Flying Pigs, *FaR* and Beyond: Evaluating LLM Reasoning in Counterfactual Worlds

Anonymous ACL submission

Abstract

A fundamental challenge in reasoning is navigating hypothetical, counterfactual worlds where logic may conflict with ingrained knowledge. We investigate this frontier for Large Language Models (LLMs) by asking: Can LLMs reason logically when the context contradicts their parametric knowledge? To facilitate a systematic analysis, we first introduce *CounterLogic*, a benchmark specifically designed to disentangle logical validity from knowledge alignment. Evaluation of 11 LLMs across six diverse reasoning datasets reveals a consistent failure: model accuracy plummets by an average of 14% in counterfactual scenarios compared to knowledge-aligned ones. We hypothesize that this gap stems not from a flaw in logical processing, but from an inability to manage the cognitive conflict between context and knowledge. Inspired by human metacognition, we propose a simple yet powerful intervention: *Flag & Reason (FaR)*, where models are first prompted to *flag* potential knowledge conflicts *before* they reason. This metacognitive step is highly effective, narrowing the performance gap to just 7% and increasing overall accuracy by 4%. Our findings diagnose and study a critical limitation in modern LLMs' reasoning and demonstrate how metacognitive awareness can make them more robust and reliable thinkers[1].

1 Introduction

LLMs have demonstrated remarkable reasoning capabilities across diverse domains, exhibiting proficiency in tasks ranging from elementary problem solving to complex multi-step reasoning challenges (Brown et al., 2020; Kojima et al., 2022; Schick and Schütze, 2021; Wei et al., 2023; Zhou et al., 2023a; Patel et al., 2024). Despite these advances, they often exhibit a significant performance degradation

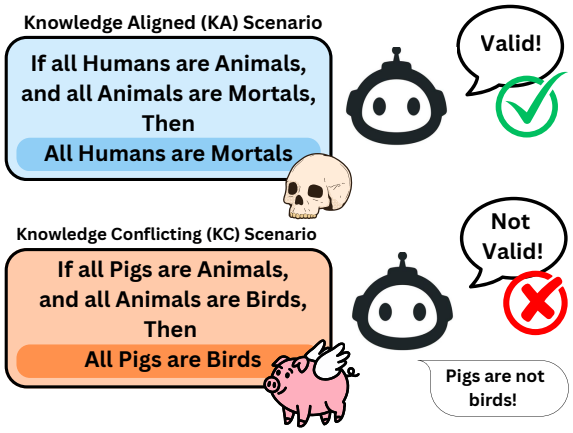


Figure 1: When asked to reason solely based on the given premises, LLMs reliably validate arguments that are **knowledge-aligned** (top), but their performance degrades on logically identical arguments that are **knowledge-conflicting** (bottom). This is because their parametric knowledge interferes with the reasoning process (e.g., "Pigs are not birds").

(see Fig 1) when reasoning with information that conflicts with their parametric knowledge (knowledge conflicts) obtained during training (Dasgupta et al., 2024; Jin et al., 2024; Lampinen et al., 2024; Su et al., 2024; Wang et al., 2024; Wu et al., 2024; Xu et al., 2024b).

The ability to reason effectively in scenarios with potentially conflicting information is crucial for deploying LLMs in real-world applications where they must process information that may be novel, unexpected, or even contradictory to their training data (Wang et al., 2024). For instance, engaging in scientific exploration or understanding historical debates requires reasoning from a counterfactual premise, such as: "Assuming the sun revolves around the Earth, describe the predicted path of Mars in the night sky." Such situations are central to creative and scientific thought (Pearl and Mackenzie, 2018), and failure to reason with them could lead to unreliable performance (Zhang

[1]Our data and code are available at https://anonymous.4open.science/r/CounterLogic-812C

et al., 2024b). Additionally, prior research also suggests that evaluating reasoning in counterfactual situations may serve as a more robust assessment of a model's reasoning capabilities (Wu et al., 2024), as standard reasoning tasks can potentially be *hacked* through pattern matching (Lewis and Mitchell, 2024; Liu et al., 2024; Kaushik et al., 2020; McCoy et al., 2019; Wu et al., 2024).

While knowledge conflicts are actively studied, prior investigations have focused on relatively simple tasks involving information extraction or single-step reasoning (example: Who is the current president of the USA?) (Xie et al., 2024b). Consequently, existing benchmarks either test complex logical reasoning without controlling for knowledge conflicts, or test simple knowledge conflicts with shallow, single-step reasoning. As summarized in Table 1, no benchmark has been designed to systematically probe multi-step logical reasoning under the stress of counterfactuals.

To address this gap, we introduce *CounterLogic*, a benchmark designed to disentangle logical validity from knowledge alignment. Our framework is easily extendable and currently implements 9 logical schemas with problems that systematically scale in complexity from 1 to 9 reasoning steps. Using *CounterLogic* and five other datasets (see Table 1), our large-scale evaluation of 11 LLMs reveals a pervasive failure: model accuracy plummets by an average of 14% in counterfactual scenarios. Further analysis reveals that this performance gap is not exacerbated by the number of reasoning steps, but is instead critically dependent on the logical schema. Specifically, the gap soars past 30% on tasks requiring negation, while being minimal on simpler forms.

We hypothesize that this failure stems not just from an inability to reason, but from an inability to manage the cognitive conflict that arises when premises contradict parametric knowledge. Inspired by human metacognition, the ability to reflect on one's own thinking to resolve such conflicts (Fletcher and Carruthers, 2012), we propose a simple yet powerful intervention called *Flag & Reason (FaR)*. Before reasoning, we prompt the model to first *flag* whether a conclusion is factually believable, forcing it to confront the knowledge conflict. Our experiments show *FaR* is highly effective, narrowing the performance gap to just 7% and increasing overall accuracy by 4% on average. Our analysis of the models' chain-of-thought reveals that this intervention works by anchoring the reasoning process in the initial acknowledgment of the conflict, preventing knowledge bias from derailing the subsequent logical steps.

Dataset	Size	# Steps	Conflict	Balance
LogicBench (Parmar et al., 2024)	2,020	$1 \sim 5$	×	×
FOLIO (Han et al., 2024)	1,435	$0 \sim 7$	×	×
KNOT (Liu et al., 2024)	5,500	$1 \sim 2$	✓	×
Syllogistic (Bertolazzi et al., 2024)	2,120	2	✓	×
CounterLogic (Ours)	1,800*	$2 \sim 9$*	✓	✓

Table 1: Comparison of logical reasoning benchmarks. *CounterLogic* uniquely combines multi-step reasoning with knowledge-conflicting scenarios while maintaining balance in labels. This enables rigorous evaluation of how parametric knowledge affects LLMs' logical reasoning capabilities, addressing limitations in existing benchmarks that typically lack proper balance across important evaluation dimensions. **We provide a dataset generation framework and the numbers can be easily changed as per the requirements.*

In this paper, we first use *CounterLogic* (Section 3) and other benchmarks to diagnose a pervasive failure in LLM reasoning (Section 4). We then propose a metacognitive intervention, *FaR*, to address this failure (Section 5) and analyse its effectiveness.

Our contributions can be summarized as follows:

1. We introduce *CounterLogic*, a balanced benchmark for evaluating multi-step logical reasoning in knowledge-conflicting scenarios.

2. We quantify and analyse a pervasive reasoning failure in LLMs: a significant and systematic performance drop when logical validity clashes with a model's parametric knowledge.

3. We propose and discuss *Flag & Reason (FaR)*, a simple, metacognitive intervention that significantly mitigates this failure, narrowing the identified performance gap and increasing overall accuracy.

2 Related Work

2.1 Logical Reasoning in LLMs

Recent advancements have significantly improved the reasoning capabilities of LLMs . Techniques such as chain-of-thought prompting guide models to articulate intermediate reasoning steps (Wei et al., 2023), zero-shot reasoning allows them to solve problems without task-specific examples (Kojima et al., 2022), and tree-of-thought methods enable the exploration of multiple reasoning paths

(Shinn et al., 2023). While these approaches have improved performance on various benchmarks (Clark et al., 2020; Parmar et al., 2024), studies reveal that LLMs still exhibit systematic errors that often mirror human cognitive biases (Dasgupta et al., 2024; Eisape et al., 2024).

Specifically, research into the limitations of logical reasoning shows that models struggle with operations involving negation, quantifiers, and abstract variables (Bertolazzi et al., 2024; Dasgupta et al., 2024; Singh et al., 2024). Performance particularly degrades when reasoning with counterfactual information that contradicts their training data (Chen et al., 2025; Wu et al., 2024), and they often handle logically equivalent problems inconsistently when presented in different formats (Estermann et al., 2025). Current approaches to address these issues include integrating symbolic representations into the reasoning chain (Xu et al., 2024a) and using verification mechanisms to validate reasoning against formal rules (Toroghi et al., 2024; Vacareanu and Ballesteros, 2024).

2.2 Knowledge Conflicts and Counterfactual Reasoning

LLMs store vast amounts of factual knowledge within their parameters, which can create challenges when they encounter conflicting information in a given context (Petroni et al., 2019; Roberts et al., 2020). These knowledge conflicts have been categorized based on their origin, such as between the context and the model's parametric memory (Xu et al., 2024b). Larger models often default to their parametric knowledge over conflicting contextual evidence, though this can vary based on the coherence and perceived reliability of the source (Longpre et al., 2021; Wang et al., 2023).

This tension is especially pronounced in counterfactual reasoning, where models must follow premises that explicitly contradict established facts (Chen et al., 2025; Wei et al., 2023). The performance drop in these scenarios is primarily attributed to the conflict between the model's stored knowledge and the counterfactual assertions it is asked to accept (Lin et al., 2025). Existing mitigation strategies often involve data augmentation with counterfactual examples or specialized prompting techniques (Chen et al., 2023b; Neeman et al., 2023; Xie et al., 2024b). However, these methods have not fully resolved the core issue, particularly in tasks requiring multi-step logical deduction (Patel et al., 2024).

2.3 Metacognition and Human Reasoning

Human reasoning is subject to well-documented cognitive biases, such as the "belief bias," where the believability of a conclusion influences the judgment of an argument's logical validity (Markovits and Nantel, 1989). This bias can create an "illusion of objectivity," where individuals are confident in their biased reasoning (Kunda, 1990; Trippas et al., 2014). LLMs exhibit similar patterns, performing better when a statement's content aligns with the knowledge they gain when training i.e. their parametric knowledge and struggling with implausible scenarios, thus mirroring the systematic errors seen in humans (Dasgupta et al., 2024; Eisape et al., 2024; Macmillan-Scott and Musolesi, 2024).

In humans, metacognitive strategies can improve logical reasoning by creating a distinction between evaluating a belief and making a logical inference. In essence, separating "what I know" from "what follows logically" (Douven et al., 2018). Similar metacognitive-like capabilities are an emerging area of research for LLMs. These include prompting models to estimate their own uncertainty (Zhou et al., 2023b), perform self-evaluation to critique their own reasoning (Wang et al., 2024), or explicitly identify when a premise conflicts with their knowledge (Chen et al., 2023a). This parallel suggests a promising direction for enhancing the robustness of logical reasoning in language models, which is the approach our work investigates.

3 The CounterLogic Dataset

Existing benchmarks in evaluating LLM reasoning (Kojima et al., 2022; Wei et al., 2023) fail to systematically disentangle logical validity from knowledge alignment. As highlighted in Table 1, current datasets either focus on complex logical structures while ignoring knowledge conflicts like LogicBench (Parmar et al., 2024) and FOLIO (Han et al., 2024), or test simple knowledge conflicts single-step reasoning like KNOT (Liu et al., 2024) and Reasoning & Reciting (Wu et al., 2024).

To fill this critical gap, we introduce *CounterLogic*, a benchmark built upon a synthetic generation framework designed to enable a controlled evaluation of how parametric knowledge interferes with mulit-step reasoning. The benchmark is easily extendable, contains 1,800 examples across 9 logical schemas shown in Table 2, and features problems that scale in complexity from 1 to 9 reasoning steps.

