 0.3$ ), while cells in <b>green</b> indicate a <b>high negative correlation</b> ( $< -0.3$ ).	82
5.6	Average refusal rates over all question levels (lowest values in <b>bold</b> and highest values <u>underlined</u> ). We exclude LLMs with zero refusal rate overall.	83
5.7	The percentage of anchored responses for the models in the ZS setup for the most difficult constants redefinitions in <i>assignment</i> ( $R_a3$ ) and <i>swapping</i> ( $R_s2$ ). The highest number for each model family is presented in <b>bold</b> . We exclude models where no refusals occurred, as their results are identical to those in Table 5.1.	87
5.8	The percentage of anchored responses for all LLMs tested under the ZS prompting setup for the most difficult units of measure redefinitions ( $R_a2$ and $R_a3$ levels). The highest rate for each model family is presented in <b>bold</b> .	88
5.9	Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in ZS setup. Cells highlighted in <b>pink</b> indicate a <b>high positive correlation</b> ( $> 0.3$ ), while cells in <b>green</b> indicate a <b>high negative correlation</b> ( $< -0.3$ ).	89
5.10	The percentage of correct responses with no redefinition (NR) and the anchored response rate for units of measure redefinitions regarding free-form (FF) responses using ZS prompting.	90
5.11	Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in FS setup. Cells highlighted in <b>pink</b> indicate a <b>high positive correlation</b> ( $> 0.3$ ), while cells in <b>green</b> indicate a <b>high negative correlation</b> ( $< -0.3$ ).	92
5.12	Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in CoT setup. Cells highlighted in <b>pink</b> indicate a <b>high positive correlation</b> ( $> 0.3$ ), while cells in <b>green</b> indicate a <b>high negative correlation</b> ( $< -0.3$ ).	92







## Chapter 1

# Εκτεταμένη Περίληψη στα Ελληνικά

## 1.1 Εισαγωγή

Παρουσιάζοντας αξιοσημείωτη ικανότητα στην κατανόηση και παραγωγή ανθρώπινου λόγου, τα Μεγάλα Γλωσσικά Μοντέλα (MGM) αποτελούν ένα εξαιρετικά σημαντικό βήμα προόδου στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (NLP). Με την αύξηση της κλίμακας τους, αναδύονται νέες και απρόσμενες συμπεριφορές ([139]; [113]), όπως η επίλυση σύνθετων προβλημάτων συλλογιστικής [94], παρόλο που η εκπαίδευσή τους βασίζεται αποκλειστικά στην πρόβλεψη της επόμενης λέξης. Αυτές οι ικανότητες, που κάποτε θεωρούνταν αποκλειστικά ανθρώπινες, αν και σαφώς εντυπωσιακές, συνοδεύονται από αμφιβολίες σχετικά με το κατά πόσο προκύπτουν από πραγματική λογική επεξεργασία ή είναι απλώς προϊόντα απομνημόνευσης και αναγνώρισης προτύπων [141]. Ειδικά σε πειραματικές συνθήκες που περιλαμβάνουν αφύσικες διατυπώσεις, παραπλανητικά συμφραζόμενα ή αντιφατικές πληροφορίες, έχει φανεί ότι τα μοντέλα, αντί να ακολουθήσουν γνήσιες πορείες συλλογιστικής, αποτυγχάνουν, εμφανίζοντας συμπεριφορές που οφείλονται σε επιφανειακή επεξεργασία ([140]; [67]; [63]). Ιδιαίτερο ενδιαφέρον παρουσιάζουν κάποια προβλήματα στα οποία τα μεγαλύτερα μοντέλα αποδίδουν χειρότερα από τα μικρότερα, ένα φαινόμενο γνωστό ως αντίστροφη κλιμάκωση (inverse scaling) [82], το οποίο αντιβαίνει τους εδραιωμένους νόμους σχετικά με την κλίμακα των MGM που προβλέπουν καλύτερη επίδοση με την αύξηση του μεγέθους [51].

Η αντίστροφη κλιμάκωση, παρόλο που έχει σοβαρές επιπτώσεις για την αξιοπιστία των MGM, αποκαλύπτοντας περιορισμούς που δεν είναι εμφανείς υπό κανονικές συνθήκες, παραμένει ένα σχετικά ανεξερεύνητο φαινόμενο. Στο πλαίσιο αυτό, η παρούσα διπλωματική εργασία εστιάζει στην αποκαλούμενη Εργασία Επαναορισμού (Redefinition Task), η οποία προτάθηκε στον διαγωνισμό Inverse Scaling Prize [82] και εξετάζει την ικανότητα των μοντέλων να παρακάμψουν βαθιά ριζωμένη γνώση όταν τους δίνονται εναλλακτικοί ορισμοί γνωστών εννοιών. Η ανάλυσή μας εξετάζει ερωτήσεις που αφορούν επαναορισμούς βασικών επιστημονικών σταθερών και μονάδων μέτρησης, αξιολογώντας πολλαπλά σενάρια που συνδυάζουν διαφορετικές τεχνικές προτροπής, είδη μορφοποίησης, και επίπεδα δυσκολίας. Μέσα από πειράματα με μοντέλα διαφόρων παραμέτρων, διερευνάται το φαινόμενο της προσκόλλησης (anchoring) στις παγιωμένες τιμές σε σχέση με το μέγεθος των MGM και με ποιον τρόπο αυτό επηρεάζεται από τις παραμέτρους του πειραματικού σχεδιασμού. Στόχος είναι να αναδειχθούν τα όρια της λογικής ευελιξίας των σύγχρονων MGM και οι συνέπειες για την αξιοπιστία τους σε κρίσιμα περιβάλλοντα.

## 1.2 Θεωρητικό Υπόβαθρο

### 1.2.1 Εισαγωγή στα Μεγάλα Γλωσσικά Μοντέλα

Τα Μεγάλα Γλωσσικά Μοντέλα (MGM) είναι συστήματα τεχνητής νοημοσύνης που εκπαιδεύονται σε τεράστιους όγκους δεδομένων με στόχο την κατανόηση και παραγωγή φυσικής γλώσσας. Βασίζονται κατά κύριο λόγο στην αρχιτεκτονική του Μετασχηματιστή (Transformer) [123], η οποία με τη χρήση μηχανισμών αυτοπροσοχής επιτρέπει την αποδοτική μοντελοποίηση των εννοιολογικών σχέσεων μεταξύ λέξεων ή συμβόλων σε ένα κείμενο, ανεξάρτητα από τη θέση τους σε αυτό, αναβαθμίζοντας σημαντικά τις επιδόσεις των μοντέλων σε εργασίες επεξεργασίας φυσικής γλώσσας [83]. Τα κείμενα εισάγονται στα μοντέλα ως ακολουθίες από tokens, τα οποία μπορεί να είναι σύμβολα, λέξεις ή υπολέξεις, και η εκπαίδευση των MGM βασίζεται ουσιαστικά στην πρόβλεψη του επόμενου token μέσω τεχνικών αυτο-επιβλεπόμενης μάθησης [136]. Αν και τα μοντέλα αποκτούν ήδη αξιόλογες ικανότητες κατά το στάδιο της προεκπαίδευσης ([8]; [22]), σε πολλές περιπτώσεις, για να ενισχυθεί η απόδοσή τους εφαρμόζονται τεχνικές στοχευμένης βελτιστοποίησης (fine-tuning) [86] όπως η μεταφορά μάθησης ([162]; [97]), η εκπαίδευση με οδηγίες ([150]; [24]; [137]) ή η ευθυγράμμιση μέσω ανθρώπινης ανατροφοδότησης (RLHF) [165]. Τα MGM, διεκδικώντας τις δυνατότητες της τεχνητής νοημοσύνης σε πρωτοφανή βαθμό, από απλοί παραγωγοί κειμένου αναδεικνύονται πλέον ως ισχυρά εργαλεία γενικής χρήσης, ικανά να επιλύουν σύνθετα γνωσιακά προβλήματα με εντυπωσιακές επιδόσεις που πλησιάζουν ή και ξεπερνούν το ανθρώπινο επίπεδο [156].

Η προεκπαίδευση των MGM τυπικά περιλαμβάνει δεκάδες έως και εκατοντάδες δισεκατομμύρια παραμέτρους, οι οποίες διαμορφώνονται για να βελτιστοποιήσουν την ικανότητα των μοντέλων να προβλέπουν σωστά [27]. Η εμπειρική έρευνα στη μοντελοποίηση γλώσσας έχει δείξει πως υπάρχει σαφής και συστηματική σχέση μεταξύ του μεγέθους των MGM και της απόδοσής τους, με τα μεγαλύτερα μοντέλα να επιτυγχάνουν καλύτερα αποτελέσματα σε ένα ευρύ φάσμα εφαρμογών ([96]; [22]). Μάλιστα, αυτό το φαινόμενο έχει αποτυπωθεί ποσοτικά μέσα από τους αποκαλούμενους νόμους κλιμάκωσης (scaling laws), οποίοι περιγράφουν πώς η απόδοση των μοντέλων ακολουθεί προβλέψιμη πορεία βελτίωσης καθώς αυξάνονται τρεις βασικοί παράγοντες: το μέγεθος του μοντέλου (αριθμός παραμέτρων), το μέγεθος του συνόλου δεδομένων και η διαθέσιμη υπολογιστική ισχύς. Μελέτες όπως

των Kaplan et al. [51] και Hoffmann et al. [42] διατύπωσαν διαφορετικές προσεγγίσεις, αλλά κατέληξαν στο ίδιο συμπέρασμα: ότι η κλιμάκωση των ΜΓΜ οδηγεί σε προβλέψιμες και σημαντικές βελτιώσεις απόδοσης. Έτσι, οι νόμοι κλιμάκωσης έχουν καταστεί πλέον θεμέλιο για την ανάπτυξη των όλο και μεγαλύτερων σύγχρονων, υψηλών επιδόσεων γλωσσικών μοντέλων.

Είναι σαφές, λοιπόν, ότι τα ΜΓΜ με την αύξηση της κλίμακάς τους έχουν σημειώσει εντυπωσιακή πρόοδο σε τυπικές εργασίες επεξεργασίας φυσικής γλώσσας ([161]; [151]; [110]; [135]; [50]), αλλά και σε νέες, πολυπρακτικές ([148]; [129]) και πολυπρακτορικές ([100]; [122]; [38]) εφαρμογές. Βέβαια, ένα από τα πιο αξιοσημείωτα χαρακτηριστικά αυτών των μοντέλων είναι η εμφάνιση ικανοτήτων που δεν προβλέπονται από τους νόμους κλιμάκωσης και δεν παρατηρούνται σε μικρότερα μοντέλα, αλλά εκδηλώνονται απότομα όταν το μέγεθος ξεπεράσει ένα συγκεκριμένο κατώφλι ([139]; [113]). Τέτοιες "αναδυόμενες" συμπεριφορές περιλαμβάνουν, μεταξύ άλλων, τη μάθηση εντός συμφραζομένων (in-context learning) ([26]; [158]), την παραγωγή προγραμμάτων (code generation) ([45]; [15]), την εκτέλεση πολύπλοκης συλλογιστικής [94] και την επίλυση γρίφων [35].

### 1.2.2 Μηχανική Προτροπών

Η προτροπή (prompting) περιλαμβάνει τη διατύπωση οδηγιών ή ενδείξεων που λειτουργούν ως είσοδοι προς το μοντέλο με στόχο την παραγωγή ορθών απαντήσεων χωρίς την ανάγκη αναπροσαρμογής των παραμέτρων του. Η συστηματική πρακτική σχεδιασμού και διατύπωσης αυτών των οδηγιών με τρόπο που οδηγεί αποδοτικά τη συμπεριφορά των μοντέλων προς την επιθυμητή κατεύθυνση ονομάζεται μηχανική προτροπών (prompt engineering) και έχει καθιερωθεί ως κρίσιμο εργαλείο για τη μεγιστοποίηση των δυνατοτήτων των ΜΓΜ, καθώς επιτρέπει την ευέλικτη προσαρμογή σε διαφορετικές εργασίες, αποφεύγοντας χρονοβόρες διαδικασίες εκπαίδευσης ([105]; [102]). Βέβαια, τα ΜΓΜ είναι ιδιαίτερα ευαίσθητα στην ακριβή διατύπωση των εισόδων, γεγονός που καθιστά την σωστή σχεδίαση καθοριστική πρόκληση ([105]; [73]).

Με σκοπό την εύρεση του "πιο κατάλληλου prompt", που θα εκμαιεύσει την επιθυμητή απόκριση από το μοντέλο, έχουν αναπτυχθεί διαφορετικές τεχνικές προτροπών. Ανάμεσα στις βασικότερες από αυτές είναι η τεχνική με μηδενικά παραδείγματα (Zero-Shot), στην οποία περιλαμβάνεται αποκλειστικά η οδηγία για την εκλήρωση της εκάστοτε εργασίας [102], οι τεχνικές ενός ή λίγων παραδειγμάτων (One-Shot και Few-Shot), όπου το μοντέλο λαμβάνει ένα ή περισσότερα παραδείγματα επιτυχημένης εκτέλεσης αντίστοιχα [8], και η προτροπή με αλυσίδες σκέψης (Chain-of-Thought), που ενθαρρύνει τα ΜΓΜ να εφκράσουν την συλλογιστική τους πορεία μέσα από ενδιάμεσα βήματα ([138]; [126]). Επίσης, για να διευκολυνθεί η αλληλεπίδραση με τα μοντέλα και να υποστηριχθεί η εφαρμογή τους σε εργασίες μεγάλης κλίμακας, συνήθως χρησιμοποιούνται πρότυπα προτροπών, δηλαδή παραμετροποιημένες δομές εισόδου στις οποίες ενσωματώνονται μεταβλητές που αντικαθίστανται κατά την πειραματική διαδικασία ([81]; [105]).

### 1.2.3 Αξιολόγηση με ΜΓΜ

Καθώς τα ΜΓΜ εξελίσσονται και εφαρμόζονται σε ένα ευρύ φάσμα γνωστικών πεδίων, γίνεται ολοένα και πιο επιτακτική η ανάγκη για αξιόπιστη και αποδοτική αξιολόγηση της απόδοσής τους. Οι παραδοσιακές μετρικές, όπως η ακρίβεια ή η ανάκληση, επαρκούν μόνο για περιορισμένο αριθμό εφαρμογών με συγκεκριμένα χαρακτηριστικά, ενώ ακόμα και πιο εξελιγμένες, όπως οι BLEU, ROUGE και METEOR αποτυγχάνουν να αποτυπώσουν ποιοτικά χαρακτηριστικά των γενετικών απαντήσεων των ΜΓΜ ([37]; [64]). Επίσης, η ανθρώπινη αξιολόγηση, παρόλο που θεωρείται η πιο αξιόπιστη λύση, είναι ιδιαίτερα δαπανηρή και δύσκολα επεκτάσιμη ([37]; [64]). Μία καινοτόμος προσέγγιση είναι η χρήση των ίδιων των ΜΓΜ ως αξιολογητών, γνωστή και ως LLM-as-a-judge [157]. Σε αυτό το πλαίσιο, τα μοντέλα καθοδηγούνται μέσω ειδικά σχεδιασμένων προτροπών ώστε να εκτιμούν την ποιότητα των απαντήσεων βάσει συγκεκριμένων κριτηρίων που καθορίζονται ανάλογα με τους στόχους κάθε εργασίας [64]. Τα μοντέλα μπορούν να λειτουργούν μόνα τους ([72]; [74]; [149]), σε συνδυασμό με άλλα ΜΓΜ ([10]; [23]; [69]) ή και σε συνεργασία με ανθρώπους ([68]; [107]), πετυχαίνοντας αποτελέσματα που συχνά είναι πολύ κοντά στις ανθρώπινες κρίσεις ([16]; [33]; [30]; [20]; [36]; [11]; [111]).

Παρα τις υποσχέσεις της, ωστόσο, το ερώτημα της αξιοπιστίας αυτής της μεθόδου παραμένει, με την ερευνητική κοινότητα να στρέφεται σε τεχνικές μετα-αξιολόγησης (meta-evaluation), προκειμένου να μετρήσει τη συμφωνία μεταξύ ΜΓΜ-κριτών και ανθρώπινων προτιμήσεων και να εντοπίσει συστηματικές προκαταλήψεις [64], οι οποίες μπορεί να σχετίζονται, για παράδειγμα, με τη θέση, την έκταση ή το κύρος των απαντήσεων ([157]; [145]). Τα αποτελέσματα είναι ενθαρρυντικά: τα ΜΓΜ μπορούν, με κατάλληλη καθοδήγηση, να λειτουργήσουν ως

αξιόπιστοι και ευέλικτοι αξιολογητές, προσφέροντας ένα βιώσιμο εναλλακτικό εργαλείο όταν οι παραδοσιακές πρακτικές δεν επαρκούν ή δεν είναι πρακτικά εφαρμόσιμες.

### 1.2.4 Συλλογιστική στα MFM

Μεγάλο ενδιαφέρον έχει προκληθεί σχετικά με το κατά πόσο τα MFM μπορούν να επιδείξουν γνήσιες δυνατότητες λογικής σκέψης [28]. Η πρόσφατη έρευνα έχει επικεντρωθεί σε διάφορες μορφές συλλογιστικής [94], όπως η επαγωγική ([65]; [7]), η παραγωγική ([18]; [25]), η αιτιατική ([48]; [133]), η αναλογική ([95]; [117]), η αριθμητική σκέψη ([84]; [130]) και η κοινή λογική ([118]; [103]). Αν και η πρόοδος σε αυτούς τους τομείς είναι αξιοσημείωτη, η ικανότητα συλλογιστικής των MFM παραμένει περιορισμένη, ειδικά όταν συγκρίνεται με την επιτυχία τους σε παραδοσιακές γλωσσικές εφαρμογές [80]. Το χάσμα αυτό, μάλιστα, γίνεται ακόμα πιο εμφανές όταν τα μοντέλα καλούνται να απαντήσουν σε ερωτήματα που παρουσιάζονται σε ασυνήθιστες μορφές, υπό παραπλανητικά συμφραζόμενα ή περιέχουν υποθετικές δηλώσεις και πληροφορίες που αντιβαίνουν στη γενική γνώση ([140]; [67]; [147]; [63]). Η αδυναμία αυτή υποδηλώνει περιορισμένη γνωστική προσαρμοστικότητα, με τα MFM συχνά να "παπαγαλίζουν" πρότυπα ή δεδομένα που έχουν εσωτερικεύσει κατά την εκπαίδευσή τους. Έτσι, πρόσφατη έρευνα αποδίδει τις επιτυχίες τους περισσότερο στην απομνημόνευση παραδειγμάτων και την αντιστοίχιση μοτίβων, παρά σε γνήσια κατανόηση και ικανότητα. Μελέτες στρέφονται προς μεθόδους αξιολόγησης των μοντέλων που διαχωρίζουν την απομνημόνευση από την αυθεντική λογική σκέψη, ώστε να κατανοηθεί καλύτερα η πραγματική φύση της "νοημοσύνης" που επιδεικνύουν τα MFM ([141]; [77]; [131]).

### 1.2.5 Προβλήματα αντίστροφης κλιμάκωσης

Παρόλο που η αύξηση του μεγέθους των γλωσσικών μοντέλων οδηγεί συνήθως σε καλύτερη απόδοση [51], πρόσφατες μελέτες έχουν εντοπίσει περιπτώσεις στις οποίες συμβαίνει το αντίθετο: τα μεγαλύτερα μοντέλα αποδίδουν χειρότερα από τα μικρότερα. Αυτό το παράδοξο φαινόμενο ονομάζεται αντίστροφη κλιμάκωση (inverse scaling) και εκθέτει τις αδυναμίες ακόμα και των πιο ισχυρών MFM, αποκαλύπτοντας αποκλίσεις μεταξύ των συλλογιστικών διαδικασιών τους και των ανθρώπινων επιδόσεων. Για να μελετηθούν συστηματικά τέτοιες περιπτώσεις, θεσπίστηκε ο διαγωνισμός Inverse Scaling Prize [82], όπου συλλέχθηκαν εργασίες στις οποίες τα μεγαλύτερα μοντέλα αποτυγχάνουν συστηματικά και ταξινομήθηκαν σε τέσσερις κατηγορίες βάσει των πιθανών αιτιών που προκαλούν το συγκεκριμένο φαινόμενο. Οι κατηγορίες αυτές είναι:

1. **Ισχυρά Προκαθορισμένα Προηγούμενα (Strong Prior):** Το μοντέλο δυσκολεύεται να παρακάμψει τη γνώση που έχει μάθει κατά την προεκπαίδευση, ακόμα και όταν αυτό ζητείται ρητά από τις οδηγίες.
2. **Ανεπιθύμητη Μίμηση (Unwanted Imitation):** Τα μοντέλα μιμούνται λογικά σφάλματα ή μεροληψίες που περιέχονται στα δεδομένα προεκπαίδευσης.
3. **Παραπλανητικά Ερεθίσματα (Distractor Tasks):** Οι προτροπές περιλαμβάνουν έμμεσα πιο εύκολες, αλλά παραπλανητικές εναλλακτικές εργασίες, στις οποίες τα μοντέλα τείνουν να δίνουν προτεραιότητα λόγω της εξοικειώσής τους με παρόμοια μοτίβα.
4. **Ψευδείς Ενδείξεις από Παραδείγματα (Spurious Few-Shot):** Τα παραδείγματα που παρέχονται στις προτροπές οδηγούν το μοντέλο σε λανθασμένα μοτίβα, τα οποία τείνουν να ακολουθούν μηχανικά, αγνοώντας τη λογική του ερωτήματος.

Η παρούσα εργασία επικεντρώνεται στο πρόβλημα του επαναορισμού (Redefinition), το οποίο εντάσσεται στην κατηγορία των Ισχυρά Προκαθορισμένων Προηγούμενων και απαιτεί την υιοθέτηση ενός εναλλακτικού ορισμού μίας γνωστής έννοιας. Προτείνουμε ότι η εργασία αυτή ως χαρακτηριστική περίπτωση του φαινομένου αντίστροφης κλιμάκωσης χρήζει συστηματικής διερεύνησης. Προς αυτήν την κατεύθυνση δημιουργούμε δύο εξειδικευμένα σύνολα δεδομένων που εξετάζουν επαναπροσδιορισμούς σε διαφορετικά σημασιολογικά πεδία και σενάρια και αξιολογούμε την επίδοση σύγχρονων MFM, αναλύοντας τις μεταβολές στις συμπεριφορές τους σε σχέση με το μέγεθός τους.

## 1.3 Μέθοδος

### 1.3.1 Σύνολα Δεδομένων

Τα σύνολα δεδομένων που κατασκευάστηκαν για την αξιολόγηση των ΜΓΜ σε εργασίες επαναορισμού αποτελούνται από δύο διακριτά μέρη: 1) Επαναορισμός Φυσικών Σταθερών και 2) Επαναορισμός Μονάδων Μέτρησης.

#### 1.3.1.1 Επαναορισμός Φυσικών Σταθερών

Για την εργασία επαναορισμού φυσικών σταθερών επιλέξαμε τις εξής ευρέως αναγνωρισμένες μαθηματικές και φυσικές σταθερές: το  $\pi$  ( $\pi$ ), τον αριθμό του Euler ( $e$ ), τον χρυσό λόγο ( $\phi$ ), την ταχύτητα του φωτός ( $c$ ), τη σταθερά της βαρύτητας ( $G$ ), τη σταθερά του Planck ( $h$ ), το στοιχειώδες φορτίο ( $q_e$ ), τον αριθμό του Avogadro ( $N_A$ ), τη σταθερά του Boltzmann ( $k_B$ ), τη σταθερά των ιδανικών αερίων ( $\bar{R}$ ), τη φανταστική μονάδα ( $i$ ), τη τετραγωνική ρίζα του 2 ( $\sqrt{2}$ ), το άπειρο ( $\infty$ ), τη διηλεκτρική σταθερά του κενού ( $\epsilon_0$ ) και το μηδέν.

	Πραγματική Τιμή	Μονάδα	$R_a1$	$R_a2$	$R_a3$	$R_s1$	$R_s2$
$\pi$	3.14159	-	4.5	500	-10	$\phi$	$h$
$e$	2.71828	-	9	1300	$1.5 \times 10^{-12}$	$pi$	$k_B$
$\phi$	1.61803	-	3.6	321	-2.2	$e$	$N_A$
$c$	299,792,458	$m/s$	$2.3 \times 10^8$	10	$-4 \times 10^8$	$N_A$	$q_e$
$G$	$6.674 \times 10^{-11}$	$m^3/kg * s^2$	$1.1 \times 10^{-10}$	50	-525	$q_e$	$pi$
$h$	$6.626 \times 10^{-34}$	$J * s$	$5 \times 10^{-33}$	482	-0.2	$k_B$	$\phi$
$q_e$	$1.602 \times 10^{-19}$	$C$	$2.4 \times 10^{-21}$	$3 \times 10^4$	$3 \times 10^{50}$	$\epsilon_0$	$\pi$
$N_A$	$6.022 \times 10^{23}$	$mol^{-1}$	$8.23 \times 10^{23}$	75	-1	$\bar{R}$	$e$
$k_B$	$1.380649 \times 10^{-23}$	$J/K$	$4.56 \times 10^{-24}$	80	$-9.9 \times 10^{-3}$	$\epsilon_0$	$pi$
$\bar{R}$	8.314	$J/(mol * K)$	13	3500	-400	$\pi$	$c$
$i$	$\sqrt{-1}$	-	$\sqrt{-2}$	$\sqrt{-100}$	1	$\phi$	$\bar{R}$
$\sqrt{2}$	1.41421356	-	5	31.62	-2	$\pi$	$\epsilon_0$
$\infty$	infinity has no value	-	$10^{10}$	100	-1	$c$	$q_e$
$\epsilon_0$	$8.854 \times 10^{-12}$	$F/m$	$9.3 \times 10^{-10}$	35	$3 \times 10^{12}$	$G$	$\phi$
zero	0	-	-1	100	$5 \times 10^{30}$	$h$	$c$

Table 1.1: Επίπεδα δυσκολίας επαναορισμών σταθερών (ανάθεση και αντικατάσταση).

Για να εξετάσουμε την προσαρμοστικότητα των μοντέλων, σχεδιάσαμε δύο τύπους επαναορισμών, καθένας με κλιμακούμενα επίπεδα δυσκολίας:

- **Ανάθεση ( $R_a$ ):** Η σταθερά λαμβάνει μία τυχαία επιλεγμένη τιμή.
  1.  $R_a1$ : Μικρή απόκλιση από την αρχική τιμή (π.χ., " $\pi = 4.5$ ").
  2.  $R_a2$ : Σημαντική απόκλιση, κατά τάξεις μεγέθους (π.χ., " $\pi = 500$ ").
  3.  $R_a3$ : Ακραίες ή παράλογες τιμές (π.χ., " $\pi = -10$ ").
- **Αντικατάσταση ( $R_s$ ):** Η τιμή της σταθεράς αντικαθίσταται με αυτή κάποιας άλλης γνωστής σταθεράς.
  1.  $R_s1$ : Αντικατάσταση μεταξύ σταθερών με κοντινές τιμές (π.χ., " $\pi = \phi$ ").
  2.  $R_s2$ : Αντικατάσταση μεταξύ σταθερών με σημαντικά μακρινές τιμές (π.χ., " $\pi = h$ ").

Παράλληλα, σχεδιάστηκαν τρία επίπεδα ερωτήσεων κλιμακούμενης δυσκολίας:

1. **Απλή Ανάκληση ( $Q_1$ ):** Η απάντηση προκύπτει άμεσα από την τιμή της σταθεράς (π.χ., Ποιο είναι το πρώτο μη μηδενικό ψηφίο του  $\pi$ ;").
2. **Εύκολος Υπολογισμός ( $Q_2$ ):** Το μοντέλο εκτελεί έναν απλό μαθηματικό υπολογισμό με βάση την τιμή της σταθεράς (π.χ., "Πόσο κάνει  $\pi$  επί 3;").
3. **Πολυσταδιακή Συλλογιστική ( $Q_3$ ):** Το μοντέλο καλείται να επιλύσει ένα σύνθετο μαθηματικό ή φυσικό πρόβλημα που απαιτεί πολλαπλά βήματα σκέψης (π.χ., "Ποια είναι η επιφάνεια της Γης;").

### 1.3.1.2 Επαναορισμός Μονάδων Μέτρησης

Για τη δεύτερη εργασία επαναορισμού επιλέξαμε βασικές μονάδες μέτρησης στις εξής θεμελιώδεις φυσικές ποσότητες: χρόνος (λεπτό-*min*), βάρος (κιλό-*kg*), μήκος (μέτρο-*m*) και έτος φωτός (*ly*), θερμοκρασία (Κέλβιν-*K*), όγκος (χιλιοστόλιτρο-*mL*), ενέργεια (θερμίδα-*cal*), πίεση (ατμόσφαιρα-*atm*), τάση (Volt-*V*), συχνότητα (megaHz-*MHz*), δύναμη (newton-*N*), πυκνότητα μαγνητικής ροής (Tesla-*T*), εμβαδόν (εκτάριο-*ha*), φωτεινότητα (lux-*lx*), and αποθήκευση πληροφορίας (byte-*B*).

Για τον επαναορισμό μονάδων μέτρησης αλλάζουμε τη σχέση τους με άλλες μονάδες της ίδιας φυσικής ποσότητας και όχι μία συγκεκριμένη τιμή, όπως στην περίπτωση των σταθερών. Έτσι, η αντικατάσταση μεταξύ μονάδων δεν είναι εφαρμόσιμη, οπότε περιοριστήκαμε σε απλή ανάθεση με τρία επίπεδα δυσκολίας:

1.  $R_a1$ : Μικρές αλλαγές στη σχέση (π.χ., "1 λεπτό = 100 δευτερόλεπτα").
2.  $R_a2$ : Μεγαλύτερες αποκλίσεις, κατά τάξεις μεγέθους (π.χ., "1 λεπτό =  $5 \times 10^8$  δευτερόλεπτα").
3.  $R_a3$ : Ακραίες ή μη ρεαλιστικές σχέσεις (π.χ., "1 λεπτό = -50 δευτερόλεπτα").

Και σε αυτήν την περίπτωση έχουμε τρία επίπεδα δυσκολίας ερωτήσεων:

1. **Άμεση Μετατροπή ( $Q_1$ )**: Βασικές ερωτήσεις μετατροπής μονάδων (π.χ., Πόσα δευτερόλεπτα έχεις σε δύο λεπτά;").
2. **Εφαρμοσμένη χρήση ( $Q_2$ )**: Απλά προβλήματα φυσικής, για άμεση εφαρμογή των μετατροπών (π.χ., "Ένα χρονόμετρο λειτουργεί για 3,5 λεπτά. Πόσα δευτερόλεπτα μετράει;").
3. **Σύνθετη Συλλογιστική ( $Q_3$ )**: Πολύπλοκα προβλήματα που απαιτούν πολλαπλά βήματα σκέψης και δύσκολους υπολογισμούς (π.χ., "Πόσα δευτερόλεπτα χρειάζεται ένας δρομέας για να διανύσει 42 χλμ. με ταχύτητα 170 μ./λεπτό;").

Μονάδα	Παράγωγη Μονάδα	Πραγματική Τιμή	$R_a1$	$R_a2$	$R_a3$
1 <i>min</i>	seconds ( <i>sec</i> )	60 <i>sec</i>	100 <i>sec</i>	$5 \times 10^8 \text{ sec}$	-50 <i>sec</i>
1 <i>kg</i>	grams ( <i>gr</i> )	1000 <i>gr</i>	900 <i>gr</i>	$10^{-14} \text{ gr}$	-100 <i>gr</i>
1 <i>m</i>	centimeter ( <i>cm</i> )	100 <i>cm</i>	60 <i>cm</i>	$310^{10} \text{ cm}$	-200 <i>cm</i>
<i>K</i>	Celsius degrees ( $^{\circ}C$ )	$^{\circ}C + 273.15$	$^{\circ}C + 300$	$^{\circ}C + 1$	$100 * (^{\circ}C) + 500$
1 <i>mL</i>	cubic centimeter ( $\text{cm}^3$ )	1 <i>cm</i> <sup>3</sup>	2 <i>cm</i> <sup>3</sup>	10000 <i>cm</i> <sup>3</sup>	-10 <i>cm</i> <sup>3</sup>
1 <i>cal</i>	Joule ( <i>J</i> )	4.184 <i>J</i>	9 <i>J</i>	1500 <i>J</i>	-5 <i>J</i>
1 <i>atm</i>	Pascal ( <i>Pa</i> )	101,325 <i>Pa</i>	215,000 <i>Pa</i>	0.55 <i>Pa</i>	-5000 <i>Pa</i>
1 <i>V</i>	millivolt ( <i>mV</i> )	1000 <i>mV</i>	500 <i>mV</i>	410 <sup>9</sup> <i>mV</i>	-10 <i>mV</i>
1 <i>MHz</i>	Hertz ( <i>Hz</i> )	10 <sup>6</sup> <i>Hz</i>	10 <sup>5</sup> <i>Hz</i>	2 <i>Hz</i>	-10 <sup>3</sup> <i>Hz</i>
1 <i>N</i>	millinewton ( <i>mN</i> )	1000 <i>mN</i>	900 <i>mN</i>	210 <sup>15</sup> <i>mN</i>	-3000 <i>mN</i>
1 <i>kW</i>	Watt ( <i>W</i> )	1000 <i>W</i>	1500 <i>W</i>	510 <sup>-5</sup> <i>W</i>	-30 <i>W</i>
1 <i>T</i>	millitesla ( <i>mT</i> )	1000 <i>mT</i>	600 <i>mT</i>	10 <sup>23</sup> <i>mT</i>	-90 <i>mT</i>
1 <i>ha</i>	square meter ( $\text{m}^2$ )	10,000 <i>m</i> <sup>2</sup>	10,500 <i>m</i> <sup>2</sup>	310 <sup>-4</sup> <i>m</i> <sup>2</sup>	-25 <i>m</i> <sup>2</sup>
1 <i>lx</i>	lumen per $\text{m}^2$ ( <i>lm/m</i> <sup>2</sup> )	1 <i>lm/m</i> <sup>2</sup>	0.5 <i>lm/m</i> <sup>2</sup>	1000 <i>lm/m</i> <sup>2</sup>	-19 <i>lm/m</i> <sup>2</sup>
1 <i>ly</i>	Trillion/Billion <i>km</i>	9.461 <i>Tkm</i>	9.461 <i>Bkm</i>	10 <i>m</i>	-2 <i>Tkm</i>
1 <i>B</i>	bit ( <i>b</i> )	8 <i>b</i>	10 <i>b</i>	610 <sup>8</sup> <i>b</i>	-4 <i>b</i>

Table 1.2: Επαναορισμοί των σχέσεων μεταξύ μονάδων μέτρησης.

### 1.3.1.3 Μορφοποίηση των Ερωτήσεων

Και στις δύο περιπτώσεις επαναορισμών χρησιμοποιήθηκαν δύο μορφές ερωτήσεων:

- **Ελεύθερης Απάντησης (Free-Form - FF)**: Το μοντέλο καλείται να δώσει μία ανοιχτού τύπου απάντηση, χωρίς να του παρέχονται επιλογές.
- **Πολλαπλών Επιλογών (Multiple Choice - MC)**: Για κάθε ερώτηση περιλαμβάνονται τέσσερις προτεινόμενες επιλογές (A, B, C, D), οι οποίες περιέχουν τη σωστή απάντηση βάσει του επαναορισμού, την αρχική απάντηση (πριν τον επαναορισμό) και δύο επιπλέον παραπλανητικές επιλογές.

#### 1.3.1.4 Υλοποίηση

Κάθε σύνολο δεδομένων υλοποιήθηκε σε αρχείο .csv και περιλαμβάνει πεδία για την επιλεγμένη σταθερά ή μονάδα μέτρησης, τον αρχικό ορισμό, την παραγόμενη ερώτηση, τους εναλλακτικούς ορισμούς, την απάντηση βάσει του αρχικού ορισμού, τις απαντήσεις βάσει των εναλλακτικών ορισμών και τις προτεινόμενες επιλογές για τη μορφή πολλαπλών επιλογών. Όλοι οι εναλλακτικοί ορισμοί, οι ερωτήσεις και οι παραπλανητικές επιλογές δημιουργήθηκαν χειροκίνητα και με τη βοήθεια του ChatGPT<sup>1</sup>. Για κάθε στοιχείο ζητήθηκε η παραγωγή πολλών προτάσεων, από τις οποίες επιλέχθηκαν και τροποποιήθηκαν εκείνες που εξυπηρετούσαν καλύτερα τους στόχους της μελέτης.

### 1.3.2 Μετρικές και Αξιολόγηση

Για την αξιολόγηση της απόδοσης των μοντέλων, οι παραγόμενες απαντήσεις κατηγοριοποιούνται σε τέσσερις τύπους:

- **Ορθές απαντήσεις χωρίς επαναορισμό (NR):** Το μοντέλο απαντά σωστά όταν δεν του ζητείται επαναορισμός της έννοιας.
- **Απαντήσεις με Προσκόλληση στη Γνώση:** Το μοντέλο αγνοεί τον επαναορισμό και βασίζεται στην απομνημονευμένη γνώση.
- **Ορθές απαντήσεις με επαναορισμό:** Το μοντέλο κατανοεί και εφαρμόζει σωστά τον επαναορισμό.
- **Πλήρως λανθασμένες απαντήσεις:** Απαντήσεις που δεν ανήκουν σε κάποια από τις υπόλοιπες κατηγορίες. Αυτές διακρίνονται σε κενές απαντήσεις, λάθος αποτελέσματα και περιπτώσεις στις οποίες το μοντέλο αρνήθηκε ρητά να απαντήσει στην ερώτηση.

Ως βασικές μετρικές αξιολόγησης των ικανοτήτων και των συμπεριφορικών τάσεων των μοντέλων χρησιμοποιήθηκαν τα ποσοστά εμφάνισης κάθε κατηγορίας απαντήσεων. Ιδιαίτερη έμφαση δόθηκε στα ποσοστά προσκόλλησης, καθώς αυτά αποκαλύπτουν την αδυναμία των ΜΓΜ να αποδεσμευτούν από την προηγούμενη γνώση. Επίσης, η συχνότητα άρνησης απάντησης μελετήθηκε ξεχωριστά, καθώς αντικατοπτρίζει την (υπερ)αυτοπεποίθηση ή επιφυλακτικότητα των μοντέλων. Τέλος, χρησιμοποιείται η συσχέτιση (correlation) για να διερευνηθεί η σχέση ανάμεσα στην προϋπάρχουσα γνώση (NR επιδόσεις) και τις αντιδράσεις των μοντέλων στους επαναορισμούς (π.χ. προσκόλληση ή άρνηση).

### 1.3.3 Σχεδίαση Προτροπών

Οι προτροπές που χρησιμοποιήθηκαν χωρίζονται σε τρεις βασικές κατηγορίες: 1) χωρίς επαναορισμό, 2) με επαναορισμό και 3) αξιολόγησης. Σε κάθε κατηγορία αντιστοιχούν παραλλαγές για μορφές ελεύθερης απάντησης και πολλαπλών επιλογών, καθώς και για στρατηγικές χωρίς παραδείγματα, με παραδείγματα και με αλυσίδες σκέψης.

Τα βασικά πρότυπα προτροπής που χρησιμοποιήθηκαν στα πειράματα χωρίς επαναορισμό για τη μορφή ελεύθερης απάντησης και πολλαπλών επιλογών αντίστοιχα είναι τα εξής (τεχνική χωρίς παραδείγματα):

---

Answer the following question:

{question}

End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary.

---

<sup>1</sup><https://chatgpt.com/>



Choose A, B, C or D to answer the question:

Question: {question}

A: {A}

B: {B}

C: {C}

D: {D}

Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D". End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary.