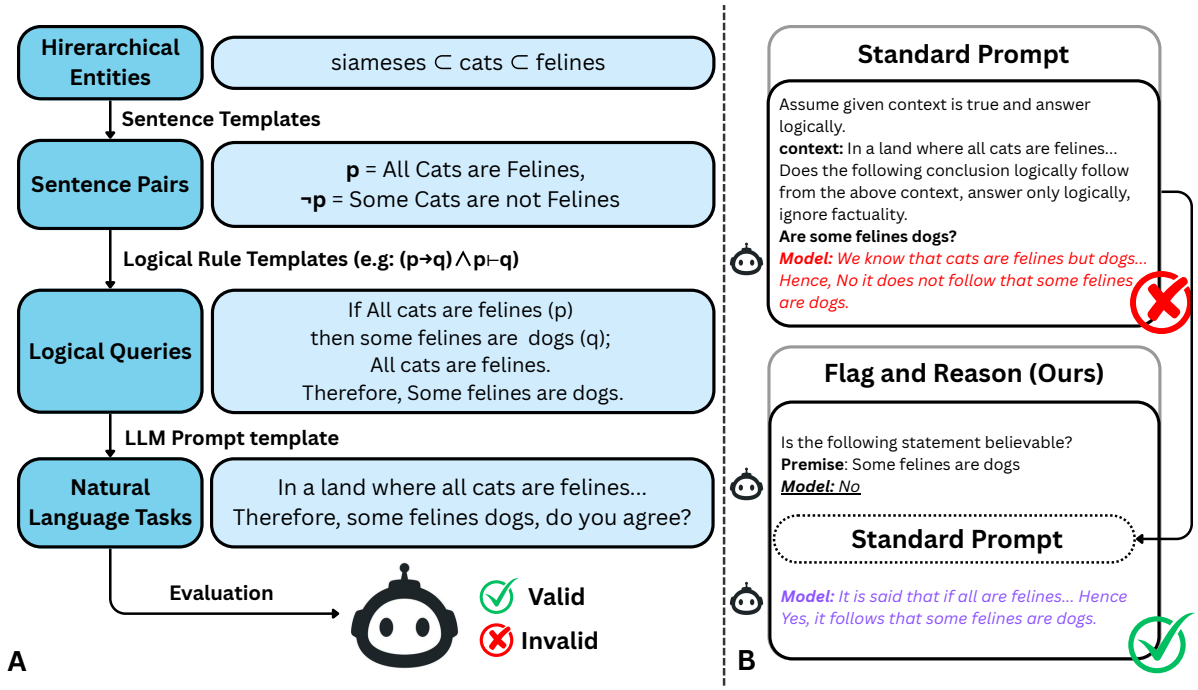


Figure 2: (A) **Dataset Preparation**: The dataset features hierarchical entity triples (e.g., siameses ⊂ cats ⊂ felines) mapped to 8 logical sentence templates across 9 logical schemas (see Appendix A). Each example is balanced across validity (50% valid/invalid) and believability (50% aligned/conflicting), with ground truth annotations for both dimensions. The dataset construction combines subset relationships with propositional logic forms such as Modus Ponens, Hypothetical Syllogism, etc. to systematically evaluate knowledge-logic interactions. (B) **FaR method**: While the standard prompt simply presents LLMs with a counterfactual context followed by related questions, our *FaR* approach first engages the model metacognitively by eliciting its responses to knowledge-alignment questions. This could be as simple as asking whether a given statement is true.

3.1 Dataset Construction

Hierarchical Entity Perturbation: We begin with a set of hierarchical entity triples (a, b, c) representing strict subset relationships ($a \subset b \subset c$), such as *siameses* $\subset$ *cats* $\subset$ *felines*.

Sentence Generation: Entities are inserted into four templates that form complementary logical pairs (S and $\neg S$), such as "All $\{A\}$ are $\{B\}$" and "Some $\{A\}$ are not $\{B\}$". This process yields atomic propositions with controlled factual accuracy. To ensure diversity, we balance the entity relationships: 50% use a correct hierarchy (e.g., *siameses* $\subset$ *cats*) and 50% use an inverted one (e.g., *cats* $\subset$ *siameses*). More details about the sentence generation are in Appendix A.1.

Logical Schema Integration: Inspired by LogicBench (Parmar et al., 2024), the sentence pairs generated are then integrated as propositions into 9 formal logical schemas: Modus Ponens (MP), Modus Tollens (MT), Hypothetical Syllogism (HS), Disjunctive Syllogism (DS), Constructive Dilemma (CD), Destructive Dilemma (DD), Bidirectional Dilemma (BD), Commutation (CT), and Material Implication (MI) (see Table 2 for formal definitions).

Balanced Label Assignment: To enable a rigorous and unbiased evaluation, we ensure the dataset is balanced by creating binary question-answer tasks that systematically control two key dimensions: **(1) Logical Validity:** We generate logically *Valid* instances where the conclusion logically follows from the premises, and *Invalid* instances where the conclusion is a non-sequitur, violating the logical form. **(2) Knowledge Alignment:** Similar to the examples in Fig 1, we generate *Knowledge-Aligned* instances where the conclusion is factually correct (aligned with parametric knowledge) based on the ground-truth entity hierarchies, and *Knowledge-Conflicting* instances where the conclusion is factually incorrect (contradicting parametric knowledge).

Examples for the above are given in the Appendix A.3. To prevent models from simply memorizing logical forms (Xie et al., 2024a; Wu et al., 2024), we use GPT-4o to paraphrase the final logical queries into natural language questions (Appendix A.5).

Varying Reasoning Depth: For schemas involving implication ($A \rightarrow B$), we create deeper reasoning chains by introducing intermediate steps:

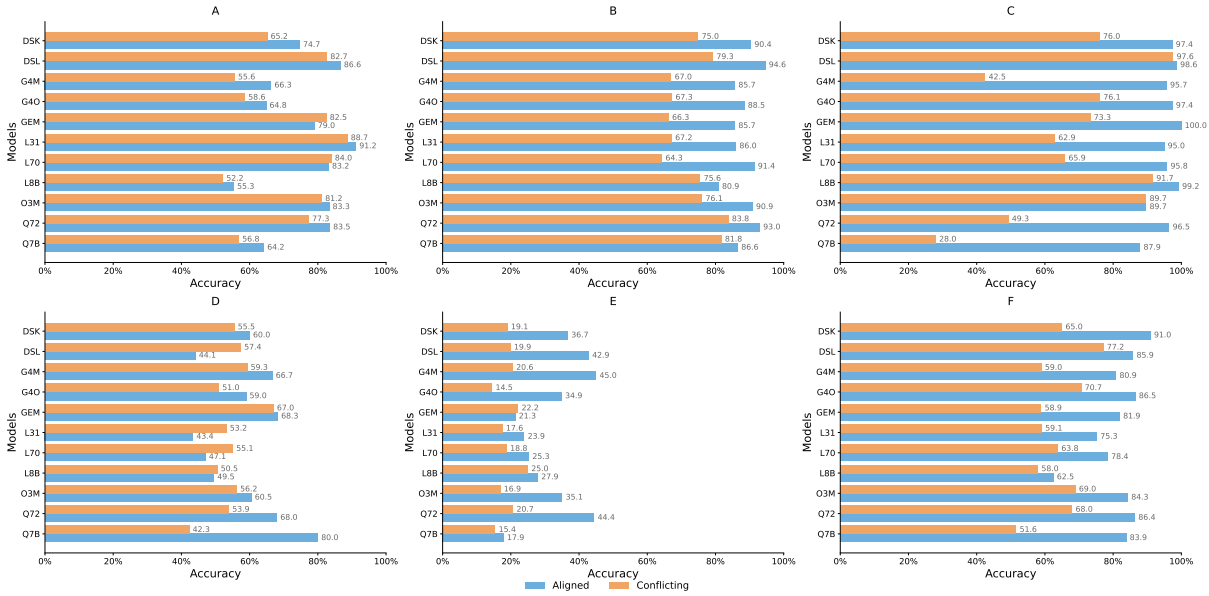


Figure 3: Average accuracy comparison for the Aligned and Conflicting Scenarios. This plot shows comparison between Knowledge-Aligned and Knowledge-Conflicting scenarios where the ground truth is valid. Blue bars represent Knowledge-Aligned example accuracy, while orange bars indicate Knowledge-Conflicting ones. We can see a clear performance gap between these scenarios across all the models and datasets with the average accuracy difference of 14% across all models and datasets with Hierarchical Syllogisms (C) showing the maximum difference of 27% (Table 6). y-axis represents the following models - DSK: DeepSeek Chat, DSL: Deepseek R1 Distill Llama 70B, G4M: GPT-4o Mini, G4O: GPT-4o, GEM: Google Gemini Flash 1.5, L8B: Llama 3.1 8B Instruct, L31: Llama 3.1 70B Instruct, L70: Llama 3.3 70B Instruct, O3M: o3-mini, Q7B: Qwen 2.5 7B Instruct, and Q72: Qwen 2.5 72B Instruct. Each of the above graphs represents the performance on the following datasets - A: CounterLogic, B: FOLIO, C: Hierarchical Syllogisms, D: KNOT-Explicit, E: KNOT-Implicit, F: Logic Bench.

$A \rightarrow X_1 \rightarrow \cdots \rightarrow X_i \rightarrow B$. This allows for a fine-grained analysis of how reasoning depth interacts with the challenges posed by knowledge conflicts, a feature absent in prior benchmarks (see Appendix A.4).

4 LLM Reasoning in Counterfactuals

4.1 Experimental Setup

Our investigation evaluates 11 LLMs spanning different architectures, parameter scales, and training paradigms (see Appendix E). Evaluations were conducted across our *CounterLogic* and five other diverse reasoning datasets: *Hierarchical Syllogisms*, *KNOT (Implicit & Explicit)*, *FOLIO*, and *LogicBench* (see Table 1). To ensure robust performance measures, we employ self-consistency checks accounting for the generation variability common in LLMs (Bonagiri et al., 2024). More deatils regarding the datasets and their evaluations is given in Appendix D.

Results presented in this sections establish a baseline using a standard CoT prompting strategy (Wei et al., 2023) without our *FaR* intervention, as this allows us to isolate and quantify the interference of parametric knowledge on logical deduction.

4.2 Consistent Performance Gap

As shown in Figure 3, our evaluation revealed a pervasive performance gap across all models and tasks. When reasoning in counterfactual scenarios, where the logical context conflicts with their parametric knowledge, model accuracy plummets by an average of 14%. Models achieve a high accuracy of 72% on average for knowledge-aligned problems, but this value drops to just 58% for counterfactual problems (Table 6). Notably, this failure occurs even when models are explicitly instructed to reason based solely on the provided premises and ignore factual correctness (see Appendix B). This highlights a challenge in managing cognitive conflict rather than a simple failure to follow instructions. Even the most capable models exhibit this gap; for instance, Qwen-2-72B's accuracy difference is 20% in the baseline setup. This suggests the interference from parametric knowledge is a fundamental challenge for current LLM reasoning.

4.3 Deconstructing the Failure

Is this performance degradation a superficial artifact of problem complexity or specific logical forms? We leveraged the controlled design of the *CounterLogic* benchmark to investigate this, find-

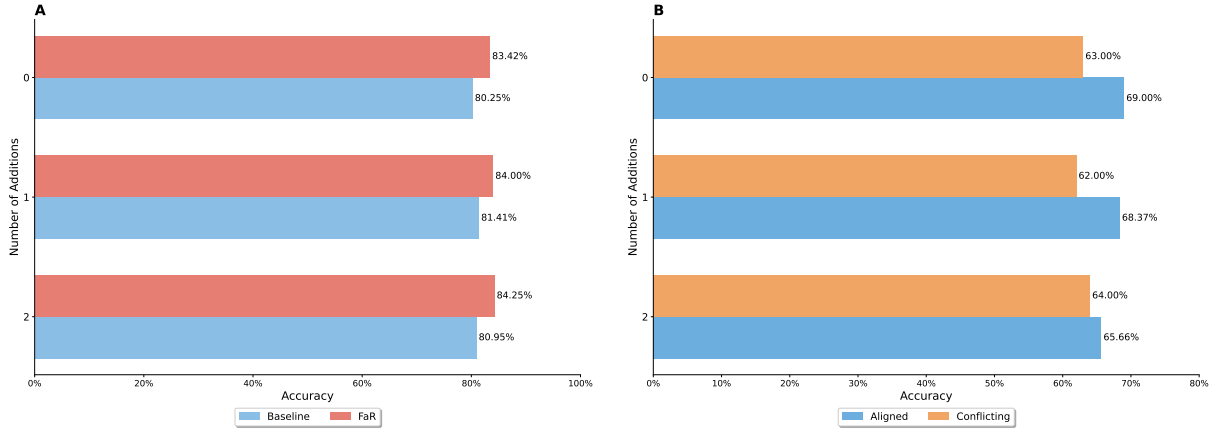


Figure 4: Variation across number of additions. Shows the performance across the number of additions (y-axis) that are done to all the schemas that have an implication statement either in their premise or in their conclusion (Section 3.1). The x-axis represents the accuracy values. **A** shows the comparison between *FaR* and baseline approaches. We can see that *FaR* improves performance in all the cases like in the case of 0 additions where performance improves from 80.22% to 83.4%. **B** shows the variation of Aligned and Conflicting scenarios across number of additions on valid data. We can see a pervasive performance gap between and Knowledge-Aligned and Knowledge-Conflicting data in all cases like in the case of 0 additions where the gap is about 6%. However increasing the depth of reasoning doesn't result in a significant performance drop.

ing that the failure is more fundamental, affecting reasoning regardless of depth or logical structure.

A common hypothesis (Wu et al., 2024) is that models falter simply because counterfactual problems are more complex. Our analysis, however, refutes this by showing that the performance gap between knowledge-aligned and counterfactual examples is largely independent of the reasoning depth. As shown in Figure 4, the core issue persists even as we increase the reasoning depth by introducing additional reasoning steps to the logical chain. We therefore hypothesize that the failure is not caused by models losing track of longer deductions, but rather by the initial, unresolved conflict with their parametric knowledge.

Another possibility is that the performance drop is isolated to specific, difficult logical rules. Our results, shown in Figure 5(B), suggest a more nuanced reality. While the failure is a universal phenomenon across all 9 logical schemas in *Counter-Logic*, the degree of degradation varies dramatically, revealing critical vulnerabilities.

On simple, affirmative logical forms like Modus Ponens (MP) and Hypothetical Syllogism (HS), models are not only accurate but also more robust; the performance gap between knowledge-aligned and conflicting scenarios is minimal. However, this stability shatters when schemas introduce negation or greater structural complexity. On these tasks, the performance gap widens significantly, indicating that the model's parametric knowledge is interfering more strongly. For instance, the gap soars to over 30 percentage points for Disjunctive Syllogism (DS) and Destructive Dilemma (DD), both of which require reasoning over negated propositions. This suggests that logical complexity and negation don't just make the task harder—they specifically amplify the model's inability to inhibit its ingrained knowledge.

This systemic weakness in handling negation is also reflected in overall performance. Schemas involving negation have a baseline accuracy of only 73%, 11% lower than the 84% achieved on those without (Figure 5(A)), aligning with findings that negation remains a core challenge for LLMs (Vrabcová et al., 2025).

Ultimately, our analysis shows the failure is not confined to one type of logic. Instead, it is a foundational issue where the interference from parametric knowledge is severely exacerbated by structural complexity and the presence of negation.

5 Flag & Reason (FaR): A MetaCognitive Intervention

Findings in Section 4 reveal a critical vulnerability in LLM reasoning: a systematic failure not of logic, but of cognitive management. In this section, we articulate our hypothesis for this failure, introduce the *FaR* intervention designed to mitigate it, and present the empirical results that validate its effectiveness.

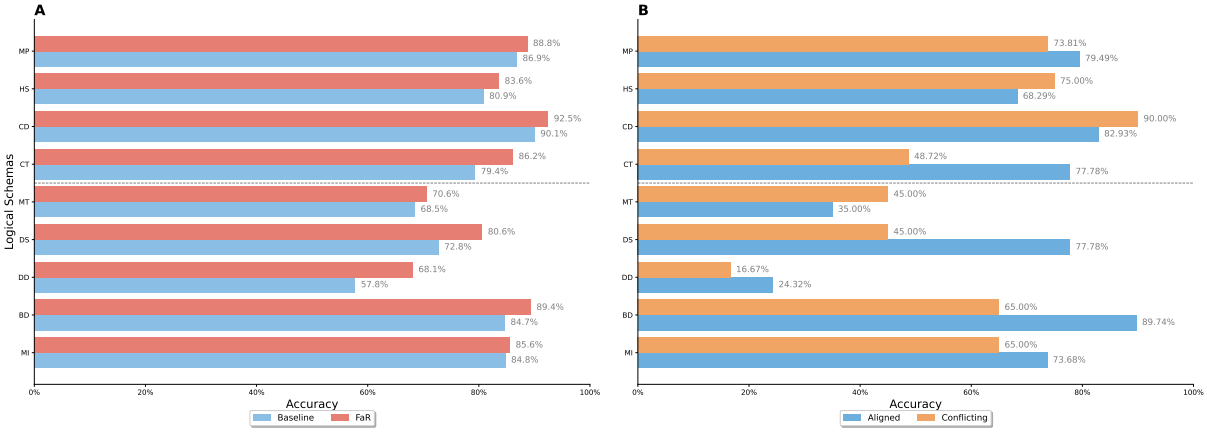


Figure 5: Variation across logical schemas. We can see the logical schemas (Table 2) on the y-axis and the accuracies on the x-axis. Dotted line separates the schemas with (below) and without (top) negations. **A** shows the comparison between the *FaR* and baseline accuracies. Pink bars represent *FaR* method and blue bars represent the baseline. We can see that *FaR* works across all schemas and improvement is more for the schemas that are harder (lower initial scores) like schema DD where the performance improves by about 11%. We can also see how LLMs perform noticeably worse on tasks requiring understanding of negations with average performance dropping from 84.29% above the line to 73.72% below the line on the baseline method. **B** shows the performance difference of Knowledge-Aligned and Knowledge-Conflicting scenarios for valid datapoints. The blue bars represent aligned accuracies and the orange bars represent the conflicting accuracies. We see can see a clear performance gap between aligned and conflicting datapoints most noticeable in schema DS where the gap is about 32%.

5.1 The Belief Inhibition Hypothesis

The consistent performance gap between knowledge-aligned and counterfactual scenarios suggests that the primary issue is not a flaw in the LLMs' capacity for logical operations. Instead, we hypothesize that they suffer from a failure of *belief inhibition*: an inability to suppress ingrained parametric knowledge when it contradicts the premises of a given task, leading to an unresolved cognitive conflict (Harmon-Jones, 2019) that ultimately derails the deductive process.

This phenomenon closely mirrors a well-known cognitive bias in humans known as "belief bias", where arguments are often judged on the believability of their conclusion rather than the logical validity (Markovits and Nantel, 1989). Humans, however, can overcome this bias through metacognition—the ability to reflect on one's own thinking. This involves consciously recognizing a conflict between a belief and a premise and choosing to set that belief aside to follow the rules of logic (Fletcher and Carruthers, 2012; Ma et al., 2025; Wang and Zhao, 2024).

5.2 The Flag & Reason (FaR) Intervention

Inspired by the human capability to recognize and set aside knowledge conflicts, we propose a simple yet powerful metacognitive intervention: **Flag & Reason (FaR)**. Instead of asking the model to solve a logical problem at once, *FaR* introduces a preliminary metacognitive step. The process, illustrated in Figure 2(B), consists of two phases:

1. **Flag:** The model is prompted to first explicitly *flag* whether a key premise or conclusion aligns with its parametric knowledge (i.e., whether it is factually believable).

2. **Reason:** After this initial assessment, the model proceeds with the standard reasoning process, now primed by the "Flag" step to treat the premises as a self-contained logical context.

The seperation of these two steps is critial. When we collapsed this process into a single prompt (asking the model to flag the conflict and reason in the same turn), the performance benefits disappeared entirely, yielding no improvement over the baseline (see Appendix H.4). This finding supports our central hypothesis: FaR works by inducing a form of **epistemic compartmentalization** (Thomas et al., 2013). The initial "Flag" step forces the model to consciously acknowledge a boundary between its internal "knowledge" and the hypothetical "rules" of the task. This deliberate, separate, metacognitive act is what enables the model to suppress its belief bias and faithfully apply logic in the subsequent "Reason" step.

5.3 Results

Our experiments confirm that the *FaR* intervention is highly effective. As detailed in Figure 6,

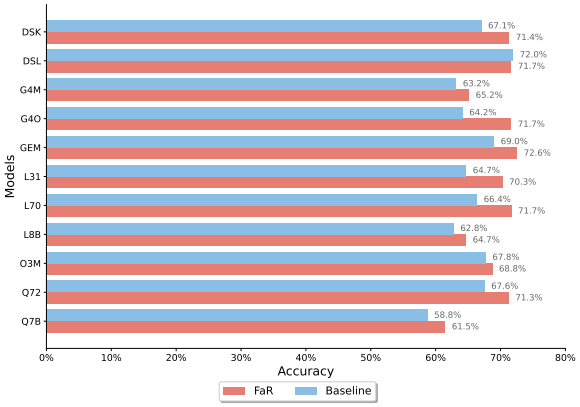


Figure 6: Accuracy comparision between the baseline setup and our *FaR* setup averaged across all the datasets. The x-axis represents accuracy and the y-axis represents the models. Pink bars represent the *FaR* method and the blue bars represent the baseline. We can see that our method *FaR* improves performance across all the models with the most noticeable improvement on GPT-4o (G4O) of about 7%. For Dataset wise results refer to figure 8 in the appendix.

FaR narrows the average performance gap between knowledge-aligned and counterfactual scenarios from 14% to just 7%, while also increasing overall accuracy by an average of 4% across all models and tasks (Table 5, 7).

The most dramatic gains are observed on datasets like Hierarchical Syllogisms and KNOT. In contrast, tasks involving deeper chains of reasoning, such as FOLIO, showed less improvement, emphasizing the need for better conflict resolution strategies for complex narrative tasks (Han et al., 2024). On our *CounterLogic* benchmark specifically, *FaR* increased overall valid accuracy by 5%.

Furthermore, this improvement is pervasive across all logical schemas in *CounterLogic*. As shown in Figure 5(A), *FaR* provides a consistent boost, proving particularly effective for more complex schemas (e.g., Destructive Dilemma) that exhibit lower baseline performance. While schemas involving negation remain challenging for LLMs in general, *FaR* improves average accuracy on these tasks from 73% to 79%.

5.4 Thought Anchor Analysis

To understand the mechanism behind these improvements, we analyzed the models' CoT generations using *thought anchors* - critical reasoning steps that disproportionately influence the final outcome (Bogdan et al., 2025). Our analysis focuses on the more human interpretable functional dynamics of the reasoning process, as opposed to typical mechanistic interpretabilty methods focusing on neuron/circuit level attributions. The full details of our analysis are presented in Appendix J.

Without the *FaR* intervention, our analysis reveals that the reasoning process is heavily anchored in *Fact Retrieval* and *Uncertainty Management*. This observation provides strong evidence for our belief inhibition hypothesis: models appear to be sidetracked by retrieving conflicting parametric knowledge and then struggle to manage the resulting cognitive dissonance. With the *FaR* intervention, we observe a decisive repositioning of these thought anchors. The reasoning process becomes dominated by *Plan Generation* and *Active Computation*. Crucially, the model's reliance on retrieving external facts and expressing uncertainty is significantly diminished. This observed shift suggests a more stable and direct path for logical deduction. By prompting the model to first acknowledge the counterfactual nature of the premises, *FaR* appears to effectively compartmentalize the conflict, allowing the model to proceed with the logical task with less interference from its ingrained knowledge. This fundamental change in the reasoning process offers a compelling explanation for the observed gains in accuracy.

6 Conclusion and Future Work

Our work confirms that LLMs exhibit a critical reasoning failure when logic conflicts with their parametric knowledge, showing a consistent 14% accuracy drop in counterfactual scenarios, as measured on six reasoning datasets including our new *CounterLogic* benchmark. We attribute this to a failure of *belief inhibition*, a phenomenon similar to human cognitive biases. To address this, we introduced *Flag & Reason (FaR)*, a simple metacognitive intervention that not only halves this performance gap to 7% but also boosts overall accuracy by 4%. Our analysis of the models' reasoning chains reveals the mechanism for this success: *FaR* helps the model to compartmentalize the knowledge conflict, thereby anchoring its deduction in the provided logical context instead of its contradictory parametric knowledge.

Future work could explore the internal model representations of these knowledge conflicts, extend metacognitive strategies to more complex, real-world reasoning domains, and develop methods to fundamentally resolve the tension between parametric memory and in-context logic, furthering the development of truly robust AI reasoners.

7 Limitations

While our study provides valuable insights, several limitations should be noted. First, our *CounterLogic* dataset cannot capture all real-world logical conflicts, as it focuses primarily on foundational structures like categorical syllogisms and propositional logic. Second, our findings are specific to the models tested. The rapid pace of architectural innovation means future models may handle counterfactual information differently. Third, our evaluation was limited to English and common-knowledge concepts. The interventions' effectiveness may vary across other languages, cultures, and specialized domains. Fourth, parallels drawn to human cognition are conceptual frameworks and should not be interpreted as evidence of identical cognitive processes in LLMs. Finally, using LLMs to synthetically generate queries may have introduced unnoticed inconsistencies or biases from the generator models themselves.

Despite these limitations, our findings demonstrate substantial improvements in counterfactual reasoning, suggesting that our core insights will remain relevant as models evolve.

8 Ethics Statement

This research adheres to the ACL Ethics Policy and addresses the following ethical considerations:

Research Integrity: We prioritize transparency and reproducibility by documenting experiments in sufficient detail for replication and clearly identifying the limitations of our work (Section 8).

Attribution and Contribution: While large language models were used as tools for writing and implementation, all research ideas, experimental design, and scientific conclusions are solely attributable to the authors. All prior work is properly credited.

Data and Resource Considerations: The *CounterLogic* dataset was synthetically generated and contains no personally identifiable information or data from human subjects. Our evaluation utilized standardized commercial APIs to ensure fair comparison.

Release of Materials: Upon acceptance, we will release all research artifacts to ensure reproducibility. This includes:

- The complete *CounterLogic* dataset and evaluation code (already provided anonymously).
- Implementation details of our *FaR* method.
- Prompt templates and model configurations.
- Comprehensive documentation.

Environmental Considerations: To minimize environmental impact, we designed an efficient evaluation that reused model instances and limited inference runs to the minimum necessary for statistically significant results.

Potential Applications and Impact: Our work aims to improve the robustness of logical reasoning in AI systems. These improvements can benefit applications like education, scientific exploration, and decision support by minimizing the risk of logical errors stemming from knowledge conflicts.

References

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. *Preprint*, arXiv:2406.11341.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which llm reasoning steps matter? *Preprint*, arXiv:2506.19143.

Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. 2024. Sage: Evaluating moral consistency in large language models. *arXiv preprint arXiv:2402.13709*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023a. Say What You Mean! Large Language Models Speak Too Positively about Negative Commonsense Knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9890–9908, Toronto, Canada. Association for Computational Linguistics.

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruixiang Tang. 2025. Counterbench: A benchmark for counterfactual reasoning in large language models. *arXiv preprint arXiv:2502.11008*.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023b. Disco: Distilling counterfactuals with large language models. *Preprint*, arXiv:2212.10534.

Elizabeth Clark, Oyvind Tafjord, Kyle Richardson, Ashish Sabharwal, and Hannaneh Hajishirzi. 2020. Transformers as soft reasoners over language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3882–3894.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. Language models show human-like content effects on reasoning tasks. *arXiv preprint*. ArXiv:2207.07051 [cs].

Igor Douven, Shira Elqayam, and Henrik Singmann. 2018. Conditionals and inferential connections: Toward a new semantics. *Cognition*, 178:31–45.

Tiwalayo Eisape, M. H. Tessler, Ishita Dasgupta, Fei Sha, Sjoerd van Steenkiste, and Tal Linzen. 2024. A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models. *arXiv preprint*. ArXiv:2311.00445 [cs].

Benjamin Estermann, Luca A. Lanzendörfer, and Roger Wattenhofer. 2025. Reasoning effort and problem complexity: A scaling analysis in large language models. *arXiv preprint arXiv:2503.15113*.

Logan Fletcher and Peter Carruthers. 2012. Metacognition and reasoning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1366–1378.

Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Vau, Matthew Caldwell, and Augustine Marvor-Parker. 2023. Susceptibility to influence of large language models.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025. Hipporag: Neurobiologically inspired long-term memory for large language models. *Preprint*, arXiv:2405.14831.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alex Wardle-Solano, Hannah Szabo, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. Folio: Natural language reasoning with first-order logic. *Preprint*, arXiv:2209.00840.