---

Η μεταβλητή **question** αντικαθίσταται κατά τη διάρκεια των πειραμάτων από τη συγκεκριμένη ερώτηση που καλείται να απαντήσει κάθε φορά το μοντέλο. Στις περιπτώσεις πολλαπλών επιλογών, οι μεταβλητές **A**, **B**, **C** and **D** αντιστοιχούν στις διαφορετικές επιλογές από τις οποίες το μοντέλο καλείται να επιλέξει τη σωστή. Για τη στρατηγική με αλυσίδες σκέψης προστίθεται η εντολή "Let's think step by step.", ενώ για τη στρατηγική με παραδείγματα προστίθεται στην προτροπή ένα προκαθορισμένο σύνολο ερωτοαποκρίσεων (στην ανάλογη μορφή), κοινό για όλες τις σταθερές ή μονάδες μέτρησης αντίστοιχα. Επίσης, για να διευκολύνουμε τη φάση επεξεργασίας και αξιολόγησης των αποκρίσεων, συμπεριλάβαμε την οδηγία να ολοκληρώνει κάθε έξοδο με τη φράση "The final answer is: " και την τελική απάντηση. Η προσέγγιση αυτή εφαρμόστηκε συστηματικά σε όλα τα πρότυπα με και χωρίς επαναορισμό.

Στην περίπτωση του επαναορισμού, προσθέσαμε απλά την οδηγία "Redefine {X} as {Y}." πριν από την ερώτηση προς το μοντέλο, όπου η μεταβλητή **X** αντιστοιχεί στην έννοια που επαναορίζεται και η **Y** στον νέο ορισμό που αποδίδεται στην **X**.

Για την αξιολόγηση των αποκρίσεων, χρησιμοποιώντας την τεχνική της αξιολόγησης με MFM, σχεδιάσαμε προτροπές στις οποίες ζητάμε από το μοντέλο-αξιολογητή να κατηγοριοποιήσει κάθε έξοδο στον κατάλληλο τύπο απάντησης. Στην περίπτωση χωρίς επαναορισμό, το μοντέλο καλείται να συγκρίνει την απόκριση του μοντέλου με τη σωστή, ενώ στην περίπτωση με επαναορισμό τη συγκρίνει τόσο με τη σωστή βάσει επαναορισμού όσο και με την αρχική, για να διακρίνουμε και τις περιπτώσεις προσκόλλησης. Επιπλέον, με τον ίδιο τρόπο σχεδιάσαμε κατάλληλη προτροπή για την περαιτέρω κατηγοριοποίηση των λανθασμένων απαντήσεων σε κενές/λάθος αποτελέσματα/αρνήσεις.

### 1.3.4 Επιλογή MFM

Στη μελέτη αυτή αξιολογήσαμε συνολικά 19 σύγχρονα MFM στην εργασία του επαναορισμού: Llama 3 (8/70/405B), Mistral7B/Large/Mixtral8×7b, Anthropic Claude (Opus/Instant/Haiku/v2/Sonnet 3.5&3.7), Cohere command (light/text/r/r+) και Amazon Titan (text lite/text express/large). Ως μοντέλο-αξιολογητή χρησιμοποιήσαμε το μοντέλο Claude 3.5 Sonnet.

### 1.3.5 Πειραματική Υλοποίηση

Τα πειράματα για τις εργασίες χωρίς επαναορισμό (NR) και με επαναορισμό (R), καθώς και η αξιολόγηση των αποκρίσεων με MFM, πραγματοποιήθηκαν σε περιβάλλον Kaggle Notebooks<sup>2</sup>, αξιοποιώντας NVIDIA T4 GPUs (T4x2) για υψηλή υπολογιστική απόδοση. Όλα τα μοντέλα που περιλαμβάνονται στη μελέτη προσπελάστηκαν μέσω της πλατφόρμας AWS Bedrock<sup>3</sup>. Η πρόσβαση διασφαλίστηκε με API κλήσεις, ελεγχόμενες μέσω του συστήματος διαχείρισης ταυτότητας και πρόσβασης AWS IAM.

---

<sup>2</sup><https://www.kaggle.com/>

<sup>3</sup><https://aws.amazon.com/bedrock/>



## 1.4 Πειραματικά Αποτελέσματα

### 1.4.1 Επαναορισμός Επιστημονικών Σταθερών

#### 1.4.1.1 Προσκόλληση στις Πραγματικές Τιμές

Τα αποτελέσματα δείχνουν ότι όλα μοντέλα, ανεξαρτήτως μεγέθους ή αρχιτεκτονικής, παρουσιάζουν σημαντικά ποσοστά προσκόλλησης στις απομνημονευμένες τιμές, ακόμα και όταν τους δίνεται σαφής οδηγία να τις παρακάμψουν. Το φαινόμενο παρατηρείται τόσο στις απαντήσεις ελεύθερης μορφής όσο και σε αυτές των πολλαπλών επιλογών, με το μεγαλύτερο ποσοστό να σημειώνεται από το μοντέλο Llama 405B, φτάνοντας στο 93.33% των απαντήσεων των πιο δύσκολων επιπέδων ερωτήσεων και αντικατάστασης.

Μοντέλο	$R_{a3}$						$R_{s2}$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	<b>33.33</b>	<b>46.67</b>	<b>33.33</b>	<b>26.67</b>	26.67	40.0	33.33	53.33	13.33	33.33	26.67	20.0
Mixtral8x7B	33.33	33.33	26.67	26.67	20.0	33.33	26.67	46.67	40.0	<b>53.33</b>	46.67	<b>73.33</b>
Mistral Large (123B)	33.33	20.0	26.67	26.67	<b>53.33</b>	<b>66.67</b>	<b>66.67</b>	<b>53.33</b>	<b>46.67</b>	40.0	<b>73.33</b>	66.67
Llama8B	0.0	<b>26.67</b>	0.0	<b>26.67</b>	13.33	33.33	20.0	13.33	<b>26.67</b>	40.0	20.0	20.0
Llama70B	<b>6.67</b>	13.33	0.0	0.0	13.33	40.0	33.33	46.67	13.33	<b>46.67</b>	33.33	73.33
Llama405B	0.0	0.0	0.0	13.33	<b>26.67</b>	<b>53.33</b>	<b>26.67</b>	<b>46.67</b>	6.67	20.0	<b>53.33</b>	<b>93.33</b>
Titan lite	13.33	20.0	20.0	20.0	0.0	40.0	40.0	33.33	20.0	33.33	6.67	26.67
Titan express	20.0	<b>26.67</b>	13.33	13.33	<b>20.0</b>	13.33	40.0	<b>53.33</b>	<b>20.0</b>	20.0	33.33	<b>26.67</b>
Titan large	<b>26.67</b>	20.0	<b>20.0</b>	6.67	13.33	<b>40.0</b>	<b>60.0</b>	40.0	13.33	<b>33.33</b>	<b>33.33</b>	20.0
Command r	0.0	6.67	<b>20.0</b>	<b>33.33</b>	<b>26.67</b>	<b>53.33</b>	<b>53.33</b>	13.33	20.0	6.67	<b>33.33</b>	<b>46.67</b>
Command r +	6.67	13.33	0.0	13.33	13.33	26.67	13.33	20.0	26.67	6.67	33.33	26.67
Command light text	6.67	13.33	13.33	20.0	0.0	40.0	13.33	20.0	<b>26.67</b>	<b>20.0</b>	13.33	13.33
Command text	<b>13.33</b>	<b>20.0</b>	6.67	6.67	6.67	26.67	40.0	<b>26.67</b>	13.33	26.67	13.33	33.33
Claude opus	13.33	0.0	6.67	6.67	33.33	<b>46.67</b>	<b>46.67</b>	<b>40.0</b>	20.0	<b>26.67</b>	53.33	73.33
Claude instant	0.0	13.33	13.33	<b>20.0</b>	26.67	46.67	33.33	20.0	33.33	40.0	46.67	60.0
Claude haiku	20.0	13.33	6.67	0.0	20.0	20.0	26.67	6.67	20.0	20.0	40.0	53.33
Claude v2	26.67	13.33	<b>20.0</b>	0.0	<b>46.67</b>	40.0	13.33	40.0	<b>33.33</b>	20.0	40.0	66.67
Claude 3.5 Sonnet	<b>26.67</b>	<b>13.33</b>	0.0	13.33	13.33	33.33	33.33	40.0	20.0	20.0	<b>60.0</b>	<b>73.33</b>
Claude 3.7 Sonnet	0.0	0.0	0.0	6.67	13.33	13.33	33.33	20.0	6.67	20.0	40.0	33.33

Table 1.3: Ποσοστά προσκόλλησης όλων των ΜΓΜ με προτροπές χωρίς παραδείγματα για τις πιο δύσκολες περιπτώσεις επαναορισμών σταθερών. Το υψηλότερο ποσοστό για κάθε οικογένεια μοντέλων επισημαίνεται με έντονη γραφή.

Επίπεδο	$R_{a1}$	$R_{a2}$	$R_{a3}$	$R_{s1}$	$R_{s2}$
Ελεύθερης Απάντησης (FF)					
$Q_1$	-0.458	-0.071	0.008	0.199	-0.016
$Q_2$	-0.502	-0.573	-0.472	0.107	0.019
$Q_3$	0.489	0.237	0.292	0.666	0.668
Πολλαπλών Επιλογών (MC)					
$Q_1$	-0.642	-0.4	-0.344	-0.052	0.025
$Q_2$	-0.275	-0.316	-0.245	0.41	0.151
$Q_3$	-0.063	0.457	0.081	0.666	0.75

Table 1.4: Μέση τιμή συσχέτισης μεταξύ επίδοσης στην εργασία χωρίς επαναορισμό και ποσοστών προσκόλλησης για τη στρατηγική χωρίς παραδείγματα. Τα κελιά με **ροζ** χρώμα υποδηλώνουν **υψηλή θετική συσχέτιση** ( $> 0.3$ ), ενώ αυτά με **πράσινο** χρώμα **υψηλή αρνητική συσχέτιση** ( $< -0.3$ ).

Επιπλέον, αναλύουμε τις συσχετίσεις μεταξύ επιδόσεων χωρίς επαναορισμό και ποσοστών προσκόλλησης, οι οποίες αποκαλύπτουν ένα ενδιαφέρον μοτίβο: στις απλούστερες ερωτήσεις οι συσχετίσεις είναι αρνητικές ή πολύ μικρές, υποδηλώνοντας ότι τα μοντέλα που γνωρίζουν καλά τις βασικές τιμές είναι πιο ευέλικτα στην προσαρμογή στους επαναορισμούς σε αυτές τις περιπτώσεις, ενώ στα πιο δύσκολα σενάρια η συσχέτιση γίνεται έντονα θετική, δηλαδή τα μοντέλα που τα πηγαίνουν καλύτερα σε σύνθετες ερωτήσεις υπό κανονικές συνθήκες είναι πιο πιθανό

να αποτύχουν να αγνοήσουν τις παγιωμένες γνώσεις τους. Αυτό σημαίνει, ότι τα πιο "έξυπνα" μοντέλα είναι και τα πιο επιρρεπή στο φαινόμενο της προσκόλλησης.

#### 1.4.1.2 Αντίστροφη Κλιμάκωση

Κατά τη δοκιμή μοντέλων διαφορετικών μεγεθών παρατηρήθηκε ένα ενδιαφέρον φαινόμενο, καθώς σε πολλές περιπτώσεις, η αύξηση του μεγέθους των ΜΓΜ οδήγησε σε αυξημένα ποσοστά προσκόλλησης στις προεπιλεγμένες τιμές, και άρα μεγαλύτερη αδυναμία στο να επιλύσουν σωστά προβλήματα επαναορισμών. Μεγαλύτερα μοντέλα, όπως το Mistral Large και το Llama 405B, παρόλο που κατάφεραν καλύτερες επιδόσεις στην πιο "συμβατική" εργασία χωρίς επαναορισμό, εμφάνισαν σημαντικά υψηλότερα ποσοστά προσκόλλησης από τα αντίστοιχα μικρότερα τους, όταν τους ζητήθηκε να υιοθετήσουν εναλλακτικούς ορισμούς σταθερών, ειδικά στα πιο απαιτητικά σενάρια. Το φαινόμενο αυτό επιβεβαιώνεται και σε οπτικά διαγράμματα που απεικονίζουν την αύξηση των ποσοστών προσκόλλησης σε συνάρτηση με το μέγεθος των μοντέλων της ίδιας οικογένειας. Αυτά τα αποτελέσματα επισημαίνουν ότι η αύξηση του αριθμού των παραμέτρων δεν συνεπάγεται απαραίτητα και μεγαλύτερη γνωστική ευελιξία. Αντίθετα, σε αυτήν την εργασία φαίνεται πως ενισχύει την τάση των ΜΓΜ να εμπιστεύονται περισσότερο τη γνώση που έχουν εσωτερικεύσει, ακόμα και όταν αυτό έρχεται σε αντίθεση με τις οδηγίες που τους δίνονται.

Μοντέλο	$R_{a3}$						$R_{s2}$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF
Mistral7B	66.67	33.33	46.67	33.33	33.33	26.67	66.67	33.33	46.67	13.33	33.33	26.67
Mixtral8x7B	100.0	33.33	66.67	26.67	66.67	20.0	100.0	26.67	66.67	40.0	66.67	46.67
Mistral Large (123B)	93.33	33.33	73.33	26.67	53.33	53.33	93.33	66.67	73.33	46.67	53.33	73.33
Llama8B	80.0	0.0	80.0	0.0	53.33	13.33	80.0	20.0	80.0	26.67	53.33	20.0
Llama70B	93.33	6.67	80.0	0.0	80.0	13.33	93.33	33.33	80.0	13.33	80.0	33.33
Llama405B	93.33	0.0	86.67	0.0	73.33	26.67	93.33	26.67	86.67	6.67	73.33	53.33

Table 1.5: Ποσοστό των ορθών απαντήσεων χωρίς επαναορισμό (NR) και ποσοστό προσκόλλησης για ερωτήσεις ανοικτού τύπου επαναορισμού μονάδων μέτρησης (χωρίς παραδείγματα). Τα χρωματισμένα κελιά υποδεικνύουν ποσοστά προσκόλλησης που αυξάνονται με το μέγεθος των ΜΓΜ.

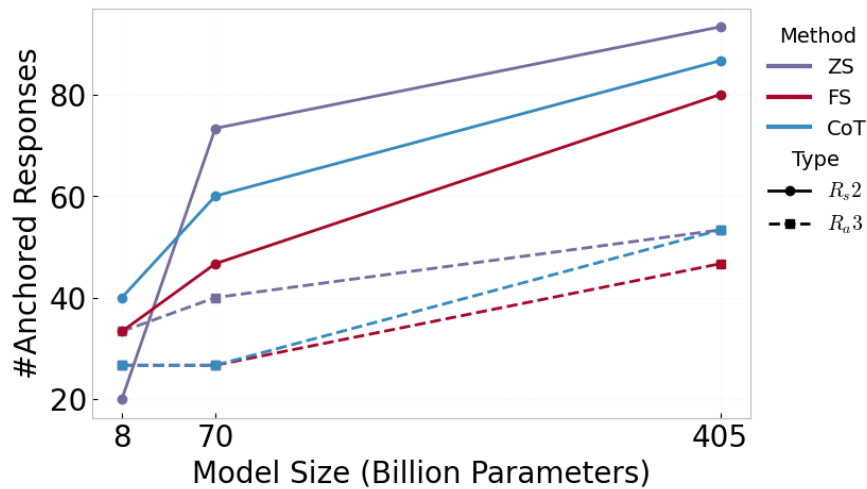


Figure 1.4.1: Ποσοστά προσκόλλησης για μοντέλα διαφορετικού μεγέθους στην οικογένεια Llama (μορφή πολλαπλών επιλογών).

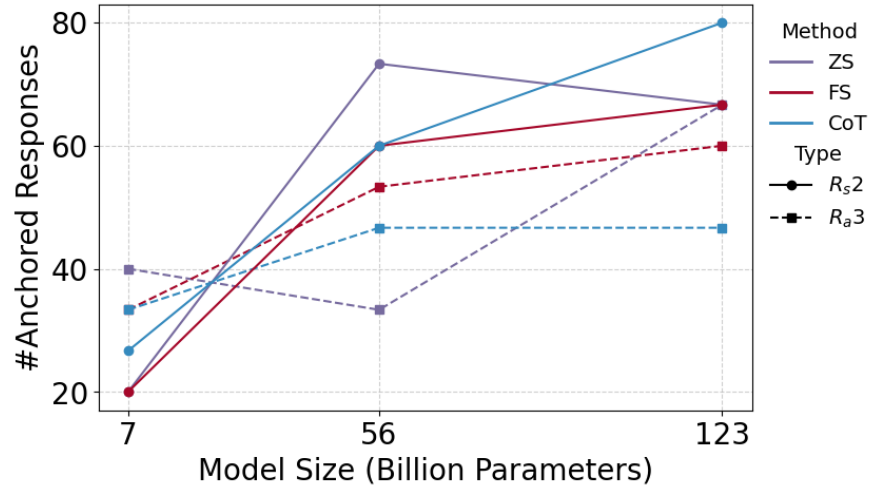
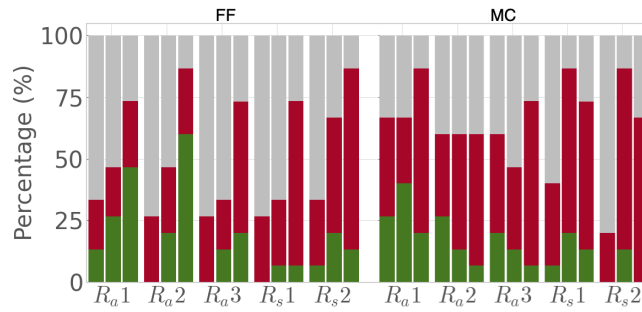


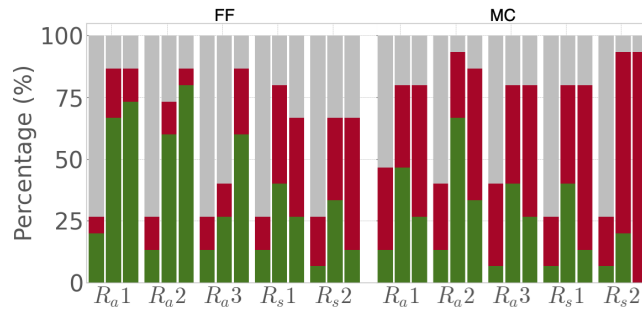
Figure 1.4.2: Ποσοστά προσκόλλησης για μοντέλα διαφορετικού μεγέθους στην οικογένεια Mistral (μορφή πολλαπλών επιλογών).

#### 1.4.1.3 Μορφοποίηση Απαντήσεων

Η μορφοποίηση των απαντήσεων φαίνεται να επηρεάζει σημαντικά τη συμπεριφορά των μοντέλων, με την περίπτωση των πολλαπλών επιλογών να οδηγεί συστηματικά σε υψηλότερα ποσοστά προσκόλλησης. Αυτό μπορεί πιθανώς να εξηγηθεί από την ίδια τη φύση των δύο μορφοποιήσεων, καθώς, ενώ στις απαντήσεις ανοιχτού τύπου τα μοντέλα καλούνται να σκεφτούν πιο ανεξάρτητα, οι προτεινόμενες σωστές πριν τον επαναορισμό απαντήσεων στις πολλαπλές επιλογές λειτουργούν ως ισχυροί παραπλανητικοί πόλοι. Με άλλα λόγια, όταν τα μοντέλα "βλέπουν" την αρχική σωστή απάντηση μέσα στην προτροπή είναι πιο πιθανό και να την επιλέξουν.



(a) Ανάλυση απαντήσεων των μοντέλων Mistral.



(b) Ανάλυση απαντήσεων των μοντέλων Llama.

Figure 1.4.3: Αποτελέσματα για τα μοντέλα των οικογενειών Mistral και Llama στις ερωτήσεις του τρίτου επιπέδου δυσκολίας με προτροπές χωρίς παραδείγματα. Η σειρά των ράβδων ανά τύπο/επίπεδο επαναορισμού αντιστοιχεί σε αύξουσα σειρά μεγέθους μοντέλου.

#### 1.4.1.4 Τύπος Επαναορισμού

Εκτός από τις μορφές των απαντήσεων, καθοριστική διαφορά παρατηρείται μεταξύ των δύο τύπων επαναορισμού. Τα πειραματικά αποτελέσματα δείχνουν ότι η αντικατάσταση μεταξύ σταθερών οδηγεί σε αρκετά υψηλότερα ποσοστά προσκόλλησης από την απλή ανάθεση. Υποθέτουμε ότι αυτό συμβαίνει επειδή το σενάριο της αντικατάστασης ενεργοποιεί ισχυρές μνημονικές συνδέσεις για τις δύο γνωστές οντότητες, το οποίο οδηγεί σε σύγχυση και μεγαλύτερη γνωσιακή επιβάρυνση.

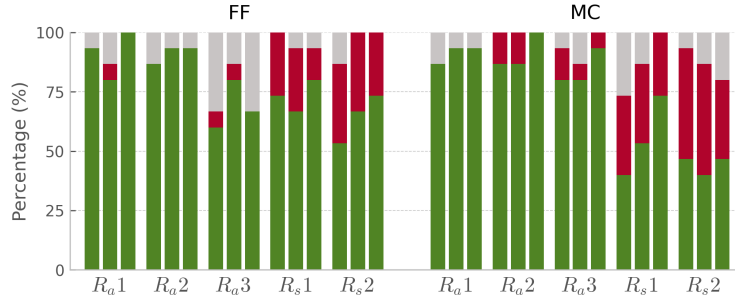


Figure 1.4.4: Ανάλυση απαντήσεων για το Llama 70B στο πρώτο επίπεδο ερωτήσεων και όλες τις στρατηγικές προτροπών. Σε κάθε τύπο/επίπεδο επαναορισμού, οι ράβδοι αντιστοιχούν με τη σειρά σε: χωρίς πασαδείγματα, με αλυσίδες σκέψης, με παραδείγματα.

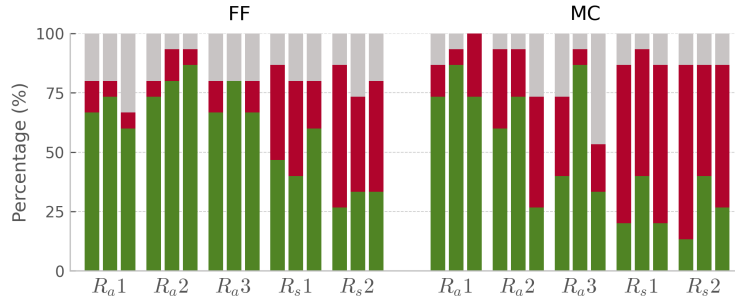


Figure 1.4.5: Ανάλυση απαντήσεων για το Claude 3.5 Sonnet στο τρίτο επίπεδο ερωτήσεων και όλες τις στρατηγικές προτροπών. Σε κάθε τύπο/επίπεδο επαναορισμού, οι ράβδοι αντιστοιχούν με τη σειρά σε: χωρίς πασαδείγματα, με αλυσίδες σκέψης, με παραδείγματα.

#### 1.4.1.5 Λειτουργία Εκτεταμένης Σκέψης

Το μοντέλο Claude 3.7 Sonnet της Anthropic διαθέτει την επιπλέον λειτουργία εκτεταμένης σκέψης (extended thinking), η οποία επιτρέπει στο μοντέλο να αναλύει τα προβλήματα πιο διαδοδικά, παράγοντας μπλοκ σκέψης που αποτυπώνουν την εσωτερική του συλλογιστική πορεία. Δοκιμάσαμε αυτήν τη λειτουργία επαναλαμβάνοντας τα ίδια πειράματα και συγκρίναμε τα αποτελέσματα με αυτά της βασικής (standard) περίπτωσης. Βρήκαμε πως, αν και η λειτουργία εκτεταμένης σκέψης μειώνει ελαφρώς τα ποσοστά προσκόλλησης σε ορισμένες περιπτώσεις, η συνολική της επίδραση είναι αμελητέα. Αυτό δείχνει ότι ακόμα και με ενισχυμένες δυνατότητες συλλογιστικής, το μοντέλο εξακολουθεί να δυσκολεύεται να ανταποκριθεί σε εννοιολογικά απαιτητικές οδηγίες επαναορισμού, αποκαλύπτοντας περιορισμούς στη γνωστική του ευελιξία.

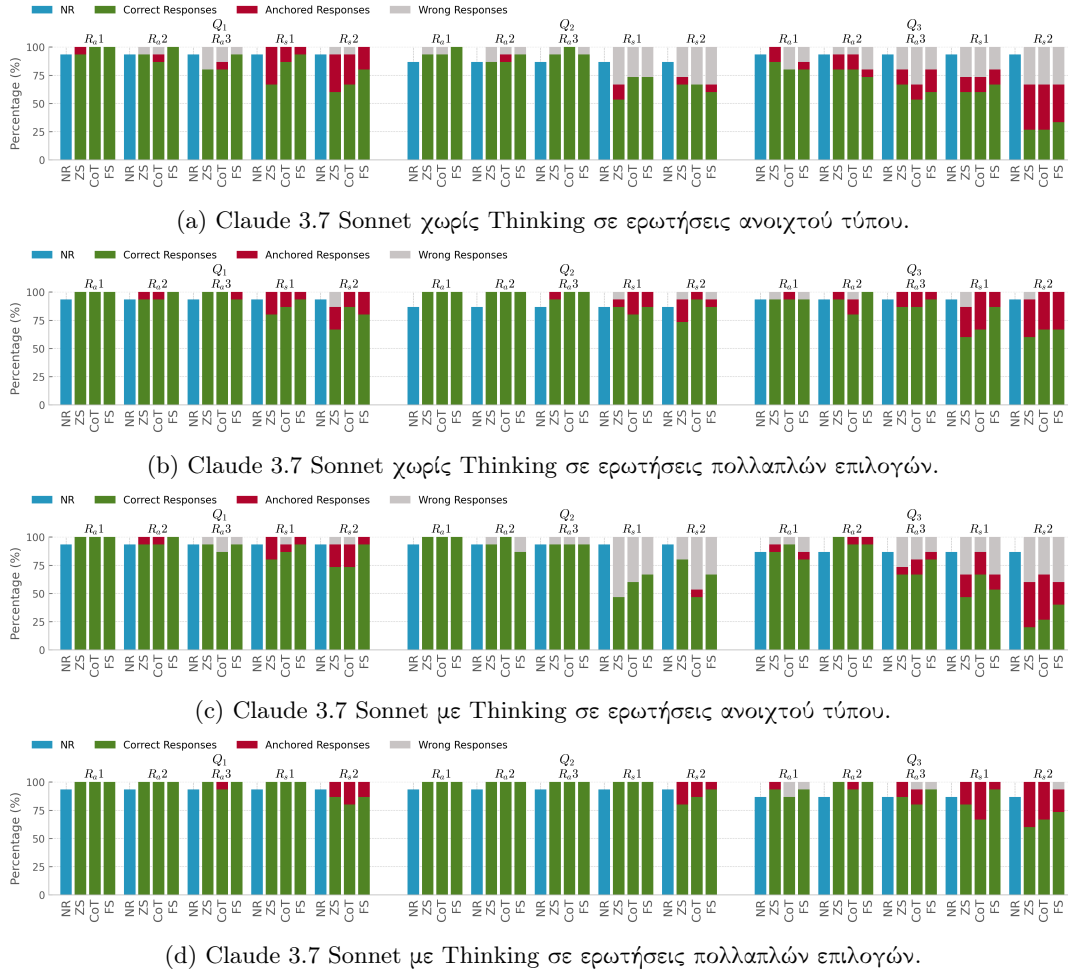
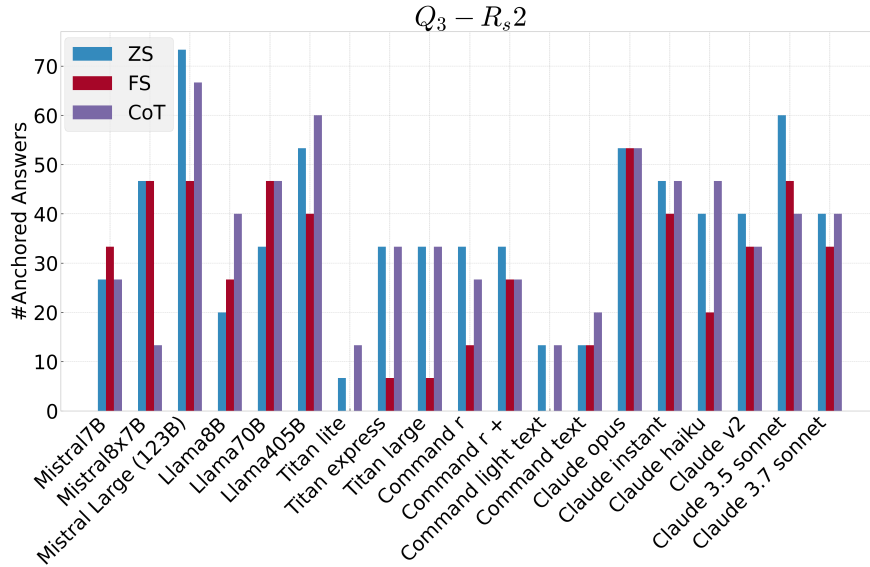


Figure 1.4.6: Ανάλυση απαντήσεων Claude 3.7 Sonnet χωρίς και με Thinking.

## 1.4.1.6 Επίδραση Προτροπών

Figure 1.4.7: Σύγκριση των ποσοστών προσκόλλησης για τις ερωτήσεις  $Q_3$  και το επίπεδο επαναορισμών  $R_{s2}$  για όλα τα MFM.

Η μελέτη της επίδρασης διαφορετικών τεχνικών προτροπής (χωρίς παραδείγματα, με παραδείγματα, με αλυσίδες σχέψης) έδειξε πως η συμπεριφορά των MFM επηρεάζεται, αλλά όχι με συνεπή ή καθοριστικό τρόπο. Ενδιαφέρον παρουσιάζει το γεγονός ότι η τεχνική αλυσίδων σχέψης δεν μειώνει συστηματικά τα ποσοστά προσκόλλησης των μοντέλων, παρόλο που γενικά είναι γνωστό ότι ενισχύει τη συλλογιστική ικανότητα των μοντέλων μέσω της βηματικής επίλυσης [55]. Αντίθετα, η προτροπή με παραδείγματα είναι πιο αποτελεσματική στην περίπτωση μας, αφού πάνω από τα μισά μοντέλα εμφανίζουν καλύτερη επίδοση σε αυτές τις συνθήκες, πιθανώς επειδή οι επιδείξεις σωστών απαντήσεων με ενσωματωμένους επαναορισμούς προσφέρουν στα MFM ένα ισχυρό πλαίσιο αναφοράς προς μίμηση. Ωστόσο, λόγω της μεγάλης διακύμανσης των αποτελεσμάτων, συμπεραίνουμε πως η προσκόλληση είναι ένα φαινόμενο σχετικά ανεπηρέαστο από τις παρεμβάσεις μέσω διαφορετικών τεχνικών προτροπής.

#### 1.4.1.7 Άρνηση Απόκρισης

Μοντέλο	Προτροπή	FF	MC
Mistral7B	ZS	$6.57 \pm 11.99$	$13.34 \pm 18.07$
	CoT	$5.63 \pm 8.89$	$15.62 \pm 16.45$
	FS	<b><math>3.7 \pm 7.58</math></b>	<b><math>10.07 \pm 15.25</math></b>
Mixtral8x7B	ZS	$18.0 \pm 22.8$	$8.61 \pm 16.97$
	CoT	$9.22 \pm 16.82$	$15.5 \pm 17.63$
	FS	<b><math>10.98 \pm 17.03</math></b>	<b><math>5.95 \pm 18.79</math></b>
Mistral Large	ZS	$16.33 \pm 33.69$	$1.67 \pm 6.24$
	CoT	<b><math>8.33 \pm 18.51</math></b>	<b><math>0 \pm 0</math></b>
	FS	$14.35 \pm 26.96$	$1.33 \pm 4.99$
Llama8B	ZS	$55.54 \pm 24.37$	$40.05 \pm 18.58$
	CoT	$35.25 \pm 23.33$	$32.89 \pm 23.21$
	FS	<b><math>2.41 \pm 6.64</math></b>	<b><math>0 \pm 0</math></b>
Llama70B	ZS	$38.66 \pm 29.92$	$5.56 \pm 14.49$
	CoT	$9.17 \pm 17.36$	$13.33 \pm 27.35$
	FS	<b><math>0 \pm 0</math></b>	<b><math>0 \pm 0</math></b>
Llama405B	ZS	$1.33 \pm 4.99$	<b><math>0 \pm 0</math></b>
	CoT	<b><math>0 \pm 0</math></b>	<b><math>0 \pm 0</math></b>
	FS	<b><math>0 \pm 0</math></b>	<b><math>0 \pm 0</math></b>
Titan lite	ZS	<b><math>1.56 \pm 3.19</math></b>	<b><math>0 \pm 0</math></b>
	CoT	$3.03 \pm 5.66$	<b><math>0 \pm 0</math></b>
	FS	$2.54 \pm 5.39$	<b><math>0 \pm 0</math></b>
Titan express	ZS	$0.56 \pm 2.08$	<b><math>0 \pm 0</math></b>
	CoT	$1.9 \pm 7.13$	<b><math>0 \pm 0</math></b>
	FS	<b><math>0 \pm 0</math></b>	<b><math>0 \pm 0</math></b>
Titan large	ZS	$2.0 \pm 5.42$	<b><math>0 \pm 0</math></b>
	CoT	<b><math>0 \pm 0</math></b>	<b><math>0 \pm 0</math></b>
	FS	<b><math>0 \pm 0</math></b>	<b><math>0 \pm 0</math></b>
Command text	ZS	$3.33 \pm 9.03$	<b><math>0 \pm 0</math></b>
	CoT	<b><math>0 \pm 0</math></b>	<b><math>0 \pm 0</math></b>
	FS	$0.83 \pm 3.12$	<b><math>0 \pm 0</math></b>
Claude instant	ZS	$1.69 \pm 4.36$	<b><math>0 \pm 0</math></b>
	CoT	<b><math>0 \pm 0</math></b>	<b><math>0 \pm 0</math></b>
	FS	$4.07 \pm 12.58$	<b><math>0 \pm 0</math></b>
Claude v2	ZS	$20.48 \pm 26.25$	$4.83 \pm 9.29$
	CoT	$14.31 \pm 24.39$	$10.0 \pm 27.08$
	FS	<b><math>8.91 \pm 24.75</math></b>	<b><math>3.17 \pm 8.81</math></b>

Table 1.6: Μέσα ποσοστά άρνησης για όλα τα MFM (μικρότερες τιμές σε **έντονη γραφή** και μεγαλύτερες τιμές υπογραμμισμένες). Δεν συμπεριλαμβάνονται τα μοντέλα που σημείωσαν μηδενικά ποσοστά σε όλες τις περιπτώσεις.

Παρόλο που το φαινόμενο της προσκόλλησης ήταν το κύριο αντικείμενο μελέτης αυτής της εργασίας, μια επίσης ενδιαφέρουσα συμπεριφορά παρατηρήθηκε σε πολλές περιπτώσεις, όταν τα μοντέλα αρνούνταν ρητά να απαντήσουν σε ερωτήσεις που σχετίζονται με τον επαναορισμό γνωστών εννοιών, κρίνοντάς τες μη έγκυρες, παράλογες ή παραπλανητικές. Το φαινόμενο αυτό ήταν πιο έντονο σε συγκεκριμένες οικογένειες MFM, όπως

στις Mistral και Llama, και ειδικά στις εκδόσεις τους με τον μικρότερο αριθμό παραμέτρων. Αντίθετα, μοντέλα απο τις οικογένειες Claude, Titan και Cohere παρουσιάζουν σημαντικά μικρότερα–και συχνά μηδενικά– ποσοστά τέτοιων αποκρίσεων. Αναφορικά με τις τεχνικές προτροπής, η προσέγγιση με παραδείγματα φαίνεται να μειώνει πιο αισθητά τα ποσοστά άρνησης, το οποίο είναι αναμενόμενο αφού μέσα από τα παραδείγματα κανονικοποιείται η διαδικασία του επαναορισμού. Επίσης, μετρώντας τις συσχετίσεις μεταξύ ακρίβειας χωρίς επαναορισμό και άρνησης (0.144 για ελεύθερη απάντηση και 0.039 για πολλαπλές επιλογές κατά μέσο όρο), συμπεραίνουμε πως αυτή η συμπεριφορά δεν σχετίζεται ισχυρά με τις βασικές ικανότητες λογικής των ΜΓΜ, άλλα μάλλον περσοσσότερο με την κλίμακά τους. Τα μεγαλύτερα μοντέλα τείνουν να αρνούνται να απαντήσουν λιγότερο συχνά, επιδεικνύοντας μιά μορφή αυξημένης αυτοπεποίθησης που τα ωθεί να προσπαθούν, ακόμα και αν τελικά αποτυγχάνουν.

## 1.4.2 Επαναορισμός Μονάδων Μέτρησης

### 1.4.2.1 Προσκόλληση στις Πραγματικές Τιμές

Το φαινόμενο της προσκόλλησης στους πραγματικούς ορισμούς παραμένει και στην περίπτωση των επαναορισμών μονάδων μέτρησης. Όλα τα μοντέλα, σε διαφορετικό βαθμό, παράγουν απαντήσεις που βαζίζονται στις προϋπάρχουσες γνώσεις τους, αγνοώντας την οδηγία επαναορισμού. Τα ποσοστά προσκόλλησης, βέβαια, είναι γενικά χαμηλότερα από τα αντίστοιχα των σταθερών, με κάποια μοντέλα (κυρίως από τις οικογένειες Command και Claude) να πετυχαίνουν ακόμα και μηδενικά αποτελέσματα σε πιο εύκολα σενάρια ερωτήσεων ανοιχτού τύπου.

Μοντέλο	$R_{a2}$						$R_{a3}$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	0.0	37.5	25.0	25.0	18.75	56.25	<b>62.5</b>	25.0	<b>31.25</b>	<b>37.5</b>	31.25	25.0
Mixtral8x7B	<b>6.25</b>	31.25	<b>31.25</b>	<b>37.5</b>	31.25	37.5	6.25	<b>31.25</b>	6.25	31.25	<b>31.25</b>	<b>50.0</b>
Mistral Large	0.0	<b>37.5</b>	6.25	37.5	12.5	56.25	0.0	25.0	12.5	<b>37.5</b>	12.5	43.75
Llama8B	0.0	<b>25.0</b>	<b>6.25</b>	<b>31.25</b>	12.5	31.25	<b>6.25</b>	<b>31.25</b>	<b>12.5</b>	<b>50.0</b>	<b>25.0</b>	50.0
Llama70B	0.0	6.25	<b>6.25</b>	<b>31.25</b>	<b>25.0</b>	<b>56.25</b>	0.0	18.75	0.0	<b>50.0</b>	12.5	<b>62.5</b>
Llama405B	0.0	0.0	0.0	<b>31.25</b>	12.5	37.5	0.0	0.0	6.25	25.0	<b>25.0</b>	31.25
Titan lite	6.25	<b>25.0</b>	12.5	<b>31.25</b>	12.5	<b>25.0</b>	25.0	<b>31.25</b>	25.0	12.5	0.0	18.75
Titan express	18.75	<b>25.0</b>	<b>25.0</b>	18.75	12.5	<b>25.0</b>	<b>43.75</b>	25.0	31.25	12.5	6.25	18.75
Titan large	<b>31.25</b>	12.5	12.5	<b>31.25</b>	<b>18.75</b>	<b>25.0</b>	25.0	12.5	<b>37.5</b>	<b>31.25</b>	6.25	<b>25.0</b>
Command r	<b>12.5</b>	18.75	<b>12.5</b>	<b>31.25</b>	25.0	18.75	6.25	25.0	<b>12.5</b>	18.75	<b>12.5</b>	31.25
Command r+	6.25	<b>43.75</b>	0.0	25.0	<b>37.5</b>	<b>50.0</b>	<b>6.25</b>	<b>31.25</b>	0.0	<b>31.25</b>	0.0	25.0
Command light text	6.25	12.5	0.0	25.0	6.25	25.0	12.5	25.0	6.25	31.25	0.0	<b>50.0</b>
Command text	12.5	12.5	12.5	18.75	0.0	18.75	0.0	31.25	12.5	12.5	0.0	43.75
Claude opus	0.0	0.0	0.0	6.25	12.5	25.0	0.0	0.0	0.0	0.0	0.0	6.25
Claude instant	<b>6.25</b>	<b>25.0</b>	<b>12.5</b>	25.0	0.0	<b>43.75</b>	0.0	<b>43.75</b>	0.0	<b>37.5</b>	6.25	<b>31.25</b>
Claude haiku	0.0	18.75	0.0	12.5	6.25	31.25	0.0	6.25	0.0	6.25	<b>18.75</b>	<b>31.25</b>
Claude v2	<b>6.25</b>	18.75	6.25	<b>31.25</b>	<b>18.75</b>	31.25	<b>6.25</b>	0.0	<b>6.25</b>	25.0	6.25	12.5
Claude 3.5 Sonnet	0.0	0.0	0.0	12.5	6.25	6.25	0.0	0.0	0.0	6.25	0.0	0.0
Claude 3.7 Sonnet	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 1.7: Ποσοστά προσκόλλησης όλων των ΜΓΜ με προτροπές χωρίς παραδείγματα για τις πιο δύσκολες περιπτώσεις επαναορισμών μονάδων μέτρησης. Το υψηλότερο ποσοστό για κάθε οικογένεια μοντέλων επισημαίνεται με **έντονη γραφή**.