Eddie Harmon-Jones. 2019. *Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology*, 2 edition. American Psychological Association.

Guan Zhe Hong, Nishanth Dikkala, Enming Luo, Cyrus Rashtchian, Xin Wang, and Rina Panigrahy. 2025. A implies b: Circuit analysis in llms for propositional logical reasoning. *Preprint*, arXiv:2411.04105.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Liming Huang, Yulei Wu, and Niccolò Tempini. 2024. A knowledge flow empowered cognitive framework for decision making with task-agnostic data regulation. *IEEE Transactions on Artificial Intelligence*, 5(5):2304–2318.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of LREC-COLING*, pages 10142–10151.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Janani Kugarajeevan, Kokul Thanikasalam, Amirthalingam Ramanan, and Subha Fernando. 2025. Selective information flow for transformer tracking. *Expert Systems with Applications*, 259:125381.

Ziva Kunda. 1990. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*.

Haowei Lin, Xiangyu Wang, Ruilin Yan, Baizhou Huang, Haotian Ye, Jianhua Zhu, Zihao Wang, James Zou, Jianzhu Ma, and Yitao Liang. 2025. Generative reasoning with large language models.

Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Untangle the KNOT: Interweaving Conflicting Knowledge and Reasoning Skills in Large Language Models. *arXiv preprint*. ArXiv:2404.03577 [cs].

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063. Association for Computational Linguistics.

Jie Ma, Weijia Lv, and Xuezhu Ren. 2025. Neural correlates of belief-bias reasoning as predictors of critical thinking: Evidence from an fnirs study. *Journal of Intelligence*, 13(9):106.

Olivia Macmillan-Scott and Mirco Musolesi. 2024. (ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(3):240255.

Henry Markovits and Guilaine Nantel. 1989. The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, 17(1):11–17.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy.

Monica K. Miller, Jordan D. Clark, and Alayna Jehle. 2015. *Cognitive Dissonance Theory (Festinger)*. John Wiley & Sons, Ltd.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangled question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. In *Proceedings of EMNLP*.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? *arXiv preprint*. ArXiv:2002.08910 [cs].

Amit Rozner, Barak Battash, Lior Wolf, and Ofir Lindenbaum. 2024. Knowledge editing in language models via adapted direct preference optimization. *Preprint*, arXiv:2406.09920.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Nataniel Shinn, Shinn Yao, Kaixuan Zhao, Dian Yu, Eric Zhao, Dan Zhao, and Dragomir Radev. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shashwat Singh, Shashwat Goel, Saujas Vaduguru, and Ponnurangam Kumaraguru. 2024. Probing negation in language models.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm.

Jenna Thomas, Christopher Ditzfeld, and Carolin Showers. 2013. Compartmentalization: A window on the defensive self 1. *Social and Personality Psychology Compass*, 10:719–7311111.

Armin Toroghi, Willis Guo, and Scott Sanner. 2024. Right for right reasons: Large language models for verifiable commonsense knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Dries Trippas, Simon Handley, and Michael Verde. 2014. Fluency and belief bias in deductive reasoning: new indices for old effects. *Frontiers in Psychology*, 5:631.

Robert Vacareanu and Miguel Ballesteros. 2024. General purpose verification for chain of thought prompting. *arXiv preprint arXiv:2405.00204*.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025. Understanding reasoning in thinking language models via steering vectors. *Preprint*, arXiv:2506.18167.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA. Association for Computing Machinery.

Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. 2025. Negation: A pink elephant in the large language models' room? *Preprint*, arXiv:2503.22395.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *Preprint*, arXiv:2310.07521.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. Resolving Knowledge Conflicts in Large Language Models. *arXiv preprint*. ArXiv:2310.00935 [cs].

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. *Preprint*, arXiv:2308.05342.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. *arXiv preprint*. ArXiv:2307.02477 [cs].

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024a. On memorization of large language models in logical reasoning. *arXiv preprint arXiv:2410.23123*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024b. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. *arXiv preprint*. ArXiv:2305.13300 [cs].

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, M. Lee, and W. Hsu. 2024a. Faithful logical reasoning via symbolic chain-of-thought. *Annual Meeting of the Association for Computational Linguistics*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge Conflicts for LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Xu Zhang, Xunjian Yin, and Xiaojun Wan. 2024a. Contrasolver: Self-alignment of language models by resolving internal preference contradictions. *Preprint*, arXiv:2406.08842.

Yujia Zhang, Wei Wang, Xiaojie Liu, Yiming Chen, and Zhenyu Li. 2024b. Bridging the gap between llms and human intentions. *arXiv preprint arXiv:2502.09101*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023b. Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. *arXiv preprint*. ArXiv:2302.13439 [cs].

A CounterLogic Dataset Details

A.1 Hierarchical Entity Triples

The CounterLogic dataset uses the following hierarchical entity triples, where each tuple (a, b, c) denotes a strict subset relationship: $a \subset b \subset c$.

siameses	*cats*	*felines*
labradors	*dogs*	*canines*
sedans	*cars*	*vehicles*
humans	*animals*	*mortals*
cruisers	*warships*	*watercrafts*
chickadees	*birds*	*winged_animals*
boeings	*planes*	*aircrafts*
pines	*evergreens*	*trees*
anguses	*cows*	*mammals*
daisies	*flowers*	*plants*

A.2 Sentence Templates

From each entity triplet, we generate sentence pairs corresponding to one of four logical sentence templates of the form S and $\neg S$, capturing contradictory or complementary quantifier relations:

1. *All {A} are {B}*, *Some {A} are not {B}*
2. *No {A} are {B}*, *Some {A} are {B}*
3. *Some {A} are {B}*, *No {A} are {B}*
4. *Some {A} are not {B}*, *All {A} are {B}*

A.3 Dataset Statistics

The final *CounterLogic* dataset consists of 1,800 examples, with 200 instances for each of the 9 logical schemas listed in Table 2. To ensure a comprehensive and balanced design, four criteria were enforced during dataset construction:

Name	Propositional Logic Form
Modus Ponens (MP)	$((p \to q) \land p) \vdash q$
Modus Tollens (MT)	$((p \to q) \land \neg q) \vdash \neg p$
Hypothetical Syllogism (HS)	$((p \to q) \land (q \to r)) \vdash (p \to r)$
Disjunctive Syllogism (DS)	$((p \lor q) \land \neg p) \vdash q$
Constructive Dilemma (CD)	$(p \to q) \land (r \to s) \land (p \lor r) \vdash (q \lor s)$
Destructive Dilemma (DD)	$((p \to q) \land (r \to s) \land (\neg q \lor \neg s)) \vdash (\neg p \lor \neg r)$
Bidirectional Dilemma (BD)	$((p \to q) \land (r \to s) \land (p \lor \neg s)) \vdash (q \lor \neg r)$
Commutation (CT)	$p \vdash (q \lor p)$
Material Implication (MI)	$(p \to q) \vdash (\neg p \lor q)$

Table 2: Formal propositional logical schemas used in the *CounterLogic* dataset. Each row presents a canonical logical inference rule and its structure in propositional form. Believability is achieved when both the premises and the conclusion are true independently. Invalid datapoints are created by replacing conclusion statements (e.g., p, q, r) with unrelated ones (e.g., p', q', r') making the logical rule invalid.

- **Knowledge alignment Balance**: Each logical schema contains 50% of examples where the conclusion and the premises are knowledge-consistent under human priors and 50% where it is not. eg.: "All humans are animals" is a Knowledge-Aligned statement whereas "Some cats are not felines" is a Knowledge-Conflicting statement.

- **Validity Balance**: Half the examples per schema are logically valid, while the remaining half intentionally violate the logical structure. eg.: "Premise: All cats are dogs; Conclusion: Some cats are dogs" is a valid conclusion whereas "Premise: All cats are dogs; Conclusion: Some cats are not dogs" is an invalid conclusion.

- **Entity Relationship Balance**: 50% of examples involve an entity A that is a subset of entity B (e.g., *siameses* $\subset$ *cats*) and 50% feature B as a subset of A (e.g., *cats* $\subset$ *siameses*).

- **Sentence Template Balance**: All 8 sentence-pair templates are applied evenly across examples within each logical schema, promoting lexical and syntactic diversity.

A.4 Varying Reasoning Depth

To systematically evaluate how logical reasoning performance scales with complexity, we implemented a method for varying the reasoning depth for certain logical schemas. For any schema that includes an implication rule of the form $p \to q$, we can increase the number of reasoning steps by replacing it with a longer chain of implications. For instance, a single implication can be expanded by introducing i intermediate propositions $(x_1, x_2, \ldots, x_i)$ to form a chain: $p \to x_1 \to x_2 \to \cdots \to x_i \to q$. This transformation increases the number of premises in the problem, requiring the model to perform a longer series of deductions to arrive at the final conclusion.

To preserve the original knowledge alignment of the implication, the intermediate propositions $(x_1, x_2, \ldots)$ introduced are always sampled from a pool of knowledge-aligned (factually true) sentences. This ensures that if a base implication $p \to q$ was knowledge-conflicting (e.g., where a true premise p leads to a false conclusion q), at least one of the new implications in the extended chain will also be knowledge-conflicting. For example, in the chain $p \to x_1 \to q$, the link $x_1 \to q$ would be factually false, as its premise (x_1) is true while its conclusion (q) is false, thereby preserving the overall knowledge-conflicting nature of the reasoning task.

This technique is applied to all schemas in Table 2 that contain one or more implication rules. Notably, this excludes the *Disjunctive Syllogism (DS)* and *Commutation (CT)* schemas, as their propositional logic forms do not involve implication.

Below are two examples illustrating how base rules are modified to increase reasoning depth.

Example 1: Modus Ponens with One Addition
The base Modus Ponens (MP) rule requires one step of deduction from two premises. By adding one intermediate step (x_1), the rule now requires two steps of deduction from three premises.

- **Base MP Rule:**

$$((p \to q) \land p) \vdash q$$

- **MP Rule with One Addition:**

$$((p \to x_1) \land (x_1 \to q) \land p) \vdash q$$

Example 2: Destructive Dilemma with Two Additions The Destructive Dilemma (DD) rule contains two separate implication rules ($p \to q$ and $r \to s$). We add one intermediate step to each, for a total of two additions, increasing the number of premises from three to five.

- **Base DD Rule:**

$$((p \to q) \land (r \to s) \land (\neg q \lor \neg s)) \vdash (\neg p \lor \neg r)$$

- **DD Rule with Two Additions:**

$$((p \to x_1) \land (x_1 \to q) \land (r \to y_1) \\ \land (y_1 \to s) \land (\neg q \lor \neg s)) \vdash (\neg p \lor \neg r)$$

A.5 Prompt for Natural Language Reformulation of Logical Structures

To prevent language models from relying on memorized patterns of formal logic structures, we reformulate logical premises and conclusions into natural language contexts and questions. Specifically, we prompt **GPT-4o** to carry out this transformation. The model is instructed to rewrite the given premise into a natural language context and the conclusion into a straightforward question, without preserving the surface structure of the original logical form. The transformation ensures the resulting question does not include meta-references like "in this context," and the phrasing is natural and intuitive:

```
premise: [premise]
conclusion: [conclusion]
premise list: [premise list]
Make the premise into a context which is
     like a natural language way of
    writing the premises. Make
    conclusion into a question.
The context/questions shouldn't be too
    complicated but shouldn't directly
    be like premise/conclusion either.
    The question must be asked normally
    without stating things like "in this
     context" or "with this information
    ".
premise list is only given for your
    better understanding.
Reply ONLY with a json with two keys '
    context' and 'question'
```

B Prompting Strategies

We evaluate three distinct prompting strategies across all tasks:

1. Standard Condition: In this baseline condition, models receive direct questions with minimal guidance, instructed to consider only the logical validity of arguments regardless of premise believability:

```
Based on the following premises,
    determine if the conclusion
    logically follows. Consider only
    the logical validity based on the
    given premises, regardless of
    whether the premises themselves
    are factually true.

Premises:
  1. [Premise 1]
  2. [Premise 2]

  Conclusion: [Conclusion]

  Does the conclusion logically follow
       from the premises? Answer with
      "Yes" or "No" and explain your
      reasoning step by step.
```

2. Metacognitive Condition: In this condition, we introduce a preliminary reflection step (asking the model what it thinks about a statement) before the reasoning task.

```
Prompt 1:

Is the following statement factually
    correct?

statement: [Conclusion]

Answer only with Yes or No.

Prompt 2(in the same context):

Now, based on the  following premises,
    determine if the conclusion
    logically follows. Consider only the
     logical validity based on the
given premises, regardless of whether
    the premises themselves are
    factually true.
Premises:
    1. [Premise 1]
    2. [Premise 2]

    Conclusion: [Conclusion]

    Does the conclusion logically follow
         from the premises? Answer with
        "Yes" or "No" and explain your
        reasoning step by step.
```

The above is the general structure of our prompting method. The prompts are modified according to the dataset we evaluate.

Task	Dataset Content	Initial Reflection Input	Reasoning Input
Syllogisms	Two premises + conclusion	Conclusion statement	Full syllogism
KNOT	Passage + Q/A pair	Answer without passage	Full passage + Q/A
FOLIO	Narrative + claim	Isolated claim	Complete narrative
LogicBench	Context + Q/A	Q/A without context	Full context + Q/A
Arithmetic	Base equation	Equation without base	Base-specified equation

Table 3: Task-Specific Implementation of Two-Stage Reflection Approach. This table outlines how our reflective prompting strategy is applied across different task types. **Initial Reflection Input** refers to the isolated information presented for knowledge alignment assessment. **Reasoning Input** shows the complete information provided in the second stage for logical assessment. This separation helps models distinguish between plausibility assessment and formal logical analysis.

C Logical Inference Schemas

The *CounterLogic* dataset uses various formal propositional logic inference schemas to generate reasoning examples, as detailed in Table 2.

D Task-Specific Reflection Approaches

Our metacognitive intervention is implemented with task-specific adaptations to ensure appropriate reflection across different reasoning formats.

Hierarchical Syllogisms: Derived from classical syllogistic reasoning and adapted from Bertolazzi et al. (2024)'s work, this task presents logically structured arguments where the conclusion may conflict with world knowledge. Each example contains two premises and a conclusion, with models evaluating logical validity. For *FaR*, models first assess the conclusion statement in isolation for its alignment with parametric knowledge, then evaluate the full syllogism's logical validity(see Figure 2B).

KNOT: Adapted from the Knowledge Conflict Resolution benchmark (Liu et al., 2024), this task evaluates reasoning through explicit (KNOT-E) and implicit (KNOT-I) conflict resolution. Each instance contains a passage with counterfactual information, a question, and an answer. The *FaR* implementation first presents the answer in isolation for plausibility assessment, then provides the full passage and question-answer pair for contextual reasoning. This separation tests models' ability to distinguish between prior knowledge and contextual truth.

FOLIO: Using long-form deductive reasoning problems from FOLIO (Han et al., 2024), this task requires evaluating whether conclusions logically follow from multi-step narratives. Our *FaR* approach first presents the conclusion for isolated plausibility judgment, then provides the complete narrative for logical analysis.

LogicBench: This reasoning dataset (Parmar et al., 2024) combines first-order, non-monotonic, and propositional logic problems. It tests models' ability to follow formal logical rules while overriding potentially conflicting parametric knowledge. The *FaR* implementation presents questions and answers without supporting context for initial plausibility assessment, followed by complete logical contexts for formal evaluation.

CounterLogic: For our novel benchmark, described in Section 3, we apply the same two-stage reflection approach used in the Hierarchical Syllogisms task, first assessing conclusion plausibility in isolation before evaluating logical validity within the full syllogistic context.

Knowledge alignment labels were available only for the *CounterLogic* and *Hierarchical Syllogisms* datasets. For the remaining datasets, we get alignment judgments by prompting the LLMs to assess the factuality of the conclusions, and we use their responses as proxy labels. This procedure results in a heavily imbalanced label distributions, most notably in the *KNOT* dataset, where the ratio of Aligned to Conflicting cases is approximately 1:15.

E Models

In our study, we evaluated 11 state-of-the-art large language models from various organizations, spanning different architectures, parameter scales, and training paradigms. Below we provide details about each model, including their version, size, and key characteristics:

E.1 Model Access

All models were accessed through the OpenRouter API to ensure consistent evaluation conditions. This approach allowed us to standardize the inference parameters across different model providers, including temperature settings (0.7), top-p (0.95),

Table 4: Details of the evaluated models

Model	Developer	Parameters	Release Date
GPT-4o	OpenAI	Unknown	May 2024
GPT-4o-mini	OpenAI	Unknown	July 2024
o3-mini	OpenAI	Unknown	April 2025
Gemini-Flash-1.5	Google DeepMind	Unknown	May 2024
Llama-3.3-70B	Meta AI	70B	December 2024
Llama-3.1-70B	Meta AI	70B	July 2024
Llama-3.1-8B	Meta AI	8B	July 2024
Qwen-2.5-72B	Alibaba	72B	September 2024
Qwen-2.5-7B	Alibaba	7B	September 2024
DeepSeek-V3	DeepSeek AI	671B	Jan 2025
Deepseek-R1-distill-Llama	DeepSeek AI	671B	January 2025

and maximum token length (4096 tokens).

E.2 Model Selection Criteria

We selected these models based on the following criteria:

1. **State-of-the-art performance:** All selected models represent the cutting edge of LLM development at the time of our study.

2. **Architectural diversity:** We included models with different architectural designs to examine whether the observed patterns generalize across various model architectures.

3. **Parameter scale variation:** The selection spans from relatively smaller models (7B parameters) to much larger ones (72B+ parameters) to investigate how model size correlates with counterfactual reasoning abilities.

4. **Training paradigm diversity:** The models employ various training approaches, including different pretraining datasets, fine-tuning strategies, and alignment techniques.

E.3 Model Specifications

E.3.1 OpenAI Models

GPT-4o, **GPT-4o-mini** and **o3-mini** represents the OpenAI's multimodal model designed for both text and imageprocessing. GPT-4o is the standard non reasoning model provided by OpenAI. GPT-4o mini is a lightweight and faster version of 4o. o3 mini is the reasoning model provided by openAI that uses specialized token to internally do CoT before answering.

E.3.2 Google Models

Gemini-Flash-1.5 is the lightweight version of Google's Gemini 1.5 model family, optimized for quick responses while maintaining strong reasoning capabilities.

E.3.3 Meta AI Models

Llama-3.3-70B, **Llama-3.1-70B**, and **Llama-3.1-8B** represent Meta AI's open-source LLM efforts. The 3.1 series is an upgrade to the Llama-3 models, with enhanced instruction-following and reasoning capabilities. We include both larger (70B) and smaller (8B) parameter variants to examine scaling effects.

E.3.4 Alibaba Models

Qwen-2.5-72B and **Qwen-2.5-7B** are Alibaba's latest generation language models, known for their strong performance across various benchmarks, particularly in multilingual reasoning tasks.

E.3.5 DeepSeek Models

DeepSeek-V3 is a 671B parameter model developed by DeepSeek AI, designed specifically for dialogue applications with strong reasoning capabilities.

DeepSeek-R1-Distill-Llama is reasoning model that is finetuned version of DeepSeek-R1(671b) using the outputs of Llama-3.3-70b-instruct model.

E.4 Model Inference Parameters

For all evaluations, we used consistent inference parameters across models. We use openRouter for all the non OpenAI models and OpenAI API for the 3 openAI models. All the models that support system prompts have standard prompt asking for instruction following.

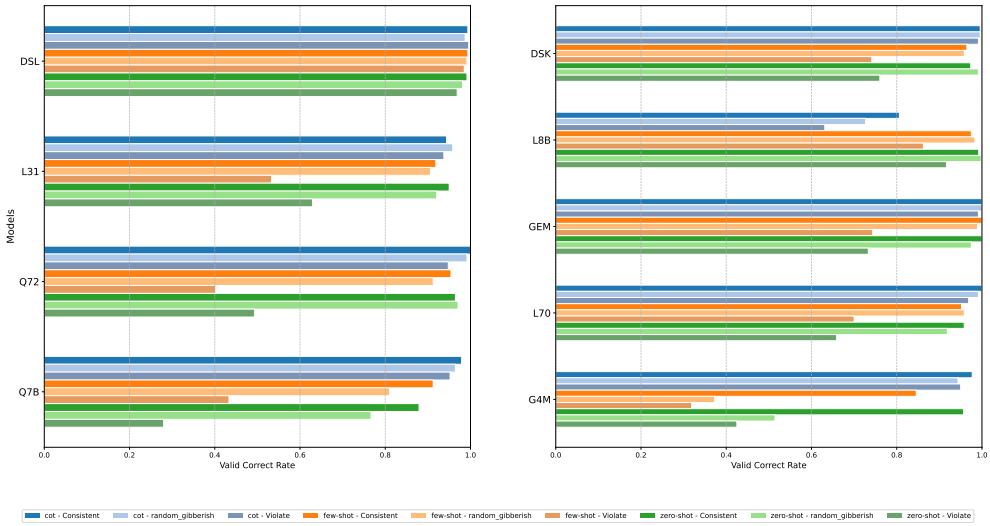


Figure 7: Ablation study comparing the various prompting strategies. Shows that CoT always outperforms zero-shot but few-shot fails to do so in some models for the Syllogistic dataset.

E.5 Cost of the Evaluations

Overall 83.2$ were spend on openRouter for all non-OpenAI models across all tasks.
About 2000$ of OpenAI credits were used for running the evaluation, most of which was used by the o1-preview model.

F Full Results

This section provides the full experimental results for all models and evaluation settings. Table 5 reports accuracies with and without *FaR*, while Tables 6 and 7 present a comparison between Knowledge Aligned and Knowledge Conflicting accuracies on the baseline and *FaR* subsets, respectively. Each value denotes the mean $\pm$ standard deviation. These tables serve as a comprehensive reference supporting the main findings discussed in paper.

G Model-wise Results and Performance Patterns

Our evaluation demonstrates that the *FaR* method provides a consistent and significant improvement in reasoning accuracy across all 11 evaluated models when averaged across the datasets. This universal performance boost is clearly illustrated in Figure 6, where the FaR method (pink bar) outperforms the baseline (blue bar) for every model tested. The most noticeable average improvement was observed for the GPT-4o (G4O) model, which saw an accuracy increase of approximately 7%.

G.1 Key Performance Patterns

A deeper analysis of the results, particularly from the detailed breakdown in Table 5, reveals several important patterns in how different models respond to the FaR intervention.

- **Variable Impact Across Datasets:** While the FaR method is universally beneficial, its impact varies significantly depending on the reasoning task.
  - The most dramatic performance gains are seen on the **Hierarchical** syllogisms dataset, where accuracy for many models, such as L70 (80.9% to 98.1%) and G4O (86.8% to 99.7%), improved by over 15 percentage points.
  - In contrast, the improvement on the **Folio** datasets is more modest, with FaR providing a smaller, though still consistent, accuracy boost emphasizing the need for better conflict resolution strategies for complex narrative tasks
- **Effectiveness Across Model Types:** The benefits of FaR are not limited to a specific class or size of model.
  - Specialized **Reasoning Models** like O3M and DSL, which already have strong baseline performance, still achieve slightly higher accuracy with the FaR intervention. For example, O3M's

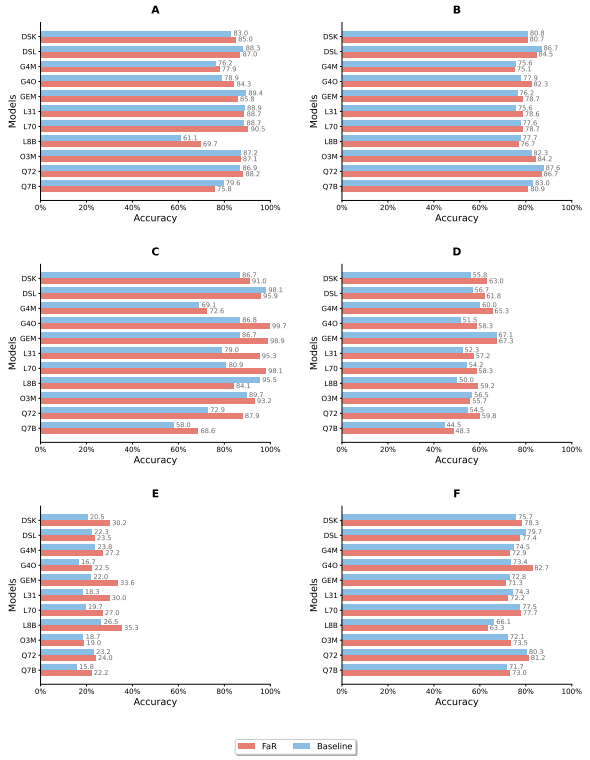


Figure 8: Accuracy comparision between the baseline setup and our *FaR* setup averaged across all the datasets. The x-axis represents accuracy and the y-axis represents the models. We can see that our method *FaR* improves performance across all the tasks. The y-axis represents the following models - DSK: DeepSeek Chat, DSL: Deepseek R1 Distill Llama 70B, G4M: GPT-4o Mini, G4O: GPT-4o, GEM: Google Gemini Flash 1.5, L8B: Llama 3.1 8B Instruct, L31: Llama 3.1 70B Instruct, L70: Llama 3.3 70B Instruct, O3M: o3-mini, Q7B: Qwen 2.5 7B Instruct, and Q72: Qwen 2.5 72B Instruct. Each of the above graphs represents the performance on the following datasets - A: CounterLogic, B: FOLIO, C: Hierarchical Syllogisms, D: KNOT-Explicit, E: KNOT-Implicit, F: Logic Bench

performance on the Hierarchical task improves from 89.7% to 93.2%.

– Similarly, **Small Models** such as L8B and Q7B show significant gains, indicating that the method is a broadly applicable technique for enhancing logical robustness, regardless of model scale.

In summary, the FaR intervention proves to be a robust and widely effective method for improving logical reasoning. The degree of improvement, however, is task-dependent, highlighting its particular strength in mitigating the types of errors prevalent in tasks with strong knowledge conflicts.

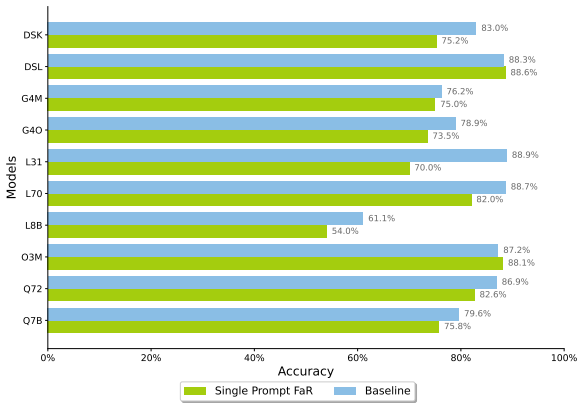


Figure 9: Accuracy comparison between the baseline setup and the Single prompt *FaR* setup averaged across all the datasets. The x-axis represents accuracy and the y-axis represents the models. We can see that this modification of *FaR* does not lead to any significant performance gains.