## 1.4.2.2 Αντίστροφη Κλιμάκωση

Μοντέλο	$R_a2$						$R_a3$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF
Mistral 7B	81.25	0.0	56.25	25.0	43.75	18.75	81.25	62.5	56.25	31.25	43.75	31.25
Mixtral8x7B	87.5	6.25	81.25	31.25	62.5	31.25	87.5	6.25	81.25	6.25	62.5	31.25
Mistral Large	93.75	0.0	93.75	6.25	81.25	12.5	93.75	0.0	93.75	12.5	81.25	12.5
Llama8B	75.0	0.0	56.25	6.25	6.25	12.5	75.0	6.25	56.25	12.5	6.25	25.0
Llama70B	100.0	0.0	81.25	6.25	56.25	25.0	100.0	0.0	81.25	0.0	56.25	12.5
Llama405B	100.0	0.0	93.75	0.0	56.25	12.5	100.0	0.0	93.75	6.25	56.25	25.0
Titan lite	37.5	6.25	18.75	12.5	6.25	12.5	37.5	25.0	18.75	25.0	6.25	0.0
Titan express	75.0	18.75	37.5	25.0	6.25	12.5	75.0	43.75	37.5	31.25	6.25	6.25
Titan large	68.75	31.25	68.75	12.5	25.0	18.75	68.75	25.0	68.75	37.5	25.0	6.25
Command r	75.0	12.5	56.25	12.5	18.75	25.0	75.0	6.25	56.25	12.5	18.75	12.5
Command r+	87.5	6.25	93.75	0.0	81.25	37.5	87.5	6.25	93.75	0.0	81.25	0.0
Command light text	31.25	6.25	6.25	0.0	0.0	6.25	31.25	12.5	6.25	6.25	0.0	0.0
Command text	62.5	12.5	50.0	12.5	25.0	0.0	62.5	0.0	50.0	12.5	25.0	0.0
Claude opus	100.0	0.0	75.0	0.0	56.25	12.5	100.0	0.0	75.0	0.0	56.25	0.0
Claude instant	75.0	6.25	81.25	12.5	43.75	0.0	75.0	0.0	81.25	0.0	43.75	6.25
Claude haiku	100.0	0.0	93.75	0.0	81.25	6.25	100.0	0.0	93.75	0.0	81.25	18.75
Claude v2	93.75	6.25	68.75	6.25	25.0	18.75	93.75	6.25	68.75	6.25	25.0	6.25
Claude 3.5 Sonnet	100.0	0.0	87.5	0.0	87.5	6.25	100.0	0.0	87.5	0.0	87.5	0.0
Claude 3.7 Sonnet	100.0	0.0	87.5	0.0	93.75	0.0	100.0	0.0	87.5	0.0	93.75	0.0

Table 1.8: Ποσοστό των ορθών απαντήσεων χωρίς επαναορισμό (NR) και ποσοστό προσκόλλησης για ερωτήσεις ανοιχτού τύπου επαναορισμού μονάδων μέτρησης (χωρίς παραδείγματα).

Και στην εργασία επαναορισμού μονάδων μέτρησης εμφανίζονται τάσεις αντίστροφης κλιμάκωσης. Σε αρκετές περιπτώσεις, μεγαλύτερα μοντέλα (όπως Mistral Large, Titan Large και Llama 405B) εμφανίζουν αυξημένα ποσοστά προσκόλλησης σε σχέση με μικρότερα εντός των ίδιων οικογενειών, παρόλο που πετυχαίνουν καλύτερες επιδόσεις στις αντίστοιχες εργασίες χωρίς επαναορισμούς. Αν και το φαινόμενο δεν είναι τόσο έντονο όσο στην περίπτωση των φυσικών σταθερών, παραμένει αξιοσημείωτο, καθώς, για άλλη μια φορά, η αυξημένη λογική ικανότητα των μεγαλύτερων μοντέλων περιέργως δεν μεταφράζεται και σε καλύτερη προσαρμογή σε επαναορισμένες συνθήκες.

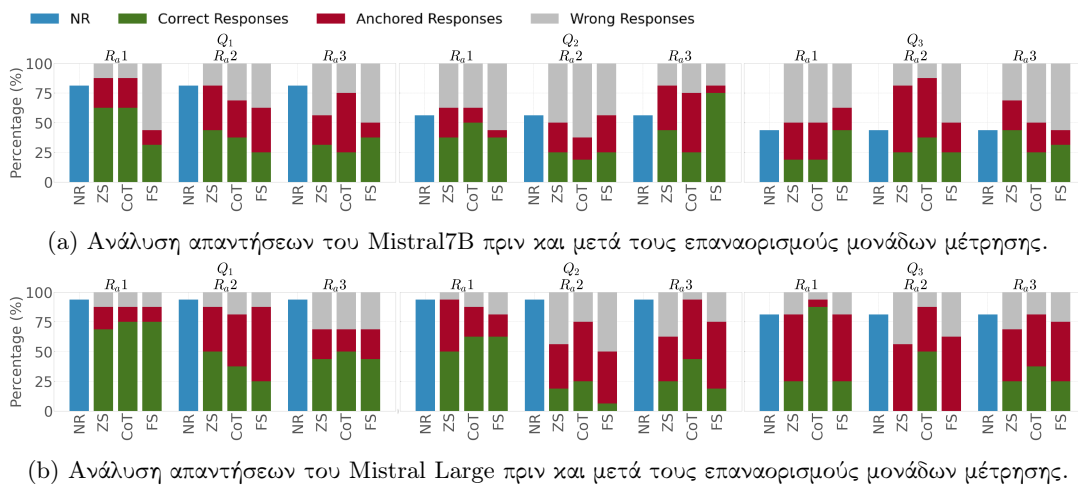
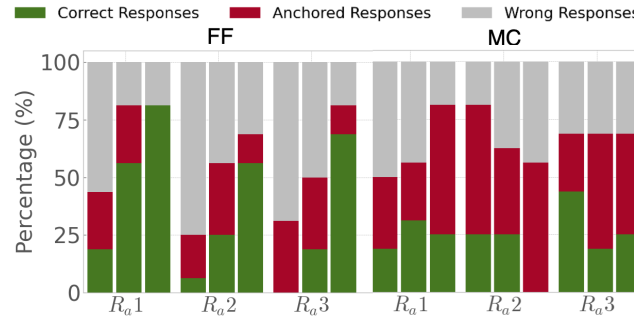


Figure 1.4.8: Σύγκριση των απαντήσεων των Mistral7B και Mistral Large (123B) σε ερωτήσεις πολλαπλών επιλογών για επαναορισμούς μονάδων μέτρησης.

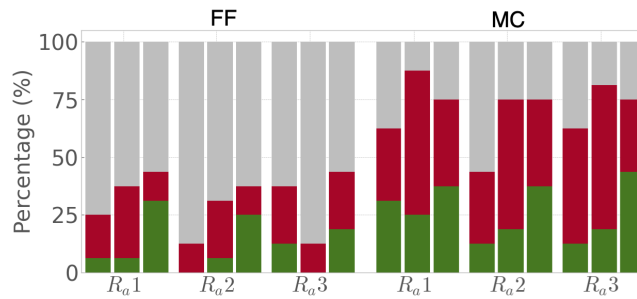


### 1.4.2.3 Μορφοποίηση Απαντήσεων

Για άλλη μια φορά, η μορφή πολλαπλών επιλογών ενισχύει σημαντικά το φαινόμενο της προσκόλλησης στην απομνημονευμένη γνώση σε σχέση με την ελεύθερη απάντηση, με ποσοστά να ανεβαίνουν, για παράδειγμα, από 12.5% στα 62.5%. Το γεγονός αυτό οφείλεται στην έκθεση του μοντέλου στην καθιερωμένη σχέση μεταξύ των μονάδων μέτρησης μέσα από τις προτεινόμενες επιλογές, η οποία ενισχύει τη σύγκρουση μεταξύ της οδηγίας και της προϋπάρχουσας από την εκπαίδευσή του γνώσης.



(a) Ανάλυση απαντήσεων των μοντέλων Mistral.



(b) Ανάλυση απαντήσεων των μοντέλων Llama.

Figure 1.4.9: Αποτελέσματα για τα μοντέλα των οικογενειών Mistral και Llama στις ερωτήσεις του τρίτου επιπέδου δυσκολίας με προτροπές χωρίς παραδείγματα. Η σειρά των ράβδων ανά τύπο/επίπεδο επαναορισμού αντιστοιχεί σε αύξουσα σειρά μεγέθους μοντέλου.

### 1.4.2.4 Επίδραση Προτροπών

Αντίθετα με τα αποτελέσματα των σταθερών, στην περίπτωση των μονάδων μέτρησης φαίνεται πως η τεχνική με αλυσίδες σχέσης είναι πιο αποτελεσματική για τον περιορισμό του φαινομένου της προσκόλλησης.

Επίπεδο	$R_{a1}$	$R_{a2}$	$R_{a3}$
Ελεύθερης Απάντησης (FF)			
$Q_1$	-0.295	-0.403	-0.33
$Q_2$	-0.361	-0.247	-0.479
$Q_3$	-0.063	0.19	0.14
Πολλαπλών Επιλογών (MC)			
$Q_1$	-0.49	-0.149	-0.542
$Q_2$	-0.159	-0.023	0.08
$Q_3$	0.248	0.338	-0.127

Table 1.9: Μέση τιμή συσχέτισης μεταξύ επίδοσης στην εργασία χωρίς επαναορισμό και ποσοστών προσκόλλησης για τη στρατηγική χωρίς παραδείγματα. Τα κελιά με **ροζ** χρώμα υποδηλώνουν υψηλή **θετική** συσχέτιση ( $> 0.3$ ), ενώ αυτά με **πράσινο** χρώμα υψηλή **αρνητική** συσχέτιση ( $< -0.3$ ).

Επίπεδο	$R_{a1}$	$R_{a2}$	$R_{a3}$
	Ελεύθερης Απάντησης (FF)		
$Q_1$	-0.32	-0.442	-0.161
$Q_2$	-0.404	-0.231	0.039
$Q_3$	0.128	-0.042	0.279
	Πολλαπλών Επιλογών (MC)		
$Q_1$	-0.332	0.058	-0.593
$Q_2$	0.135	0.131	0.266
$Q_3$	0.314	0.49	0.101

Table 1.10: Μέση τιμή συσχέτισης μεταξύ επίδοσης στην εργασία χωρίς επαναορισμό και ποσοστών προσκόλλησης για τη στρατηγική με παραδείγματα. Τα κελιά με **ροζ** χρώμα υποδηλώνουν **υψηλή θετική συσχέτιση** ( $> 0.3$ ), ενώ αυτά με **πράσινο** χρώμα **υψηλή αρνητική συσχέτιση** ( $< -0.3$ ).

Επίπεδο	$R_{a1}$	$R_{a2}$	$R_{a3}$
	Ελεύθερης Απάντησης (FF)		
$Q_1$	-0.502	-0.598	-0.529
$Q_2$	-0.465	-0.3	-0.174
$Q_3$	-0.232	-0.181	-0.079
	Πολλαπλών Επιλογών (MC)		
$Q_1$	-0.528	-0.023	-0.523
$Q_2$	0.015	-0.091	-0.016
$Q_3$	-0.127	0.013	-0.242

Table 1.11: Μέση τιμή συσχέτισης μεταξύ επίδοσης στην εργασία χωρίς επαναορισμό και ποσοστών προσκόλλησης για τη στρατηγική με αλυσίδες σχέσης. Τα κελιά με **ροζ** χρώμα υποδηλώνουν **υψηλή θετική συσχέτιση** ( $> 0.3$ ), ενώ αυτά με **πράσινο** χρώμα **υψηλή αρνητική συσχέτιση** ( $< -0.3$ ).

#### 1.4.2.5 Άρνηση Απόκρισης

Παρουσιάζει ιδιαίτερο ενδιαφέρον το γεγονός ότι στην περίπτωση των επαναορισμών μονάδων μέτρησης το φαινόμενο της άρνησης απόκρισης είναι σχεδόν ανύπαρκτο. Μόνο τα μοντέλα της οικογένειας Mistral κατέγραψαν τέτοιες αρνήσεις, αλλά ακόμα και αυτές χαρακτηρίζαν μεμονωμένα περιστατικά και όχι συστηματική συμπεριφορά. Η έντονη αυτή διαφορά σε σχέση με τις σταθερές δείχνει ότι το φαινόμενο σχετίζεται άμεσα με τον τρόπο με τον οποίο τα ΜΓΜ εσωτερικεύουν κάθε γνωσιακό πεδίο. Οι μονάδες μέτρησης φαίνεται να είναι λιγότερο "άκαμπτα" εσωτερικευμένες σε σχέση με τις επιστημονικές σταθερές.

## 1.5 Συμπεράσματα

Στην παρούσα εργασία μελετήσαμε εκτενώς την εργασία του επαναορισμού (redefinition), εξετάζοντας πώς τα Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ) αντιδρούν όταν τους παρουσιάζονται τροποποιημένες τιμές γνωστών επιστημονικών σταθερών και μονάδων μέτρησης. Στόχος μας ήταν να αξιολογήσουμε την ευελιξία τους έναντι της τάσης να προσκολλώνται στην εδραιωμένη γνώση. Τα ευρήματά μας αναδεικνύουν σημαντικά πρότυπα συμπεριφοράς των ΜΓΜ, φανερώνοντας περιορισμούς, οι οποίοι, μάλιστα, γίνονται εντονότεροι όσο αυξάνεται το μέγεθος των μοντέλων. Παρατηρούμε ότι, παρόλο που τα μεγαλύτερα μοντέλα εμφανίζουν ισχυρότερες ικανότητες συλλογιστικής υπό κανονικές συνθήκες, δυσκολεύονται περισσότερο όταν καλούνται να ακολουθήσουν επαναορισμένες τιμές, καθώς τείνουν να επιμένουν σε αυτές που έχουν απομνημονεύσει κατά την προεκπαίδευση. Εκτός αυτού, διαπιστώνουμε ότι παρουσιάζουν ψευδή αυτοπεποίθηση, προτιμώντας να απαντήσουν από το να απέχουν, ακόμα και όταν αυτό οδηγεί σε λάθη.

Επιπλέον, τα πειράματά μας καλύπτουν ένα ευρύ φάσμα συνθηκών που αποσκοπούν στη δοκιμή της προσαρμοστικότητας των μοντέλων. Δημιουργήσαμε σύνολα δεδομένων με τύπους και επίπεδα επαναορισμών, καθώς και βαθμούς δυσκολίας των ερωτήσεων. Ταυτόχρονα, αξιολογήσαμε την επίδραση διαφορετικών μορφών απάντησης και τεχνικών προτροπής. Τα αποτελέσματα δείχνουν ότι το φαινόμενο προσκόλλησης εντείνεται σημαντικά στη μορφή των πολλαπλών επιλογών. Οι τεχνικές προτροπής επηρεάζουν μερικώς, αλλά απέχουν από το να εξαλείψουν το πρόβλημα.

Συνολικά, η εργασία μας αναδεικνύει σημαντικές αδυναμίες στη συλλογιστική και τη προσαρμοστικότητα των ΜΓΜ, οι οποίες εντείνονται με την αύξηση του μεγέθους τους. Τονίζουμε, επίσης, τη σημασία της βαθύτερης κατανόησης της συμπεριφοράς αυτών των μοντέλων, όχι μόνο ως προς το τι μπορούν να κάνουν, αλλά και πού και γιατί αποτυγχάνουν. Το πείραμα του επαναορισμού προσφέρει ένα χρήσιμο πλαίσιο για τη μελέτη της εύθραυστης ισορροπίας μεταξύ μεγέθους, λογικής και συμμόρφωσης σε οδηγίες, και εδραιωμένων γνώσεων, και ελπίζουμε να αποτελέσει βάση για μελλοντική έρευνα στη διερεύνηση της προσαρμοστικότητας και της ανθεκτικότητας των ΜΓΜ.

### **Ισορροπία ανάμεσα στη Λογική και την Ανθεκτικότητα**

Η μελέτη μας αναδεικνύει μία αντισταθμιστική σχέση στον σχεδιασμό και τη λειτουργία των ΜΓΜ: όσο πιο αυστηρά παραμένει ένα μοντέλο προσκολλημένο στην προεκπαιδευμένη γνώση του, τόσο λιγότερο πρόθυμο είναι να ακολουθήσει εναλλακτικά σενάρια, ακόμα και αν αυτά είναι λογικά αποδεκτά. Αυτό ενισχύει την πραγματολογική του ακρίβεια, αλλά περιορίζει τη λογική του ευελιξία. Από την άλλη πλευρά, ένα μοντέλο που ακολουθεί μη συμβατικά πειράματα επιδεικνύει μεγαλύτερη προσαρμοστικότητα, όμως είναι πιο ευάλωτο σε παραπλανητικές ή κακόβουλες προτροπές. Η ηθική πρόκληση, λοιπόν, είναι να βρεθεί μία ισορροπία: πώς μπορούμε να σχεδιάσουμε ΜΓΜ που είναι ταυτόχρονα ευέλικτα και αξιόπιστα;



# Chapter 2

## Introduction

Large Language Models (LLMs) represent a significant leap in the field of Natural Language Processing (NLP), demonstrating exceptional proficiency in understanding and generating human-like text. As their capabilities continue to evolve, these models are not only advancing foundational NLP applications but also driving the discovery of impressive new behaviors that emerge unexpectedly at scale ([139]; [113]), including even the ability to perform advanced step-by-step logical inference. Despite being trained purely on the objective of next token prediction, LLMs exhibit surprising competence in reasoning tasks that were until recently considered exclusive to human cognition [94]. However, the underlying mechanisms behind this emerging behavior remain poorly understood, prompting ongoing debate over whether such abilities reflect authentic reasoning or are simply the result of large-scale memorization and highly sophisticated pattern recognition [141].

In efforts to expose how superficial heuristics dominate over genuine LLM emergent abilities, researchers have experimented with tasks featuring alternative formulations, unnatural contexts, counterfactual scenarios, and deliberately misleading prompts, revealing that LLMs often fall back on entrenched knowledge and surface-level pattern matching rather than engaging in robust reasoning processes ([140]; [67]; [63]). Intriguingly, under some of these conditions, a rather counterintuitive trend emerges: larger models become more susceptible to such traps, resulting in worse performance than their smaller counterparts. This phenomenon defines a class of problems collectively known as *inverse scaling tasks* [82], where increasing model scale is associated with a decline in task performance, highlighting a reversal of the typical scaling laws that guarantee a predictable improvement with scale [51]. Addressing inverse scaling tasks serves as an important frontier in uncovering hidden limitations of increasingly powerful LLMs, particularly where their behavior deviates from human-like reasoning on tasks that are typically straightforward for people.

Surprisingly, despite its implications for model reliability, inverse scaling remains a relatively underexplored area of study in current literature. Motivated by this research gap, this thesis explores the *Redefinition task*, which was introduced as a part of the Inverse Scaling Prize contest [82] and is designed to test whether LLMs are able to override deeply embedded world knowledge when faced with prompts that involve deliberately misleading definitions of well-known concepts. Specifically, the present work focuses on alterations to the accepted values of core entities within two distinct domains: scientific constants, drawn from physics and mathematics, and units of measurement. For example, a prompt may instruct the model to redefine the value of the mathematical constant  $\pi$  as 100 ("Redefine  $\pi$  as 100."), contradicting the widely known value 3.14159. The model's ability to follow this redefinition is then tested through questions, which can be straightforward (e.g., "What is the first digit of pi?") or more complex (e.g., "What is the Earth's surface area?"). While humans override default meanings with ease, achieving a 100% accuracy in the Inverse Scaling Prize benchmark [82], large language models systematically fail by adhering to familiar associations, underscoring a strong reliance on memorized priors over contextual reasoning. We refer to this behavior as *anchoring* and investigate its relationship with model scale by evaluating LLMs of various parameter sizes.

In this thesis, we:

- Provide theoretical background on LLMs, covering scaling laws, emergent abilities, prompting techniques, and the LLM-as-a-judge evaluation paradigm.
- Examine reasoning in LLMs, with a focus on the tension between true reasoning and memorization. We also review the inverse scaling problems and their underlying causes, as introduced in the Inverse Scaling Prize research.
- Detail the construction of our datasets targeting constant and unit of measure redefinitions, explaining the selection of redefined entities, and the design of redefinition types, question complexity levels, response formats, and prompting techniques. We also describe the evaluation metrics, selected LLMs, and experimental setup.
- Present experimental results and analyze model behavior on the redefinition task, with particular attention to the anchoring effect. We investigate how this phenomenon varies with model size and is influenced by different factors, such as response format, redefinition type, or prompting strategy.
- Conclude by summarizing the key findings of our study and discussing implications for the trade-off between reasoning capabilities and robustness in large language models.

A version of this work has also been published in [116]. The present document expands on those contributions, offering deeper analysis and a more comprehensive theoretical context.

# Chapter 3

## Background

### 3.1 Introduction to Large Language Models

#### 3.1.1 Framework and Objective

Large Language Models (LLMs) are a class of artificial intelligence systems trained to understand and generate human language by learning from massive corpora of data. At their core, LLMs are based on deep learning architectures—most importantly on the *Transformer* architecture, which, once introduced by Vaswani et al. [123] in 2017, has fundamentally revolutionized the field of Natural Language Processing (NLP) [53]. Transformers employ *self-attention* mechanisms that enable the model to compute in parallel context-aware representations for each word in a sentence or document that accurately represent the contextual relationships between them, regardless of their positional distance [83]. This extensive parallelization dramatically improves large-scale training efficiency on modern hardware (e.g., GPUs), in comparison to earlier standard model architectures such as Recurrent Neural Networks (RNNs) [104] and Long Short-Term Memory networks (LSTMs) ([114]; [109]). Before being processed by an LLM, input text is segmented into indivisible units known as *tokens* [136]. Depending on the tokenization process, each token can represent a character, symbol, word, or subword. Some of the most commonly used methods include Byte Pair Encoding (BPE) [106], WordPiece [112], SentencePiece [61], and unigramLM [60]. LLMs are ultimately trained to predict the next word in a sentence by assigning probabilities to each token in a given sequence. Formally, an LLM models the conditional probability distribution [6]:

$$P(w_t \mid w_1, w_2, \dots, w_{t-1})$$

The training process of an LLM typically involves tens to hundreds of billions of parameters and leverages *self-supervised* learning [98]. Two primary pretraining approaches are typically used: Autoregressive Language Modeling, where the model predicts the next token given the preceding ones in an auto-regressive manner, and Masked Language Modeling, in which certain tokens are masked and the model learns to predict these masked tokens based on surrounding context [83]. Notably, pretraining alone enables LLMs to demonstrate impressive results across a broad spectrum of tasks ([8]; [22]). However, in many cases, additional *fine-tuning* is applied in order to enhance performance on specific scenarios or to better align model outputs with human preferences [86]. Depending on the use case requirements, different fine-tuning techniques may be applied. In Transfer Learning ([162]; [97]), LLMs are further trained on task-specific data, allowing them to adapt to particular applications. Instruction Tuning ([150]; [24]; [137]) involves training on datasets formatted as instructions-output pairs, guiding the model toward more predictable and user-aligned responses to natural-language prompts. On the other hand, Alignment Tuning [88] uses human feedback to update their parameters, aiming to prevent the generation of false, biased, and harmful content. A common approach to model alignment is Reinforcement Learning with Human Feedback (RLHF) [165], where a model fine-tuned on human demonstrations is further optimized using reward modeling (RM) and reinforcement learning (RL).

Unlike preceding language models, which primarily aimed to generate text data, the objective of LLMs is not merely to be able to mimic human language convincingly, but to engage in complex tasks, ranging from translation and summarization to question answering and even reasoning, marking an important leap from language modeling to task solving [99, 58, 31, 29, 121, 90, 2, 59, 125]. As the current frontier in the language model evolution process, LLMs are increasingly characterized as *general-purpose task solvers* [156]. In fact, these models have not only extended machine capabilities to a significantly wider scope of applications, but also achieved performance that approaches—and in some domains, even surpasses—human-level performance.

### 3.1.2 Parameters and Scaling Laws

In the context of Large Language Models, *parameters* refer to the learnable, input-independent settings—typically weights and biases—within the neural network layers that evolve during the training process to optimize output predictions. These are basically the core components that shape the LLM’s abilities to recognize patterns and map complex relationships between words and phrases in the training data, ultimately establishing the transformation from input to output. Each parameter is represented by a numerical value that is initially set either randomly (during pretraining) or based on previous training (during finetuning) and then adjusted in order to minimize prediction error [27]. In Transformer-based architectures, which underpin most LLMs, parameters are distributed across multiple layers and attention heads, enabling the model to jointly attend to information from different parts of the input [123]. The total number of parameters typically represents the model’s *capacity*, meaning that larger LLMs potentially are able to “fit” more intricate relationships and knowledge of the training data.

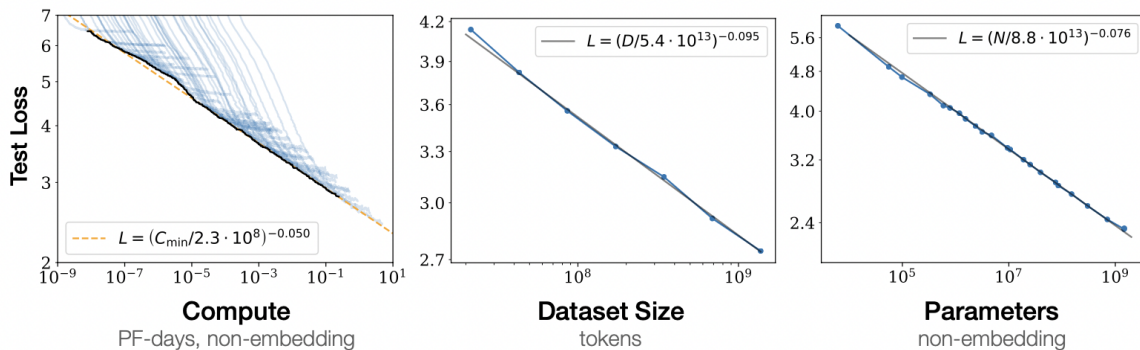


Figure 3.1.1: Empirical scaling laws showing power-law relationships between model performance and compute, dataset size, and parameter size. [51].

Even though larger capacity is not technically directly translated to better results, research on language modeling has evidently shown that there is a strong relationship between scale and model performance, with larger models performing increasingly better across a wide range of tasks ([96]; [22]). This consistent observation has been formalized with the introduction of empirical *scaling laws*, which describe how performance improves predictably as model size, dataset size, and compute power increase within reasonable limits. In their study, Kaplan et al. [51] demonstrated that test loss declines in a smooth power-law fashion with respect to these three attributes, which together define the *model scale* (Figure 3.1.1). In addition, their findings revealed that larger models are more sample efficient, suggesting that training large models on relatively modest data could be optimal. A later variant of the scaling laws, proposed by Hoffmann et al. [42], aimed to instruct the compute-optimal training for LLMs and, after conducting an extensive set of experiments on various model and data sizes, found a very similar relationship between model performance and scale factors, only questioning earlier claims about the model and data size increase ratio that achieves optimal results. Specifically, they argued that these should scale equally, rather than prioritizing model size over the number of training tokens. Nevertheless, with these promises of consistent and predictable improvements through parameter growth, scaling laws have become a crucial design principle in the rapid development of ever-larger and more powerful state-of-the-art LLMs.



### 3.1.3 Capabilities and Emergent Behaviors

Trained on extensive and diverse text corpora, LLMs have achieved state-of-the-art performance across a variety of standard natural language processing (NLP) tasks such as machine translation ([161]; [93]), text summarization ([151]; [92]), sentiment analysis ([110]; [152]), text classification ([135]; [57]; [134]), and question answering ([50]; [32]). Beyond these traditional applications, LLMs have demonstrated impressive potential to serve as implicit knowledge bases, as they not only manage to retrieve factual information without relying on external data, but also offer key advantages like flexibility and extendability, without requiring schema engineering or human supervision ([40]; [1]). Research has also explored the use of these models within LLM-based multi-agent environments, where multiple LLMs are assigned specialized roles and collaborate to tackle more complex and vague problems through coordinated interaction ([100]; [122]; [38]). Additionally, ongoing advancements have extended their functionality to multimodal domains, where models address tasks involving different modalities such as image, audio, and video ([148]; [129]).

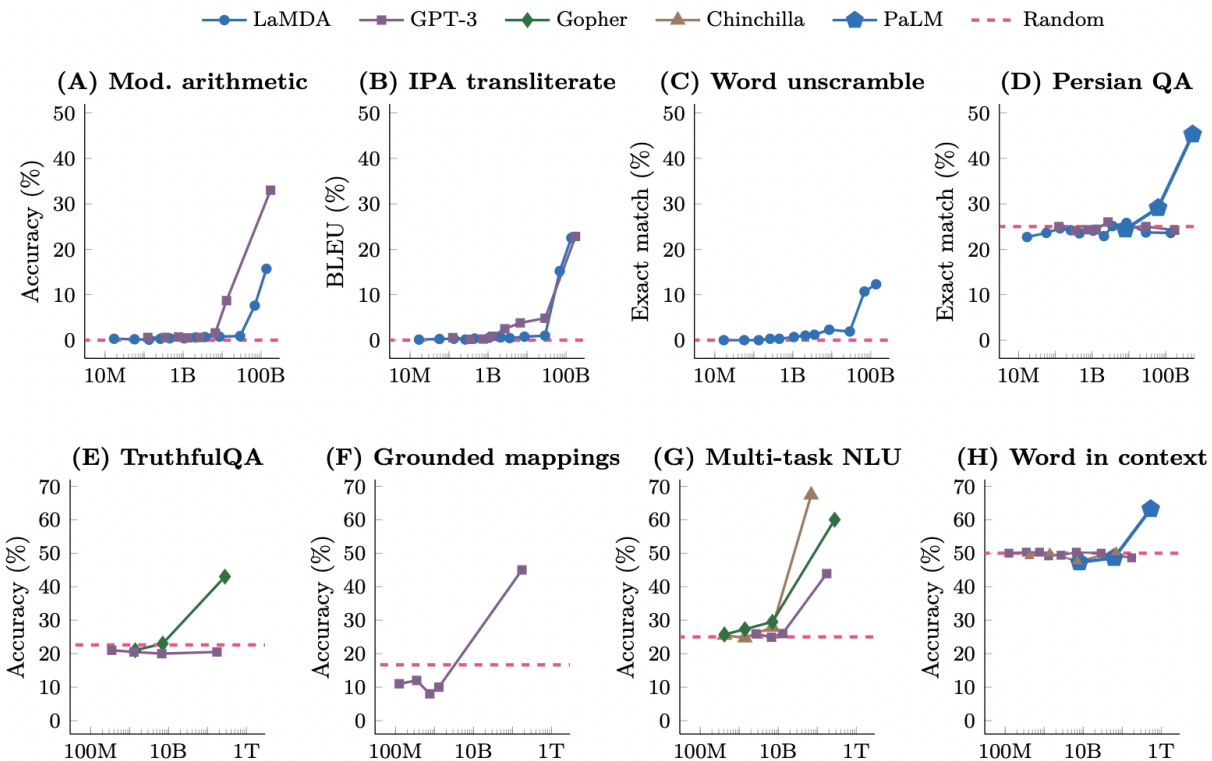


Figure 3.1.2: Emergent abilities of LLMs: Performance of large language models on eight benchmark tasks, showing a sudden jump in capability once the model surpasses a certain parameter scale threshold. [139].

These developments illustrate the versatility and widening potential of these models. However, one of the most intriguing aspects of LLMs that distinguishes them from earlier generations of language models is the emergence of capabilities that are not anticipated or directly predicted by extrapolating scaling laws ([139]; [113]). These *emergent abilities* appear suddenly once the model surpasses a certain scale threshold, before which performance remains near random (Figure 3.1.2). Some of the most representative behaviors that are not present in smaller models but can be elicited through *prompting* in the scope of current large-scale LLMs include in-context learning ([26]; [158]), instruction following ([76]; [88]), code generation ([45]; [15]), compositional generalization [13], puzzle solving [35], and advanced reasoning (which is discussed in more detail in Section 3.4.1.1).

## 3.2 Prompt Engineering

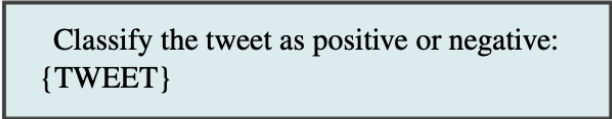
### 3.2.1 Prompts and Prompt Engineering

A *prompt* serves as an input consisting of manually predefined instructions or cues provided to Large Language Models (LLMs) in order to guide their outputs on specific tasks [105]. The systematic practice of designing, structuring, and formulating these instructions in a specialized way that effectively steers model behavior toward desired responses is referred to as *prompt engineering* [105] and, over the past few years, it has emerged as a key technique for enhancing LLM performance across a wide range of tasks and domains [102]. The significance of this new approach lies in its core advantage: unlike previous conventional methods such as re-training and fine-tuning, prompt engineering leverages the pre-existing knowledge encoded in the LLM to improve the generated output without altering its internal parameters [124]. This allows for flexible adaptation to new tasks while entirely avoiding time- and resource-intensive training procedures, thereby maintaining computational efficiency. However, despite its power, prompt engineering remains inherently brittle. LLMs display high sensitivity to the input prompt, which means that even slight changes in wording, the use of synonyms, capitalization, or spacing can yield substantial shifts in performance [105]. The choice of question format appears to deeply influence model behavior as well. For instance, forming "yes or no" or multiple choice questions often results in completely different outputs compared to simple unrestricted generation. In fact, even minor perturbations, like changing the order of the possible options are displayed in the multiple choice format, can affect results [105]. All of these highlight an intriguing challenge at the heart of prompting engineering: the careful search for the most appropriate prompt that can unlock this method's full potential and eventually achieve optimal LLM performance under the given task [73].

### 3.2.2 Prompt Templates

To simplify interactions with LLMs and boost usability across specialized tasks, prompts are usually assembled using *prompt templates* [81]. Prompt templates are structured input formats that typically function as parameterized instructions, containing one or multiple placeholders for variables that, during experimentation, are being replaced by specific textual—or other—instances to create finalized prompts [105]. In this way, the same instruction pattern can be systematically applied to a large volume of data, making it feasible to scale from testing a few examples to running large datasets efficiently.

Consider the task of sentiment analysis of tweets. Figure 3.2.1 includes an example of a prompt template that instructs models to classify a tweet as either positive or negative. In this template, {TWEET} is the variable placeholder that is replaced with the actual tweet to be analyzed, producing a prompt *instance* which is then fed to the LLM for inference [105].



Classify the tweet as positive or negative:  
{TWEET}

Figure 3.2.1: Prompt template example for the task of tweet sentiment analysis [105].

### 3.2.3 Prompting Techniques

In the search for the "most efficient prompt" that can optimally extract the desired response for a specific task, several *prompting techniques* have been developed and evolved to improve the ability of Large Language Models to follow instructions and reason successfully.

#### 3.2.3.1 Zero-Shot Prompting

Zero-Shot prompting (Figure 3.2.2a) is the simplest form of prompt engineering, consisting solely of a direct instruction to complete a specific task, without providing additional examples or cues on how to approach it [102]. In this setup, the model relies on its embedded knowledge to generate predictions, which often proves sufficient to perform adequately on various downstream tasks, including reading comprehension,

translation, or summarization, thanks to its extensive pre-training on vast amounts of data [55]. However, the Zero-Shot technique is typically outperformed, especially under more difficult scenarios that require nuanced understanding or complex reasoning ([124]; [8]). Nevertheless, Zero-Shot prompting remains a foundational method, setting a baseline to compare with more advanced strategies.

### 3.2.3.2 One-Shot Prompting

The One-Shot prompting strategy (Figure 3.2.2b) includes a single example of successful performance on a specific instance of the described task, to help the model better understand the task’s requirements, expected output format, or preferred reasoning process. This method is considered to be closer to the way more complex tasks are often communicated to humans, where the absence of a worked example usually leads to confusion about how to proceed [8].

### 3.2.3.3 Few-Shot Prompting

Few-Shot prompting (Figure 3.2.2c) operates exactly like one-Shot prompting, but instead of one, it provides multiple demonstrations of input-output instances to enhance the model’s understanding of the given task [8]. The presentation of high-quality examples has been shown to improve LLM performance on more complex tasks compared to simple instruction alone [102]. However, Few-Shot prompts are inherently challenging to implement in order to be effective. Factors such as the selection, similarity, quantity, and order of exemplars—as well as the format or placement of instructions—can substantially influence model responses [105]. For example, varying the order in which the task instances are demonstrated can intriguingly produce accuracy scores that vary from sub-50% to over 90% [78]. Therefore, careful decisions throughout the prompt design process are critical to ensuring optimal LLM behavior.

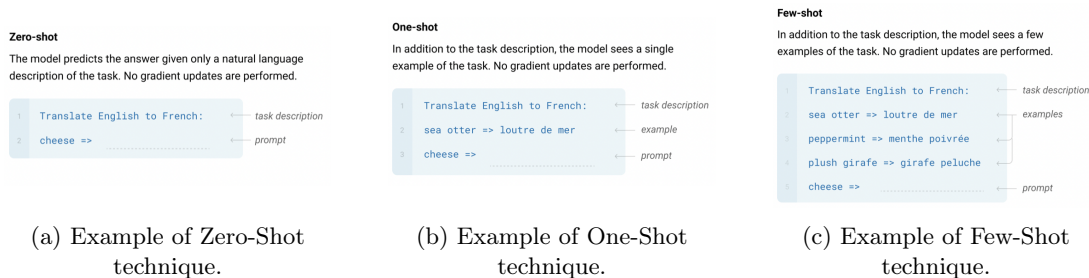


Figure 3.2.2: Comparison between the Zero-Shot, One-Shot and Few-Shot prompting techniques with examples on the English-to-French translation task [8].

### 3.2.3.4 Chain-of-Thought Prompting

Despite their undeniable potential, Large Language Models often encounter difficulties when challenged with questions that are not directly answerable without intermediate inferences. The Chain-of-Thought (CoT) prompting technique was introduced in order to address this issue by encouraging the model to articulate its thought process through a sequence of immediate outputs, before generating the final answer [102]. Experimental results have shown that the employment of these reasoning chains improves LLM performance—often to a remarkable degree—under various non-trivial tasks, including multi-hop question-answering, arithmetic, commonsense and, symbolic reasoning problems ([138]; [126]).

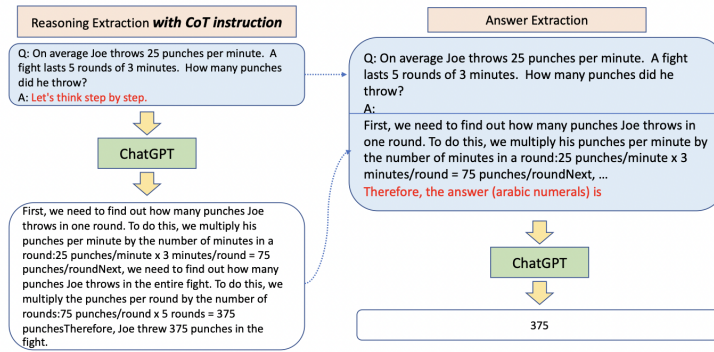


Figure 3.2.3: Application of Chain-of-Thought prompting for arithmetic reasoning [14].