H Ablation Studies

H.1 Comparison of Prompting Strategies

Figure 10 presents an ablation study evaluating model performance under three prompting strategies: zero-shot, few-shot, and chain-of-thought (CoT) across three belief consistency conditions; aligned, conflicting, and random gibberish. Here random gibberish datapoints are obtained by replacing entities in aligned and conflicting scenarios with random strings (such as 'cat' with 'nsjf'). Aligned and conflicting datapoints are from the Syllogistic Dataset (Bertolazzi et al., 2024).

Across all models, we see that aligned datapoints perform better than gibberish datapoints which perform better than violate datapoints, highlighting the reliance of model on its internal knowledge and inability to reason purely based on logical rules. CoT prompting consistently improves accuracy without altering the general trend observed in belief sensitivity. Interestingly, few-shot prompting does not universally help: OpenAI models (e.g., gpt-4o-mini) actually show degraded performance in few-shot settings across all belief types, suggesting potential sensitivity to in-context demonstrations or prompt formatting. In contrast, most other models maintain or slightly improve their performance under few-shot.

H.2 Perturbing with Model-Generated Evidence

To evaluate whether language models reason purely based on logical structure or are influenced by surface-level content, we perform an experiment involving model-generated evidence. Specifically,

	Explicit		Implicit		Folio		Hierarchical		Counterlogic		Logic Bench	
Models	FaR	Baseline	FaR	Baseline	FaR	Baseline	FaR	Baseline	FaR	Baseline	FaR	Baseline
Large models												
L70	**58.3** $\pm$ 0.7	54.2 $\pm$ 0.2	**27.0** $\pm$ 0.7	19.7 $\pm$ 1.3	**78.7** $\pm$ 2.2	77.6 $\pm$ 1.1	**98.1** $\pm$ 0.1	80.9 $\pm$ 0.1	**90.5** $\pm$ 0.2	88.7 $\pm$ 0.1	**77.7** $\pm$ 0.4	77.5 $\pm$ 0.2
L31	**57.2** $\pm$ 0.2	52.3 $\pm$ 0.3	**30.0** $\pm$ 1.0	18.3 $\pm$ 0.3	**78.6** $\pm$ 0.1	75.6 $\pm$ 2.0	**95.3** $\pm$ 0.5	79.0 $\pm$ 0.6	88.7 $\pm$ 0.2	**88.9** $\pm$ 0.4	72.2 $\pm$ 0.6	**74.3** $\pm$ 0.3
GEM	**67.3** $\pm$ 0.3	67.1 $\pm$ 0.3	**33.6** $\pm$ 0.6	22.0 $\pm$ 0.0	**78.7** $\pm$ 0.6	76.2 $\pm$ 0.6	**98.9** $\pm$ 0.0	86.7 $\pm$ 0.0	85.8 $\pm$ 0.2	**89.4** $\pm$ 0.2	71.3 $\pm$ 0.2	**72.8** $\pm$ 0.3
Q72	**59.8** $\pm$ 0.5	54.5 $\pm$ 0.8	**24.0** $\pm$ 0.3	23.2 $\pm$ 0.2	86.7 $\pm$ 0.4	**87.6** $\pm$ 1.2	**87.9** $\pm$ 0.2	72.9 $\pm$ 0.5	**88.2** $\pm$ 0.8	86.9 $\pm$ 0.4	**81.2** $\pm$ 0.5	80.3 $\pm$ 0.1
DSK	**63.0** $\pm$ 1.0	55.8 $\pm$ 0.2	**30.2** $\pm$ 0.5	20.5 $\pm$ 1.8	80.7 $\pm$ 1.1	**80.8** $\pm$ 1.0	**91.0** $\pm$ 0.8	86.7 $\pm$ 0.6	**85.0** $\pm$ 0.6	83.0 $\pm$ 0.4	**78.3** $\pm$ 0.0	75.7 $\pm$ 0.0
G4O	**58.3** $\pm$ 0.3	51.5 $\pm$ 0.2	**22.5** $\pm$ 0.2	16.7 $\pm$ 0.0	**80.8** $\pm$ 3.9	78.1 $\pm$ 1.7	**99.7** $\pm$ 0.1	86.8 $\pm$ 0.1	**84.3** $\pm$ 0.9	78.9 $\pm$ 0.7	**82.7** $\pm$ 0.3	73.4 $\pm$ 0.4
Small models												
L8B	**59.2** $\pm$ 1.8	50.0 $\pm$ 2.3	**35.3** $\pm$ 2.7	26.5 $\pm$ 1.8	76.7 $\pm$ 0.5	**77.7** $\pm$ 2.1	84.1 $\pm$ 0.7	**95.5** $\pm$ 1.2	**69.7** $\pm$ 1.2	61.1 $\pm$ 1.7	63.3 $\pm$ 0.7	**66.2** $\pm$ 0.6
Q7B	**48.3** $\pm$ 0.3	44.5 $\pm$ 1.2	**22.2** $\pm$ 0.8	15.8 $\pm$ 0.8	80.9 $\pm$ 2.3	**83.0** $\pm$ 1.3	**68.6** $\pm$ 0.1	58.0 $\pm$ 0.4	75.8 $\pm$ 0.2	**79.6** $\pm$ 0.4	**73.0** $\pm$ 0.3	71.7 $\pm$ 1.0
G4M	**65.3** $\pm$ 2.0	60.0 $\pm$ 0.3	**27.2** $\pm$ 0.5	23.8 $\pm$ 0.2	75.1 $\pm$ 1.6	**75.1** $\pm$ 1.5	**72.6** $\pm$ 0.2	69.1 $\pm$ 0.7	**77.9** $\pm$ 0.4	76.2 $\pm$ 0.3	72.9 $\pm$ 0.4	**74.4** $\pm$ 0.4
Reasoning models												
DSL	**61.8** $\pm$ 0.2	56.7 $\pm$ 0.3	**23.5** $\pm$ 0.5	22.3 $\pm$ 0.7	84.5 $\pm$ 1.5	**86.7** $\pm$ 0.2	95.9 $\pm$ 1.2	**98.1** $\pm$ 0.1	87.0 $\pm$ 0.3	**88.3** $\pm$ 0.8	77.4 $\pm$ 0.0	**79.7** $\pm$ 0.7
O3M	55.7 $\pm$ 0.0	**56.5** $\pm$ 0.2	**19.0** $\pm$ 0.3	18.7 $\pm$ 0.3	**84.2** $\pm$ 0.0	82.3 $\pm$ 0.1	**93.2** $\pm$ 0.6	89.7 $\pm$ 0.5	86.9 $\pm$ 0.4	**87.2** $\pm$ 0.4	**73.5** $\pm$ 0.3	72.1 $\pm$ 1.1

Table 5: Model performance (FaR vs. Baseline) across datasets, showing mean accuracy and standard deviation (%)

	Implicit		Explicit		Folio		Hierarchical		Counterlogic		Logic Bench	
Models	Aligned	Conflicting	Aligned	Conflicting	Aligned	Conflicting	Aligned	Conflicting	Aligned	Conflicting	Aligned	Conflicting
Large models												
L70	**25.4** $\pm$ 2.9	18.8 $\pm$ 1.1	47.0 $\pm$ 0.4	**55.1** $\pm$ 0.3	**91.4** $\pm$ 1.0	64.3 $\pm$ 1.7	**95.8** $\pm$ 0.6	65.9 $\pm$ 0.4	83.2 $\pm$ 0.5	**84.0** $\pm$ 0.4	**78.4** $\pm$ 1.2	63.8 $\pm$ 0.3
L31	**24.3** $\pm$ 3.8	17.6 $\pm$ 0.0	43.2 $\pm$ 3.2	**53.2** $\pm$ 0.1	**86.0** $\pm$ 0.5	67.2 $\pm$ 3.3	**95.0** $\pm$ 0.7	62.9 $\pm$ 0.6	**91.2** $\pm$ 0.2	88.7 $\pm$ 1.0	**75.2** $\pm$ 1.3	59.1 $\pm$ 0.1
GEM	21.3 $\pm$ 1.3	**22.2** $\pm$ 0.2	**68.3** $\pm$ 0.0	67.0 $\pm$ 0.3	**85.7** $\pm$ 1.7	66.3 $\pm$ 0.5	**100.0** $\pm$ 0.0	73.3 $\pm$ 0.0	79.0 $\pm$ 0.4	**82.5** $\pm$ 0.4	**81.9** $\pm$ 1.3	58.9 $\pm$ 0.5
Q72	**44.4** $\pm$ 2.5	20.7 $\pm$ 0.1	**68.3** $\pm$ 6.7	53.9 $\pm$ 0.6	**93.0** $\pm$ 1.2	83.8 $\pm$ 1.9	**96.5** $\pm$ 0.3	49.3 $\pm$ 1.0	**83.5** $\pm$ 0.7	77.3 $\pm$ 2.3	**86.4** $\pm$ 1.8	68.0 $\pm$ 0.1
DSK	**36.8** $\pm$ 0.8	19.1 $\pm$ 2.0	**60.0** $\pm$ 5.0	55.5 $\pm$ 0.5	**90.4** $\pm$ 1.3	75.0 $\pm$ 1.4	**97.4** $\pm$ 0.2	76.0 $\pm$ 1.2	**74.7** $\pm$ 1.5	65.2 $\pm$ 1.0	**91.0** $\pm$ 0.7	65.0 $\pm$ 0.9
G4O	**34.9** $\pm$ 0.6	14.5 $\pm$ 0.0	**59.1** $\pm$ 2.0	51.0 $\pm$ 0.3	**88.1** $\pm$ 2.7	68.1 $\pm$ 2.5	**97.4** $\pm$ 0.3	76.1 $\pm$ 0.1	**64.8** $\pm$ 0.8	58.6 $\pm$ 1.6	**86.4** $\pm$ 2.6	70.7 $\pm$ 0.3
Small models												
L8B	**27.9** $\pm$ 0.8	24.9 $\pm$ 4.6	49.6 $\pm$ 1.4	**50.3** $\pm$ 3.4	**80.8** $\pm$ 4.3	75.5 $\pm$ 1.3	**99.2** $\pm$ 0.2	91.7 $\pm$ 2.5	**55.3** $\pm$ 1.2	52.2 $\pm$ 1.5	**62.5** $\pm$ 0.0	57.9 $\pm$ 2.3
Q7B	**17.8** $\pm$ 2.2	15.4 $\pm$ 1.4	**82.1** $\pm$ 10.7	42.3 $\pm$ 1.4	**86.6** $\pm$ 2.0	81.8 $\pm$ 1.1	**87.9** $\pm$ 0.5	28.0 $\pm$ 0.4	**64.2** $\pm$ 1.2	56.8 $\pm$ 0.2	**83.9** $\pm$ 6.5	51.6 $\pm$ 1.3
G4M	**45.1** $\pm$ 3.6	20.6 $\pm$ 0.3	**66.7** $\pm$ 4.7	59.3 $\pm$ 0.8	**85.5** $\pm$ 1.0	66.4 $\pm$ 2.5	**95.7** $\pm$ 0.3	42.5 $\pm$ 1.2	**66.3** $\pm$ 0.7	55.6 $\pm$ 1.0	**80.8** $\pm$ 1.3	59.0 $\pm$ 1.3
Reasoning models												
DSL	**42.7** $\pm$ 2.7	19.9 $\pm$ 1.2	43.0 $\pm$ 9.6	**57.4** $\pm$ 0.1	**94.6** $\pm$ 0.0	79.3 $\pm$ 0.6	**98.6** $\pm$ 0.5	97.6 $\pm$ 0.7	**86.6** $\pm$ 0.6	82.7 $\pm$ 2.0	**85.9** $\pm$ 1.3	77.2 $\pm$ 1.4
O3M	**35.2** $\pm$ 0.8	16.9 $\pm$ 0.5	**60.5** $\pm$ 2.6	56.2 $\pm$ 0.4	**90.9** $\pm$ 0.5	76.1 $\pm$ 0.5	89.7 $\pm$ 0.1	89.7 $\pm$ 0.9	**83.3** $\pm$ 0.6	81.2 $\pm$ 1.0	**84.3** $\pm$ 0.8	69.0 $\pm$ 1.2

Table 6: Model performance (Aligned vs. Conflicting) on valid datapoints with standard CoT method (baseline).

we prompt the models to generate evidences for a given conclusions and premises: one that supports the conclusion and one that neagtes it. The model is given complete freedom in how it constructs this evidence, encouraging creativity and variability in content. This is adapted from (Xie et al., 2024b)

We then construct a separate task: given a premise, above generated evidence, and a conclusion, we ask whether the conclusion follows from the premise purely logically. The correct answer is determined solely based on the logical relation between the premise and conclusion, independent of the evidence. However, our results reveal a clear pattern: models show an increase in accuracy for logically valid datapoints when the supporting evidence aligns with the conclusion, and a drop in accuracy when the evidence contradicts it. This behavior suggests that models are not performing strict logical reasoning, but are instead heavily influenced by the factuality of premises and conclusions, even when it is explicitly stated to be potentially fabricated. This indicates a reliance on heuristic signals rather than formal logical inference.

H.3 Quantifying Counterfactual Bias of Models

We quantify the Counterfactual Bias of a model by defining Statistical Parity Difference(SPD) (Verma and Rubin, 2018). It is the difference between the likelihood that a model will give a positive answer in a scenario consistent with its factual knowledge (believable scenario) and the scenario inconsistent with its factual knowledge (unbelievable), disregarding the Logical Reasoning.

$$SPD = (\mathbb{P}(\text{Yes} \mid \text{Believable}) - \mathbb{P}(\text{Yes} \mid \text{Unbelievable})) \times 100 \quad (1)$$

A zero value for SPD implies no bias and a positive value denotes the likelihood of the model answering factually while disregarding the premises. The value of the metric is independent of the underlying ground truth distribution, which evaluates logical correctness. Hence, this metric is a good fit to evaluate all the different types of datasets.