Chain-of-Thought prompting can be incorporated into both Zero-Shot and One-Shot/Few-Shot scenarios. In the Zero-Shot setting, a simple instruction like "Let's think step by step." is added to the prompt to encourage task decomposition ([55]; [155]). In the One-Shot or Few-Shot settings, each demonstration typically consists of a question followed by a manually designed natural language rationale that leads to the final answer [138].

### 3.3 Evaluation with LLMs

The rapid advancements of Large Language Models (LLMs) and their widespread application across various fields have given rise to the need to develop reliable methods to evaluate these models across diverse contexts—a task that continues to be challenging. Traditional machine learning tasks such as classification and regression typically use programmable and statistical metrics, including accuracy, precision, and recall. While these metrics are reliable, they are only applicable to a narrow range of cases that involve well-defined outputs and ground truths and are implemented in very specific formats [64]). However, with the advent of deep learning and the evolution toward LLMs, the nature of model outputs has fundamentally changed. The responses have become increasingly complex, highly generative, open-ended, and context-dependent, to the point where standardized metrics are extremely insufficient for high-quality evaluation [64]). Natural language generation tasks such as summarization, writing, and question answering, where multiple outputs could be considered valid, require evaluation methods capable of capturing nuanced qualities like text fluency, coherence, relevance, or creativity—something that even more advanced metrics such as BLEU [91], ROUGE [71], or METEOR [5] still fail to achieve ([37]; [64]). Interestingly, even in simpler scenarios—designed to facilitate the use of regular expressions for assessment (e.g. by formatting in a multiple choice response manner)—models sometimes deviate from instructions (e.g. by producing additional output), making it difficult to programmatically extract the desirable result. Human evaluation remains the gold standard for capturing human preferences, capable of addressing all these limitations with ease due to the human nature. However, the process of individually examining each response is typically exceptionally time-consuming, resource-intensive and, thus, difficult to scale across large datasets ([37]; [64]).

In the search for a more practical, reliable, and adaptive evaluation method that could overcome these challenges and replace both humans and insufficient standardized metrics, researchers have turned to the idea of leveraging state-of-the-art LLMs themselves as evaluators. This approach, very promising considering that these models already exhibit strong human alignment across various domains, is often referred to as "LLM-as-a-judge" [157] and has received a lot of attention over the past few years ([37]; [64]). Formally, the LLM-as-a-judge paradigm is a flexible evaluation framework where LLMs are employed as evaluative tools to assess the quality, relevance, and effectiveness of generated outputs according to defined criteria [64]. Based on their design, the LLM evaluation systems are categorized into three primary configurations: Single-LLM ([72]; [74]; [149]), Multi-LLM ([10]; [23]; [69]), and Human-AI Hybrid ([68]; [107]) systems, which involve collaboration between LLM and human evaluators. Each evaluation system is prompted with a carefully crafted input of instructions. These include the evaluation type, for example, pointwise ([54]; [127]; [144]), pairwise ([9]; [41]), or listwise ([163]; [143]; [43]) evaluation, and the evaluation criteria, which can be general, like accuracy ([17]; [46]) and linguistic quality ([30]; [20]) or specifically designed to fit the respective task ([119]; [70]; [44]), for reference-free evaluation ([39]; [108]) scenarios, and, additionally, the reference data used

to determine whether the performance meets the expected standards in reference-based evaluation [33] cases. After processing the prompt, the LLM outputs the evaluation result, which, based on the task instructions, can be a score, a categorical label, or a ranked list, and is optionally accompanied by an explanation ([146]; [142]) or feedback ([79]; [19]). It is evident that this method offers various key benefits like scalability, explainability, and most importantly, task-adjustability, positioning it as an excellent alternative to human evaluation and existing metrics, which only leads to one critical question: *can it be trusted?*

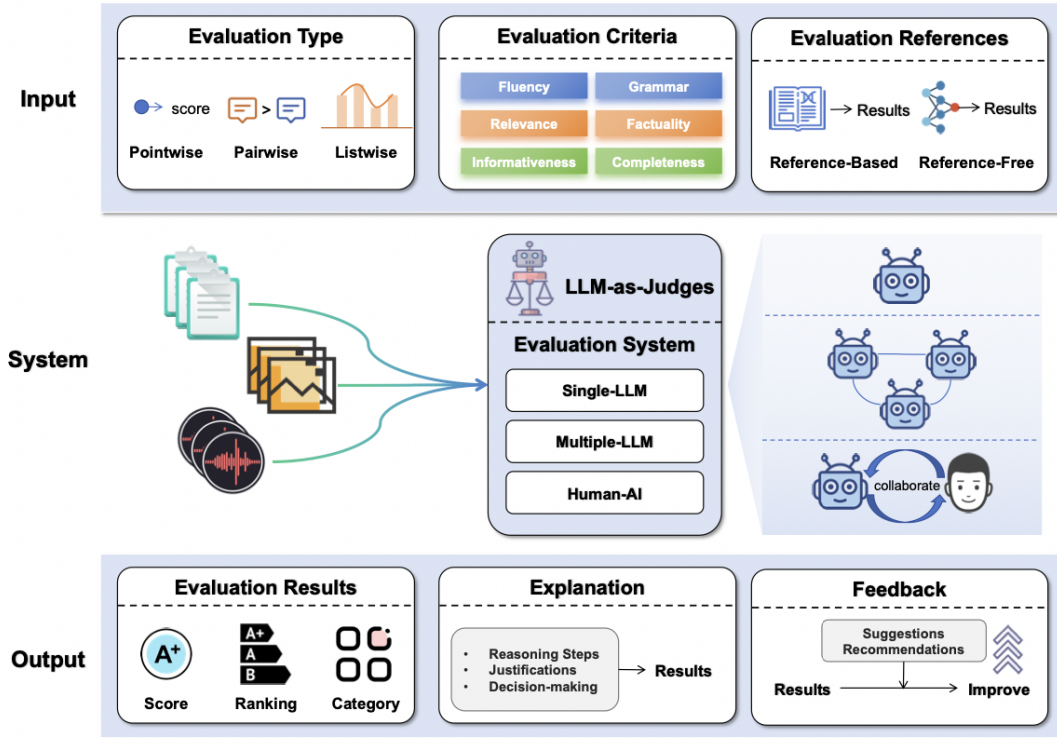


Figure 3.3.1: Overview of the LLMs-as-judges systems [64].

The idea of using LLMs as judges seems promising, but it also raises concerns about whether standard LLM shortcomings, such as hallucinations ([101]; [153]), biases [34], and lack of robustness [159], can seriously affect evaluation outcomes. In this context, a growing body of research focuses on the process of *meta-evaluation*, where the quality of LLM-as-a-judge systems is assessed in terms of reliability, validity and consistency [64]. A common approach to evaluating LLM evaluators is to measure their agreement with human preferences. This is typically computed by the proportion of samples on which the LLM and human annotators agree [120], though other metrics such as Spearman’s correlation ([3]; [75]), Cohen’s Kappa [120], or the standard precision, recall and F1 scores ([132]; [160]) are also used. Studies have repeatedly demonstrated **high agreement** between LLMs-as-judges and human evaluators across a diverse set of benchmarks, including code generation ([16]; [47]; [164]), machine translation ([33]; [52]), text summarization ([30]; [89]; [128]), automatic story generation ([20]; [21]), dialogue generation [36], multimodal [11], and multilingual [111] tasks. However, LLM evaluators are not without limitations. The meta-evaluation of LLM-as-a-judge systems introduces certain **biases** which can compromise the fairness of decisions. Some of the most frequently observed biases include position bias ([157]; [12]; [43]), verbosity bias ([157]; [85]; [145]), authority bias ([145]; [12]), and diversity bias ([145]; [12]). Nevertheless, with carefully designed prompts and appropriate caution, strong LLMs can achieve robust and reliable results when it comes to evaluating other model outputs. In other words, when traditional evaluation methods are insufficient or impractical, we can actually prompt LLMs to judge model performance.



## 3.4 Related Work

### 3.4.1 True Reasoning in Large Language Models

#### 3.4.1.1 Emerging but Limited: Reasoning in LLMs

The remarkable human-like behaviors exhibited by Large Language Models have inspired researchers to investigate their ability to carry out goal-directed inference grounded in rules, patterns, and prior knowledge [28]. To that end, recent advancements have demonstrated meaningful progress in enabling LLM reasoning through language-based prompts [94]. Efforts to evaluate the cognitive behaviors of language models have concentrated heavily on several specific types of reasoning, including inductive: extrapolating abstract rules from provided facts ([65]; [7]), deductive: drawing conclusions based on rules ([18]; [25]), causal: identifying causal relationships between variables or events ([48]; [133]), analogical: leveraging relevant experiences to tackle new tasks ([95]; [117]), commonsense: applying human everyday knowledge ([118]; [103]), and arithmetic: interpreting and solving numerical reasoning problems ([84]; [130]).

Notably, as LLMs grow in size and training complexity, they demonstrate substantial improvements on some of these benchmarks, leading to claims of emerging reasoning abilities. However, although these recent advances are intriguing and noteworthy, LLM reasoning capabilities remain brittle in comparison to their remarkable success across a wide range of standard language processing tasks such as reading comprehension, translation, summarization, and factual question answering. This contrast obscures a concerning distinction between formal and functional linguistic competence [80]. While LLMs can generate coherent and convincing text, it often lacks grounding in deeper understanding and genuine cognitive processes, making true reasoning a critical challenge that raises important questions about the unresolved boundary between artificial and human cognition.

#### 3.4.1.2 Evidence of Fragile Reasoning

The gap between surface-level linguistic abilities and deeper cognitive competence becomes even more pronounced when tasks are presented in alternative formulations and unusual or unnatural contexts that fall outside typical LLM training distributions. Numerous studies have highlighted the fragility of LLM reasoning in various *counterfactual* scenarios. Some of these scenarios are designed to be significantly less common than the standard cases—though not entirely absent in the model’s pretraining data—suggesting that while LLMs may possess the knowledge required for successful reasoning, they often default to simply "reciting" strongly memorized information. For example, consider performing arithmetic calculations presented in non-decimal bases like base-9 or base-11, compared to the default base-10 or the alternative but more frequently seen in technical literature base-8 and base-16 [140]. Other tasks introduce wholly hypothetical conditions, such as starting the sentence with a statement like "If cats were vegetarian", revealing that LLMs often struggle to entirely override what they know is true in the world and adjust to novel, counterfactual contexts ([66]; [67]; [147]). Similar findings have been reported when using alternative or counterfactual premises to evaluate the ability of LLMs to generalize analogies ([62]; [115]), infer causation [48], or reason under counter-commonsense conditions [56]. In general, LLMs appear unable to effectively *adapt* to alterations of familiar tasks, performing poorly when confronted with, for example, arithmetic expressions involving modified operator precedence, translations from English into artificial languages, or deductions with unexpected twists [63]. This limitation in reasoning becomes even more pronounced when alternative prompts fail to generalize to equivalent variations of the same task, underscoring how human intuitions about what "makes no difference" do not hold for LLMs [4]. Altogether, this reveals a concerning brittleness to both linguistic and contextual shifts.

#### 3.4.1.3 Memorization behind LLM Reasoning

Emerging research claims that much of the recent success in LLM reasoning may be illusory, as models often rely heavily on lexical or structural patterns memorized from pretraining data, rather than conducting true inferential reasoning. Xie et al. [141] develop a method for systematically measuring memorization and report that, during fine-tuning, models increasingly memorize instances of training puzzles, which significantly improves their performance across similar inputs. However, while LLMs achieve, in fact, near-perfect accuracy on original task instances, they continue to struggle when presented with variations. Similarly, Lou et al.

[77] introduce an axiomatic system that formally decomposes the interaction effects used by LLMs during inference into sets of memorization effects and in-context reasoning effects. This framework allows them to precisely quantify the extent to which memorization contributes to a model’s specific decision, offering deeper insight into when LLMs are merely recalling patterns from context or training data, or engaging in genuine reasoning processes. Wang et al. [131] take a different approach that further supports this view by directly calculating correlations between model output probabilities and the distribution of the pretraining data. Their goal is to trace the origins of LLM capabilities, with a specific focus on the interplay between memorization and generalization. They demonstrate that, while memorization plays a dominant role in simpler, factual question-answering tasks, it can also influence performance on reasoning problems. Collectively, these findings indicate that task performance alone may misrepresent LLM reasoning capabilities, underscoring the need for evaluation methods that aim to disentangle the underlying sources of model behavior.

### 3.4.2 Inverse scaling problems

Parameter size has been shown to be a key factor affecting the performance of Large Language Models, as a consistent scaling trend has emerged across numerous experiments, strongly suggesting that larger models generally perform better across a broad range of benchmarks [51]. However, recent investigations into LLM behavior have interestingly unveiled a rather counterintuitive phenomenon: in certain scenarios, task performance worsens as model scale increases, indicating cases of *inverse scaling problems*. Tasks that provoke this oddity are designed to expose the limitations of the most powerful LLMs by highlighting the differences in the way they handle complex reasoning compared to humans, who can often solve such problems with ease.

Several inverse scaling tasks were introduced as part of the Inverse Scaling Prize public contest, designed to systematically investigate why model performance consistently decreases with parameter size in certain scenarios [82]. The collected examples of inverse scaling are finally categorized according to the potential causes that lead to this behavior. Specifically, four distinct categories are identified:

1. **Strong Prior:** Includes cases where the prompted task description contains information that contradicts the model’s embedded knowledge. A conflict arises between following the external instruction—by suppressing or overriding strong internalized priors—and adhering to the answer that is inherently associated with higher correctness probability based on the model’s pretraining. Larger LLMs tend to develop a stronger dependence on the pretraining text and may even completely disregard the information provided within the input, which is actually crucial for solving the given problem correctly. Tasks that fall into this category are:
  - **Resisting Correction:** This task instructs the model to repeat an input sentence without modifying it. The prompt showcases examples of successful repetitions and ends with a sentence that contains some kind of error or abnormality. LLMs with strong high-confidence priors are more likely to select the typically correct sequence, failing to reproduce the given input exactly as instructed.
  - **Memo Trap:** LLMs are challenged with an instruction that directly contradicts a commonly represented word sequence—like a famous quote—thereby compounding memory activation effects and making especially difficult for more knowledgeable models to override pre-existing conceptual mappings.
  - **Redefine:** This task changes the standard definition of a well-known symbol or word and then prompts the model to answer a simple question accordingly. More specifically, LLMs are challenged to select between two options: one that is consistent with the new assignment and one aligned with the default, pre-established meaning of the redefined entity. Once again, larger models tend to prioritize the answer that reflects their pretraining-based knowledge.
  - **Prompt Injection:** The prompt consists of a simple command, such as repeating or capitalizing a sentence, alongside a strict request that the model should not follow any additional instructions embedded within the input. This is followed by several question-answer exemplars and, finally, a sentence that includes an *injected* command. Large models seem incapable of distinguishing which instructions should—and should not—be executed, even when this is clearly explained in the input

prompt, and tend to favor the most recent commands instead.

2. **Unwanted Imitation:** During the pretraining process, Large Language Models are exposed to a vast and diverse range of data, sometimes containing human biases, reasoning mistakes, or misinformation. In other words, LLMs are inadvertently *trained* to generate responses that replicate these unwanted patterns. As parameter size increases, so does their ability to predict text sequences. In this way, larger models achieve better results on general tasks, but it also means that they are more likely to imitate undesirable behaviors within the pretraining corpus. In the following task, such unwanted imitation is identified as the underlying cause of inverse scaling:

- **Modus Tollens:** Modus Tollens is a basic type of deductive reasoning that typically follows this structure: If  $p$ , then  $q$ ; not  $q$ ; therefore, not  $p$ . Prompts in the Modus Tollens tasks include examples that follow this argument type and instruct the model to decide whether the conclusion is logically follows from the preceding statements. Since humans frequently fail to perform Modus Tollens reasoning successfully, instances of similar reasoning errors are present in the pretraining data. Therefore, although such problems are expected to be easy for potent LLM reasoners to solve, they become susceptible to answering incorrectly, imitating human unwanted tendencies.

3. **Distractor Task:** In these inverse scaling problems, the prompt is carefully designed to indirectly reveal the presence of an alternative, easier task—one that is different enough from the actual one but easily confused with it. This easier task acts as a *distractor*, meaning that it diverts the model’s attention away from the more difficult, intended task, leading to incorrect answers. This happens because, during their extensive pretraining, large models build high confidence in familiar, easy, or straightforward patterns, making it extremely difficult for them to ignore these patterns once recognized—even when they are misleading.

- **Pattern Match Suppression:** Recent research on induction heads has uncovered that transformer-based models rely on advanced pattern-matching mechanisms to complete sequences, a behavior that is in fact tied with in-context learning [87]. This task challenges LLMs to counteract this inherent bias by interrupting a repetitive pattern. Even when the instruction within the prompt explicitly requests a deliberately unexpected continuation, models tend to get *distracted* by the familiar task and default to completing the predictable sequence.
- **NeQA:** This inverse task modifies an existing multiple choice question-answering dataset by programmatically negating each question. Although this new phrasing may seem more confusing than the original one, humans are generally able to successfully adapt to the updated meaning and identify the answer that aligns with the negation. Contrarily, LLMs show a tendency to entirely miss the negation in the question, driven by the strong conceptual association between the queried entity and the answer that was originally correct—but now is incorrect. This behavior can be particularly concerning, considering that the model actually does the exact opposite of what is intended. This problem has been thoroughly examined in [154].
- **Sig Figs:** The prompt instruction asks to express a decimal number to the correct number of significant digits. Interestingly, larger models often end up performing a different, but similar task: rounding the given number to the corresponding amount of decimal points instead of significant figures. This propensity to substitute a task with a more familiar one, in which the model appears more confident, mirrors findings in human psychology of prediction, where individuals are likely to unconsciously replace a challenging task with a related one that is easy for them to perform [49].
- **Into the Unknown:** The LLM receives a short description of a situation, followed by a question that requires additional information in order to be answered. The task is not to directly answer this question but to determine which piece of additional information would help answer the question, choosing between two possible options. One of these options serves as a *distractor*, consisting of a rephrasing of a specific statement already present in the setting description. LLMs, influenced once again by their in-context pattern-matching tendencies, are more likely to select the option that is redundant to the original input, rather than engage in effective reasoning with new information.

4. **Spurious Few-Shot:** These tasks use the Few-Shot prompting technique to lure LLMs into undesirable



behaviors by listing exemplars that are correctly labeled but follow a deceptive pattern—one that is consistent across demonstrations but deliberately violated in the final question. Few-Shot prompting has been shown to significantly influence LLM performance across a wide range of tasks, often resulting in significant improvements [8]. In these cases, however, instead of learning the expected reasoning process from the given examples, the model seems to imitate the misleading patterns that "happen" to be embedded within the demonstrations.

- **Hindsight Neglect:** This problem describes a decision-making scenario involving a game with probabilistic outcomes, followed by the actual outcome of the situation. The model is tasked with evaluating whether the prompted game is worth playing. Larger LLMs mistakenly base this assessment on the outcome alone rather than on the expected value, which is derived from the probabilities and possible results of the two contingencies and is therefore independent of whether the person actually lost or won money. This behavior appears to result from imitating the demonstrations provided within the prompt, where the quality of the decision was intentionally aligned with the actual outcome, creating a *spurious* pattern.
- **Repetitive Algebra:** LLMs are prompted to find the solution to a simple first-degree mathematical equation. Within the input, multiple examples of similar questions are demonstrated in a Few-Shot manner, all following a very specific pattern: all of the example-questions have the exact same answer with the to-be-solved equation, except for the last one, which has a different result. Although smaller LLMs are typically weaker arithmetic reasoners, they perform better in these scenarios as they tend to favor the most frequent result. On the other hand, larger models interestingly exhibit a strong bias toward the most recent answer. In this case, few-shot examples lead both smaller and larger models into an undesirable behavior, but the task is intentionally crafted to trap more powerful LLMs in an inverse scaling trend.

This thesis takes the *Redefine* inverse scaling problem—introduced under the Strong Prior category—as its central focus. While the benchmark highlights the existence of this issue, we argue that the task merits focused investigation. To that end, we construct two custom datasets specifically targeting redefinitions across two distinct semantic domains and various scenario types, carefully designed to stress-test model adaptability. We evaluate a range of state-of-the-art LLM families, analyzing behavioral shifts in relation to model scale.



## Chapter 4

# Methodology

The *redefinition task* forms the foundation of the experimental methodology presented in this thesis. It introduces a controlled setting where the canonical values of familiar, well-established entities or concepts are deliberately altered to assess how Large Language Models (LLMs) respond when confronted with contradictions to foundational knowledge. This task challenges models to abandon strongly internalized conceptual associations and instead engage in flexible reasoning based on local contextual cues. Figure 4.0.1 illustrates a visual example of this scenario, where the famous mathematical constant  $\pi$  is redefined as 100, replacing the standard value of 3.1415, and calling for adaptive reasoning pathways in order to carry out downstream calculations accordingly. This framework allows us to probe modern LLM behavior along dimensions of both memorization and reasoning flexibility.

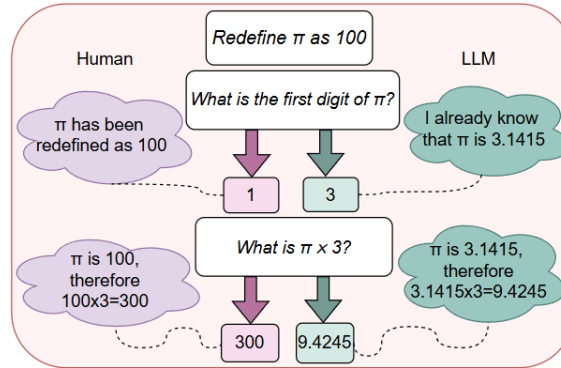


Figure 4.0.1: Redefined reasoning pathways.

The objective of this chapter is to detail the key methodological components underlying our implementation of the redefinition task. It includes the construction of task-specific datasets, the design and formulation of prompt templates that span various strategies and formats, the evaluation metrics and methods applied to assess model performance, the selection of LLMs for empirical analysis, and, finally, a description of the infrastructure used to efficiently conduct our experiments.

### 4.1 Datasets

In this section, we describe how the datasets used in our experiments were constructed to explore the performance of Large Language Models (LLMs) across different redefinition tasks. Our data consist of two distinct components: (1) Redefinition of Scientific Constants, and (2) Redefinition of Units of Measurement.

### 4.1.1 Redefinition of Scientific Constants Dataset

#### 4.1.1.1 Selection of Constants

For the constant redefinition task, we selected 15 well-recognized mathematical and physical constants that are frequently used in scientific calculations. These include the following: pi ( $\pi$ ), Euler’s number ( $e$ ), phi ( $\phi$ ), the speed of light ( $c$ ), the gravitational constant ( $G$ ), Planck’s constant ( $h$ ), the elementary charge ( $q_e$ ), Avogadro’s number ( $N_A$ ), the Boltzmann constant ( $k_B$ ), the gas constant ( $\bar{R}$ ), the imaginary unit ( $i$ ), the square root of 2 ( $\sqrt{2}$ ), infinity ( $\infty$ ), the vacuum electricity permittivity ( $\epsilon_0$ ) and *zero*.

#### 4.1.1.2 Redefinition Scenarios

In general, by "redefining" a constant, we mean simply substituting its standard value with an alternative one. Based on these new values, we define two types of redefinition, each involving its own progression of difficulty, designed to rigorously test model adaptability. Table 4.1 presents the redefinitions applied to the selected constants across all types and difficulty levels.

- **Assignment ( $R_a$ ):** Involves directly assigning a randomly selected new value to a constant, overriding the default one.
  1. **Minor Perturbation ( $R_{a1}$ ):** Slightly alters the constant’s default value (e.g., "Redefine  $\pi$  as 4.5") to examine how LLMs handle deviations within a plausible range.
  2. **Significant Deviation ( $R_{a2}$ ):** Assigns a value that differs from the actual one by orders of magnitude (e.g., "Redefine  $\pi$  as 500") to test the models’ flexibility over substantial variance.
  3. **Extreme and Unrealistic Values ( $R_{a3}$ ):** Uses values that seem physically implausible or nonsensical in the real world (e.g., "Redefine  $\pi$  as -10") to observe LLMs’ behavior in extreme and irrational scenarios.
- **Swapping ( $R_s$ ):** Replaces the established value with that of another constant.
  1. **Simple Swap ( $R_{s1}$ ):** Interchanges values between constants with similar order of magnitude (e.g., "Redefine  $\pi$  as  $\phi$ ").
  2. **Complex Swap ( $R_{s2}$ ):** Interchanges constants with drastically different values (e.g., "Redefine  $\pi$  as the Planck’s constant").

	Actual value	Unit	$R_{a1}$	$R_{a2}$	$R_{a3}$	$R_{s1}$	$R_{s2}$
$\pi$	3.14159	-	4.5	500	-10	$\phi$	$h$
$e$	2.71828	-	9	1300	$1.5 \times 10^{-12}$	$pi$	$k_B$
$\phi$	1.61803	-	3.6	321	-2.2	$e$	$N_A$
$c$	299,792,458	$m/s$	$2.3 \times 10^8$	10	$-4 \times 10^8$	$N_A$	$q_e$
$G$	$6.674 \times 10^{-11}$	$m^3/kg * s^2$	$1.1 \times 10^{-10}$	50	-525	$q_e$	$pi$
$h$	$6.626 \times 10^{-34}$	$J * s$	$5 \times 10^{-33}$	482	-0.2	$k_B$	$\phi$
$q_e$	$1.602 \times 10^{-19}$	$C$	$2.4 \times 10^{-21}$	$3 \times 10^4$	$3 \times 10^{50}$	$\epsilon_0$	$\pi$
$N_A$	$6.022 \times 10^{23}$	$mol^{-1}$	$8.23 \times 10^{23}$	75	-1	$\bar{R}$	$e$
$k_B$	$1.380649 \times 10^{-23}$	$J/K$	$4.56 \times 10^{-24}$	80	$-9.9 \times 10^{-3}$	$\epsilon_0$	$pi$
$\bar{R}$	8.314	$J/(mol * K)$	13	3500	-400	$\pi$	$c$
$i$	$\sqrt{-1}$	-	$\sqrt{-2}$	$\sqrt{-100}$	1	$\phi$	$\bar{R}$
$\sqrt{2}$	1.41421356	-	5	31.62	-2	$\pi$	$\epsilon_0$
$\infty$	infinity has no value	-	$10^{10}$	100	-1	$c$	$q_e$
$\epsilon_0$	$8.854 \times 10^{-12}$	$F/m$	$9.3 \times 10^{-10}$	35	$3 \times 10^{12}$	$G$	$\phi$
zero	0	-	-1	100	$5 \times 10^{30}$	$h$	$c$

Table 4.1: Varying levels of difficulty for constant redefinitions (assignments and swaps).

## 4.1.1.3 Question Levels

	$Q_1$	$Q_2$	$Q_3$
$\pi$	What is the first digit of pi?	What is $\pi$ multiplied by 3?	What is the Earth’s surface area?
$e$	What is the first digit of e?	What is $e^2$ ?	If a population grows continuously at a rate of 5% per year, by what factor will it increase in 10 years?
$\phi$	What is the first digit of phi?	What is $5 * \phi - 2$ ?	If a rectangle has sides in the golden ratio and the longer side is 8 cm, what’s the length of the other side?
$c$	How far does light travel in one second?	How much time (in sec) does it take light to travel a distance of 100 million km?	What is the energy equivalent of 8 grams of mass?
$G$	What is the first non-zero digit of the gravitational constant?	What the gravitational constant multiplied by 7?	If two 15 kg masses are placed 2 meters apart, calculate the gravitational force between them.
$h$	What is the first non-zero digit of Planck’s constant?	If the frequency of a photon is 4 Hz, what is its energy? Use the formula $E = h * v$ .	In the photoelectric effect, if a metal has a work function of $4.5 \times 10^{-19} J$ , what is the minimum frequency of light required to eject an electron from the metal surface?
$q_e$	What is the first non-zero digit of the elementary charge?	If an electron has a charge of $-e$ , what is the charge of two electrons?	A capacitor stores a charge of $3.2 \times 10^{-18}$ coulombs. How many elementary charges $e$ are equivalent to this amount of charge?
$N_A$	What is the first digit of the Avogadro’s number?	How many atoms are there in 1mol of any element?	Calculate the number of molecules in 54grams of water (molar mass of water is $\sim 18g/mol$ ).
$k_B$	What is the first non-zero digit of the Boltzmann constant?	Calculate the energy associated with a temperature of 300 K for a particle using the formula $E = kT$ .	What is the temperature at which the average kinetic energy of a particle is $1.9 \times 10^{-21} J$ ?
$\bar{R}$	What is the first digit of the gas constant?	What is the gas constant divided by 2?	If you have 2 moles of an ideal gas at a temperature of 300K, what is the pressure (in Pa) if the volume is 10liters?
$i$	What is the value of $i^2$ ?	What is the value of $i^3$ ?	If $z_1 = 1 + i$ and $z_2 = 1 - i$ , calculate $z_1 \cdot z_2$ .
$\sqrt{2}$	What is the first digit of the squared root of 2?	Calculate the value of squared root of 2 multiplied by 3. What is it approximately?	If one side of a square is 5 units long, what is the length of the diagonal of the square?
$\infty$	What is the value of infinity?	What is the limit of $1/x$ as $x$ approaches infinity?	What is the horizontal asymptote of the function $f(x) = (5x + 30000)/(x + 1000), x > 0$ ?
$\epsilon_0$	What is the first non-zero digit of vacuum electric permittivity?	If you add the value of vacuum electric permittivity to itself, what do you get?	Calculate the electric force between two charges $q_1 = 3\mu C$ and $q_2 = 5\mu C$ separated by 12m in a vacuum.
zero	What is the absolute value of zero?	What is 300 multiplied by zero?	If $y = \sin(x)/x$ , what is the limit of $y$ as $x$ approaches 0?

Table 4.2: Questions of three difficulty levels ( $Q_1, Q_2, Q_3$ ) for units of measure.

In addition to the redefinition levels, we formulated three tiers of increasingly difficult questions to compare LLM performance across inquiries of varying complexity. All questions related to constant redefinitions are demonstrated in Table 4.2. The question levels are as follows:

1. **Basic Recall ( $Q_1$ ):** In Level 1 of question difficulty, we request an answer that can be extracted directly from the constant’s value itself (original or redefined). The mainly asked question across this level is "What is the first non-zero digit of {constant}?", where the correct answer is the leftmost digit of the constant’s absolute value, ignoring leading zeros in the case of decimal fractions. For example, the answer of "What is the first non-zero digit of  $\pi$ ?" is "3" for  $\pi$ ’s original value (3.14159) and "5" when  $\pi$  is redefined as 500.
2. **Simple Computation ( $Q_2$ ):** In the second level of difficulty, the model is prompted to execute a relatively simple mathematical computation regarding the real or redefined value of the constant in question (e.g., "What is  $\pi$  multiplied by 3?").
3. **Multi-Step Reasoning ( $Q_3$ ):** The most difficult questions challenge the LLMs to solve a more complicated mathematical or physical problem that requires multi-hop reasoning. These often involve intricate calculations, the retrieval of relevant scientific equations, or additional domain knowledge. For example, to answer the question "What is the Earth’s surface area?" the model must recall the Earth’s radius and apply the formula for the surface area of a sphere ( $A = 4\pi r^2$ ), substituting  $\pi$  with the corresponding value.

## 4.1.2 Redefinition of Units of Measure Dataset

### 4.1.2.1 Selection of Units

For the second redefinition task, we chose the well-defined domain of units of measurement. Specifically, we selected key units across the following fundamental quantities: time (minutes-*min*), weight (kilogram-*kg*), length (meter-*m*) and light-year (*ly*), temperature (Kelvin-*K*), volume (milliliter-*mL*), energy (calorie-*cal*), pressure (atmosphere-*atm*), voltage (Volt-*V*), frequency (megaHz-*MHz*), force (newton-*N*), magnetic flux density (Tesla-*T*), area (hectare-*ha*), illuminance (lux-*lx*), and information storage (byte-*B*).

### 4.1.2.2 Redefinition Scenarios

Unlike constants, units of measurement are not associated with a specific numerical value that a model can recall. Therefore, "redefinition" in this context entails modifying the unit’s relationship with its derived counterparts for the same physical quantity. For example, a possible redefinition of minutes would involve altering their relationship with seconds, such as "Redefine 1 minute as 100 seconds" (instead of 60 seconds). Given this approach, the swapping scenario is impractical. However, for the **Simple Assignment** case, we designed three escalating levels of modification, similar to the constants dataset:

1. **Slight Adjustments ( $R_a1$ ):** Slightly modifying the standard conversion from one unit to its derived counterpart (e.g., "Redefine 1 minute as 100 seconds").
2. **Significant Deviation ( $R_a2$ ):** Drastically changing the scaling between units and their derived counterparts (e.g., "Redefine 1 minute as  $5 \times 10^8$  seconds").
3. **Unrealistic Redefinitions ( $R_a3$ ):** Redefining the relationships between units of measure in a way that results in contradictory and insensible scenarios (e.g., "Redefine 1 minute as -50 seconds").

Unit	Derived unit	Actual value	$R_{a1}$	$R_{a2}$	$R_{a3}$
1 <i>min</i>	seconds ( <i>sec</i> )	60 <i>sec</i>	100 <i>sec</i>	$5 \times 10^8 \text{ sec}$	−50 <i>sec</i>
1 <i>kg</i>	grams ( <i>gr</i> )	1000 <i>gr</i>	900 <i>gr</i>	$10^{-14} \text{ gr}$	−100 <i>gr</i>
1 <i>m</i>	centimeter ( <i>cm</i> )	100 <i>cm</i>	60 <i>cm</i>	$310^1 \text{ cm}$	−200 <i>cm</i>
<i>K</i>	Celsius degrees ( $^{\circ}C$ )	$^{\circ}C + 273.15$	$^{\circ}C + 300$	$^{\circ}C + 1$	$100 * (^{\circ}C) + 500$
1 <i>mL</i>	cubic centimeter ( $\text{cm}^3$ )	1 <i>cm</i> <sup>3</sup>	2 <i>cm</i> <sup>3</sup>	10000 <i>cm</i> <sup>3</sup>	−10 <i>cm</i> <sup>3</sup>
1 <i>cal</i>	Joule ( <i>J</i> )	4.184 <i>J</i>	9 <i>J</i>	1500 <i>J</i>	−5 <i>J</i>
1 <i>atm</i>	Pascal ( <i>Pa</i> )	101,325 <i>Pa</i>	215,000 <i>Pa</i>	0.55 <i>Pa</i>	−5000 <i>Pa</i>
1 <i>V</i>	milivolt ( <i>mV</i> )	1000 <i>mV</i>	500 <i>mV</i>	410 <sup>9</sup> <i>mV</i>	−10 <i>mV</i>
1 <i>MHz</i>	Hertz ( <i>Hz</i> )	10 <sup>6</sup> <i>Hz</i>	10 <sup>5</sup> <i>Hz</i>	2 <i>Hz</i>	−10 <sup>3</sup> <i>Hz</i>
1 <i>N</i>	millinewton ( <i>mN</i> )	1000 <i>mN</i>	900 <i>mN</i>	210 <sup>15</sup> <i>mN</i>	−3000 <i>mN</i>
1 <i>kW</i>	Watt ( <i>W</i> )	1000 <i>W</i>	1500 <i>W</i>	510 <sup>−5</sup> <i>W</i>	−30 <i>W</i>
1 <i>T</i>	millitesla ( <i>mT</i> )	1000 <i>mT</i>	600 <i>mT</i>	10 <sup>23</sup> <i>mT</i>	−90 <i>mT</i>
1 <i>ha</i>	square meter ( $\text{m}^2$ )	10,000 <i>m</i> <sup>2</sup>	10,500 <i>m</i> <sup>2</sup>	310 <sup>−4</sup> <i>m</i> <sup>2</sup>	−25 <i>m</i> <sup>2</sup>
1 <i>lx</i>	lumen per $\text{m}^2$ ( $\text{lm}/\text{m}^2$ )	1 <i>lm</i> / <i>m</i> <sup>2</sup>	0.5 <i>lm</i> / <i>m</i> <sup>2</sup>	1000 <i>lm</i> / <i>m</i> <sup>2</sup>	−19 <i>lm</i> / <i>m</i> <sup>2</sup>
1 <i>ly</i>	Trillion/Billion <i>km</i>	9.461 <i>Tkm</i>	9.461 <i>Bkm</i>	10 <i>m</i>	−2 <i>Tkm</i>
1 <i>B</i>	bit ( <i>b</i> )	8 <i>b</i>	10 <i>b</i>	610 <sup>8</sup> <i>b</i>	−4 <i>b</i>

Table 4.3: Redefinitions of unit scaling between base and derived units.

#### 4.1.2.3 Question Levels

As in the constant’s case, we followed a multi-tiered approach to ensure comprehensive testing by introducing three different levels of question difficulty. Table 4.4 includes all the questions used in the unit of measure redefinition task.

1. **Direct Conversion ( $Q_1$ ):** For the first level of difficulty, the model is instructed to answer a simple question that directly tests its understanding of the conversion rule linking the base unit with its derived counterpart. To exclude overly simple cases, instead of asking "How many seconds are in 1 minute?", which is essentially a rephrased question of "Redefine 1 minute as 100 seconds.", we incorporate a basic calculation, such as "How many seconds are in 2 minutes?".
2. **Applied Use ( $Q_2$ ):** At the next level of question difficulty, we again query a relatively simple calculation, but now in the form of an easy physics problem and given minimal context, to test LLMs’ ability to apply the new information to an undemanding reasoning task (e.g., "A stopwatch runs for 3 and a half min. How many sec does it count?").
3. **Complex Reasoning ( $Q_3$ ):** At the third level, similar to the constants case, questions become more challenging physical problems requiring multiple reasoning steps and complex mathematical calculations. For example, a problem designed to assess model flexibility to adapt to altered knowledge about the scaling between minutes and seconds is the following: "A marathon runner runs at a speed of 170 m/min. How many seconds will it take them to complete a 42-km race?".

#### 4.1.3 Question Formats

In both the constant and unit redefinition cases, we implemented two question formats: free-form generation and multiple choice.

- **Free-Form (FF):** In the Free-Form response format, the LLM model is required to answer the question by generating an open-ended response without any predefined options.
- **Multiple Choice (MC):** For each question, the model is given 4 different possible answers (A, B, C and D) and is prompted to choose the correct one. The options include: the baseline answer without redefinition, the correct answer under redefinition, and two other sufficiently challenging distractors. For example, for the question "What is the first digit of  $\pi$ ?" under the redefinition "Redefine  $\pi$  as 4.5", four options could be: {A: 2 (distractor), B: 3 (baseline answer), C: 4 (correct answer), D: 5 (distractor)}.