SPD values on various datasets are given in Figure 11. We can see large positive values across all the datasets indicating a large bias towards aligning

	Implicit		Explicit		Folio		Hierarchical		Counterlogic		Logic Bench	
Models	Aligned	Conflicting	Aligned	Conflicting	Aligned	Conflicting	Aligned	Conflicting	Aligned	Conflicting	Aligned	Conflicting
Large models												
L70	**31.7** $\pm$ 0.8	26.3 $\pm$ 0.7	51.3 $\pm$ 1.3	**59.2** $\pm$ 0.8	**90.6** $\pm$ 1.0	67.5 $\pm$ 4.6	**98.4** $\pm$ 0.2	97.5 $\pm$ 0.1	87.7 $\pm$ 0.9	**92.0** $\pm$ 0.8	81.5 $\pm$ 0.5	**84.3** $\pm$ 0.2
L31	**34.7** $\pm$ 1.4	29.4 $\pm$ 1.3	38.9 $\pm$ 3.2	**58.8** $\pm$ 0.8	**87.4** $\pm$ 2.8	72.1 $\pm$ 2.2	**95.4** $\pm$ 0.3	95.0 $\pm$ 0.9	92.2 $\pm$ 0.6	**94.3** $\pm$ 0.2	70.6 $\pm$ 2.8	**84.5** $\pm$ 0.9
GEM	**58.0** $\pm$ 1.4	29.7 $\pm$ 0.5	**72.6** $\pm$ 2.5	66.9 $\pm$ 0.5	**85.1** $\pm$ 0.7	72.0 $\pm$ 0.9	**99.6** $\pm$ 0.0	98.0 $\pm$ 0.1	**81.2** $\pm$ 0.5	80.3 $\pm$ 0.6	**78.0** $\pm$ 3.2	62.2 $\pm$ 0.6
Q72	**37.5** $\pm$ 0.0	22.4 $\pm$ 0.4	**75.7** $\pm$ 4.3	59.0 $\pm$ 0.3	**93.8** $\pm$ 1.2	81.5 $\pm$ 1.3	**94.2** $\pm$ 0.5	80.7 $\pm$ 0.3	**89.2** $\pm$ 0.2	86.8 $\pm$ 0.8	**88.3** $\pm$ 2.0	81.8 $\pm$ 0.6
DSK	**38.0** $\pm$ 12.0	29.4 $\pm$ 0.6	58.7 $\pm$ 0.4	**63.4** $\pm$ 1.0	**93.5** $\pm$ 0.7	72.7 $\pm$ 1.8	**95.0** $\pm$ 0.4	86.2 $\pm$ 1.2	**82.7** $\pm$ 0.2	74.8 $\pm$ 1.0	**82.8** $\pm$ 1.8	75.5 $\pm$ 0.5
G4O	**39.3** $\pm$ 2.6	20.6 $\pm$ 0.5	**60.7** $\pm$ 7.7	58.2 $\pm$ 0.1	**89.9** $\pm$ 3.8	72.1 $\pm$ 4.4	99.7 $\pm$ 0.1	**99.8** $\pm$ 0.0	**76.2** $\pm$ 1.9	73.5 $\pm$ 4.2	**83.0** $\pm$ 0.9	82.6 $\pm$ 0.2
Small models												
L8B	30.3 $\pm$ 2.4	**40.8** $\pm$ 3.2	58.1 $\pm$ 0.6	**60.5** $\pm$ 3.4	**79.1** $\pm$ 0.8	75.3 $\pm$ 1.2	**91.8** $\pm$ 1.5	76.1 $\pm$ 2.5	66.5 $\pm$ 1.7	**67.2** $\pm$ 0.3	42.7 $\pm$ 1.7	**87.6** $\pm$ 1.8
Q7B	**26.5** $\pm$ 0.5	21.3 $\pm$ 0.9	**82.8** $\pm$ 0.5	46.2 $\pm$ 0.4	**90.0** $\pm$ 3.0	77.8 $\pm$ 2.1	**90.2** $\pm$ 0.3	45.6 $\pm$ 0.2	**72.8** $\pm$ 0.8	63.2 $\pm$ 0.6	**84.6** $\pm$ 2.9	59.3 $\pm$ 0.7
G4M	**49.2** $\pm$ 4.5	23.8 $\pm$ 1.4	**73.2** $\pm$ 1.8	64.7 $\pm$ 2.4	**87.1** $\pm$ 1.6	65.4 $\pm$ 1.6	**87.5** $\pm$ 0.8	54.6 $\pm$ 1.3	**77.2** $\pm$ 1.2	58.0 $\pm$ 0.7	76.6 $\pm$ 1.7	**85.6** $\pm$ 1.0
Reasoning models												
DSL	**38.7** $\pm$ 5.4	21.6 $\pm$ 0.1	**66.6** $\pm$ 1.9	61.5 $\pm$ 0.0	**93.0** $\pm$ 0.5	76.4 $\pm$ 2.4	**96.0** $\pm$ 1.1	95.8 $\pm$ 1.3	80.0 $\pm$ 0.8	**80.7** $\pm$ 1.3	**83.3** $\pm$ 0.1	75.1 $\pm$ 0.1
O3M	**30.3** $\pm$ 3.0	17.8 $\pm$ 0.6	**67.8** $\pm$ 5.9	54.8 $\pm$ 0.4	**91.7** $\pm$ 0.3	78.4 $\pm$ 0.1	**93.8** $\pm$ 0.1	92.4 $\pm$ 1.4	**83.7** $\pm$ 1.2	80.3 $\pm$ 1.6	**85.1** $\pm$ 1.8	70.7 $\pm$ 0.1

Table 7: Model performance (Aligned vs. Conflicting) on valid datapoints with standard *FaR* method.

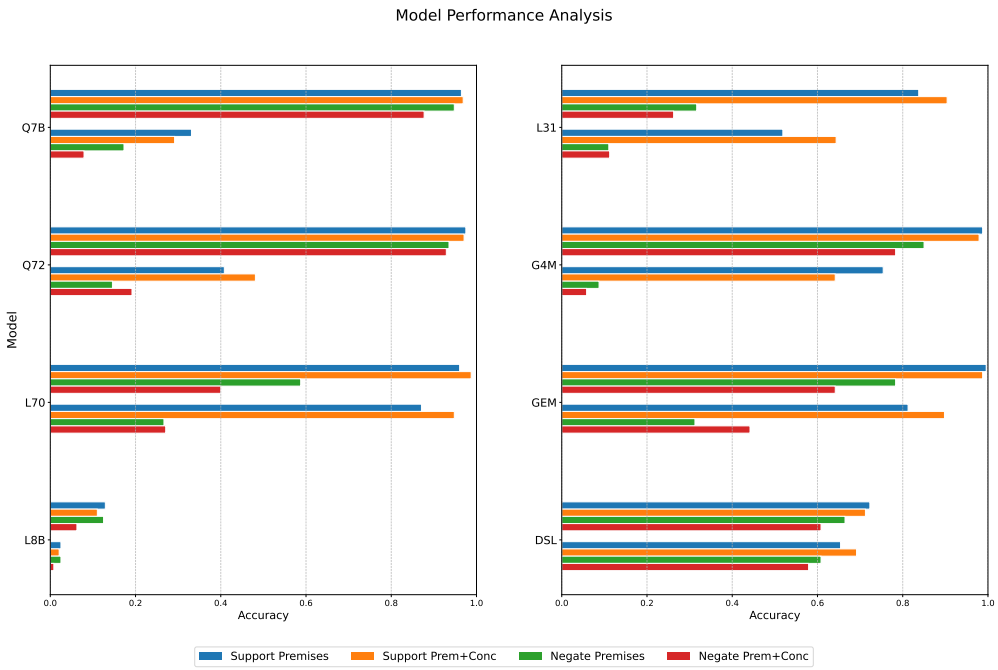


Figure 10: Effect of evidences on logical reasoning. Shows that providing evidences for premises or conclusions (or both) effect the outputs generated by the models even when prompted not to pay attention to factuality.

with the factually correct answer even when it is not necessarily true in the given context.

H.4 Single Prompt *FaR*

In figure 9, we evaluate results on a modified version of *FaR*, where the *FaR* prompt and the evaluation prompt are presented together. This setup does not exhibit any performance gains over the baseline unlike the standard *FaR*, in which the *FaR* prompt and the reasoning prompt are provided separately.

I Open Questions in Knowledge-Conflict Resolution

Our empirical findings raise fundamental questions about the mechanisms of knowledge interference in LLMs and potential resolution strategies. While our interventions show promise, they highlight critical gaps in understanding how logical reasoning interacts with parametric knowledge. Below we outline key open problems informed by our results and existing literature.

I.1 Competing Pathways in Knowledge Interference

The observed performance patterns suggest a potential tension between two systems that warrants further investigation:

Hypothesis 1 (Dual-Path Interference): Reasoning failures in counterfactual scenarios may stem from competition between: *Parametric*

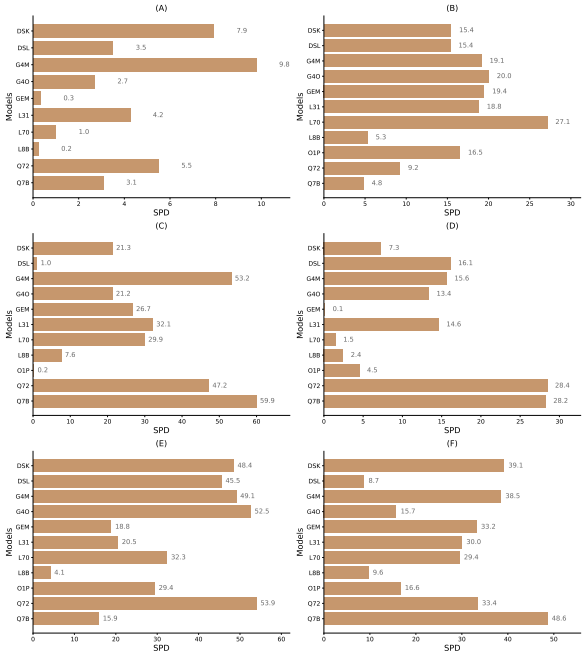


Figure 11: SPD values for various datasets. The Plot shows SPD values of Different models across various Datasets. We can see that the values are largely positive indicating an inherent bias towards factual answers over belief in the premises.

Knowledge Activation: Automatic retrieval of factual information acquired during pre-training (as observed in knowledge utilization studies (Huang et al., 2024)) and *Contextual Reasoning Processes*: Controlled application of logical operations to current premises.

This hypothesized conflict aligns with findings from mechanistic analyses of LLM reasoning (Hong et al., 2025), though the exact neural implementations remain unknown. Future work could employ pathway analysis techniques (Kugarajeevan et al., 2025) to test whether distinct neural substrates mediate these processes.

I.2 Metacognitive Dissonance in Artificial Systems

Our believability assessment intervention parallels human cognitive dissonance resolution strategies (Miller et al., 2015), raising questions about whether LLMs experience analogous conflicts. Recent work demonstrates that LLMs exhibit susceptibility to influence patterns resembling human cognitive dissonance (Griffin et al., 2023), suggesting:

Open Question: Do LLMs develop internal conflict states when parametric knowledge contradicts task premises, and if so, how do resolution strategies differ from human cognition?

Preliminary evidence from our error analysis shows models occasionally produce self-contradictory explanations (e.g., "Although the premises state X, we know Y..."), hinting at unresolved internal conflicts. Systematic investigation using contradiction detection paradigms (Mündler et al., 2024) could quantify this phenomenon.

I.3 Epistemic Compartmentalization Mechanisms

The effectiveness of metacognitive prompts suggests LLMs might transiently suppress parametric knowledge during reasoning tasks. This observation intersects with emerging research on Knowledge editing techniques that modify specific model parameters (Rozner et al., 2024) and Hippocampal indexing approaches for contextual knowledge isolation (Gutiérrez et al., 2025)

Research Direction: Can we identify neural mechanisms that enable temporary belief suspension? Layer-wise relevance propagation studies (Hong et al., 2025) could reveal whether metacognitive prompts induce distinct activation patterns in higher transformer layers.

I.4 Parametric Knowledge as Malleable Belief Systems

Our evidence-based reinforcement results hint at an unexpected plasticity in LLM "beliefs." This aligns with findings that LLMs exhibit illusory truth effects (Griffin et al., 2023) and can transiently adopt fictional premises when provided authoritative citations (Huang et al., 2025).

Open Question: To what extent does parametric knowledge function as a belief system versus a fixed retrieval database? Controlled studies using contradiction learning paradigms (Zhang et al., 2024a) could probe the conditions under which LLMs update or maintain conflicting knowledge.

J Thought Anchor Analysis

More insights on how reasoning changes with *FaR* Intervention. The average response is 17.9 sentences long (95% CI: [7.0, 28.8]; this corresponds to 889 tokens [95% CI: 625, 1154]).

Based on the framework presented by (Venhoff et al., 2025; Bogdan et al., 2025) we label sentences into the sentence categories that capture reasoning functions in logical reasoning as shown in Table 8. We use Algorithm 1 to measure sentence to sentence importance.