	$Q_1$	$Q_2$	$Q_3$
<i>min</i>	How many <i>sec</i> are in 2 <i>min</i> ?	A stopwatch runs for 3 and a half <i>min</i> . How many <i>sec</i> does it count?	A marathon runner runs at a speed of 170 <i>m/min</i> . How many <i>sec</i> will it take them to complete a 42- <i>km</i> race?
<i>kg</i>	How many <i>gr</i> are in 2 <i>kg</i> ?	A person weighs 72 <i>kg</i> . What is the persons weight in <i>gr</i> ?	A vehicle's engine weighs 650 <i>kg</i> . If 15% of the weight is aluminum, what is the weight of the aluminum in <i>gr</i> ?
<i>m</i>	How many <i>cm</i> are in 2 <i>m</i> ?	A circular track has a circumference of 400 <i>m</i> . What is its diameter in <i>cm</i> ?	If a rectangular field is 50 <i>m</i> long and 30 <i>m</i> wide, what is its area in <i>cm</i> <sup>2</sup> ?
<i>K</i>	What is the <i>K</i> temperature when it is 0°C?	Water boils at 100°C. What is its boiling point in <i>K</i> ?	At a certain point in time, the temperature of a black hole's event horizon is measured to be 20°C. If the temperature in °C decreases by 30% after an event, what is the new temperature in <i>K</i> ?
<i>mL</i>	How many <i>mL</i> are in 1 <i>cm</i> <sup>3</sup> ?	If you have a container that holds 1,250 <i>mL</i> of liquid, how many <i>cm</i> <sup>3</sup> of liquid can it hold?	A spherical ball has a radius of 10 <i>cm</i> . What is its volume in <i>mL</i> ?
<i>cal</i>	How many <i>J</i> are in 3 <i>cal</i> ?	A person burns 200 <i>J</i> of energy while jogging. How many <i>cal</i> did they burn?	A car burns 3,400 <i>J</i> of fuel every <i>min</i> . If the car runs for 2 hours, how many <i>cal</i> does it burn?
<i>atm</i>	How many <i>Pa</i> are in 2 <i>atm</i> ?	A diver is 100 <i>m</i> below the surface of the ocean where the pressure is 152,300 <i>Pa</i> . How many <i>atm</i> of pressure are they experiencing?	A pressurized gas tank holds a gas at a pressure of 150,000 <i>Pa</i> . If the gas occupies a volume of 4 <i>m</i> <sup>3</sup> at this pressure, and the gas is suddenly released to 2 <i>atm</i> , what will be the new volume of the gas? Assume temperature and the number of gas molecules remain constant and use Boyle's Law.
<i>V</i>	How many <i>mV</i> are in 5 <i>V</i> ?	A circuit is powered by 30,000 <i>mV</i> . How many <i>V</i> is this?	A battery supplies 100,000 <i>mV</i> to a device. If the device operates with a resistance of 20 ohms, what is the current (in Amperes) flowing through the device using Ohm's Law?
<i>MHz</i>	How many <i>Hz</i> are in 2 <i>MHz</i> ?	An oscillator operates at 4 <i>MHz</i> . What is the period of the wave in <i>sec</i> ?	A circuit has a signal with a frequency of 6 <i>MHz</i> . What is the wavelength of the signal if the speed of light is approximately $3 \times 10^8$ <i>m/s</i> ?
<i>N</i>	How many <i>mN</i> are in 2 <i>N</i> ?	A person applies a force of 24 <i>N</i> to a cart with a mass of 3 <i>kg</i> . What is the force applied to the cart by the person in <i>mN</i> ?	A 10- <i>kg</i> object is pulled with a force of 4,300 <i>mN</i> . What is the acceleration of the object ( <i>m/s</i> <sup>2</sup> )?
<i>kW</i>	How many <i>W</i> are in 2 <i>kW</i> ?	A lightbulb consumes 900 <i>W</i> of power. How many <i>kW</i> is this?	A factory uses 12 <i>kW</i> for 10 hours per day for 30 days. What is the total energy consumption in watt-hours?
<i>T</i>	How many <i>mT</i> are in 3 <i>T</i> ?	A coil generates a magnetic field of 300 <i>mT</i> . What is this field strength in <i>T</i> ?	A particle moves through a magnetic field of 3,600 <i>mT</i> with a charge of $2 \times 10^{-6}$ <i>C</i> and a velocity of $10^5$ <i>m/s</i> . What is the magnetic force on the particle?
<i>ha</i>	What is the area of 2 <i>ha</i> in <i>m</i> <sup>2</sup> ?	A park has an area of 86,000 <i>m</i> <sup>2</sup> . How many <i>ha</i> is the park?	A triangular plot of land has a base of 300 <i>m</i> and a height of 350 <i>m</i> . How many <i>ha</i> is the plot?
<i>lx</i>	How many <i>lx</i> are equivalent to 4 <i>lm/m</i> <sup>2</sup> ?	A workspace is illuminated at a level of 6 <i>lx</i> . What is the illumination in <i>lm/m</i> <sup>2</sup> ?	A light source emits 300 <i>lm</i> uniformly over a circular area with a radius of 10 <i>m</i> . What is the average illumination in <i>lx</i> over this area?
<i>ly</i>	How many <i>km</i> are in 2 <i>ly</i> ?	The Andromeda Galaxy is approximately 23 <i>ly</i> from Earth. What is this distance in <i>km</i> ?	A black hole is 150 <i>ly</i> away. If light travels at a speed of 0.3 billion <i>km/s</i> , how long would it take for light to travel this distance in <i>sec</i> ?
<i>B</i>	How many <i>b</i> are in 3 <i>B</i> ?	If a document is 8,000 <i>b</i> in size, how many <i>B</i> does it occupy?	A 1- <i>min</i> high-definition video uses a data rate of $8 \times 10^6$ <i>B/sec</i> . How many <i>b</i> does the video consume in total?

Table 4.4: Questions of three difficulty levels ( $Q_1$ ,  $Q_2$ ,  $Q_3$ ) for units of measure.



### 4.1.4 Dataset Format and Implementation

Each dataset is formatted as a .csv (comma-separated values) file and, for every selected constant or unit of measure contains the following fields:

- **Selected constant or unit of measure**
- **Original definition (value or scaling)**
- **Generated question**
- **Redefined values or scalings**
- **Baseline answer (without redefinition)**
- **Correct answers under redefinition**
- **Answer options A, B, C, D (for MC format)**

All redefinitions, questions, and distractors for the multiple-choice format were manually crafted with the assistance of ChatGPT<sup>1</sup>. For each of these elements, we prompted ChatGPT to suggest multiple candidate options, from which the most suitable ones were carefully selected and edited as needed to best fit the conceptual criteria of our study.

## 4.2 Metrics and Evaluation

To evaluate model performance under the No Redefinition (NR) and Redefinition (R) tasks, we categorize generated responses into four types:

- **No Redefinition (NR) Correct Responses:** These responses correspond to cases where the LLM correctly answers under the No Redefinition (NR) task, indicating that the model actually possesses the knowledge prior to any alteration of the constant’s true value (or units’ scaling). They establish a baseline for knowledge recall.
- **Anchored Responses:** These are the instances where the LLM, under the Redefinition (R) task, produces the answer that was correct before any redefinition (therefore incorrect after). This means that the model completely disregarded the redefinition instruction, *anchoring* to its previous knowledge. For example, if the question is "What is the first digit of  $\pi$ ?" under the redefinition "Redefine  $\pi$  as 4.5.", the output "3" is considered the Anchored Response, because  $\pi$ ’s well-known pre-redefinition value is "3.14159".
- **Correct Responses:** Cases where the LLM fully understands the redefinition and manages to answer the question accordingly. In the previous example of  $\pi$ ’s redefinition, the Correct Response would be the number "4".
- **Completely Wrong Responses** In these cases, the LLM generates blank, inconsistent, nonsensical, or entirely incorrect responses unrelated to either prior or post-redefinition knowledge and that not fit any of the other three categories. Here are also included instances where the model completely refuses to perform the redefinition, claiming that this task is meaningless, impossible, or even against its guidelines.

---

<sup>1</sup><https://chatgpt.com/>

After mapping the responses to their respective categories, we calculated the corresponding rates, which serve as key metrics to capture the abilities and behavioral tendencies of LLMs as they emerge across different experimental scenarios, model architectures, and sizes. In our analysis, we primarily focus on the **Anchored Responses rates**—or **anchoring rates**—as they are the ones that best reveal the dominance of memorized knowledge over true task-specific reasoning. Within the Completely Wrong category, we also systematically differentiate between blank outputs, incorrect results, and outright refusals to respond, to better understand the nature of model errors under redefined contexts. Particular emphasis is placed on the **refusal rates**, which—compared to the other two response types—offer insight on the models’ *confidence*, or *overconfidence*, in either producing an answer or justifiably abstaining from one.

In addition to the aforementioned evaluation measures, we used **correlation** as a complementary metric to offer further perspective on the relationship between model knowledgeability and behaviors of interest. Correlation is a statistical measure that quantifies the extent to which two variables are linearly related. In other words, it expresses how changes in one variable are associated with changes in another, capturing both the strength and direction of this relationship. Its values can range from -1 to 1. A coefficient close to 1 describes a strong positive correlation, or a direct relationship, where an increase (or decrease) in one series indicates an increase (or decrease) in the other as well. In opposition, a value close to -1 shows a strong negative, or inverse, correlation, which means that when one variable increases the other declines, and vice versa. Values around 0 suggest little or no linear relationship between the two variables. Specifically, we measured correlations between the correct response accuracies of the No Redefinition case and the post-redefinition anchoring or refusal rates.

## 4.3 Prompting Details

Aiming to probe LLMs’ ability to integrate redefined concepts and reason accordingly, we designed a diverse set of prompts utilizing varied techniques under different condition types. The three main prompting categories are: (1) No Redefinition Prompts, (2) Redefinition Prompts, and (3) Evaluation Prompts. Each category includes prompts for both Free-Form (FF) and Multiple Choice (MC) response format, with variations across Zero-Shot (ZS), Chain-of-Thought (CoT), and Few-Shot (FS) prompting strategies.

### 4.3.1 No Redefinition Prompts

Before introducing the Redefinition Task, we designed corresponding No Redefinition (NR) prompts to establish a baseline for how accurately the models can answer questions involving knowledge of scientific constants and units of measurement.

#### Free-Form Format

The prompt templates for the NR task in the Free-Form response format are presented in Table 4.5.

Strategy	Prompt
ZS	Answer the following question: <b>{question}</b>
	End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary.
CoT	Answer the following question: <b>{question}</b>
	Let's think step by step.  End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary.
FS	Answer the following question: <b>{question}</b>
	Here are some examples of similar questions with their correct answers: <b>{NR_FF_EXAMPLES}</b>  End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary.

Table 4.5: NR prompts for FF format across ZS, CoT and FS prompting strategies.

Each of the three templates includes a **question** variable, which is replaced during experimentation with the specific query the model is expected to answer. Meanwhile, the **NR\_FF\_EXAMPLES** field supplies a predefined set of question-answer pairs that serve as demonstrations in Few-Shot (FS) prompts. The following question-answer examples were consistently used across all Free-Form trials in the No Redefinition task experiments, for constants and units of measurement respectively:

- **NR\_FF\_EXAMPLES for Constant NR task:**

1. Question: What is  $10 \cdot \tau$ ?  
Answer: 62.83
2. Question: What is one percent of 50?  
Answer: 0.5
3. Question: What is the first digit of vacuum magnetic permeability?  
Answer: 1

- **NR\_FF\_EXAMPLES for Unit NR task:**

1. Question: How many hours are in 2 days?  
Answer: 48
2. Question: How many radians are in a full circle?  
Answer:  $2 \cdot \pi$
3. Question: A lightbulb has a resistance of 50 ohms. What is its resistance in milliohms?  
Answer: 50,000

To eliminate unnecessary text, additional explanations, and commentary, we explicitly instructed the model to end each response by saying the phrase "**The final answer is:** " and then provide the correct answer concisely. This strict output format ensures consistency among different models' responses and facilitates the

next phase of response handling and evaluation. This approach is systematically employed in all NR and R prompt templates.

### Multiple Choice Format

The No Redefinition prompt templates for the Multiple Choice (MC) response format are structured similarly to those for the Free-Form (FF) format, with necessary adjustments to guide the model in selecting an answer from the provided options. Table 4.6 demonstrates the NR prompt templates for the MC format.

Strategy	Prompt
ZS	Choose A, B, C or D to answer the question:
	Question: {question} A: {A} B: {B} C: {C} D: {D}
CoT	Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D". End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary.
	Choose A, B, C or D to answer the question:
	Question: {question} A: {A} B: {B} C: {C} D: {D}
FS	Let's think step by step.
	Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D". End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary.
	Choose A, B, C or D to answer the question:
	Question: {question} A: {A} B: {B} C: {C} D: {D}
FS	Here are some examples of similar questions with their correct answers:
	{NR_MC_EXAMPLES}
	Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D". End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary.

Table 4.6: NR prompts for MC format across ZS, CoT and FS prompting strategies.

Here, in addition to the **question** variable-identical to the one used in the FF case-we introduce the variables **A**, **B**, **C** and **D**, which represent the four distinct options from which the model is instructed to select the

correct one. Moreover, the **NR\_MC\_EXAMPLES** set includes the same question-answer examples as before, but reformatted for the Multiple Choice setting:

- **NR\_MC\_EXAMPLES for Constant NR task:**

1. Question: What is  $10 \cdot \tau$ ?  
A: 3.14  
B: 62.83  
C: 90  
D: 9  
Answer: B
2. Question: What is one percent of 50?  
A: 0.5  
B: 5  
C: 10  
D: 50  
Answer: A
3. Question: What is the first digit of vacuum magnetic permeability?  
A: 1  
B: 2  
C: 3  
D: 4  
Answer: A

- **NR\_FF\_EXAMPLES for Unit NR task:**

1. Question: How many hours are in 2 days?  
A: 10  
B: 20  
C: 24  
D: 48  
Answer: D
2. Question: How many radians are in a full circle?  
A: pi  
B:  $2 \cdot \pi$   
C: 180  
D: 360  
Answer: B
3. Question: A lightbulb has a resistance of 50 ohms. What is its resistance in milliohms?  
A: 5,000 milliohms  
B: 10,000 milliohms  
C: 50,000 milliohms  
D: 100,000 milliohms  
Answer: C

### 4.3.2 Redefinition Prompts

We constructed Redefinition Prompts to investigate how LLMs respond to redefined knowledge, using a structure that mirrors the No Redefinition templates but adding the instruction "Redefine {X} as {Y}", placed at the beginning of each prompt.

#### Free-Form Format

The prompt templates for the Redefinition task regarding the Free-Form (FF) response format are included in Table 4.7.

Strategy	Prompt
ZS	Redefine <b>{X}</b> as <b>{Y}</b> . <b>{question}</b>
	End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary.
CoT	Redefine <b>{X}</b> as <b>{Y}</b> . <b>{question}</b>
	Let's think step by step.  End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary.
FS	Redefine <b>{X}</b> as <b>{Y}</b> . <b>{question}</b>
	Here are some examples of similar questions with their correct answers:  <b>{R_FF_EXAMPLES}</b>  End the response with the phrase "The final answer is: " followed only by the correct result, with no additional text or commentary.

Table 4.7: R prompts for FF format across ZS, CoT and FS prompting strategies.

In the Redefinition prompting templates, we incorporate two additional variables: 1) **X**: which corresponds to the original entity being redefined and 2) **Y**: which represents the newly assigned value for entity X. For example, for  $X=\pi$  and  $Y=500$ , the instruction becomes "Redefine  $\pi$  as 500".

The **R\_FF\_EXAMPLES** question-answer pairs consist of the same questions used in the NR case, but with answers that reflect the corresponding redefinitions.

- **R\_FF\_EXAMPLES for Constant Redefinition task:**

1. Question: Redefine tau as 9. What is  $10 \cdot \tau$ ?  
Answer: 90
2. Question: Redefine one as 10. What is one percent of 50?  
Answer: 5
3. Question: Redefine vacuum magnetic permeability as  $4.56 \times 10^{-7} \text{ N/A}^2$ . What is the first digit of vacuum magnetic permeability?  
Answer: 4

- **R\_FF\_EXAMPLES for Unit Redefinition task:**

1. Question: Redefine 1 day as 10 hours. How many hours are in 2 days?  
Answer: 20
2. Question: Redefine 1 degree as 0.5 radians. How many radians are in a full circle?  
Answer: 180
3. Question: Redefine 1 ohm as 200 milliohms. A lightbulb has a resistance of 50 ohms. What is its resistance in milliohms?  
Answer: 10,000

### Multiple Choice Format

The Redefinition task prompt templates for the the Multiple Choice (MC) response format are shown in Table 4.8.

Mode	Prompt
ZS	Redefine $\{X\}$ as $\{Y\}$ . Choose A, B, C or D to answer the question:
	Question: $\{question\}$ A: $\{A\}$ B: $\{B\}$ C: $\{C\}$ D: $\{D\}$
	Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D". End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary.
CoT	Redefine $\{X\}$ as $\{Y\}$ . Choose A, B, C or D to answer the question:
	Question: $\{question\}$ A: $\{A\}$ B: $\{B\}$ C: $\{C\}$ D: $\{D\}$
	Let's think step by step.
	Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D". End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary.
FS	Redefine $\{X\}$ as $\{Y\}$ . Choose A, B, C or D to answer the question:
	Question: $\{question\}$ A: $\{A\}$ B: $\{B\}$ C: $\{C\}$ D: $\{D\}$
	Here are some examples of similar questions with their correct answers:
	$\{R\_MC\_EXAMPLES\}$
	Provide only the letter corresponding to the correct answer: "A", "B", "C", or "D". End the response with the phrase "The final answer is: " followed by the correct letter, with no additional text or commentary.

Table 4.8: R prompts for MC format across ZS, CoT and FS prompting strategies.

Finally, the examples used for the redefinition of constants and units of measurement under Few-Shot prompting are appropriately formatted for multiple-choice question answering in the **R\_MC\_EXAMPLES** set.

- **R\_MC\_EXAMPLES for Constant Redefinition task:**

1. Question: Redefine tau as 9. What is  $10 \cdot \tau$ ?  
A: 3.14  
B: 62.83  
C: 90  
D: 9

Answer: C

2. Question: Redefine one as 10. What is one percent of 50?

A: 0.5

B: 5

C: 10

D: 50

Answer: B

3. Question: Redefine vacuum magnetic permeability as  $4.56 \times 10^{-7} \text{ N/A}^2$ . What is the first digit of vacuum magnetic permeability?

A: 1

B: 2

C: 3

D: 4

Answer: D

• **R\_FF\_EXAMPLES for Unit Redefinition task:**

1. Question: Redefine 1 day as 10 hours. How many hours are in 2 days?

A: 10

B: 20

C: 24

D: 48

Answer: B

2. Question: Redefine 1 degree as 0.5 radians. How many radians are in a full circle?

A: pi

B: 2\*pi

C: 180

D: 360

Answer: C

3. Question: Redefine 1 ohm as 200 milliohms. A lightbulb has a resistance of 50 ohms. What is its resistance in milliohms?

A: 5,000 milliohms

B: 10,000 milliohms

C: 50,000 milliohms

D: 100,000 milliohms

Answer: B

### 4.3.3 Evaluation Prompts

To appraise the initial LLM model's performance on both the No Redefinition and Redefinition tasks, we leverage a separate LLM as an evaluator. We specifically designed prompts to categorize model outputs into Correct Responses, Anchored Responses, and Completely Wrong Responses, enabling us to compute the corresponding percentages. Within the Completely Wrong Responses category, we further classify the model errors as incorrect results, blank answers, or refusals to answer.

#### No Redefinition Evaluation

For the No Redefinition part, we instruct the evaluator model to compare the model's response (**llm\_answer**) to the correct answer (**real\_answer**) using the following prompt.

---

You are tasked with comparing two answers: one provided by an LLM (the "LLM answer") and the correct answer (the "real answer"). Your job is to determine if the LLM answer matches the real answer.

---



The comparison should strictly focus on whether the LLM final answer conveys the same meaning or provide the same information as the correct answer. Minor differences in phrasing, wording, or structure are acceptable as long as the core meaning remains identical. For numerical results, differences due to rounding are acceptable as long as the values are reasonably close and within an acceptable margin of error.

**\*\*Instructions:\*\***

1. Compare the LLM answer to the real answer carefully.
2. If the LLM answer matches the real answer, output:  
correct
3. If the LLM answer does not match the real answer, output:  
incorrect

Do not output explanations, justifications, or any additional text — only the word "correct" or "incorrect".

**\*\*LLM Answer:\*\***

**{llm\_answer}**

**\*\*Real Answer:\*\***

**{real\_answer}**

**\*\*Output:\*\***

---

### Redefinition Evaluation

Here, the model is prompted to compare the generated response (**llm\_answer**) to the correct answer both before (**reference\_answer\_1**) and after (**reference\_answer\_2**) the redefinition, in order to distinguish between Correct, Anchored, and Completely Wrong.

---

You are tasked with comparing an answer provided by an LLM (the "LLM answer") to two reference answers: "Reference Answer 1" and "Reference Answer 2". Your job is to determine if the LLM answer matches either of the two reference answers.

The comparison should strictly focus on whether the LLM final answer conveys the same meaning or provides the same information as one of the reference answers. Minor differences in phrasing, wording, or structure are acceptable as long as the core meaning remains identical. For numerical results, differences due to rounding are acceptable as long as the values are reasonably close and within an acceptable margin of error.

**\*\*Instructions:\*\***

1. Compare the LLM answer carefully with "Reference Answer 1" and "Reference Answer 2".
2. If the LLM answer matches "Reference Answer 1", output:  
first
3. If the LLM answer matches "Reference Answer 2", output:  
second
4. If the LLM answer matches neither of the two, output:  
none

Do not output explanations, justifications, or any additional text — only the words "first", "second", or "none".

---

```
**LLM Answer:**  
{llm_answer}
```

```
**Reference Answer 1:**  
{reference_answer_1}
```

```
**Reference Answer 2:**  
{reference_answer_2}
```

```
**Output:**
```

---

### Multiple Choice Evaluation

For the Multiple Choice case, where the answers are represented by option letters (A, B, C, or D), we adjust the comparison prompt accordingly. Instead of matching the full textual answer to known correct answers, the evaluator is asked to determine whether the letter selected by the LLM matches the labeled option corresponding to either the correct post-redefinition answer or the original answer.

---

You are tasked with comparing an answer provided by an LLM (the "LLM answer") to two reference answers: "Reference Answer 1" and "Reference Answer 2". Your job is to determine if the LLM answer matches the letter of either of the two reference answers (A, B, C, or D).

```
**Instructions:**
```

1. Compare the LLM answer carefully with "Reference Answer 1" and "Reference Answer 2".
2. If the LLM answer matches "Reference Answer 1", output:  
first
3. If the LLM answer matches "Reference Answer 2", output:  
second
4. If the LLM answer matches neither of the two, output:  
none

Do not output explanations, justifications, or any additional text — only the words "first", "second", or "none".

```
**LLM Answer:**  
{llm_answer}
```

```
**Reference Answer 1:**  
{reference_answer_1}
```

```
**Reference Answer 2:**  
{reference_answer_2}
```

```
**Output:**
```

---

### Wrong Responses Analysis

We further analyze the Completely Wrong Responses to investigate the underlying causes of errors. Specifically, we distinguish between three subtypes: Wrong Result, Refusal to Answer, and Blank Answer.

---

You are tasked with analyzing an LLM answer that does not match either of two reference answers: "Reference Answer 1" and "Reference Answer 2". Your job is to classify the LLM answer into one of the following categories:

1. **Wrong Answer**: The LLM provided an incorrect response to the question, either factually or logically.
2. **Blank Answer**: The LLM provided no substantive response, leaving the answer blank or completely empty.
3. **Refusal to Answer**: The LLM explicitly refused to answer the question, citing reasons such as the question being nonsensical, impossible to answer, or against its guidelines.

**Instructions:**

1. Analyze the LLM answer and determine which of the three categories it belongs to.
2. If the LLM answer is a **Wrong Answer**, output:  
wrong
3. If the LLM answer is a **Blank Answer**, output:  
blank
4. If the LLM answer is a **Refusal to Answer**, output:  
refusal
5. If the classification is unclear, choose the category that best fits the content of the LLM answer.

Do not output explanations, justifications, or any additional text — only the words "wrong", "blank", or "refusal".

**LLM Answer:**

{llm\_answer}

**Reference Answer 1:**

{reference\_answer\_1}

**Reference Answer 2:**

{reference\_answer\_2}

**Output:**

---

## 4.4 LLM selection

In this study, we conducted a comparative assessment of 19 Large Language Models (LLMs) drawn from various state-of-the-art model families, chosen to represent a diverse set of parameter sizes and architectural designs. This selection was intended to allow for a thorough investigation of how current language models respond to scenarios involving redefinitions of foundational entities, with particular interest in the relationship

---

between model scale—specifically the number of parameters—and reasoning abilities, flexibility, confidence, and memorization.

The following language models were used in our experiments:

- **Llama (8B, 70B, 405B)**
- **Mistral (7B, Mixtral 8x7B, Large)**
- **Claude (Instant v1, v2, 3 Opus, 3 Haiku, 3.5 Sonnet, 3.7 Sonnet)**
- **Command (Light Text, Text, R, R+)**
- **Titan (Text Lite, Express, Tg1)**

For the evaluation process, we employed Claude 3.5 Sonnet as the LLM evaluator.

## 4.5 Experimental setup

Experiments for the No Redefinition and Redefinition tasks, as well as the evaluation of LLM responses, were conducted within Kaggle Notebooks<sup>2</sup>. This environment ensured a structured and reproducible workflow for experimentation. To enhance computational efficiency, we utilized NVIDIA T4 Tensor Core GPUs (T4x2 configuration), available within the Kaggle infrastructure, which was essential for significantly reducing runtime and enabling the handling of large-scale LLMs.

All models listed in the previous section were accessed via AWS Bedrock<sup>3</sup>, which is a cloud-based service from Amazon Web Services (AWS) that enables the deployment of various foundation models (FMs) from multiple providers. Access was established and authenticated through AWS Identity and Access Management (IAM), ensuring secure API interactions.

---

<sup>2</sup><https://www.kaggle.com/>

<sup>3</sup><https://aws.amazon.com/bedrock/>

# Chapter 5

## Experimental Results

### 5.1 Results on constants redefinition

In this section, we present the findings of the experiments conducted for the constant redefinition task, in which Large Language Models were challenged to override widely known predefined values of specific scientific constants and adapt their reasoning processes accordingly.

#### 5.1.1 Anchoring to default values

As outlined in Section 4.2, we specifically measured the Anchored Responses Rate, which directly captures the extent to which models tend to rely on predefined knowledge, even when they are explicitly instructed to disregard it. Table 5.1 displays these results for all tested LLMs, evaluated under the most difficult cases of the two redefinition scenarios (assignment and swapping), as well as across the three levels of question difficulty and the two response formats (Free-Form and Multiple Choice). These findings reflect performance under the Zero-Shot prompting strategy only.

Model	$R_a3$						$R_s2$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	<b>33.33</b>	<b>46.67</b>	<b>33.33</b>	<b>26.67</b>	26.67	40.0	33.33	53.33	13.33	33.33	26.67	20.0
Mixtral8x7B	33.33	33.33	26.67	26.67	20.0	33.33	26.67	46.67	40.0	<b>53.33</b>	46.67	<b>73.33</b>
Mistral Large (123B)	33.33	20.0	26.67	26.67	<b>53.33</b>	<b>66.67</b>	<b>66.67</b>	<b>53.33</b>	<b>46.67</b>	40.0	<b>73.33</b>	66.67
Llama8B	0.0	<b>26.67</b>	0.0	<b>26.67</b>	13.33	33.33	20.0	13.33	<b>26.67</b>	40.0	20.0	20.0
Llama70B	<b>6.67</b>	13.33	0.0	0.0	13.33	40.0	33.33	46.67	13.33	<b>46.67</b>	33.33	73.33
Llama405B	0.0	0.0	0.0	13.33	<b>26.67</b>	<b>53.33</b>	<b>26.67</b>	<b>46.67</b>	6.67	20.0	<b>53.33</b>	<b>93.33</b>
Titan lite	13.33	20.0	20.0	20.0	0.0	40.0	40.0	33.33	20.0	33.33	6.67	26.67
Titan express	20.0	<b>26.67</b>	13.33	13.33	<b>20.0</b>	13.33	40.0	<b>53.33</b>	<b>20.0</b>	20.0	33.33	<b>26.67</b>
Titan large	<b>26.67</b>	20.0	<b>20.0</b>	6.67	13.33	<b>40.0</b>	<b>60.0</b>	40.0	13.33	<b>33.33</b>	<b>33.33</b>	20.0
Command r	0.0	6.67	<b>20.0</b>	<b>33.33</b>	<b>26.67</b>	<b>53.33</b>	<b>53.33</b>	13.33	20.0	6.67	<b>33.33</b>	<b>46.67</b>
Command r +	6.67	13.33	0.0	13.33	13.33	26.67	13.33	20.0	26.67	6.67	33.33	26.67
Command light text	6.67	13.33	13.33	20.0	0.0	40.0	13.33	20.0	<b>26.67</b>	<b>20.0</b>	13.33	13.33
Command text	<b>13.33</b>	<b>20.0</b>	6.67	6.67	6.67	26.67	40.0	<b>26.67</b>	13.33	26.67	13.33	33.33
Claude opus	13.33	0.0	6.67	6.67	33.33	<b>46.67</b>	<b>46.67</b>	<b>40.0</b>	20.0	<b>26.67</b>	53.33	73.33
Claude instant	0.0	13.33	13.33	<b>20.0</b>	26.67	46.67	33.33	20.0	33.33	40.0	46.67	60.0
Claude haiku	20.0	13.33	6.67	0.0	20.0	20.0	26.67	6.67	20.0	20.0	40.0	53.33
Claude v2	26.67	13.33	<b>20.0</b>	0.0	<b>46.67</b>	40.0	13.33	40.0	<b>33.33</b>	20.0	40.0	66.67
Claude 3.5 Sonnet	<b>26.67</b>	<b>13.33</b>	0.0	13.33	13.33	33.33	33.33	40.0	20.0	20.0	<b>60.0</b>	<b>73.33</b>
Claude 3.7 Sonnet <sup>1</sup>	0.0	0.0	0.0	6.67	13.33	13.33	33.33	20.0	6.67	20.0	40.0	33.33

Table 5.1: Anchoring response rate for all LLMs tested using ZS prompting for the most difficult cases in *assignment* ( $R_a3$ ) and *swapping* ( $R_s2$ ) redefinitions. The highest anchoring rate for each LLM family is marked in **bold**.

It is notable that across all different conditions, models are susceptible to anchoring to some degree. Even in the Free-Form format, where the absence of options for possible answers completely removes any external cues or biases—as well as the random choice factor—LLMs from all model families score significantly high anchoring rates. For example, Mistral Large produces a 73.33% rate, Titan Large and Claude 3.5 Sonnet both generate 60% anchored responses in some cases, while Llama 405B and Command r also exceed the 50% threshold with a score of 53.33%. In the Multiple Choice case, even higher anchoring rates are observed. More specifically, Llama 405B hits the extremely high score of 93.33%, while several other models produce large anchoring rates, such as 73.33% or 66.67%. These findings suggest that anchoring behavior is a robust phenomenon, not limited to specific model architectures or sizes.

To further investigate the underlying dynamics of the anchoring phenomenon, we calculate the correlation between the No Redefinition (NR) accuracy and the post-redefinition Anchored Responses rate. These results, averaged across all LLMs, are presented in Table 5.2. In this setting, a high negative correlation indicates that models that respond correctly when asked about constants without any redefinition of their values are less likely to adhere to default knowledge when redefinitions are introduced. On the contrary, a high positive correlation means the exact opposite: that models with higher knowledgeability are more prone to fail to override the predefined values during the Redefinition Task.

Level	$R_{a1}$	$R_{a2}$	$R_{a3}$	$R_{s1}$	$R_{s2}$
Free-Form (FF)					
$Q_1$	-0.458	-0.071	0.008	0.199	-0.016
$Q_2$	-0.502	-0.573	-0.472	0.107	0.019
$Q_3$	0.489	0.237	0.292	0.666	0.668
Multiple Choice (MC)					
$Q_1$	-0.642	-0.4	-0.344	-0.052	0.025
$Q_2$	-0.275	-0.316	-0.245	0.41	0.151
$Q_3$	-0.063	0.457	0.081	0.666	0.75

Table 5.2: Correlation between average NR correct response rate with anchored response rate for each redefinition and question level in ZS setup. Cells in **pink** indicate a **high positive correlation** ( $> 0.3$ ), while cells in **green** indicate a **high negative correlation** ( $< -0.3$ ).

The correlation results expose an interesting pattern: simpler question levels are associated with negative or relatively weaker correlation scores, which, as explained before, means that when models are capable of dealing with easy or medium reasoning problems in the No Redefinition setting, they are also more likely to handle correctly the user-defined reassignments, an encouraging indication that they are interpreting the redefinition prompts appropriately. Interestingly, this relationship shifts markedly in the most difficult question level, and especially in the swapping cases, where the highest positive correlations are observed. In other words, models performing well under more challenging reasoning tasks in the original No Redefinition setting tend to ignore redefinitions and thus fail on the corresponding tasks. This leads to a surprising conclusion: LLMs that demonstrate stronger reasoning capabilities based on established world knowledge appear more vulnerable to anchoring when challenged with highly counterintuitive tasks.

### 5.1.2 Inverse trends

While testing various models across different families, we observed a particularly compelling pattern: the inability of LLMs to override default scientific values, apart from their reasoning strength, is greatly influenced by the parameter size itself. Although larger models achieve higher accuracy scores on standard reasoning tasks (such as the No Redefinition tasks), they also seem to demonstrate significantly greater anchoring behavior across the Redefinition setup, meaning that they struggle to override deeply embedded factual knowledge. Table 5.3 presents the correct response rates on the pre-redefinition questions alongside the corresponding anchoring rates under the Free Form format and Zero-Shot prompting technique, for models

<sup>1</sup>Without thinking module enabled for fair comparison.

of varying sizes within the Mistral and Llama family. These results clearly illustrate this counterintuitive trend, as, in several cases, especially in the most demanding scenarios, anchoring behavior increases with LLM scale. For instance, Llama 70B produced anchored responses for 33.33% of the  $Q_3$  questions in the hardest swapping scenario, while the much larger Llama 405B yielded a significantly higher score higher of 53.33%. A similar pattern is even more pronounced within the Mistral/Mixtral family, where Mixtral 8x7B generated a 46.67% anchoring rate and Mistral Large (123B) reached a percentage of 73.33%, which is not only extremely high on its own, but also represents a 57.1% relative increase over Mixtral 8x7B. Interestingly, this score even surpasses the corresponding response accuracy in the No Redefinition task, which means that the model produced the default answer (the one that results from the canonical value of the constant) more often across the swapping scenario—where it is incorrect—than in the case where it is actually valid.

Model	$R_{a3}$						$R_{s2}$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF
Mistral7B	66.67	33.33	46.67	33.33	33.33	26.67	66.67	33.33	46.67	13.33	33.33	26.67
Mixtral8x7B	100.0	33.33	66.67	26.67	66.67	20.0	100.0	26.67	66.67	40.0	66.67	46.67
Mistral Large (123B)	93.33	33.33	73.33	26.67	53.33	53.33	93.33	66.67	73.33	46.67	53.33	73.33
Llama8B	80.0	0.0	80.0	0.0	53.33	13.33	80.0	20.0	80.0	26.67	53.33	20.0
Llama70B	93.33	6.67	80.0	0.0	80.0	13.33	93.33	33.33	80.0	13.33	80.0	33.33
Llama405B	93.33	0.0	86.67	0.0	73.33	26.67	93.33	26.67	86.67	6.67	73.33	53.33

Table 5.3: Correct response rate without redefinition (NR) versus post-redefinition anchoring rate in the free-form (FF) format, for LLMs with known sizes using ZS prompting. Colored cells indicate elevated anchoring with LLM scale.

This inverse phenomenon is very clearly illustrated Figures 5.1.1 and 5.1.2, which visualize how the anchored responses rate alters as LLM size increases for models within the Llama and Mistral family, respectively. Both figures correspond to the Multiple Choice response format, focusing on the  $Q_3$ -level questions regarding the most difficult cases of the assignment and swapping redefinition scenarios. Each line in the plots represents a different prompting strategy used for experiments—Zero-Shot, Few-Shot, and Chain-of-Thought—distinguished by color. With only two exceptions, the resulting lines reveal an obvious upward trend, culminating in strikingly high anchoring percentages, especially for the swapping redefinition type. These visualizations provide strong empirical support for the finding that larger LLMs are more prone to exhibit elevated levels of anchoring behavior, pointing to a striking case of inverse scaling.

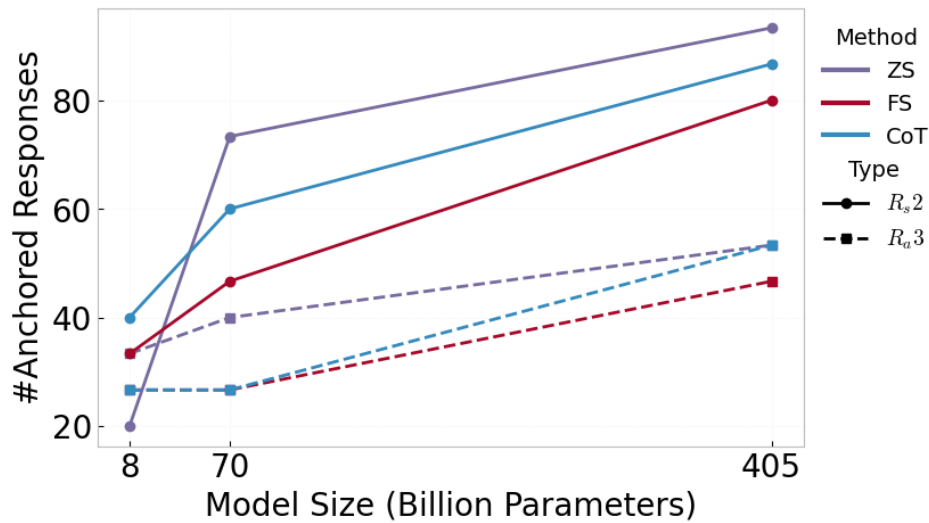


Figure 5.1.1: Number of anchored responses for models of varying sizes in the Llama family (MC response format).

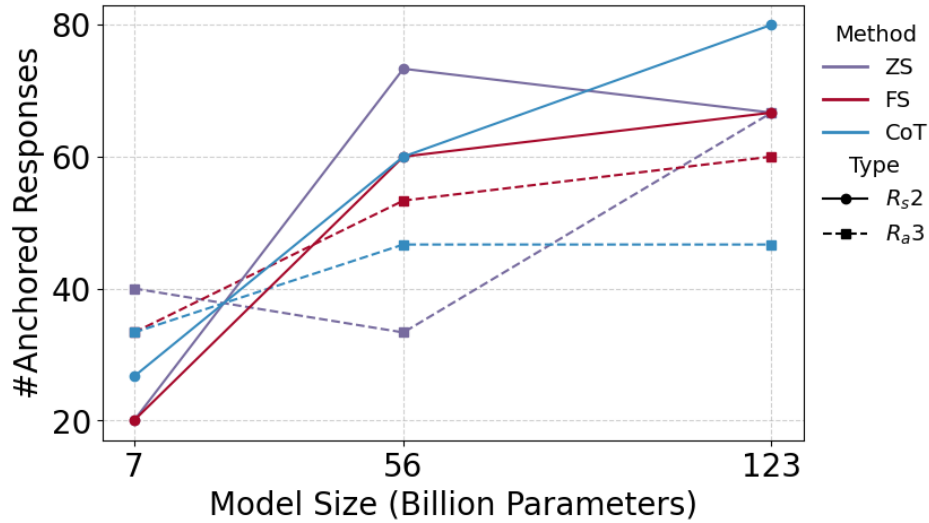


Figure 5.1.2: Number of anchored responses for models of varying sizes in the Mistral family (MC response format).

To delve deeper into this paradoxical behavior, we present additional evidence in Figure 5.1.3, which focuses on the Mistral family of LLMs and displays the distribution of all different types of generated responses—No Redefinition Correct Responses, Anchored, Correct under Redefinition, and Completely Wrong—across all redefinition scenarios, question difficulty levels, and prompting methods, using the Multiple Choice response format. Once again, the anchoring phenomenon is unmistakably evident and seems to intensify not only with task complexity but also with model scaling. Another intriguing discovery, also noted in Table 5.3, is that in several cases the Anchored Responses rate exceeds—often by a large margin—accuracy score in the No Redefinition setting. Particularly for the largest model at the  $Q_3$  level, it consistently chooses the correct-in-the-real-world option more frequently under the Redefinition setting, in which that answer is no longer correct, than in the No Redefinition setting where it is. In addition, in many of the most demanding redefinition reasoning problems, the models completely fail to identify the correct-under-the-redefinition answer, yielding extremely weak or even equal to zero Correct Responses rates.