We find that *Plan Generation* and *Fact Retrieval* sentences to be the most important sentences while

Category	Function
Problem Setup	Repeating or rephrasing the problem (initial reading or comprehension).
Plan Generation	Stating or deciding on a plan of action (often meta-reasoning).
Fact Retrieval	Recalling facts or problem details (without immediate computation).
Active Computation	Performing simplifications, deductions, or derivative sentences toward the answer.
Result Consolidation	Aggregating intermediate results, summarizing, or preparing the final answer.
Uncertainty Management	Expressing confusion, knowledge-conflicts, re-evaluating, or proposing alternative plans.
Final Answer Emission	Explicitly stating the final boxed answer or earlier chunks that contain the final answer.
Self Checking	Verifying previous steps or reconfirmations.

Table 8: Sentence taxonomy with reasoning functions in problem-solving

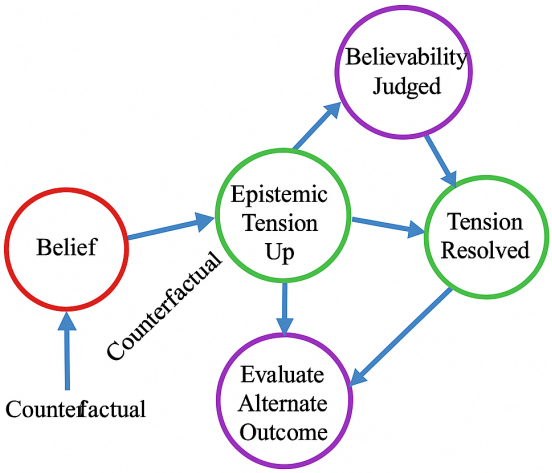


Figure 12: Believability-Guided Cognitive Dissonance Resolution in Counterfactual Reasoning. Conflict between a model's stored knowledge and given premises creates dissonance. Believability assessments help resolve this dissonance by guiding the model toward epistemically appropriate inference strategies.

reasoning normally. With *Uncertainty Management* also having significant impact. While *Plan Generation*, *Active Computation* and *Thinking* sentences to be most influential when reasoning using the *FaR* method. (Figure 13a and 13b).

We studied examples where *FaR* was helpful in turning models' incorrect answers correct. On average we see that the ration of tokens spent on *Problem Setup*, *Fact Retrieval* and *Uncertainty Management* is much higher when reasoning normally (Figure 14). On the other hand, *Active Computation* uses the maximum share of tokens and arrive at the correct answer as opposed to being confused with fact retrieval (Figure 15).

K Case Study : How does reasoning change with knowledge-conflicts?

Here we analyze an example where knowledge effects are visible on the outcome of logical reasoning even when the underlying logical structure is identical.

Consider these questions -

Question A -

```
It's mentioned that if all cars are
    considered vehicles,
then all aircraft should logically be
    classified as planes.
Additionally, it's stated that some
    aircraft are not actually planes.

Are there any cars that aren't considered
     vehicles?
```

Question B -

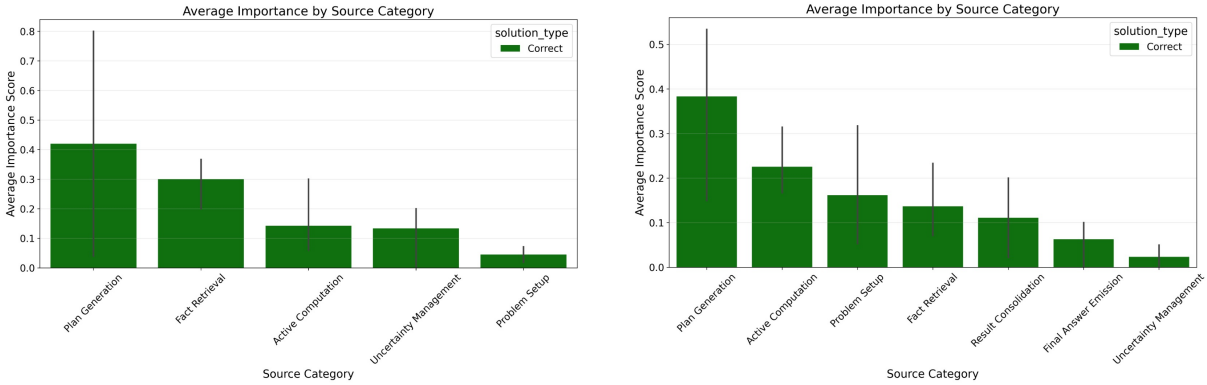
```
It's mentioned that if all vehicles are
    considered cars,
then all aircraft should logically be
    classified as planes.
Additionally, it's stated that some
    aircraft are not actually planes.

Are there any vehicles that aren't
    considered cars?
```

While the logical structures for these questions are identical, *deepseek-r1-distill-llama-8b* returns *No* for Question A while *Yes* for Question B upon multiple sampling.

Since the only thing that changed is the order in which the entities *cars* and *vehicles* occur, the supposed difference is that answers must come from their hierarchical relationship.

Reasoning Chains for A and B are listed in Tables 9 and 10. For Problem A, the top influencing steps were: Step 0 (*Problem Setup*), Step 3 (*Fact Retrieval*), and Step 4 (*Plan Generation*). For Problem B, the top influencing steps were: Step T (*Active Computation*), Step 6 (*Fact Retrieval*), and Step 4 (*Problem Setup*). We compare the reasoning processes across problems based on their semantic composition and category balance. Figure 16 shows the importances



(a) Normal reasoning

(b) Reasoning with *FaR*

Figure 13: Comparison of average importance across different categories for normal reasoning vs. reasoning with *FaR*.

Algorithm 1 Computation of Importance Score $\text{importance}(i \to j)$

Require: Sentence indices i, j where $j > i$; threshold $t = 0.8$
Ensure: Importance score $\text{importance}(i \to j)$

```
Obtain rollouts R_keep where sentence i is kept (resampling from i+1)
Obtain rollouts R_remove where sentence i is removed (resampling from i)
for each rollout r ∈ R_keep do
    Extract all sentences S_r from rollout r
    Compute sentence embeddings for all s ∈ S_r and target sentence j
    Compute cosine similarity sim(s, j) for each s ∈ S_r
    s* ← arg max_{s∈S_r} sim(s, j)
    if sim(s*, j) ≥ t then
        Add s* to matches_keep
    end if
end for
for each rollout r ∈ R_remove do
    Extract all sentences S_r from rollout r
    Compute sentence embeddings and similarities sim(s, j) for each s ∈ S_r
    s* ← arg max_{s∈S_r} sim(s, j)
    if sim(s*, j) ≥ t then
        Add s* to matches_remove
    end if
end for
Compute match rates:
match_rate_keep = |matches_keep| / |R_keep|
match_rate_remove = |matches_remove| / |R_remove|
return importance(i → j) = match_rate_keep − match_rate_remove
```

Problem A shows a balanced mix of *Problem Setup*, *Fact Retrieval*, and *Uncertainty Management*, indicating both contextual understanding and factual grounding. In contrast, Problem B is dominated by *Active Computation* and *Fact Retrieval* steps across both influential and dependent stages, suggesting a stronger focus on recalling external facts the problem correctly rather than interpreting and evaluating.

Idx	Chunk Text	Function Tags
0	Let's break this down:	problem_setup
1	1. The initial statement says, "if all cars are considered vehicles, then all aircraft should logically be classified as planes." This is a conditional statement (if-then).	problem_setup
2	It does not assert that all cars are vehicles or that all aircraft are planes; it merely states a hypothetical scenario and its logical consequence.	fact_retrieval
3	2. The statement "some aircraft are not actually planes" directly contradicts the hypothetical scenario where all aircraft are classified as planes.	fact_retrieval
4	However, this contradiction does not negate the conditional statement itself but rather indicates that the condition under which all aircraft would be planes is not met.	plan_generation
5	3. The question asks if there are any cars that aren't considered vehicles.	problem_setup
6	The information given does not directly address this question.	uncertainty_management
7	It presents a hypothetical scenario about cars and vehicles and provides a separate statement about aircraft and planes.	fact_retrieval
8	4. Logically, the truth of the statements about aircraft and planes does not affect the classification of cars as vehicles.	plan_generation
9	The initial conditional statement does not assert that all cars are vehicles; it simply states a hypothetical scenario.	fact_retrieval
10	5. Since the given information does not provide any evidence or logical pathway to conclude that there are cars that aren't considered vehicles, we cannot logically deduce an affirmative answer.	result_consolidation
11	Therefore, based on the logical analysis of the given statements, the answer to the question "Are there any cars that aren't considered vehicles?" is not supported by the information provided, leading to a conclusion that we cannot affirmatively say there are cars not considered vehicles based on the given context.	final_answer_emission

Table 9: Reasoning chain for Problem A.

Idx	Chunk Text	Function Tags
0	The statement is, if all vehicles are considered cars, then all aircraft should logically be classified as planes.	problem_setup
1	This is a conditional statement, which can be rewritten as: if all vehicles are cars, then all aircraft are planes.	problem_setup
2	Now, it's also stated that some aircraft are not actually planes.	problem_setup
3	This is the second piece of information.	problem_setup
4	The question is asking us if there are any vehicles that aren't considered cars.	problem_setup
5	In other words, we are looking for the opposite of the first part of the conditional statement.	problem_setup
6	The second piece of information is telling us that the second part of the conditional statement is false (some aircraft are not planes).	fact_retrieval
7	This means we cannot conclude the first part of the conditional statement is false (all vehicles are cars).	active_computation
8	In fact, this is an example of the logical fallacy of denying the antecedent.	thinking
9	We cannot logically conclude the opposite of the first part of the conditional statement from the opposite of the second part of the conditional statement.	active_computation
10	However, we are given that if all vehicles are considered cars, then all aircraft should logically be classified as planes.	fact_retrieval
11	Since we are given that some aircraft are not actually planes, then it must be the case that not all vehicles are cars.	active_computation

Table 10: Reasoning chain for Problem B.

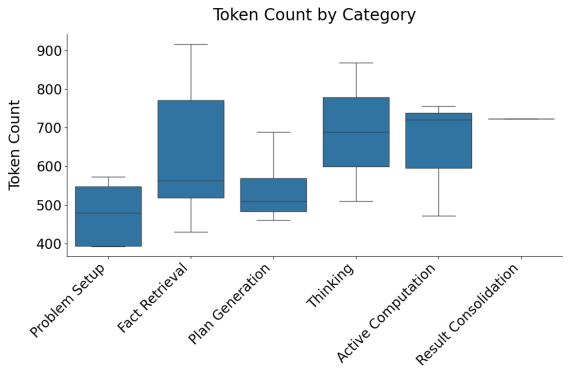


Figure 14: Average Token Count by Category when reasoning normally

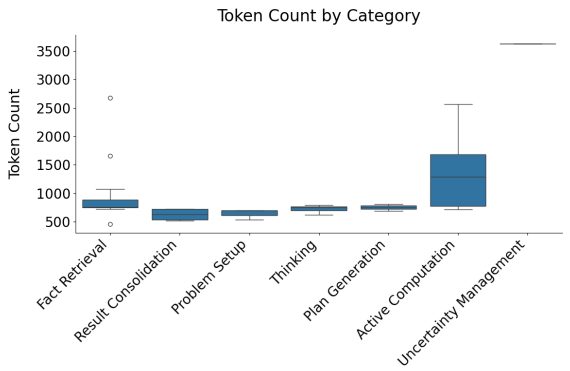
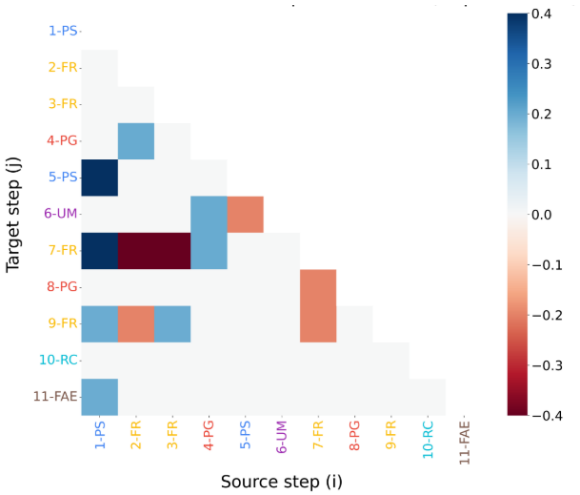
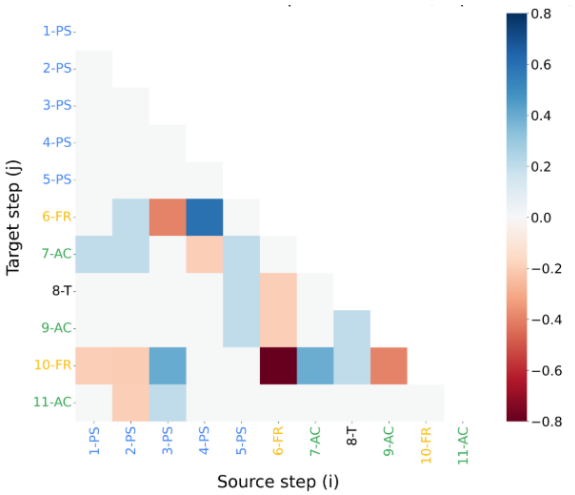


Figure 15: Average Token Count by Category when reasoning with *FaR*



(a) Steps Importance Matrix for Problem A which is *Knowledge-aligned*



(b) Steps Importance Matrix for Problem B which is *Knowledge-conflicting*

Figure 16: Comparison of step importance matrices across reasoning cases. Subfigure (a) corresponds to Problem A and (b) to Problem B. Each cell (i, j) shows the causal importance of sentence i on sentence j, calculated as the difference in the probability that sentence j semantically occurs (with cosine similarity > 0.8) when sentence i is present versus when it is resampled.