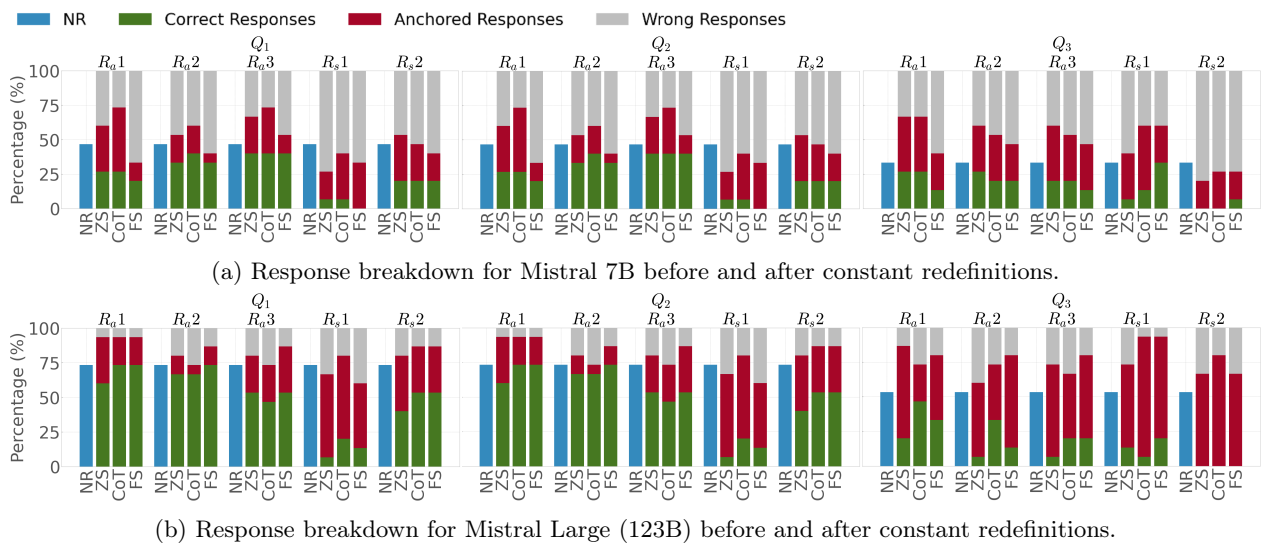


Figure 5.1.3: Comparison of Mistral 7B and Mistral Large responses on the MC response format.



A similar pattern is observed in the Llama models, as demonstrated in Figure 5.1.4, which compares the different response rates for Llama 8B and Llama 405B. Notably, the number of anchored responses is once again significantly higher in the larger model, especially under the most difficult conditions.

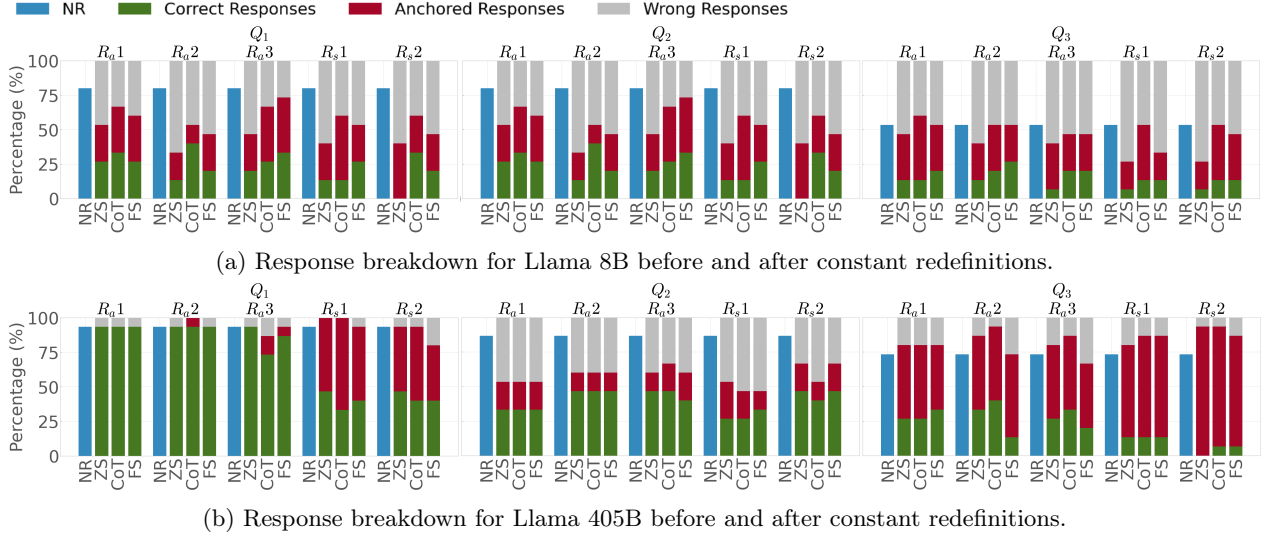


Figure 5.1.4: Comparison of Mistral 7B and Mistral Large responses on the MC response format.

All these findings complicate the logical assumption that models with larger parameter scales are universally more capable to handle all types of tasks. Even when we explicitly prompt them to adapt to new values, larger LLMs, despite their generally stronger reasoning abilities, are more likely to prioritize internalized knowledge and disregard external counterfactual instructions.

### 5.1.3 Response format

The choice between Free-Form and Multiple Choice formats for generated responses seems to deeply affect model behavior. This comparison is demonstrated in Figure 5.1.5, which shows the percentages of each model response type in the assignment and swapping redefinition scenarios of the  $Q_3$  questions, across the Mistral and Llama model families. It is clear that the Multiple Choice format systematically leads to higher anchoring rates in both cases. For instance, in the Free-Form setup Llama 70B and 405B generate anchored response percentages of 33.33% and 53.33% respectively in the second swapping redefinition type, while, under the same conditions, Multiple Choice climbs to 73.33% and 93.33%. The same pattern holds for Mistral models, where Mixtral 8x7B anchors at 26.67% and 46.67% when asked to answer in a free-form way, but at 66.67% and 73.33% when given possible options to choose from in the easier and harder swapping cases.

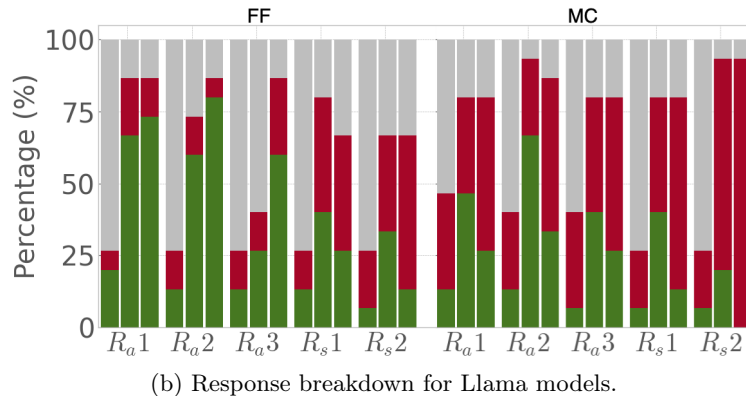
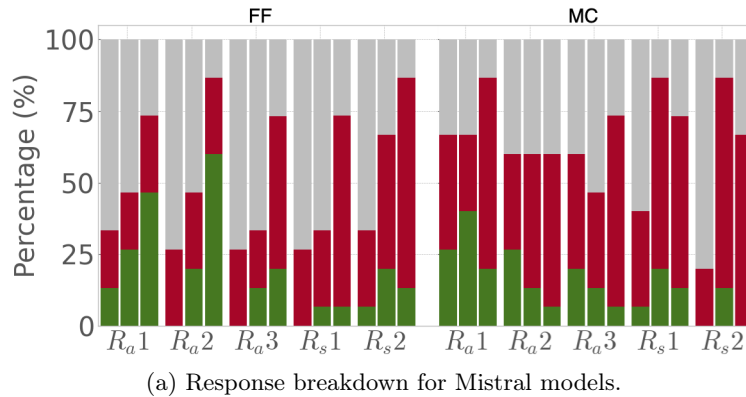


Figure 5.1.5: Results for the different Mistral and Llama models on  $Q_3$  questions using ZS prompting. The order of the bars per redefinition type/level corresponds to increasing model size. The color coding is the same as in Figure 5.1.3.

This disparity is not particularly surprising and can be attributed to the fundamental nature of each format. In Free-Form question-answering, models are challenged to independently generate responses, without any external cues or reinforcement. As a result, LLMs seem to rely more on the instructions given in the redefinition prompt, which makes them more likely to reason accordingly rather than fall back on memorized facts. On the other hand, Multiple Choice response format introduces pre-existing options that effectively serve as cognitive traps. Among these, the default, pre-redefinition correct result evidently becomes the most powerful distractor, as it is inherently associated with high correctness probability, built during model pretraining. In other words, since LLMs are optimized to select high-probability token sequences, the Multiple Choice setup amplifies their tendency to favor these familiar, statistically "safe" options when they "see" them.

#### 5.1.4 Assignment vs Swapping

In addition to the generated response format, a clear discrepancy emerges between the two redefinition types: Assignment ( $R_a$ ) and Swapping ( $R_s$ ). Experimental results indicate that swapping scenarios produce significantly higher anchoring rates compared to assignment ones. Figure 5.1.6 provides a response breakdown for all Llama 70B responses on  $Q_1$ -level questions, where it is rather obvious that swapping values triggers increased anchoring behavior across both response formats. More specifically, while this model maintains strong performance on the simple assignment tasks when tested on the easier questions about constant values—even achieving close or equal to zero anchored response rates (the highest is 13.33%)—its anchoring

behavior escalates drastically in the swapping case, reaching a 46.67% percentage. A similar pattern is observed with the Claude 3.5 Sonnet model across the most difficult question level, as shown in Figure 5.1.7. In this case, correct response accuracy seemingly drops and average anchoring rates go up from 15.93% to 51.11%, when shifting from simple assignment to value swapping.

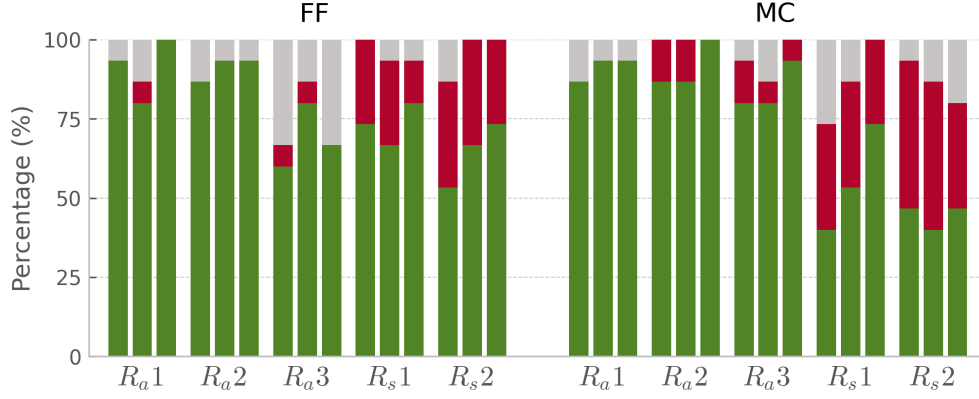


Figure 5.1.6: Response breakdown for Llama 70B on  $Q_1$ -level questions across all prompting strategies. Within each redefinition type/level, the bars are ordered as follows: Zero-Shot, Chain-of-Thought, and Few-Shot.

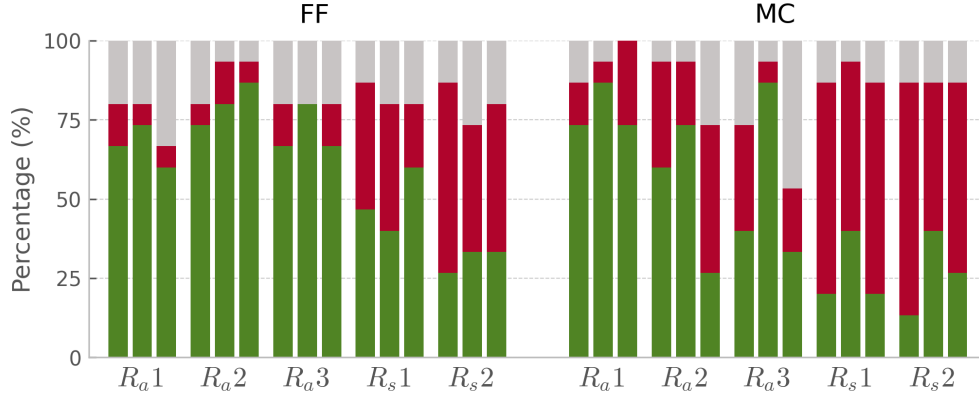


Figure 5.1.7: Response breakdown for Claude 3.5 Sonnet on  $Q_3$ -level questions across all prompting strategies. Within each redefinition type/level, the bars are ordered as follows: Zero-Shot, Chain-of-Thought, and Few-Shot.

We hypothesize that this phenomenon arises from the way memory associations are activated during redefinitions. When a straightforward assignment is expected (e.g. "Redefine  $\pi$  as 500"), a single, widely known entity is simply being replaced by a random number that does not trigger the model's prior memory mappings. This allows the model to focus on adjusting to this altered, entirely new concept of each constant. In contrast, in the value swapping scenario (e.g. "Redefine  $\pi$  as  $\phi$ "), the introduction of the second familiar constant confuses the model by causing multiple strong memory association activations simultaneously. This increases the cognitive work load of the model, which must initially retrieve the default meanings of both entities and then correctly override their relationship. As observed, the model eventually succumbs to its pretraining biases, completely ignores the instructed swapping redefinition, and simply outputs the answer based on the original value of the queried constant.

### 5.1.5 Extended Thinking Blocks

Anthropic’s Claude 3.7 Sonnet offers an additional *extended thinking* mode, which directs the model to analyze problems in greater detail by generating thinking content *blocks* that capture its internal reasoning processes. We enabled this mode and repeated the same experiments, comparing its performance to the standard mode in order to examine whether this advanced reasoning mechanism can help Claude 3.7 Sonnet correctly handle the prompted redefinitions, particularly in the most challenging scenarios where its performance in standard mode rapidly declines.

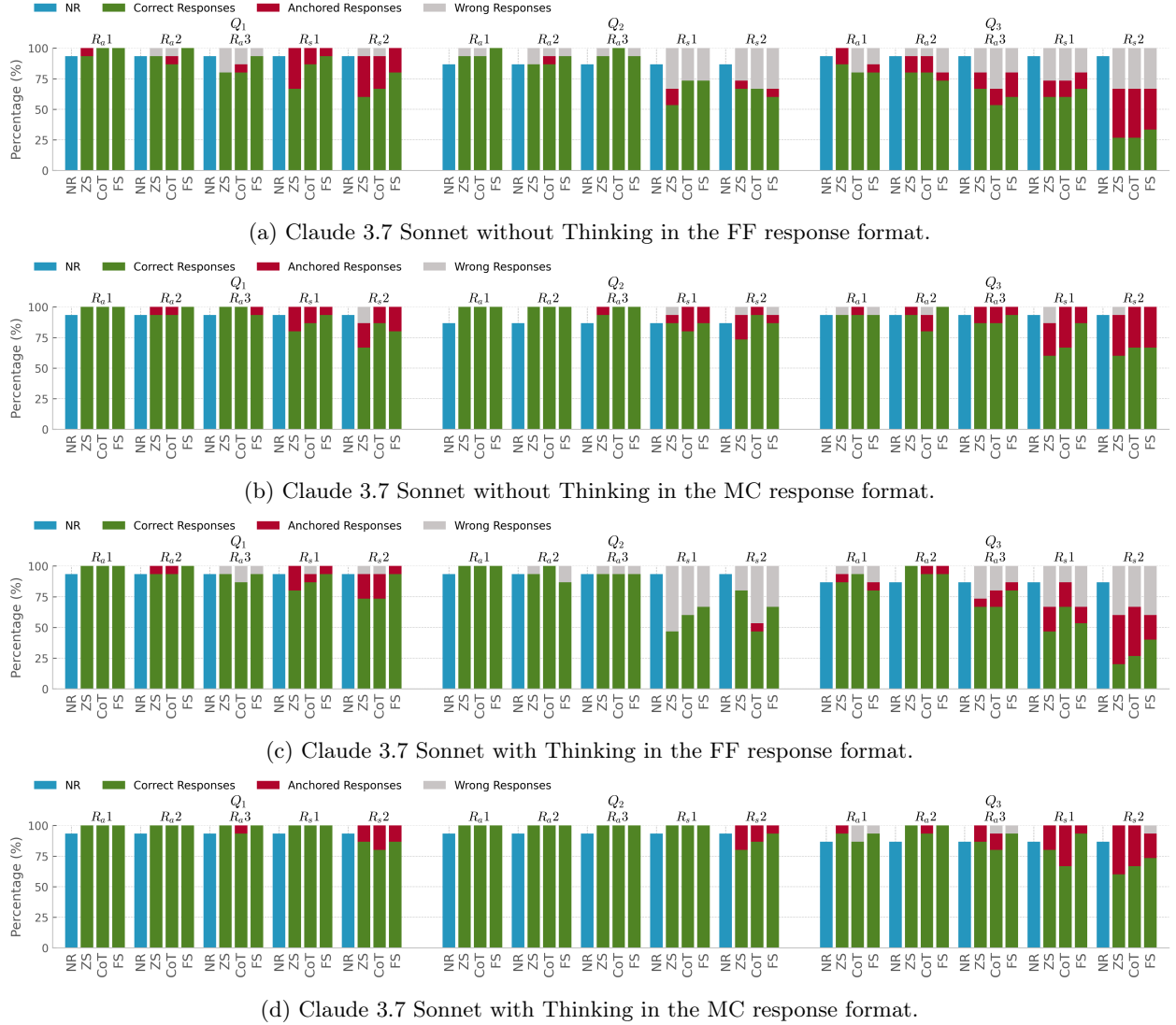


Figure 5.1.8: Claude 3.7 Sonnet results without and with Thinking.

Figure 5.1.8 summarizes the results of Claude 3.7 Sonnet for both standard and extended thinking modes. As observed, while thinking slightly reduces the anchored response rates in a few cases, its overall impact is actually insignificant. This suggests that Claude 3.7 Sonnet, even when equipped with enhanced reasoning capabilities, cannot overcome the cognitive rigidity exposed by conceptually demanding redefinition prompts, underscoring a fundamental limitation in the flexibility and reasoning capacity of current state-of-the-art models.

### 5.1.6 The influence of prompting

As described in section 4.3, we utilized different techniques to investigate how the design of our redefinition prompts influences models' ability to resist anchoring tendencies. Figure 5.1.9 visualizes the comparison of anchored response rates for  $Q_3$ -level questions and across the second swapping redefinition scenario, under Zero-Shot, Few-Shot, and Chain-of-Thought prompting. A rather surprising finding is that the Chain-of-Thought prompting strategy—although generally known to improve LLM performance on reasoning problems by decomposing them into intermediate steps [55]—fails to help models to meaningfully reduce anchoring percentages. Even the more capable LLMs of larger parameter size that typically benefit from step-by-step reasoning chains continue to adhere to their entrenched knowledge. One notable exception is Mistral 8x7b, which, in the case of  $Q_3$  questions and  $R_s2$  level, manages a substantial reduction in anchoring across the CoT setting, dropping from 46.67% to 13.33%. Few-Shot prompting, on the other hand, appears more successful in mitigating anchoring behavior. Models such as Mistral Large, Titan express, Titan large, Command r, and Claude haiku significantly improve their performance in this setting. In fact, over 50% of all evaluated LLMs achieve their lowest anchored response percentages under Few-Shot conditions. We assume that this occurs because the explicitly demonstrated instances in the FF-prompts—featuring similar redefinition scenarios accompanied with their correct solutions—provide a strong behavioral cue for the models to mimic. By "seeing" familiar entities being redefined and accepted within context, LLMs become more likely to "trust" the instruction over their pretraining priors.

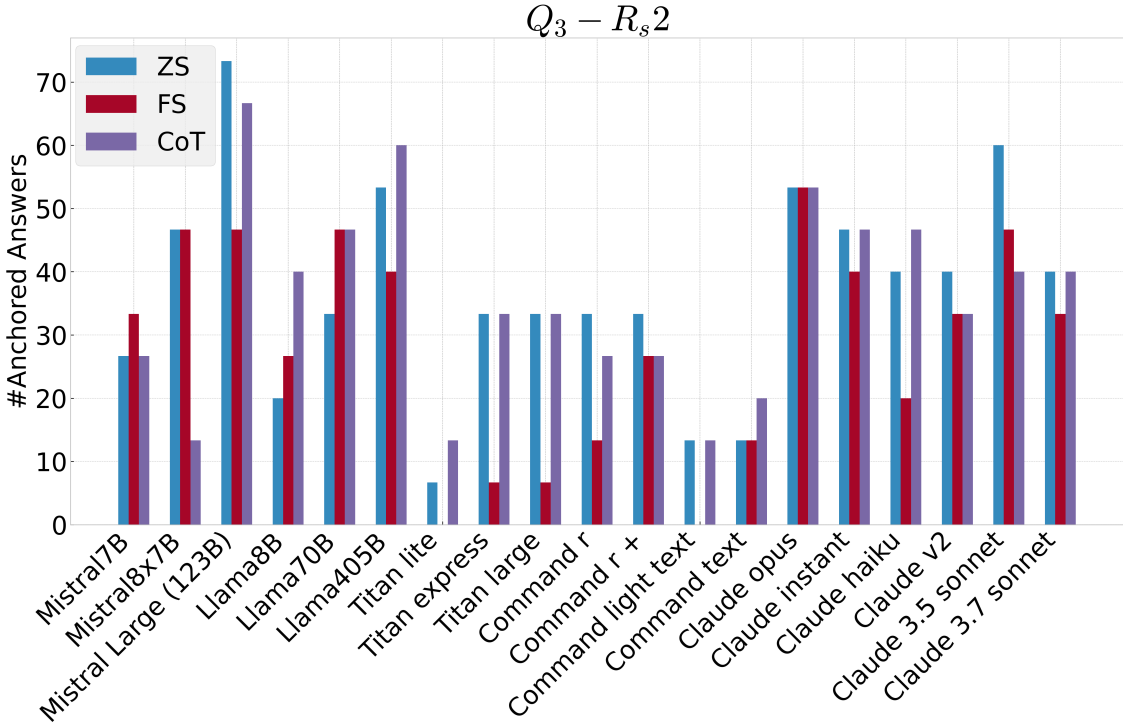


Figure 5.1.9: Comparison of the anchored response rate for  $Q_3$  questions in the  $R_{s2}$  redefinition level for all LLMs.

Although some of the previously discussed cases show that different prompting techniques can lead to substantially varying anchoring rates, we cannot claim that a consistent pattern holds across all LLMs tested. We measured the average difference between the maximum and minimum anchored responses percentages for all models to be  $16.29\% \pm 9.22$ , which indicates that anchoring is a relatively *prompt-insensitive phenomenon*. This conclusion is further supported by Tables 5.4 and 5.5, which report the correlation between performance before redefinition and the percentage of anchored responses in the Few-Shot and Chain-of-Thought setups—analogue to Table 5.2, presented in Section 5.1.1. The overall pattern remains the same across all prompting methods, with only minor variations: in the first two levels of question difficulty, the

two values are weakly correlated, while in the most difficult level, more knowledgeable models are more prone to anchoring under redefinitions, resulting in stronger correlation values, especially in swapping scenarios.

Level	$R_{a1}$	$R_{a2}$	$R_{a3}$	$R_{s1}$	$R_{s2}$
Free-Form (FF)					
$Q_1$	-0.055	-0.129	-0.472	0.235	-0.008
$Q_2$	-0.283	-0.359	-0.444	0.085	-0.148
$Q_3$	0.356	0.374	0.492	0.596	0.823
Multiple Choice (MC)					
$Q_1$	-0.71	-0.624	-0.711	-0.304	-0.28
$Q_2$	-0.258	-0.473	-0.312	0.441	-0.15
$Q_3$	0.269	0.589	0.288	0.624	0.694

Table 5.4: Correlation between model performance before redefinition with the percentage of anchored answers for each type of constant redefinition and question level in FS setup. Cells highlighted in pink indicate a **high positive correlation** ( $> 0.3$ ), while cells in green indicate a **high negative correlation** ( $< -0.3$ ).

Level	$R_{a1}$	$R_{a2}$	$R_{a3}$	$R_{s1}$	$R_{s2}$
Free-Form (FF)					
$Q_1$	-0.539	-0.542	-0.552	-0.244	-0.319
$Q_2$	-0.521	-0.626	-0.58	0.143	-0.125
$Q_3$	0.41	0.116	-0.085	0.71	0.588
Multiple Choice (MC)					
$Q_1$	-0.529	-0.483	-0.358	-0.17	0.16
$Q_2$	-0.183	-0.224	-0.202	0.329	-0.044
$Q_3$	0.134	0.366	0.009	0.679	0.657

Table 5.5: Correlation between model performance before redefinition with the percentage of anchored answers for each type of constant redefinition and question level in CoT setup. Cells highlighted in pink indicate a **high positive correlation** ( $> 0.3$ ), while cells in green indicate a **high negative correlation** ( $< -0.3$ ).

## 5.1.7 Completely Wrong Responses Analysis

### 5.1.7.1 Refusal to Respond

Even though Anchored Responses gathered most of our interest in this work, as they best highlight models' failure to suppress and override high confidence internalized priors, another intriguing behavioral tendency emerged among Completely Wrong Responses. In several cases, models not only failed to provide a specific result, but they actively refused to engage with the redefined premise altogether. Instead, they generated outputs like: "I can't assist you with that. Redefining Planck's constant is not a valid scientific approach.", "I should avoid making unsupported claims or providing potentially misleading information." or "I cannot reasonably redefine a scientific constant or answer a nonsensical question.". To systematically assess this refusal phenomenon, we further analyze Completely Wrong Responses by categorizing them into three types: 1) Actually Wrong Result: the model attempts to solve the task, but generates an answer that is neither correct under the redefinition nor the anchored one, 2) Blank Answer: the model produces a completely blank output or fails to conclude with a final result, and 3) Refusal to Answer: the model cites reasons such as the question being nonsensical, impossible to answer, or against its guidelines to explicitly refuse to perform the instructed redefinition. For each of these response categories, we calculate the corresponding rate. Table 5.6 includes the average refusal rates across the three levels of question difficulty for all LLMs that exhibited this behavior.

Model	Prompt	FF	MC
Mistral7B	ZS	<u>6.57 ± 11.99</u>	13.34 ± 18.07
	CoT	5.63 ± 8.89	<u>15.62 ± 16.45</u>
	FS	<b>3.7 ± 7.58</b>	<b>10.07 ± 15.25</b>
Mixtral8x7B	ZS	<u>18.0 ± 22.8</u>	8.61 ± 16.97
	CoT	9.22 ± 16.82	<u>15.5 ± 17.63</u>
	FS	<b>10.98 ± 17.03</b>	<b>5.95 ± 18.79</b>
Mistral Large	ZS	<u>16.33 ± 33.69</u>	1.67 ± 6.24
	CoT	<b>8.33 ± 18.51</b>	<b>0 ± 0</b>
	FS	14.35 ± 26.96	1.33 ± 4.99
Llama8B	ZS	<u>55.54 ± 24.37</u>	40.05 ± 18.58
	CoT	35.25 ± 23.33	32.89 ± 23.21
	FS	<b>2.41 ± 6.64</b>	<b>0 ± 0</b>
Llama70B	ZS	<u>38.66 ± 29.92</u>	5.56 ± 14.49
	CoT	9.17 ± 17.36	<u>13.33 ± 27.35</u>
	FS	<b>0 ± 0</b>	<b>0 ± 0</b>
Llama405B	ZS	<u>1.33 ± 4.99</u>	<b>0 ± 0</b>
	CoT	<b>0 ± 0</b>	<b>0 ± 0</b>
	FS	<b>0 ± 0</b>	<b>0 ± 0</b>
Titan lite	ZS	<b>1.56 ± 3.19</b>	<b>0 ± 0</b>
	CoT	<u>3.03 ± 5.66</u>	<b>0 ± 0</b>
	FS	2.54 ± 5.39	<b>0 ± 0</b>
Titan express	ZS	0.56 ± 2.08	<b>0 ± 0</b>
	CoT	<u>1.9 ± 7.13</u>	<b>0 ± 0</b>
	FS	<b>0 ± 0</b>	<b>0 ± 0</b>
Titan large	ZS	<u>2.0 ± 5.42</u>	<b>0 ± 0</b>
	CoT	<b>0 ± 0</b>	<b>0 ± 0</b>
	FS	<b>0 ± 0</b>	<b>0 ± 0</b>
Command text	ZS	<u>3.33 ± 9.03</u>	<b>0 ± 0</b>
	CoT	<b>0 ± 0</b>	<b>0 ± 0</b>
	FS	0.83 ± 3.12	<b>0 ± 0</b>
Claude Instant	ZS	<u>1.69 ± 4.36</u>	<b>0 ± 0</b>
	CoT	<b>0 ± 0</b>	<b>0 ± 0</b>
	FS	4.07 ± 12.58	<b>0 ± 0</b>
Claude v2	ZS	<u>20.48 ± 26.25</u>	4.83 ± 9.29
	CoT	14.31 ± 24.39	10.0 ± 27.08
	FS	<b>8.91 ± 24.75</b>	<b>3.17 ± 8.81</b>

Table 5.6: Average refusal rates over all question levels (lowest values in **bold** and highest values underlined). We exclude LLMs with zero refusal rate overall.

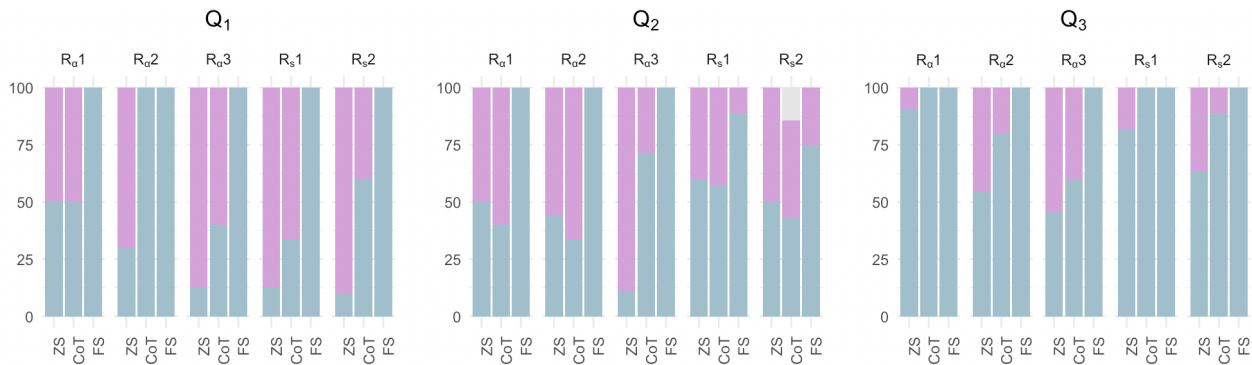
We can clearly observe a significant variation in refusal rates across different model families. Notably, Mistral and Llama models appear to exhibit markedly higher refusal tendencies, in contrast to models from the Cohere, Titan, and Claude families, which are consistently associated with lower—and frequently equal to zero—refusal rates. Within each family, interestingly, model size seems to also influence refusal appearance: larger models tend to generate lower percentages. This likely suggests that the increase in parameter count causes LLMs to become more and more confident in reasoning through the redefinition and ultimately providing a response. However, this "confidence", in many cases, proves to be false, leading to more anchored responses, which also agrees with the increased rates remarked in Section 5.1.2. On the other hand, we calculated the average correlations between No redefinition and Refusal to Answer responses and found them to be relatively weak (0.144 for the Free-Form and 0.039 for the Multiple Choice format), meaning that the refusal phenomenon is relatively independent of baseline reasoning capability. Regarding prompting strategies, the Few-Shot technique mostly achieves the lowest refusal rates. This aligns with expectations, because the demonstration of other successful redefinitions likely normalizes the task, reducing the chances that the models will judge the instruction as invalid or impossible and increasing its willingness to proceed to an attempt.

### 5.1.7.2 Case studies on Refusal and Overconfidence

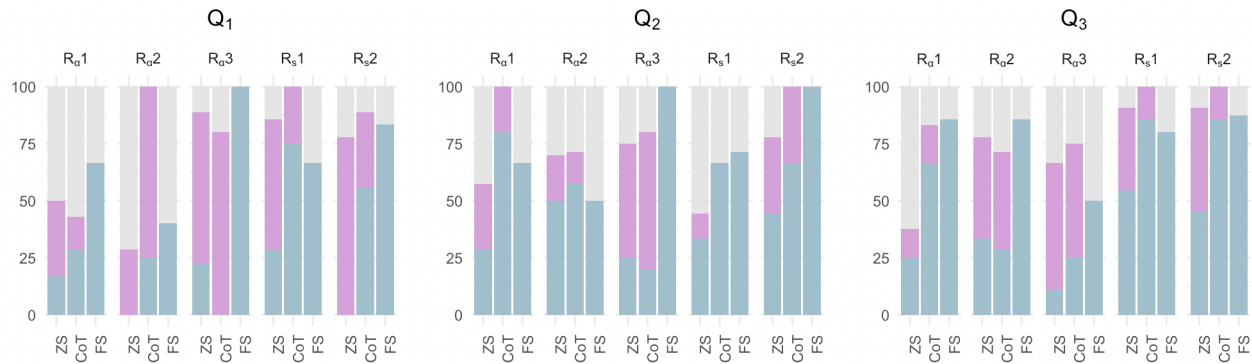
We conduct a more detailed analysis of some of the most interesting cases regarding these refusal and error dynamics, including figures for Llama 8B (5.1.10), Llama 70B (5.1.11), Mixtral 8x7B (5.1.12), and Claude v2 (5.1.13), which demonstrate a comprehensive breakdown of Completely Wrong Responses across all experimental combinations, for both question-answer formats.

#### Llama 8B

Interestingly, Llama 8B exhibits higher refusal rates on the easier question levels. Instead of recognizing its own limitations in the most challenging cases, it engages with the problem more and more frequently, even though it mostly fails, as indicated by the low Correct Responses and high Anchored Responses rates at level  $Q_3$ . The model’s tendency to respond even when lacking sufficient reasoning capabilities highlights a problematic form of *overconfidence*.



(a) Response breakdown for Llama8B FF responses.



(b) Response breakdown for Llama8B MC responses.

Figure 5.1.10: Completely wrong responses breakdown for Llama8B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses.

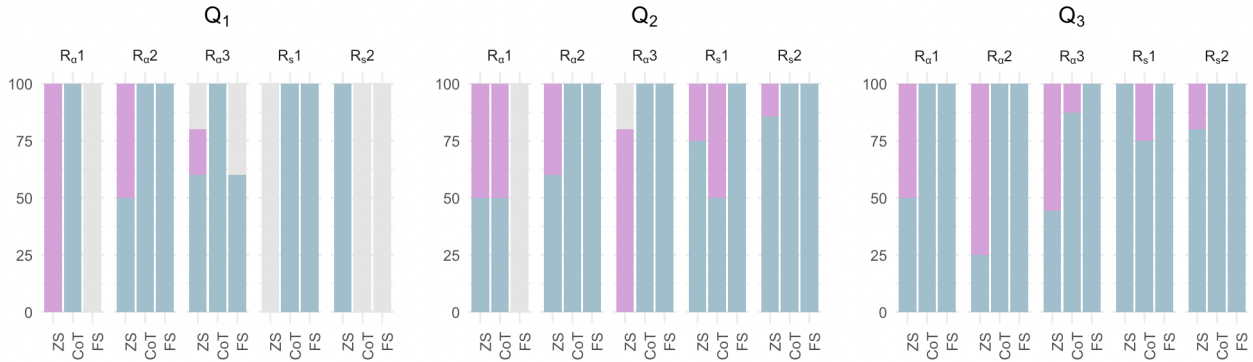
Comparing Multiple Choice to Free-Form response formats, we can clearly observe that, when asked to choose between possible options, Llama 8B generates higher Blank Response rates. In fact, across the Free-Form format these rates are almost entirely zero. This indicates that unrestricted generation may foster a false sense of confidence, while selecting between specific answers can lead to confusion when the model is unable to handle the queried reasoning task correctly. However, it is particularly interesting that refusal behavior



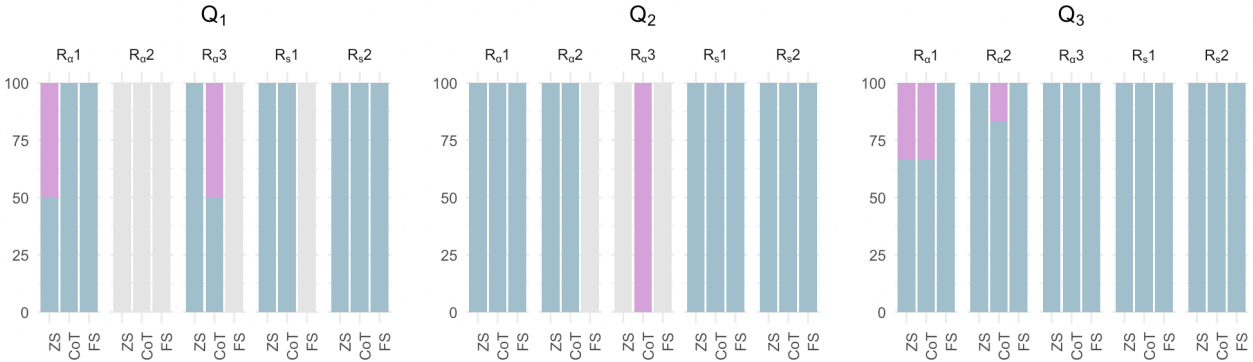
is also evident throughout Multiple Choice experiments. Even when the prompt explicitly states that the correct answer to the given problem exists and is actually included among the Multiple Choice options, Llama 8B still declines to attempt the task.

## Llama 70B

In the case of Llama 70B, these behavioral tendencies seem to follow a different trajectory. Refusal and blank answer rates are generally lower than those of its smaller counterpart, which supports the claim in the previous section that models of larger parameter size are more confident in generating a solution to the reasoning problem. Prompting strategies also appear to mitigate refusal behavior. Both Chain-of-Thought and Few-Shot techniques result in lower refusal percentages compared to the Zero-Shot case. In fact, in the Few-Shot setup, Llama 70B did not generate any Refusal-to-Answer instances across all different question/redefinition/format combinations.



(a) Response breakdown for Llama70B FF responses.

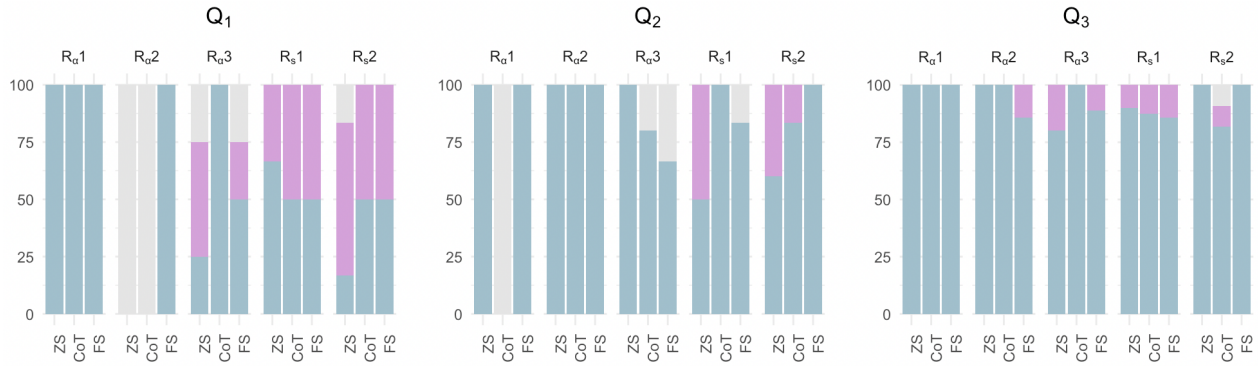


(b) Response breakdown for Llama70B MC responses.

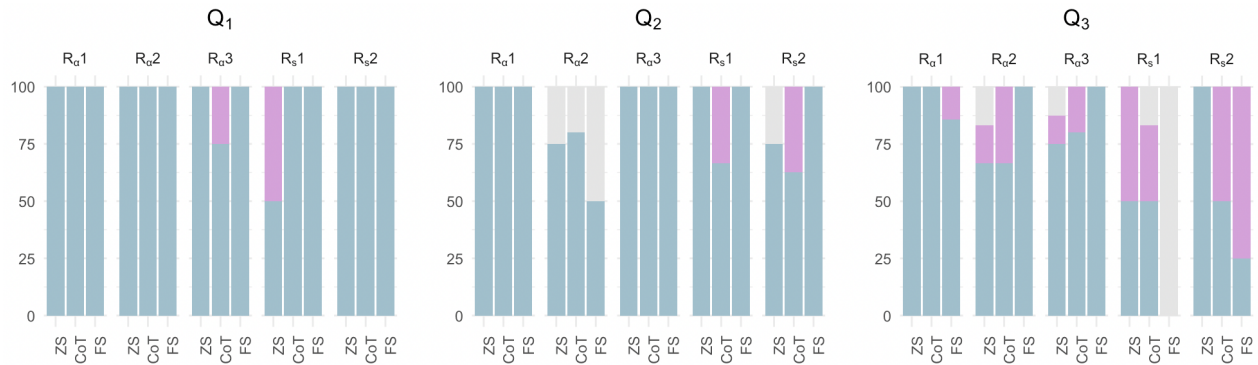
Figure 5.1.11: Completely wrong responses breakdown for Llama70B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

## Mixtral 8x7B

Mixtral 8x7B reveals an intriguing distinction between response formats. In the Few-Shot format, refusal behavior decreases with question difficulty, leading to more frequent completely wrong outputs, suggesting a form of overconfidence that counterintuitively intensifies under more challenging conditions. Conversely, in the Multiple Choice case, refusal rates increase as questions become more difficult. When restricted to specific options, Mixtral 8x7B is more likely to detect the epistemic conflict and refuse to attempt a solution in complex reasoning scenarios.



(a) Response breakdown for Mixtral8x7 FF responses.

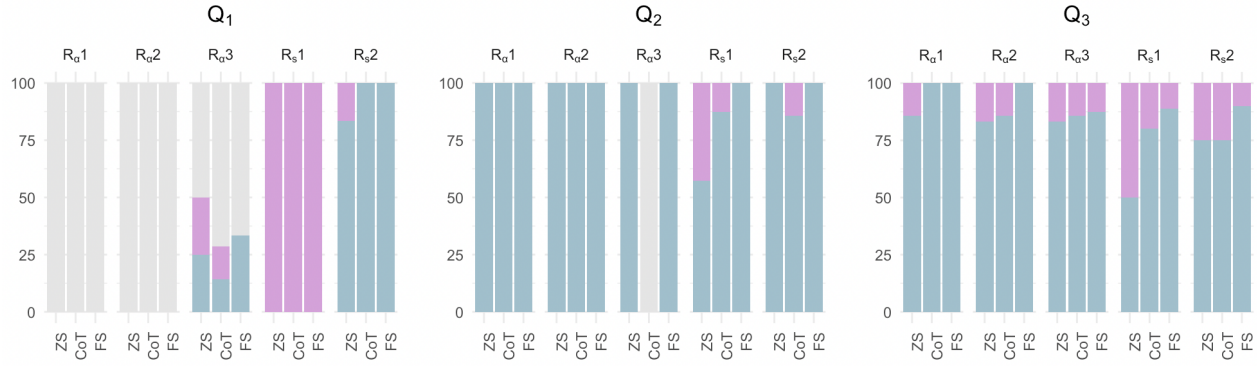


(b) Response breakdown for Mixtral8x7 MC responses.

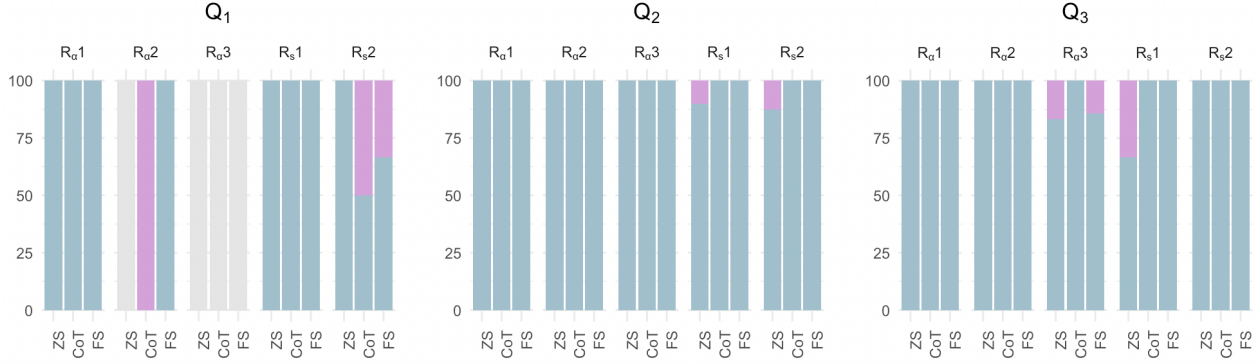
Figure 5.1.12: Completely wrong responses breakdown for Mixtral8x7. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

## Claude v2

Claude v2 exhibits an inconsistent pattern in how it reacts to the redefinition instruction. On the easiest questions, refusal and error rates often fluctuate sharply between 100% Refusal-to-Answer and 100% Actually-Wrong-Result. As difficulty increases, Claude v2 becomes more confidently willing to engage with the task, even if it eventually fails. In the Multiple Choice response format in particular, refusal behavior is significantly reduced, suggesting that the model presumably trusts that the queried problem is solvable and proceeds to identify the correct solution among the given options.



(a) Response breakdown for Claude v2 FF responses.



(b) Response breakdown for Claude v2 MC responses.

Figure 5.1.13: Completely wrong responses breakdown for Claude v2. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

### 5.1.7.3 Refusal-Adjusted Anchoring: A Refined Perspective

Model	$R_{a3}$						$R_{s2}$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	33.33	<b>50.0</b>	<b>33.33</b>	<b>28.57</b>	28.57	40.0	45.45	53.33	14.28	35.71	26.67	23.08
Mixtral8x7B	<b>38.46</b>	33.33	26.67	26.67	23.08	35.71	36.37	46.67	46.15	<b>53.33</b>	46.67	<b>73.33</b>
Mistral Large	33.33	20.0	26.67	26.67	<b>57.14</b>	<b>66.67</b>	<b>71.43</b>	<b>53.33</b>	<b>50.0</b>	40.0	<b>73.33</b>	66.67
Llama8B	0.0	<b>44.45</b>	0.0	<b>36.37</b>	22.22	50.0	<b>50.0</b>	24.99	<b>40.0</b>	<b>50.0</b>	27.27	30.0
Llama70B	<b>7.15</b>	13.33	0.0	0.0	20.0	40.0	33.33	46.67	14.28	46.67	35.71	73.33
Llama405B	0.0	0.0	0.0	13.33	<b>26.67</b>	<b>53.33</b>	26.67	<b>46.67</b>	6.67	20.0	<b>57.14</b>	<b>93.33</b>
Command text	13.33	20.0	6.67	6.67	6.67	26.67	40.0	26.67	13.33	26.67	13.33	38.46
Claude v2	28.57	13.33	20.0	0.0	50.0	42.86	14.28	40.0	33.33	21.43	46.15	66.67

Table 5.7: The percentage of anchored responses for the models in the ZS setup for the most difficult constants redefinitions in *assignment* ( $R_{a3}$ ) and *swapping* ( $R_{s2}$ ). The highest number for each model family is presented in **bold**. We exclude models where no refusals occurred, as their results are identical to those in Table 5.1.

In previous sections, we highlighted the ubiquitous role of anchoring in redefinition reasoning tasks. The uncovering of refusal behavior, however, raises an interesting question: Would anchoring rates more accurately reflect model behavior if extracted only from the outputs where the model actually attempted to solve the problem? Therefore, to isolate LLMs' "true" anchoring tendencies, we conduct a refined analysis that excludes responses in which the model refused to answer from the Anchored Responses rates calculation. These adjusted results are demonstrated in Table 5.7, covering the  $R_a3$  and  $R_s2$  redefinition scenarios and the  $Q_3$ -level of question difficulty. Models that never exhibit refusal behavior are not included in this table, as their rates are indeed unaffected by this filtering and are identical to those presented in Table 5.1.

## 5.2 Results on units redefinition

We present and further analyze the corresponding results for the task of unit of measurement redefinition. This section aims to investigate behavioral tendencies such as anchoring to predefined values and refusal to answer. At the same time, it evaluates how factors like knowledgeability, reasoning skills, parameter scale, but also prompting techniques, response formats, and levels of difficulty influence the way LLMs react to the alteration of familiar relationships between widely used units of measurement. We also compare these findings to the ones previously discussed in the scientific constant redefinition case.

### 5.2.1 Anchoring to default values

The clear presence of anchoring behavior is once again undeniable under this new redefinition task. All queried models generate responses in which, to varying degrees, they disregard the altered premise and instead adhere to their familiar priors. Detailed results of the Anchored Responses percentages for the two more extreme redefinition types across both response formats and under the Zero-Shot prompting setup are displayed in Table 5.8.

Model	$R_a2$						$R_a3$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	0.0	37.5	25.0	25.0	18.75	56.25	<b>62.5</b>	25.0	<b>31.25</b>	<b>37.5</b>	31.25	25.0
Mixtral8x7B	<b>6.25</b>	31.25	<b>31.25</b>	37.5	<b>31.25</b>	37.5	6.25	<b>31.25</b>	6.25	31.25	<b>31.25</b>	<b>50.0</b>
Mistral Large	0.0	<b>37.5</b>	6.25	<b>37.5</b>	12.5	<b>56.25</b>	0.0	25.0	12.5	<b>37.5</b>	12.5	43.75
Llama8B	0.0	<b>25.0</b>	<b>6.25</b>	<b>31.25</b>	12.5	31.25	<b>6.25</b>	<b>31.25</b>	<b>12.5</b>	<b>50.0</b>	<b>25.0</b>	50.0
Llama70B	0.0	6.25	<b>6.25</b>	<b>31.25</b>	<b>25.0</b>	<b>56.25</b>	0.0	18.75	0.0	<b>50.0</b>	12.5	<b>62.5</b>
Llama405B	0.0	0.0	0.0	<b>31.25</b>	12.5	37.5	0.0	0.0	6.25	25.0	<b>25.0</b>	31.25
Titan lite	6.25	<b>25.0</b>	12.5	<b>31.25</b>	12.5	<b>25.0</b>	25.0	<b>31.25</b>	25.0	12.5	0.0	18.75
Titan express	18.75	<b>25.0</b>	<b>25.0</b>	18.75	12.5	<b>25.0</b>	<b>43.75</b>	25.0	31.25	12.5	6.25	18.75
Titan large	<b>31.25</b>	12.5	12.5	<b>31.25</b>	<b>18.75</b>	<b>25.0</b>	25.0	12.5	<b>37.5</b>	<b>31.25</b>	6.25	<b>25.0</b>
Command r	<b>12.5</b>	18.75	<b>12.5</b>	<b>31.25</b>	25.0	18.75	6.25	25.0	<b>12.5</b>	18.75	<b>12.5</b>	31.25
Command r+	6.25	<b>43.75</b>	0.0	25.0	<b>37.5</b>	<b>50.0</b>	<b>6.25</b>	<b>31.25</b>	0.0	<b>31.25</b>	0.0	25.0
Command light text	6.25	12.5	0.0	25.0	6.25	25.0	12.5	25.0	6.25	31.25	0.0	<b>50.0</b>
Command text	12.5	12.5	12.5	18.75	0.0	18.75	0.0	31.25	12.5	12.5	0.0	43.75
Claude opus	0.0	0.0	0.0	6.25	12.5	25.0	0.0	0.0	0.0	0.0	0.0	6.25
Claude instant	<b>6.25</b>	<b>25.0</b>	<b>12.5</b>	25.0	0.0	<b>43.75</b>	0.0	<b>43.75</b>	0.0	<b>37.5</b>	6.25	<b>31.25</b>
Claude haiku	0.0	18.75	0.0	12.5	6.25	31.25	0.0	6.25	0.0	6.25	<b>18.75</b>	<b>31.25</b>
Claude v2	<b>6.25</b>	18.75	6.25	<b>31.25</b>	<b>18.75</b>	31.25	<b>6.25</b>	0.0	<b>6.25</b>	25.0	6.25	12.5
Claude 3.5 Sonnet	0.0	0.0	0.0	12.5	6.25	6.25	0.0	0.0	0.0	6.25	0.0	0.0
Claude 3.7 Sonnet	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 5.8: The percentage of anchored responses for all LLMs tested under the ZS prompting setup for the most difficult units of measure redefinitions ( $R_a2$  and  $R_a3$  levels). The highest rate for each model family is presented in **bold**.

Evidently, several models exhibit elevated anchoring rates across these experimental combinations. Among others, Llama 70B reaches a 62.5% score in the hardest scenario of Multiple-Choice-structured questions, and an only slightly lower rate of 56.25% in the second more complex redefinition level. Anchoring rates of 56.25% are also recorded for both Mistral 7B and Mistral large under the same conditions. Notably, Mistral and Titan models exhibit persistently substantial anchoring behavior even at easier question levels. However, it appears that unit of measure anchoring results are overall significantly lower in comparison to those of the constant case. In fact, LLMs from the Command and Claude families even achieve several 0% scores, especially in the Free-Form question setup and the first two difficulty levels.

A familiar trend reappears in the correlation metrics between performance accuracy in the No Redefinition task and anchoring rates across the various types of unit of measurement reassignments, as demonstrated in Table 5.9 for the Zero-Shot case. Weaker correlations are observed throughout the easier  $Q_1$  and  $Q_2$  questions, in contrast to the hardest  $Q_3$  level, where the values noticeably increase. This indicates that potent reasoners that successfully solve the most challenging tasks involving predefined unit relationships are also more likely to adhere to their entrenched knowledge when faced with conflicting instructions.

Level	$R_{a1}$	$R_{a2}$	$R_{a3}$
Free-Form (FF)			
$Q_1$	-0.295	-0.403	-0.33
$Q_2$	-0.361	-0.247	-0.479
$Q_3$	-0.063	0.19	0.14
Multiple Choice (MC)			
$Q_1$	-0.49	-0.149	-0.542
$Q_2$	-0.159	-0.023	0.08
$Q_3$	0.248	0.338	-0.127

Table 5.9: Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in ZS setup. Cells highlighted in pink indicate a **high positive correlation** ( $> 0.3$ ), while cells in green indicate a **high negative correlation** ( $< -0.3$ ).

Interestingly, a more apparent distinction occurs between the two response formats. Correlation results are substantially weaker—and more often negligible—in the Free-Form question-answering setup across all escalating levels. This suggests that, when generation is unrestricted, reasoning capacity plays a less important role in anchoring tendencies under the unit of measure redefinition task.

### 5.2.2 Inverse Trends

Table 5.10 reports the percentages of Anchored Responses in relation to the corresponding pre-redefinition performance accuracies for the  $R_{a2}$  and  $R_{a3}$  scenarios and under the Zero-Shot/Free-Form setup. Inverse scaling patterns seem to emerge in several cases under the unit redefinition task. For example, within the Titan model family, Titan lite, express, and Large produce 6.25%, 18.75%, and 31.25% anchoring rates in the  $Q_1$  level of  $R_{a2}$  unit reassignments, and 25%, 31.25%, and 37.5% in the  $Q_2$  level of  $R_{a3}$  redefinitions, respectively. In addition, Mistral 8x7B surpasses Mistral 7B in anchoring with a percentage of 31.25% over 25% in the first level of question difficulty and over 18.75% in the second, while Llama 405B also experiences a higher rate (25%) than its smaller counterpart, Llama 70B (12.5%), in the more complex redefinition scenario, generating Anchored Responses twice as many times.

Model	$R_a2$						$R_a3$					
	$Q_1$		$Q_2$		$Q_3$		$Q_1$		$Q_2$		$Q_3$	
	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF	NR	FF
Mistral 7B	81.25	0.0	56.25	25.0	43.75	18.75	81.25	62.5	56.25	31.25	43.75	31.25
Mixtral8x7B	87.5	6.25	81.25	31.25	62.5	31.25	87.5	6.25	81.25	6.25	62.5	31.25
Mistral Large	93.75	0.0	93.75	6.25	81.25	12.5	93.75	0.0	93.75	12.5	81.25	12.5
Llama8B	75.0	0.0	56.25	6.25	6.25	12.5	75.0	6.25	56.25	12.5	6.25	25.0
Llama70B	100.0	0.0	81.25	6.25	56.25	25.0	100.0	0.0	81.25	0.0	56.25	12.5
Llama405B	100.0	0.0	93.75	0.0	56.25	12.5	100.0	0.0	93.75	6.25	56.25	25.0
Titan lite	37.5	6.25	18.75	12.5	6.25	12.5	37.5	25.0	18.75	25.0	6.25	0.0
Titan express	75.0	18.75	37.5	25.0	6.25	12.5	75.0	43.75	37.5	31.25	6.25	6.25
Titan large	68.75	31.25	68.75	12.5	25.0	18.75	68.75	25.0	68.75	37.5	25.0	6.25
Command r	75.0	12.5	56.25	12.5	18.75	25.0	75.0	6.25	56.25	12.5	18.75	12.5
Command r+	87.5	6.25	93.75	0.0	81.25	37.5	87.5	6.25	93.75	0.0	81.25	0.0
Command light text	31.25	6.25	6.25	0.0	0.0	6.25	31.25	12.5	6.25	6.25	0.0	0.0
Command text	62.5	12.5	50.0	12.5	25.0	0.0	62.5	0.0	50.0	12.5	25.0	0.0
Claude opus	100.0	0.0	75.0	0.0	56.25	12.5	100.0	0.0	75.0	0.0	56.25	0.0
Claude instant	75.0	6.25	81.25	12.5	43.75	0.0	75.0	0.0	81.25	0.0	43.75	6.25
Claude haiku	100.0	0.0	93.75	0.0	81.25	6.25	100.0	0.0	93.75	0.0	81.25	18.75
Claude v2	93.75	6.25	68.75	6.25	25.0	18.75	93.75	6.25	68.75	6.25	25.0	6.25
Claude 3.5 Sonnet	100.0	0.0	87.5	0.0	87.5	6.25	100.0	0.0	87.5	0.0	87.5	0.0
Claude 3.7 Sonnet	100.0	0.0	87.5	0.0	93.75	0.0	100.0	0.0	87.5	0.0	93.75	0.0

Table 5.10: The percentage of correct responses with no redefinition (NR) and the anchored response rate for units of measure redefinitions regarding free-form (FF) responses using ZS prompting.

Notably, inverse scaling trends also appear under the Multiple Choice response format. For instance, as visualized in Figure 5.2.1, Mistral Large tends to produce more Anchored responses compared to its smaller counterpart, Mistral 7B, when faced with unit of measure redefinitions, despite significantly outperforming it in the baseline NR experiments. This reinforces the broader theme that increases in model size and stronger reasoning capabilities in standard conditions do not straightforwardly predict improvements under input perturbations, such as redefinitions, as well.

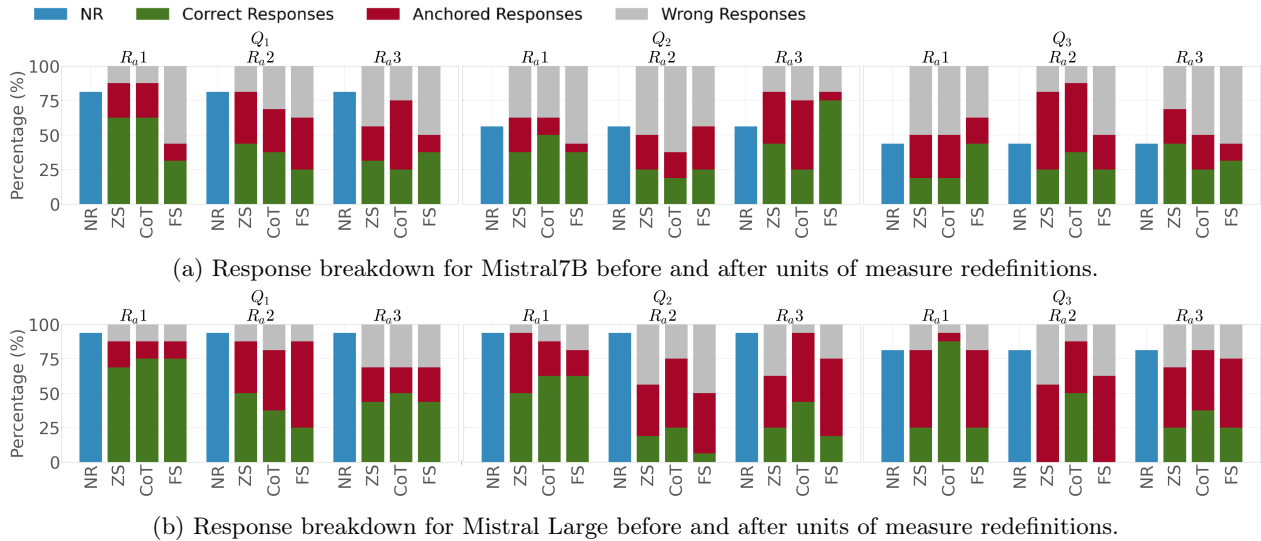


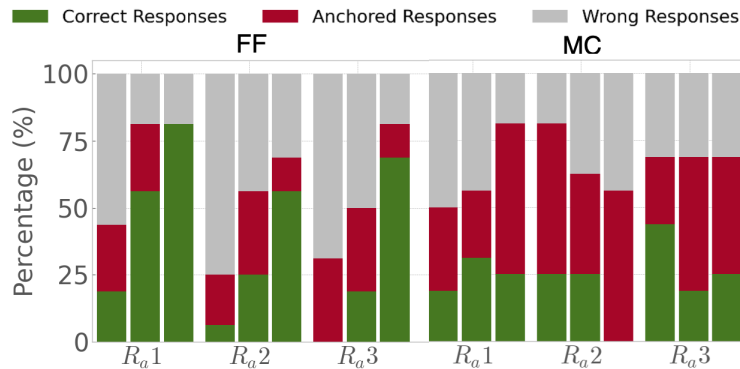
Figure 5.2.1: Comparison of Mistral7B and Mistral Large (123B) responses on the MC response format for units of measure redefinitions.



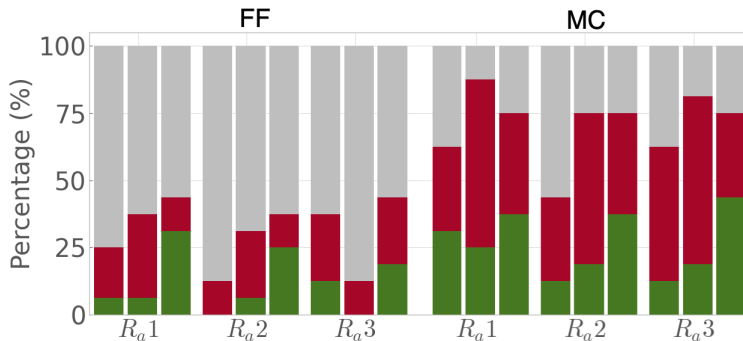
This inverse phenomenon may be less prominent under the current conditions, compared to the previously studied constants case, but it remains present nonetheless. Larger models do not produce as drastically increased anchoring rates, but they do not seem to "fix" the problem either, even though they demonstrate stronger reasoning capabilities, as evidenced by the results of the No Redefinition task. In any case, unit of measure redefinition still qualifies as an inverse scaling task.

### 5.2.3 Response Format

Once again, it is evident that the Multiple Choice response format is associated with significantly higher susceptibility to anchoring in contrast to the Free-Form one. This phenomenon is most clearly accentuated in Figure 5.2.2, which showcases all different types of responses to the redefinition query across the Free-Form and Multiple Choice setup, for Mistral and Llama models of various parameter sizes. While Anchored Responses rates are relatively low in the Free-Form case, they drastically escalate in the Multiple Choice format, sometimes experiencing an over 100% increase. Llama 70B, in particular, interestingly generates rates of 12.5% and 62.5% in FF and MC format, indicating a sharp 400% increase in anchoring behavior when shifting from answering freely to choosing between possible options. This event is a consequence of exposing the model to the default unit/counterpart relationship within the given answers, which creates a strong conflict between instruction and memorization.



(a) Response breakdown for Mistral models.



(b) Response breakdown for Llama models.

Figure 5.2.2: Results for the different Mistral and Llama models on  $Q_3$  questions using ZS prompting. The order of the bars per redefinition type/level corresponds to increasing model size.

### 5.2.4 The influence of prompting

Tables 5.11 and 5.12 demonstrate the correlations between No redefinition accuracies and post-redefinition anchoring percentages for the Few-Shot and Chain-of-Thought prompting strategies, respectively. We compare these results to the corresponding ones presented in Table 5.9 of section 5.2.1 in order to determine how the employment of different prompting techniques can affect—and potentially mitigate—the anchoring phenomenon.

Level	$R_{a1}$	$R_{a2}$	$R_{a3}$
Free-Form (FF)			
$Q_1$	-0.32	-0.442	-0.161
$Q_2$	-0.404	-0.231	0.039
$Q_3$	0.128	-0.042	0.279
Multiple Choice (MC)			
$Q_1$	-0.332	0.058	-0.593
$Q_2$	0.135	0.131	0.266
$Q_3$	0.314	0.49	0.101

Table 5.11: Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in FS setup. Cells highlighted in **pink** indicate a **high positive correlation** ( $> 0.3$ ), while cells in **green** indicate a **high negative correlation** ( $< -0.3$ ).

Level	$R_{a1}$	$R_{a2}$	$R_{a3}$
Free-Form (FF)			
$Q_1$	-0.502	-0.598	-0.529
$Q_2$	-0.465	-0.3	-0.174
$Q_3$	-0.232	-0.181	-0.079
Multiple Choice (MC)			
$Q_1$	-0.528	-0.023	-0.523
$Q_2$	0.015	-0.091	-0.016
$Q_3$	-0.127	0.013	-0.242

Table 5.12: Correlation between model performance before redefinition with the percentage of anchored answers for each type of unit of measure redefinition and question level in CoT setup. Cells highlighted in **pink** indicate a **high positive correlation** ( $> 0.3$ ), while cells in **green** indicate a **high negative correlation** ( $< -0.3$ ).

A clear pattern mirrors the Zero-Shot correlation case in the Few-Shot prompting setup. Higher correlation values are observed in the most difficult level of questions, while correlations remain weak in the easier cases, particularly in the Free-Form question-answering format. However, a stark departure occurs in the results of Chain-of-Thought conducted experiments: the CoT technique achieves a significant reduction in anchoring, especially among more potent LLM reasoners, across all levels of question and assignment difficulty. Even for  $Q_3$ -level questions, where stronger correlations generally arise, the calculated results are almost entirely negative in this case. Thus, task decomposition and reasoning chains of intermediate steps—even though ineffective for constants—seem to become beneficial methods for guiding models to reason beyond strongly memorized conceptual mappings in the unit of measurement redefinition task.



## 5.2.5 Completely Wrong Responses Analysis

### 5.2.5.1 Refusal to respond

In Section 5.1.7 we thoroughly examined the Refusal-to-Respond phenomenon, where the models explicitly refrain from attempting a solution to the instructed redefinition problem, because of its persistent presence in the responses of the majority of the queried LLMs across all different experimental combinations in the constant redefinition task. However, in striking contrast to the constants case, similar behavior is almost entirely absent under these new conditions, where the redefinitions concern relationships between units of measure. More specifically, Refusal-to-Answer response rates are entirely zero across all LLMs tested, with the exception of the three models within the Mistral family, which once again proves that this is a model family-specific tendency. Some particular outputs, for example, that these models generated when exhibiting this type of behavior are: "The question cannot be answered meaningfully with the redefined unit of time.", "This question is designed to be impossible to answer.", "The question is not valid." or "I am an assistant and may not have the ability to redefine units.". Even these instances, though, which will be further investigated in the following section, are mostly isolated throughout different scenarios.

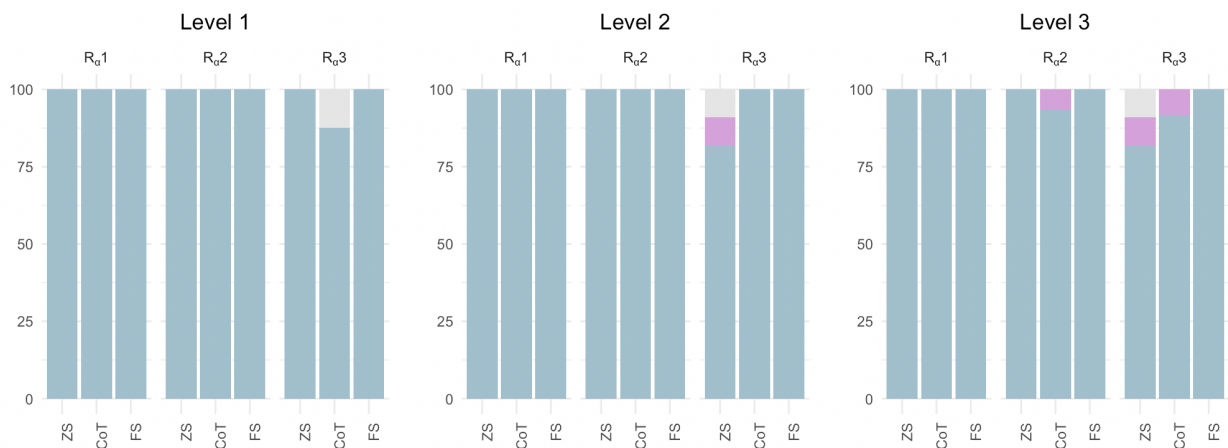
The only difference between the two redefinition tasks explored in this work lies in the distinct knowledge domains of the entities being redefined. Therefore, the sharp discrepancy in the frequency of refusal under otherwise similar experimental conditions points to a deeper distinction in the way that LLMs internalize the knowledge behind each cognitive topic.

### 5.2.5.2 Case studies on Refusal and Overconfidence

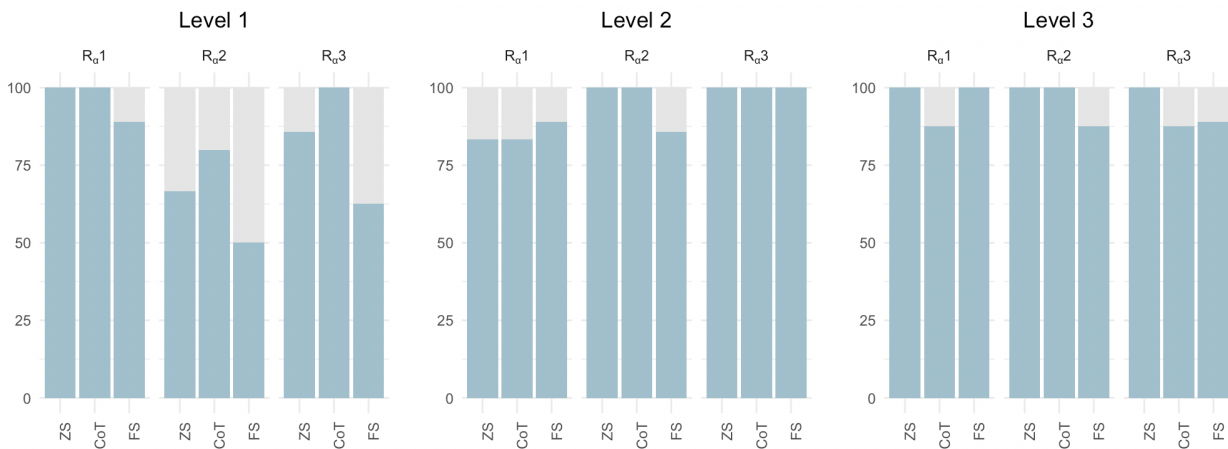
We analyze the outputs of the models within the Mistral family, which uniquely displayed refusal behavior. Figures 5.2.3, 5.2.4, and 5.2.5 provide a detailed presentation of Completely Wrong Responses for Mistral 7B, Mixtral 8x7B, and Mistral Large, respectively.

#### Mistral 7B

In the case of Mistral 7B, refusal responses are highly infrequent and appear only within the Free-Form format results. Nearly all the completely wrong outputs in this setup fall under the category of "Actually Wrong Results". On the other hand, in the Multiple Choice case, while refusal rates are entirely zero, there is a drastic increase of Blank Response occurrences, particularly at the  $Q_1$  level. This suggests that, while unconstrained generation leads to a false sense of certainty, when operating under the MC format, this model seems to suffer from indecision, as it is forced to select among conflicting possible answers.



(a) Response breakdown for Mistral 7B FF responses.

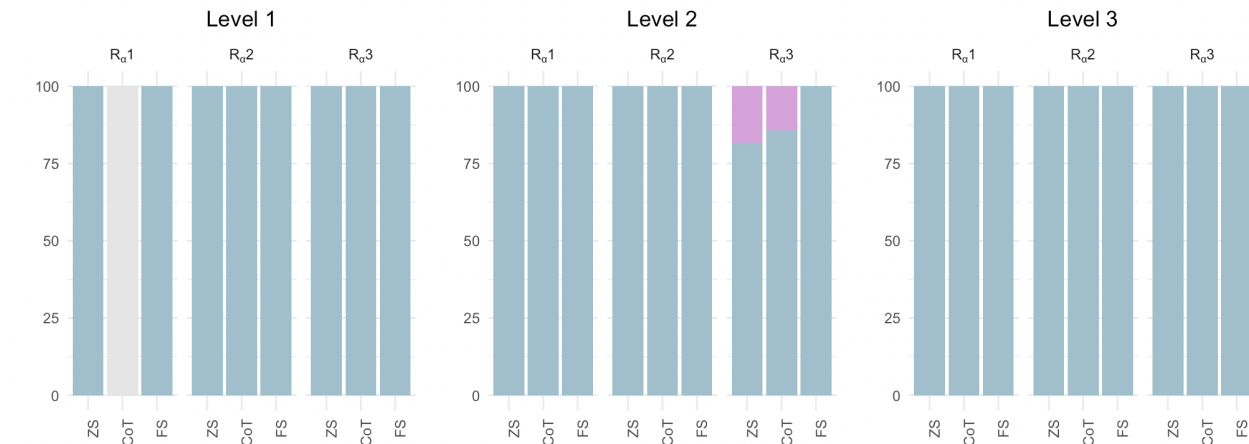


(b) Response breakdown for Mistral 7B MC responses.

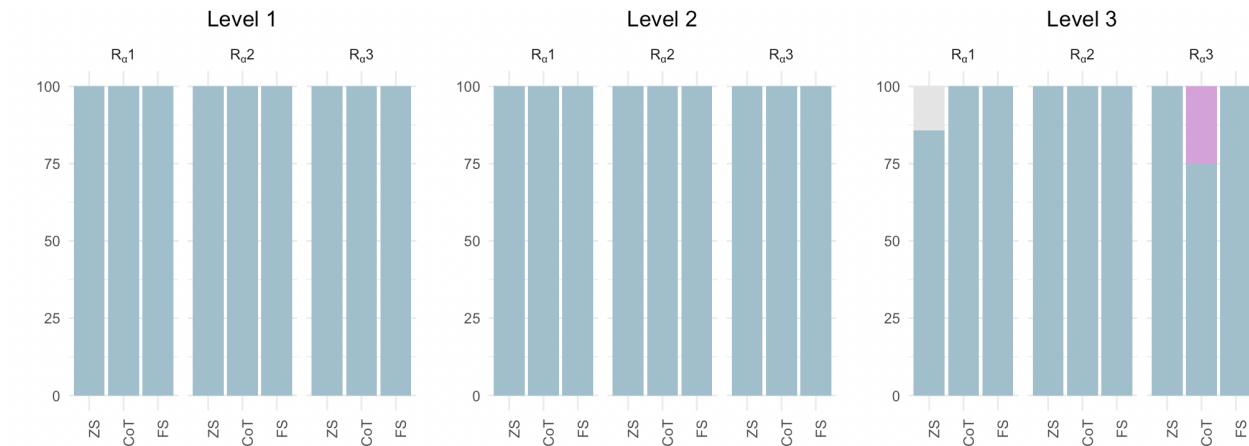
Figure 5.2.3: Completely wrong responses breakdown for Mistral 7B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

### Mixtral 8x7B

Cases of refusal behavior in the Mixtral 8x7B model are, once again, extremely isolated—this time occurring in both Free-Form and Multiple Choice settings. Interestingly, in contrast to its smaller counterpart, Blank answers are almost entirely absent across all different conditions. This implies that Mixtral 8x7B reveals a heightened level of overconfidence in responding, regardless of the question-answering format.



(a) Response breakdown for Mixtral 8x7B FF responses.

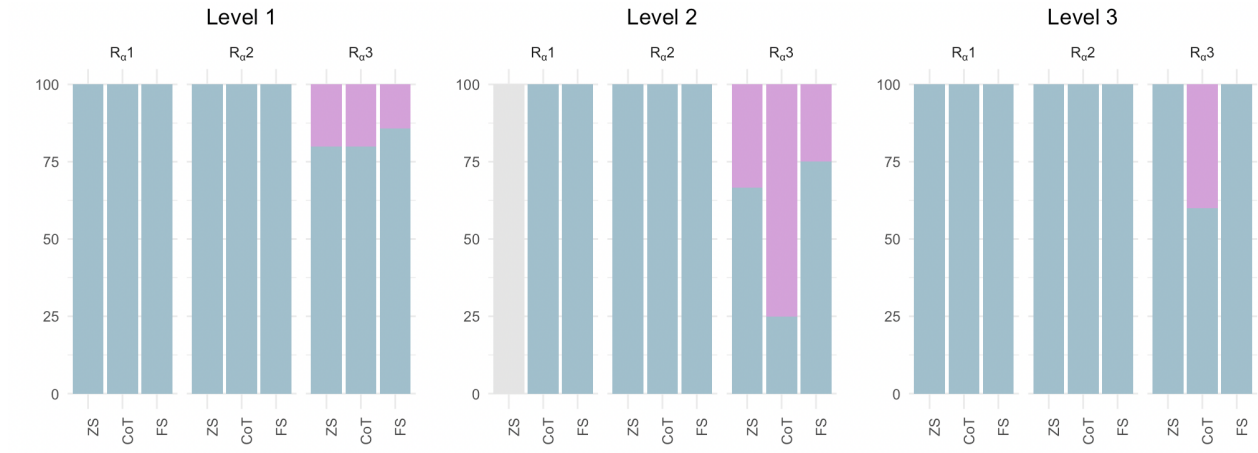


(b) Response breakdown for Mixtral 8x7B MC responses.

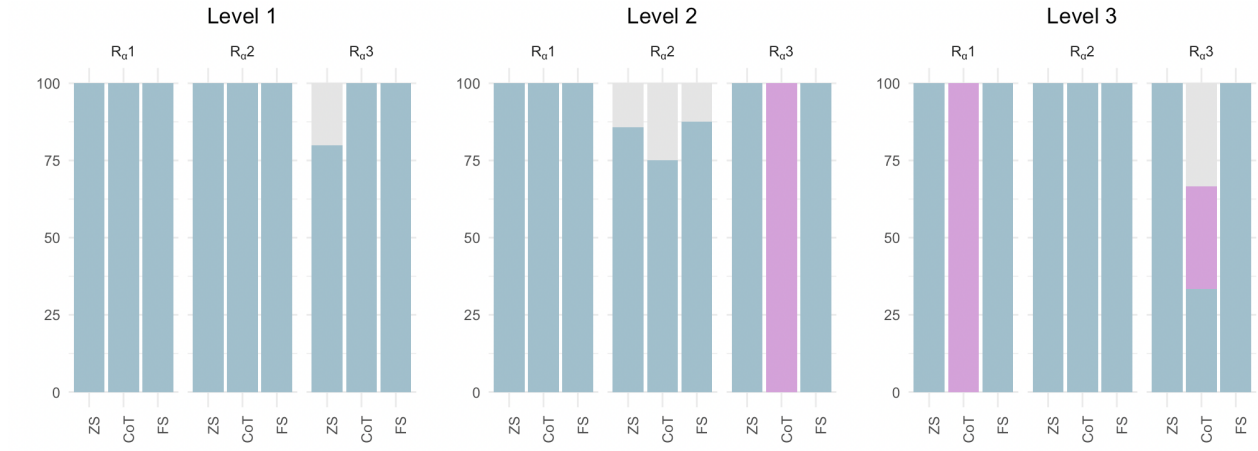
Figure 5.2.4: Completely wrong responses breakdown for Mixtral 8x7B. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

### Mistral Large

Mistral Large is the model that refuses to answer the most under the unit of measurement redefinition task. This is rather intriguing because, in the constants case, we observed that an increase in parameter size meaningfully reduces Refusal to Answer percentages. In this case, however, the largest model within the LLM family exhibits substantially high refusal rates in both response formats and especially under the most extreme assignment type scenario.



(a) Response breakdown for Mistral Large FF responses.



(b) Response breakdown for Mistral Large MC responses.

Figure 5.2.5: Completely wrong responses breakdown for Mistral Large. Blue denotes actually wrong responses, Purple indicates refusals, while Gray instances correspond to blank responses

# Chapter 6

## Conclusion

### 6.1 Conclusion

In this work, we conducted a comprehensive investigation of the *redefinition task*, examining how Large Language Models (LLMs) respond when presented with redefined values of widely known and used scientific constants and units of measurement. By prompting models of different sizes and architectures to reason under these deliberately altered premises, we aimed to assess their flexibility, as opposed to their susceptibility to anchoring on deeply internalized prior knowledge. Our findings expose critical behavioral patterns in LLMs, showcasing limitations that, paradoxically, become more pronounced with increasing model scale.

We specifically find that although larger models demonstrate stronger reasoning capabilities under standard, pre-redefinition conditions, they tend to perform worse when required to follow redefinition as they consistently revert to the original values that they memorized during their pretraining. In addition, our in-depth analysis of model outputs reveals a pattern of false confidence towards responding rather than abstaining, which leads larger LLMs in providing incorrect answers with high certainty.

Furthermore, our experiments span a diverse range of conditions designed to stress-test the adaptability of LLMs. We constructed datasets that include variations in types and levels of redefinitions, as well as three escalating tiers of question difficulty applied to both default and redefined values. Across these dimensions, we systematically evaluate how different response formats (Free-Form and Multiple Choice) and prompting strategies (Zero-Shot, Few-Shot, and Chain-of-Thought) influence model behavior. Our results show that LLMs anchor substantially more under the Multiple Choice format, often being misled by the presence of the default answer among the provided options. Prompting techniques can only partially mitigate the anchoring effect, but they fall short of fully eliminating the problem.

The general trends observed in our findings extend across both scientific constants and units of measurement. However, the anchoring phenomenon appears to be relatively less prominent under the unit of measurement redefinitions, and refusal behavior is nearly absent. This indicates that the way LLMs handle redefinitions is highly dependent on the specific knowledge domain and how that information is internalized during pretraining. We hypothesize that LLMs are more capable of successfully overriding established definitions of units of measurement, likely because their actual values are not commonly used in calculations and therefore less deeply embedded during pretraining.

Overall, our work uncovers key gaps in reasoning and flexibility that become more pronounced as LLMs grow in size. It also highlights the importance of developing a deeper understanding of LLM behavior, not only in terms of what these models can accomplish, but also in how and why they fail. The inverse task of redefinition offers a valuable lens through which to examine the fragile interplay between scale, instruction following, and entrenched priors, and we hope this framework will serve as a foundation for future research in probing model adaptability, robustness, and reasoning capacity.

## 6.2 Reasoning vs. Robustness Trade-Off

Our study exposes a fundamental tension in LLM behavior: the balance between reasoning transparency and robust adherence to factual knowledge. On one hand, models that are more robust tend to resist redefined concepts by refusing the task altogether. While this rigidity makes them less susceptible to prompt-based manipulation or misuse, it also constrains their capacity for flexible, context-sensitive reasoning, particularly in unconventional but valid scenarios. In contrast, models that effectively reason with redefined values display greater adaptability and generalization, but at the risk of becoming more vulnerable to misleading or malicious prompts that exploit this openness. This trade-off raises a critical question: Should LLMs be optimized for strict factual reliability, even when that limits their reasoning flexibility, or should they remain open to alternative premises, despite the associated risks? Striking the right balance between these competing priorities is a central ethical and design challenge in the development of trustworthy and adaptive LLMs.

# Chapter 7

## Bibliography

- [1] AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., and Ghazvininejad, M. *A Review on Language Models as Knowledge Bases*. 2022. arXiv: [2204.06031 \[cs.CL\]](#). URL:
- [2] Argyrou, G., Dimitriou, A., Lymperaio, M., Filandrianos, G., and Stamou, G. “Automatic Generation of Fashion Images Using Prompting in Generative Machine Learning Models”. In: *Computer Vision – ECCV 2024 Workshops*. Ed. by A. Del Bue, C. Canton, J. Pont-Tuset, and T. Tommasi. Cham: Springer Nature Switzerland, 2025, pp. 286–302. ISBN: 978-3-031-91569-7.
- [3] Bai, Y. et al. *Benchmarking Foundation Models with Language-Model-as-an-Examiner*. 2023. arXiv: [2306.04181 \[cs.CL\]](#). URL:
- [4] Ball, T., Chen, S., and Herley, C. *Can We Count on LLMs? The Fixed-Effect Fallacy and Claims of GPT-4 Capabilities*. 2024. arXiv: [2409.07638 \[cs.AI\]](#). URL:
- [5] Banerjee, S. and Lavie, A. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. URL:
- [6] Bengio, Y., Ducharme, R., and Vincent, P. “A Neural Probabilistic Language Model”. In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000. URL:
- [7] Bowen, C., Sætre, R., and Miyao, Y. “A Comprehensive Evaluation of Inductive Reasoning Capabilities and Problem Solving in Large Language Models”. In: *Findings of the Association for Computational Linguistics: EACL 2024*. Ed. by Y. Graham and M. Purver. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 323–339. URL:
- [8] Brown, T. B. et al. *Language Models are Few-Shot Learners*. 2020. arXiv: [2005.14165 \[cs.CL\]](#). URL:
- [9] Cao, M., Lam, A., Duan, H., Liu, H., Zhang, S., and Chen, K. *CompassJudge-1: All-in-one Judge Model Helps Model Evaluation and Evolution*. 2024. arXiv: [2410.16256 \[cs.CL\]](#). URL:
- [10] Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. *ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate*. 2023. arXiv: [2308.07201 \[cs.CL\]](#). URL:
- [11] Chen, D., Chen, R., Zhang, S., Liu, Y., Wang, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., and Sun, L. *MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark*. 2024. arXiv: [2402.04788 \[cs.CL\]](#). URL:
- [12] Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. “Humans or LLMs as the Judge? A Study on Judgement Bias”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 8301–8327. DOI: [10.18653/v1/2024.emnlp-main.474](#). URL:
- [13] Chen, J., Pan, X., Yu, D., Song, K., Wang, X., Yu, D., and Chen, J. “Skills-in-Context: Unlocking Compositionality in Large Language Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA:



- Association for Computational Linguistics, Nov. 2024, pp. 13838–13890. DOI: [10.18653/v1/2024.findings-emnlp.812](https://doi.org/10.18653/v1/2024.findings-emnlp.812). URL:
- [14] Chen, J., Chen, L., Huang, H., and Zhou, T. *When do you need Chain-of-Thought Prompting for ChatGPT?* 2023. arXiv: [2304.03262](https://arxiv.org/abs/2304.03262) [cs.AI]. URL:
  - [15] Chen, M. et al. *Evaluating Large Language Models Trained on Code*. 2021. arXiv: [2107.03374](https://arxiv.org/abs/2107.03374) [cs.LG]. URL:
  - [16] Chen, M. et al. *Evaluating Large Language Models Trained on Code*. 2021. arXiv: [2107.03374](https://arxiv.org/abs/2107.03374) [cs.LG]. URL:
  - [17] Chen, M. et al. *Evaluating Large Language Models Trained on Code*. 2021. arXiv: [2107.03374](https://arxiv.org/abs/2107.03374) [cs.LG]. URL:
  - [18] Chen, M. K., Zhang, X., and Tao, D. *JustLogic: A Comprehensive Benchmark for Evaluating Deductive Reasoning in Large Language Models*. 2025. arXiv: [2501.14851](https://arxiv.org/abs/2501.14851) [cs.CL]. URL:
  - [19] Chen, X., Lin, M., Schärli, N., and Zhou, D. *Teaching Large Language Models to Self-Debug*. 2023. arXiv: [2304.05128](https://arxiv.org/abs/2304.05128) [cs.CL]. URL:
  - [20] Chhun, C., Colombo, P., Suchanek, F. M., and Clavel, C. “Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by N. Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 5794–5836. URL:
  - [21] Chiang, C.-H. and Lee, H.-y. *Can Large Language Models Be an Alternative to Human Evaluations?* 2023. arXiv: [2305.01937](https://arxiv.org/abs/2305.01937) [cs.CL]. URL:
  - [22] Chowdhery, A. et al. *PaLM: Scaling Language Modeling with Pathways*. 2022. arXiv: [2204.02311](https://arxiv.org/abs/2204.02311) [cs.CL]. URL:
  - [23] Chu, Z., Ai, Q., Tu, Y., Li, H., and Liu, Y. *PRE: A Peer Review Based Large Language Model Evaluator*. 2024. arXiv: [2401.15641](https://arxiv.org/abs/2401.15641) [cs.IR]. URL:
  - [24] Chung, H. W. et al. *Scaling Instruction-Finetuned Language Models*. 2022. arXiv: [2210.11416](https://arxiv.org/abs/2210.11416) [cs.LG]. URL:
  - [25] Clark, P., Tafjord, O., and Richardson, K. *Transformers as Soft Reasoners over Language*. 2020. arXiv: [2002.05867](https://arxiv.org/abs/2002.05867) [cs.CL]. URL:
  - [26] Dong, Q. et al. “A Survey on In-context Learning”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1107–1128. DOI: [10.18653/v1/2024.emnlp-main.64](https://doi.org/10.18653/v1/2024.emnlp-main.64). URL:
  - [27] Drexler, J. and Hilty, R. “Technical aspects of artificial intelligence: An understanding from an intellectual property law perspective”. In: (Jan. 2019).
  - [28] Duan, N., Tang, D., and Zhou, M. “Machine Reasoning: Technology, Dilemma and Future”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. Ed. by A. Villavicencio and B. Van Durme. Online: Association for Computational Linguistics, Nov. 2020, pp. 1–6. DOI: [10.18653/v1/2020.emnlp-tutorials.1](https://doi.org/10.18653/v1/2020.emnlp-tutorials.1). URL:
  - [29] Evangelatos, A., Filandrianos, G., Lymperaio, M., Voulodimos, A., and Stamou, G. *AILS-NTUA at SemEval-2025 Task 8: Language-to-Code prompting and Error Fixing for Tabular Question Answering*. 2025. arXiv: [2503.00435](https://arxiv.org/abs/2503.00435) [cs.CL]. URL:
  - [30] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. *SummEval: Re-evaluating Summarization Evaluation*. 2021. arXiv: [2007.12626](https://arxiv.org/abs/2007.12626) [cs.CL]. URL:
  - [31] Filandrianos, G., Dimitriou, A., Lymperaio, M., Thomas, K., and Stamou, G. *Bias Beware: The Impact of Cognitive Biases on LLM-Driven Product Recommendations*. 2025. arXiv: [2502.01349](https://arxiv.org/abs/2502.01349) [cs.CL]. URL:
  - [32] Fischer, K., Fürst, D., Steindl, S., Lindner, J., and Schäfer, U. *Question: How do Large Language Models perform on the Question Answering tasks? Answer*. 2024. arXiv: [2412.12893](https://arxiv.org/abs/2412.12893) [cs.CL]. URL:
  - [33] Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Foster, G., Lavie, A., and Bojar, O. “Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain”. In: *Proceedings of the Sixth Conference on Machine Translation*. Ed. by L. Barrault et al. Online: Association for Computational Linguistics, Nov. 2021, pp. 733–774. URL:
  - [34] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Derroncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. “Bias and Fairness in Large Language Models: A Survey”. In: *Computational Linguistics* 50.3 (Sept. 2024), pp. 1097–1179. DOI: [10.1162/coli\\_a\\_00524](https://doi.org/10.1162/coli_a_00524). URL:



- 
- [35] Giadikiaroglou, P., Lymperaiou, M., Filandrianos, G., and Stamou, G. “Puzzle Solving using Reasoning of Large Language Models: A Survey”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 11574–11591. DOI: [10.18653/v1/2024.emnlp-main.646](https://doi.org/10.18653/v1/2024.emnlp-main.646). URL:
- [36] Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tur, D. *Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations*. 2023. arXiv: [2308.11995](https://arxiv.org/abs/2308.11995) [cs.CL]. URL:
- [37] Gu, J. et al. *A Survey on LLM-as-a-Judge*. 2025. arXiv: [2411.15594](https://arxiv.org/abs/2411.15594) [cs.CL]. URL:
- [38] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. *Large Language Model based Multi-Agents: A Survey of Progress and Challenges*. 2024. arXiv: [2402.01680](https://arxiv.org/abs/2402.01680) [cs.CL]. URL:
- [39] He, H., Zhang, H., and Roth, D. *SocREval: Large Language Models with the Socratic Method for Reference-Free Reasoning Evaluation*. 2024. arXiv: [2310.00074](https://arxiv.org/abs/2310.00074) [cs.CL]. URL:
- [40] He, Q., Wang, Y., and Wang, W. *Can Language Models Act as Knowledge Bases at Scale?* 2024. arXiv: [2402.14273](https://arxiv.org/abs/2402.14273) [cs.CL]. URL:
- [41] He, Y., Kang, Y., Fan, L., and Yang, Q. *FedEval-LLM: Federated Evaluation of Large Language Models on Downstream Tasks with Collective Wisdom*. 2024. arXiv: [2404.12273](https://arxiv.org/abs/2404.12273) [cs.AI]. URL:
- [42] Hoffmann, J. et al. *Training Compute-Optimal Large Language Models*. 2022. arXiv: [2203.15556](https://arxiv.org/abs/2203.15556) [cs.CL]. URL:
- [43] Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., and Zhao, W. X. *Large Language Models are Zero-Shot Rankers for Recommender Systems*. 2024. arXiv: [2305.08845](https://arxiv.org/abs/2305.08845) [cs.IR]. URL:
- [44] Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., Qiu, T., Li, B., and Yang, Y. *PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference*. 2024. arXiv: [2406.15513](https://arxiv.org/abs/2406.15513) [cs.AI]. URL:
- [45] Jiang, J., Wang, F., Shen, J., Kim, S., and Kim, S. *A Survey on Large Language Models for Code Generation*. 2024. arXiv: [2406.00515](https://arxiv.org/abs/2406.00515) [cs.CL]. URL:
- [46] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* 2024. arXiv: [2310.06770](https://arxiv.org/abs/2310.06770) [cs.CL]. URL:
- [47] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* 2024. arXiv: [2310.06770](https://arxiv.org/abs/2310.06770) [cs.CL]. URL:
- [48] Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M., and Schölkopf, B. *Can Large Language Models Infer Causation from Correlation?* 2024. arXiv: [2306.05836](https://arxiv.org/abs/2306.05836) [cs.CL]. URL:
- [49] Kahneman, D. and Tversky, A. “On the Psychology of Prediction”. In: *Psychological Review* 80.4 (1973), pp. 237–251.
- [50] Kamalloo, E., Dziri, N., Clarke, C., and Rafiei, D. “Evaluating Open-Domain Question Answering in the Era of Large Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 5591–5606. DOI: [10.18653/v1/2023.acl-long.307](https://doi.org/10.18653/v1/2023.acl-long.307). URL:
- [51] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. *Scaling Laws for Neural Language Models*. 2020. arXiv: [2001.08361](https://arxiv.org/abs/2001.08361) [cs.LG]. URL:
- [52] Karpinska, M. and Iyyer, M. “Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist”. In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by P. Koehn, B. Haddow, T. Kočmi, and C. Monz. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 419–451. DOI: [10.18653/v1/2023.wmt-1.41](https://doi.org/10.18653/v1/2023.wmt-1.41). URL:
- [53] Khurana, D., Koli, A., Khatter, K., and Singh, S. “Natural language processing: state of the art, current trends and challenges”. In: *Multimedia Tools and Applications* 82.3 (July 2022), pp. 3713–3744. ISSN: 1573-7721. DOI: [10.1007/s11042-022-13428-4](https://doi.org/10.1007/s11042-022-13428-4). URL:
- [54] Kim, S. et al. *Prometheus: Inducing Fine-grained Evaluation Capability in Language Models*. 2024. arXiv: [2310.08491](https://arxiv.org/abs/2310.08491) [cs.CL]. URL:
- [55] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. *Large Language Models are Zero-Shot Reasoners*. 2023. arXiv: [2205.11916](https://arxiv.org/abs/2205.11916) [cs.CL]. URL:
-

- [56] Kondo, K., Sugawara, S., and Aizawa, A. “Probing Physical Reasoning with Counter-Commonsense Context”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 603–612. DOI: [10.18653/v1/2023.acl-short.53](https://doi.org/10.18653/v1/2023.acl-short.53). URL:
- [57] Kostina, A., Dikaiakos, M. D., Stefanidis, D., and Pallis, G. *Large Language Models For Text Classification: Case Study And Comprehensive Review*. 2025. arXiv: [2501.08457](https://arxiv.org/abs/2501.08457) [cs.CL]. URL:
- [58] Kritharoula, A., Lymperaioi, M., and Stamou, G. *Language Models as Knowledge Bases for Visual Word Sense Disambiguation*. 2023. arXiv: [2310.01960](https://arxiv.org/abs/2310.01960) [cs.CL]. URL:
- [59] Kritharoula, A., Lymperaioi, M., and Stamou, G. “Large Language Models and Multimodal Retrieval for Visual Word Sense Disambiguation”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13053–13077. DOI: [10.18653/v1/2023.emnlp-main.807](https://doi.org/10.18653/v1/2023.emnlp-main.807). URL:
- [60] Kudo, T. *Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates*. 2018. arXiv: [1804.10959](https://arxiv.org/abs/1804.10959) [cs.CL]. URL:
- [61] Kudo, T. and Richardson, J. *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. 2018. arXiv: [1808.06226](https://arxiv.org/abs/1808.06226) [cs.CL]. URL:
- [62] Lewis, M. and Mitchell, M. *Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models*. 2024. arXiv: [2402.08955](https://arxiv.org/abs/2402.08955) [cs.AI]. URL:
- [63] Li, C., Tian, Y., Zerong, Z., Song, Y., and Xia, F. “Challenging Large Language Models with New Tasks: A Study on their Adaptability and Robustness”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 8140–8162. DOI: [10.18653/v1/2024.findings-acl.485](https://doi.org/10.18653/v1/2024.findings-acl.485). URL:
- [64] Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., and Liu, Y. *LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods*. 2024. arXiv: [2412.05579](https://arxiv.org/abs/2412.05579) [cs.CL]. URL:
- [65] Li, J., Cao, P., Jin, Z., Chen, Y., Liu, K., and Zhao, J. *MIRAGE: Evaluating and Explaining Inductive Reasoning Process in Language Models*. 2025. arXiv: [2410.09542](https://arxiv.org/abs/2410.09542) [cs.CL]. URL:
- [66] Li, J., Yu, L., and Ettinger, A. *Counterfactual reasoning: Do language models need world knowledge for causal understanding?* 2022. arXiv: [2212.03278](https://arxiv.org/abs/2212.03278) [cs.CL]. URL:
- [67] Li, J., Yu, L., and Ettinger, A. *Counterfactual reasoning: Testing language models’ understanding of hypothetical scenarios*. 2023. arXiv: [2305.16572](https://arxiv.org/abs/2305.16572) [cs.CL]. URL:
- [68] Li, Q., Cui, L., Kong, L., and Bi, W. *Exploring the Reliability of Large Language Models as Customized Evaluators for Diverse NLP Tasks*. 2025. arXiv: [2310.19740](https://arxiv.org/abs/2310.19740) [cs.CL]. URL:
- [69] Li, R., Patel, T., and Du, X. *PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations*. 2024. arXiv: [2307.02762](https://arxiv.org/abs/2307.02762) [cs.CL]. URL:
- [70] Lin, B. Y., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R. L., and Choi, Y. *WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild*. 2024. arXiv: [2406.04770](https://arxiv.org/abs/2406.04770) [cs.CL]. URL:
- [71] Lin, C.-Y. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL:
- [72] Lin, Y.-T. and Chen, Y.-N. *LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models*. 2023. arXiv: [2305.13711](https://arxiv.org/abs/2305.13711) [cs.CL]. URL:
- [73] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. 2021. arXiv: [2107.13586](https://arxiv.org/abs/2107.13586) [cs.CL]. URL:
- [74] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. “G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. DOI: [10.18653/v1/2023.emnlp-main.153](https://doi.org/10.18653/v1/2023.emnlp-main.153). URL:

- 
- [75] Liu, Y., Zhou, H., Guo, Z., Shareghi, E., Vulić, I., Korhonen, A., and Collier, N. *Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators*. 2025. arXiv: [2403.16950 \[cs.CL\]](#). URL:
- [76] Lou, R., Zhang, K., and Yin, W. *Large Language Model Instruction Following: A Survey of Progresses and Challenges*. 2024. arXiv: [2303.10475 \[cs.CL\]](#). URL:
- [77] Lou, S., Chen, Y., Liang, X., Lin, L., and Zhang, Q. *Quantifying In-Context Reasoning Effects and Memorization Effects in LLMs*. 2024. arXiv: [2405.11880 \[cs.LG\]](#). URL:
- [78] Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. *Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity*. 2022. arXiv: [2104.08786 \[cs.CL\]](#). URL:
- [79] Madaan, A. et al. *Self-Refine: Iterative Refinement with Self-Feedback*. 2023. arXiv: [2303.17651 \[cs.CL\]](#). URL:
- [80] Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. *Dissociating language and thought in large language models*. 2024. arXiv: [2301.06627 \[cs.CL\]](#). URL:
- [81] Mao, Y., He, J., and Chen, C. *From Prompts to Templates: A Systematic Prompt Template Analysis for Real-world LLMapps*. 2025. arXiv: [2504.02052 \[cs.SE\]](#). URL:
- [82] McKenzie, I. R. et al. *Inverse Scaling: When Bigger Isn't Better*. 2024. arXiv: [2306.09479 \[cs.CL\]](#). URL:
- [83] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. *Large Language Models: A Survey*. 2025. arXiv: [2402.06196 \[cs.CL\]](#). URL:
- [84] Mishra, S. et al. *Lila: A Unified Benchmark for Mathematical Reasoning*. 2023. arXiv: [2210.17517 \[cs.CL\]](#). URL:
- [85] Nasrabadi, D. *JurEE not Judges: safeguarding llm interactions with small, specialised Encoder Ensembles*. 2024. arXiv: [2410.08442 \[cs.LG\]](#). URL:
- [86] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. *A Comprehensive Overview of Large Language Models*. 2024. arXiv: [2307.06435 \[cs.CL\]](#). URL:
- [87] Olsson, C. et al. *In-context Learning and Induction Heads*. 2022. arXiv: [2209.11895 \[cs.LG\]](#). URL:
- [88] Ouyang, L. et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: [2203.02155 \[cs.CL\]](#). URL:
- [89] Pagnoni, A., Balachandran, V., and Tsvetkov, Y. “Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Online: Association for Computational Linguistics, June 2021, pp. 4812–4829. DOI: [10.18653/v1/2021.naacl-main.383](#). URL:
- [90] Panagiotopoulos, I., Filandrianos, G., Lymperaio, M., and Stamou, G. “RISCORE: Enhancing In-Context Riddle Solving in Language Models through Context-Reconstructed Example Augmentation”. In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 9431–9455. URL:
- [91] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by P. Isabelle, E. Charniak, and D. Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: [10.3115/1073083.1073135](#). URL:
- [92] Pu, X., Gao, M., and Wan, X. *Summarization is (Almost) Dead*. 2023. arXiv: [2309.09558 \[cs.CL\]](#). URL:
- [93] Qian, S., Orasan, C., Kanojia, D., and Do Carmo, F. “Are Large Language Models State-of-the-art Quality Estimators for Machine Translation of User-generated Content?” In: *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*. Ed. by T. Nakazawa and I. Goto. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 45–55. DOI: [10.18653/v1/2024.wat-1.4](#). URL:
- [94] Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., and Chen, H. *Reasoning with Language Model Prompting: A Survey*. 2023. arXiv: [2212.09597 \[cs.CL\]](#). URL:
-

- [95] Qin, C., Xia, W., Wang, T., Jiao, F., Hu, Y., Ding, B., Chen, R., and Joty, S. *Relevant or Random: Can LLMs Truly Perform Analogical Reasoning?* 2024. arXiv: [2404.12728 \[cs.CL\]](#). URL:
- [96] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. “Language Models are Unsupervised Multitask Learners”. In: 2019. URL:
- [97] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2023. arXiv: [1910.10683 \[cs.LG\]](#). URL:
- [98] Raiaan, M., Mukta, S., Fatema, K., Fahad, N., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., and Azam, S. “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges”. In: *IEEE Access* PP (Jan. 2024), pp. 1–1. DOI: [10.1109/ACCESS.2024.3365742](#).
- [99] Raptopoulos, P., Filandrianos, G., Lymperaioi, M., and Stamou, G. *PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements*. 2025. arXiv: [2506.00608 \[cs.CL\]](#). URL:
- [100] Rasal, S. and Hauer, E. J. *Navigating Complexity: Orchestrated Problem Solving with Multi-Agent LLMs*. 2024. arXiv: [2402.16713 \[cs.MA\]](#). URL:
- [101] Rawte, V., Sheth, A., and Das, A. *A Survey of Hallucination in Large Foundation Models*. 2023. arXiv: [2309.05922 \[cs.AI\]](#). URL:
- [102] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. 2025. arXiv: [2402.07927 \[cs.AI\]](#). URL:
- [103] Sap, M., Shwartz, V., Bosselut, A., Choi, Y., and Roth, D. “Commonsense Reasoning for Natural Language Processing”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Ed. by A. Savary and Y. Zhang. Online: Association for Computational Linguistics, July 2020, pp. 27–33. DOI: [10.18653/v1/2020.acl-tutorials.7](#). URL:
- [104] Schmidt, R. M. *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. 2019. arXiv: [1912.05911 \[cs.LG\]](#). URL:
- [105] Schulhoff, S. et al. *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*. 2025. arXiv: [2406.06608 \[cs.CL\]](#). URL:
- [106] Sennrich, R., Haddow, B., and Birch, A. *Neural Machine Translation of Rare Words with Subword Units*. 2016. arXiv: [1508.07909 \[cs.CL\]](#). URL:
- [107] Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A. G., and Arawjo, I. *Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences*. 2024. arXiv: [2404.12272 \[cs.HC\]](#). URL:
- [108] Shen, Y. and Wan, X. *OpinSummEval: Revisiting Automated Evaluation for Opinion Summarization*. 2023. arXiv: [2310.18122 \[cs.CL\]](#). URL:
- [109] Sherstinsky, A. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network”. In: *Physica D: Nonlinear Phenomena* 404 (Mar. 2020), p. 132306. ISSN: 0167-2789. DOI: [10.1016/j.physd.2019.132306](#). URL:
- [110] Simmering, P. F. and Huoviala, P. *Large language models for aspect-based sentiment analysis*. 2023. arXiv: [2310.18025 \[cs.CL\]](#). URL:
- [111] Son, G., Yoon, D., Suk, J., Aula-Blasco, J., Aslan, M., Kim, V. T., Islam, S. B., Prats-Cristià, J., Tormo-Bañuelos, L., and Kim, S. *MM-Eval: A Multilingual Meta-Evaluation Benchmark for LLM-as-a-Judge and Reward Models*. 2025. arXiv: [2410.17578 \[cs.CL\]](#). URL:
- [112] Song, X., Salcianu, A., Song, Y., Dopson, D., and Zhou, D. *Fast WordPiece Tokenization*. 2021. arXiv: [2012.15524 \[cs.CL\]](#). URL:
- [113] Srivastava, A. et al. “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. In: *Trans. Mach. Learn. Res.* 2023 (2023). URL:
- [114] Staudemeyer, R. C. and Morris, E. R. *Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks*. 2019. arXiv: [1909.09586 \[cs.NE\]](#). URL:
- [115] Stevenson, C. E., Pafford, A., Maas, H. L. J. van der, and Mitchell, M. *Can Large Language Models generalize analogy solving like people can?* 2025. arXiv: [2411.02348 \[cs.AI\]](#). URL:
- [116] Stringli, E., Lymperaioi, M., Filandrianos, G., Voulodimos, A., and Stamou, G. *Pitfalls of Scale: Investigating the Inverse Task of Redefinition in Large Language Models*. 2025. arXiv: [2502.12821 \[cs.CL\]](#). URL:



- 
- [117] Sultan, O. and Shahaf, D. *Life is a Circus and We are the Clowns: Automatically Finding Analogies between Situations and Processes*. 2023. arXiv: [2210.12197 \[cs.CL\]](#). URL:
- [118] Talmor, A., Herzig, J., Lourie, N., and Berant, J. “CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4149–4158. DOI: [10.18653/v1/N19-1421](#). URL:
- [119] Tan, S., Zhuang, S., Montgomery, K., Tang, W. Y., Cuadron, A., Wang, C., Popa, R. A., and Stoica, I. *JudgeBench: A Benchmark for Evaluating LLM-based Judges*. 2025. arXiv: [2410.12784 \[cs.AI\]](#). URL:
- [120] Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. *Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges*. 2025. arXiv: [2406.12624 \[cs.CL\]](#). URL:
- [121] Thomas, K., Filandrianos, G., Lymperaious, M., Zerva, C., and Stamou, G. “‘I Never Said That’: A dataset, taxonomy and baselines on response clarity classification”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5204–5233. DOI: [10.18653/v1/2024.findings-emnlp.300](#). URL:
- [122] Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O’Sullivan, B., and Nguyen, H. D. *Multi-Agent Collaboration Mechanisms: A Survey of LLMs*. 2025. arXiv: [2501.06322 \[cs.AI\]](#). URL:
- [123] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. *Attention Is All You Need*. 2023. arXiv: [1706.03762 \[cs.CL\]](#). URL:
- [124] Vatsal, S. and Dubey, H. *A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks*. 2024. arXiv: [2407.12994 \[cs.CL\]](#). URL:
- [125] Vazquez, R. et al. *SemEval-2025 Task 3: Mu-SHROOM, the Multilingual Shared Task on Hallucinations and Related Observable Overgeneration Mistakes*. Apr. 2025. DOI: [10.48550/arXiv.2504.11975](#).
- [126] Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. *Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters*. 2023. arXiv: [2212.10001 \[cs.CL\]](#). URL:
- [127] Wang, C., Zhou, H., Chang, K., Liu, T., Zhang, C., Du, Q., Xiao, T., Zhang, Y., and Zhu, J. *Learning Evaluation Models from Large Language Models for Sequence Generation*. 2025. arXiv: [2308.04386 \[cs.CL\]](#). URL:
- [128] Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., and Zhou, J. *Is ChatGPT a Good NLG Evaluator? A Preliminary Study*. 2023. arXiv: [2303.04048 \[cs.CL\]](#). URL:
- [129] Wang, J. et al. *A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks*. 2024. arXiv: [2408.01319 \[cs.AI\]](#). URL:
- [130] Wang, T. and Lu, W. “Learning Multi-Step Reasoning by Solving Arithmetic Tasks”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1229–1238. DOI: [10.18653/v1/2023.acl-short.106](#). URL:
- [131] Wang, X., Antoniadis, A., Elazar, Y., Amayuelas, A., Albalak, A., Zhang, K., and Wang, W. Y. *Generalization v.s. Memorization: Tracing Language Models’ Capabilities Back to Pretraining Data*. 2025. arXiv: [2407.14985 \[cs.CL\]](#). URL:
- [132] Wang, Y. et al. *PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization*. 2024. arXiv: [2306.05087 \[cs.CL\]](#). URL:
- [133] Wang, Z. “CausalBench: A Comprehensive Benchmark for Evaluating Causal Reasoning Capabilities of Large Language Models”. In: *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*. Ed. by K.-F. Wong, M. Zhang, R. Xu, J. Li, Z. Wei, L. Gui, B. Liang, and R. Zhao. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 143–151. URL:
- [134] Wang, Z., Pang, Y., and Lin, Y. *Large Language Models Are Zero-Shot Text Classifiers*. 2023. arXiv: [2312.01044 \[cs.CL\]](#). URL:
- [135] Wang, Z., Pang, Y., Lin, Y., and Zhu, X. *Adaptable and Reliable Text Classification using Large Language Models*. 2024. arXiv: [2405.10523 \[cs.CL\]](#). URL:
-

- [136] Webster, J. J. and Kit, C. “Tokenization as the initial phase in NLP”. In: *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*. COLING ’92. Nantes, France: Association for Computational Linguistics, 1992, pp. 1106–1110. DOI: [10.3115/992424.992434](https://doi.org/10.3115/992424.992434). URL:
- [137] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. *Finetuned Language Models Are Zero-Shot Learners*. 2022. arXiv: [2109.01652](https://arxiv.org/abs/2109.01652) [cs.CL]. URL:
- [138] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: [2201.11903](https://arxiv.org/abs/2201.11903) [cs.CL]. URL:
- [139] Wei, J. et al. *Emergent Abilities of Large Language Models*. 2022. arXiv: [2206.07682](https://arxiv.org/abs/2206.07682) [cs.CL]. URL:
- [140] Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks*. 2024. arXiv: [2307.02477](https://arxiv.org/abs/2307.02477) [cs.CL]. URL:
- [141] Xie, C., Huang, Y., Zhang, C., Yu, D., Chen, X., Lin, B. Y., Li, B., Ghazi, B., and Kumar, R. *On Memorization of Large Language Models in Logical Reasoning*. 2025. arXiv: [2410.23123](https://arxiv.org/abs/2410.23123) [cs.CL]. URL:
- [142] Xie, Y. et al. *Improving Model Factuality with Fine-grained Critique-based Evaluator*. 2025. arXiv: [2410.18359](https://arxiv.org/abs/2410.18359) [cs.CL]. URL:
- [143] Yan, L., Qin, Z., Zhuang, H., Jagerman, R., Wang, X., Bendersky, M., and Oosterhuis, H. *Consolidating Ranking and Relevance Predictions of Large Language Models through Post-Processing*. 2024. arXiv: [2404.11791](https://arxiv.org/abs/2404.11791) [cs.IR]. URL:
- [144] Ye, H. and Ng, H. T. *Self-Judge: Selective Instruction Following with Alignment Self-Evaluation*. 2024. arXiv: [2409.00935](https://arxiv.org/abs/2409.00935) [cs.CL]. URL:
- [145] Ye, J. et al. *Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge*. 2024. arXiv: [2410.02736](https://arxiv.org/abs/2410.02736) [cs.CL]. URL:
- [146] Ye, Z., Li, X., Li, Q., Ai, Q., Zhou, Y., Shen, W., Yan, D., and Liu, Y. *Beyond Scalar Reward Model: Learning Generative Judge from Preference Data*. 2024. arXiv: [2410.03742](https://arxiv.org/abs/2410.03742) [cs.CL]. URL:
- [147] Yu, W., Jiang, M., Clark, P., and Sabharwal, A. *IfQA: A Dataset for Open-domain Question Answering under Counterfactual Presuppositions*. 2023. arXiv: [2305.14010](https://arxiv.org/abs/2305.14010) [cs.CL]. URL:
- [148] Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., and Yu, D. *MM-LLMs: Recent Advances in MultiModal Large Language Models*. 2024. arXiv: [2401.13601](https://arxiv.org/abs/2401.13601) [cs.CL]. URL:
- [149] Zhang, K., Yuan, S., and Zhao, H. *TALEC: Teach Your LLM to Evaluate in Specific Domain with In-house Criteria by Criteria Division and Zero-shot Plus Few-shot*. 2024. arXiv: [2407.10999](https://arxiv.org/abs/2407.10999) [cs.CL]. URL:
- [150] Zhang, S. et al. *Instruction Tuning for Large Language Models: A Survey*. 2024. arXiv: [2308.10792](https://arxiv.org/abs/2308.10792) [cs.CL]. URL:
- [151] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. “Benchmarking Large Language Models for News Summarization”. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 39–57. DOI: [10.1162/tac1\\_a\\_00632](https://doi.org/10.1162/tac1_a_00632). URL:
- [152] Zhang, W., Deng, Y., Liu, B., Pan, S., and Bing, L. “Sentiment Analysis in the Era of Large Language Models: A Reality Check”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by K. Duh, H. Gomez, and S. Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 3881–3906. DOI: [10.18653/v1/2024.findings-naacl.246](https://doi.org/10.18653/v1/2024.findings-naacl.246). URL:
- [153] Zhang, Y. et al. *Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. 2023. arXiv: [2309.01219](https://arxiv.org/abs/2309.01219) [cs.CL]. URL:
- [154] Zhang, Y., Yasunaga, M., Zhou, Z., HaoChen, J. Z., Zou, J., Liang, P., and Yeung, S. *Beyond Positive Scaling: How Negation Impacts Scaling Trends of Language Models*. 2023. arXiv: [2305.17311](https://arxiv.org/abs/2305.17311) [cs.CL]. URL:
- [155] Zhang, Z., Zhang, A., Li, M., and Smola, A. *Automatic Chain of Thought Prompting in Large Language Models*. 2022. arXiv: [2210.03493](https://arxiv.org/abs/2210.03493) [cs.CL]. URL:
- [156] Zhao, W. X. et al. *A Survey of Large Language Models*. 2025. arXiv: [2303.18223](https://arxiv.org/abs/2303.18223) [cs.CL]. URL:
- [157] Zheng, L. et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. arXiv: [2306.05685](https://arxiv.org/abs/2306.05685) [cs.CL]. URL:
- [158] Zhou, Y., Li, J., Xiang, Y., Yan, H., Gui, L., and He, Y. “The Mystery of In-Context Learning: A Comprehensive Survey on Interpretation and Analysis”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen.

- 
- Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 14365–14378. DOI: [10.18653/v1/2024.emnlp-main.795](https://doi.org/10.18653/v1/2024.emnlp-main.795). URL:
- [159] Zhu, K. et al. *PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts*. 2024. arXiv: [2306.04528](https://arxiv.org/abs/2306.04528) [cs.CL]. URL:
- [160] Zhu, L., Wang, X., and Wang, X. *JudgeLM: Fine-tuned Large Language Models are Scalable Judges*. 2025. arXiv: [2310.17631](https://arxiv.org/abs/2310.17631) [cs.CL]. URL:
- [161] Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. “Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by K. Duh, H. Gomez, and S. Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 2765–2781. DOI: [10.18653/v1/2024.findings-naacl.176](https://doi.org/10.18653/v1/2024.findings-naacl.176). URL:
- [162] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. *A Comprehensive Survey on Transfer Learning*. 2020. arXiv: [1911.02685](https://arxiv.org/abs/1911.02685) [cs.LG]. URL:
- [163] Zhuang, S., Zhuang, H., Koopman, B., and Zuccon, G. “A Setwise Approach for Effective and Highly Efficient Zero-shot Ranking with Large Language Models”. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR 2024. ACM, July 2024, pp. 38–47. DOI: [10.1145/3626772.3657813](https://doi.org/10.1145/3626772.3657813). URL:
- [164] Zhuge, M. et al. *Agent-as-a-Judge: Evaluate Agents with Agents*. 2024. arXiv: [2410.10934](https://arxiv.org/abs/2410.10934) [cs.AI]. URL:
- [165] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. *Fine-Tuning Language Models from Human Preferences*. 2020. arXiv: [1909.08593](https://arxiv.org/abs/1909.08593) [cs.CL]. URL: