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

Anish R Joishy*
IIIT Hyderabad
anish.joishy@research.iiit.ac.in

Ishwar B Balappanawar*
IIIT Hyderabad
ishwar.balappanawar@students.iiit.ac.in

Vamshi Krishna Bonagiri*
IIIT Hyderabad
vamshi.b@research.iiit.ac.in

Manas Gaur
University of Maryland, Baltimore County
manas@umbc.edu

Krishnaprasad Thirunarayan
Wright State University
t.k.prasad@wright.edu

Ponnurangam Kumaraguru
IIIT Hyderabad
pk.guru@iiit.ac.in

***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].**

***Index Terms*—Large Language Models, Logical Reasoning, Knowledge Conflicts, Counterfactual Reasoning, Benchmarking, Belief Bias**

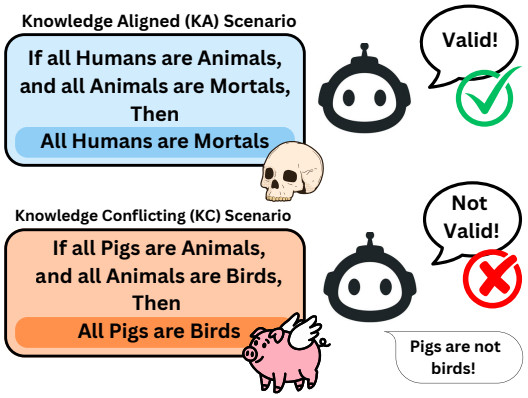

Fig. 1: When asked to reason solely based on the given premises, LLMs reliably validate arguments that are **knowledge-aligned** (top), but their performance degrades on logically identical arguments that are **knowledge-conflicting** (bottom). In the above case, the model labels "All Pigs are Birds" as false even though it follows from the given premises. This is observed despite asking the model to reason solely based on premises and ignore factuality.

I. 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 [1]–[5]. Despite these advances, they often exhibit a significant performance degradation (see Fig 1) when reasoning with information that conflicts with their parametric knowledge obtained during training [6]–[11].

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 [9]. 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 [12], and failure to reason with them could lead to unreliable performance [13]. Additionally, prior research also suggests that evaluating reasoning in counterfactual situations may serve as a more robust assessment of a model's reasoning capabilities [10], as standard reasoning tasks can potentially be *hacked* through pattern matching [10], [14]–[16].

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?) [17]. Consequently, existing benchmarks either test complex logical reasoning without controlling for knowledge conflicts, or test simple knowledge conflicts with shallow, single-step reasoning. As

[1]Our data and code are available at https://github.com/Sukin272/CounterLogic.git
* Authors contributed equally to this work

summarized in Table I, 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 I), 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 [18], 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.

In this paper, we first use *CounterLogic* (Section III) and other benchmarks to diagnose a pervasive failure in LLM reasoning (Section IV). We then propose a metacognitive intervention, *FaR*, to address this failure (Section V) 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, specifically a significant performance drop when logical validity clashes with a model's parametric knowledge; and **(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.

II. Related Work

Recent advancements in prompting techniques, such as chain-of-thought and tree-of-thought, have significantly enhanced LLM reasoning capabilities [2], [3], [22]. Despite these improvements on standard benchmarks [19], [23], models continue to exhibit systematic failures akin to human cognitive biases, particularly when handling negation, quantifiers, and abstract variables [6], [21], [24], [25]. This fragility is most evident in counterfactual scenarios where logical validity conflicts with parametric knowledge [10], [26], prompting researchers to explore hybrid symbolic methods and formal verification to improve consistency and robustness [27]–[29].

Dataset	Size	# Steps	Conflict	Balance
LogicBench [19]	2,020	$1 \sim 5$	×	×
FOLIO [20]	1,435	$0 \sim 7$	×	×
KNOT [15]	5,500	$1 \sim 2$	✓	×
Syllogistic [21]	2,120	2	✓	×
CounterLogic (Ours)	1,800*	$2 \sim 9$*	✓	✓

TABLE I: 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.*

The tension between an LLM's vast parametric memory and contradictory context creates significant "knowledge conflicts" [11], [30], [31]. Larger models often default to prior knowledge over new evidence [32], [33], a tendency that severely degrades performance in counterfactual tasks requiring the acceptance of false premises [3], [26], [34]. While strategies such as counterfactual data augmentation and specialized prompting have been proposed [17], [35], [36], they fail to fully resolve these conflicts, particularly within complex, multi-step logical deduction tasks [5].

This behavior mirrors the human "belief bias," where the believability of a conclusion distorts logical judgment [37], leading to an "illusion of objectivity" [6], [24], [38]–[40]. Just as humans utilize metacognition to separate beliefs from logical inferences [41], emerging research suggests LLMs can benefit from similar capabilities, such as estimating uncertainty, self-evaluation, or explicitly identifying knowledge conflicts [9], [42]. Our work builds on this parallel, investigating how metacognitive awareness can enhance the robustness of logical reasoning in language models.

III. The CounterLogic Dataset

Existing benchmarks in evaluating LLM reasoning [2], [3] fail to systematically disentangle logical validity from knowledge alignment. As highlighted in Table I, current datasets either focus on complex logical structures while ignoring knowledge conflicts like LogicBench [19] and FOLIO [20], or test simple knowledge conflicts single-step reasoning like KNOT [15] and Reasoning & Reciting [10]. 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 multi-step reasoning. The benchmark is easily extendable, contains 1,800 examples across 9 logical schemas, and features problems that scale in complexity from 1 to 9 reasoning steps.

A. Dataset Construction

Generation Pipeline: We begin with hierarchical triples $(a \subset b \subset c)$ to populate templates forming atomic propositions

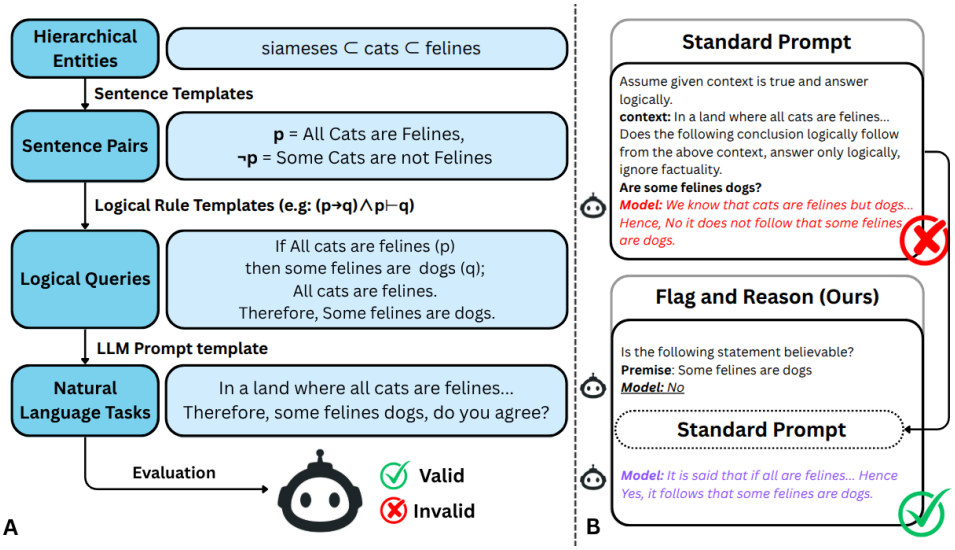

Fig. 2: (A) **Dataset Preparation**: The dataset features hierarchical entity triples (e.g., siameses ⊂ cats ⊂ felines) mapped to 8 logical sentence templates across 9 logical schemas (such as Modus Ponens, Modus Tollens etc.). Each example is balanced across validity (50% valid/invalid) and believability (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.

$(p, \neg p)$ with controlled factuality (50% correct hierarchy, 50% inverted). These propositions are mapped into 9 formal logical schemas (e.g., Modus Ponens, Disjunctive Syllogism) following LogicBench [19]. See figure 2(A) for more details.

Balancing & Depth: The dataset is balanced across two dimensions: **(1) Logical Validity** (Valid/Invalid) and **(2) Knowledge Alignment**, ensuring an equal split between factually consistent and contradictory conclusions. To prevent pattern matching [10], [43], final queries are paraphrased via GPT-4o. Finally, we scale reasoning complexity for implication schemas $(A \rightarrow B)$ by injecting intermediate steps $(A \rightarrow X_1 \rightarrow \cdots \rightarrow X_i \rightarrow B)$, enabling fine-grained depth analysis.

IV. LLM Reasoning in Counterfactuals

A. Experimental Setup

Our investigation evaluates 11 LLMs spanning different architectures, parameter scales, and training paradigms. Evaluations were conducted across our *CounterLogic* and five other diverse reasoning datasets: *Hierarchical Syllogisms*, *KNOT (Implicit & Explicit)*, *FOLIO*, and *LogicBench* (see Table I). To ensure robust performance measures, we employ self-consistency checks accounting for the generation variability common in LLMs [44]. Results presented in this section establish a baseline using a standard CoT prompting strategy [3] without our *FaR* intervention, as this allows us to isolate and quantify the interference of parametric knowledge on logical deduction. Models were explicitly instructed not to use their real world knowledge and ignore real world contradictions.

B. 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. Notably, this failure occurs even when models are explicitly instructed to reason based solely on the provided premises and ignore factual correctness. 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.

C. Deconstructing the Failure

We leveraged the controlled design of the *CounterLogic* benchmark to investigate whether this performance degradation is merely a superficial artifact of problem complexity. While a common hypothesis suggests that models falter simply because counterfactual problems are inherently more complex [10], our findings refute this, revealing the failure to be fundamental and independent of structural depth. Specifically, the performance gap between knowledge-aligned and counterfactual examples remains consistent ($\sim$ 6%) regardless of the number of intermediate steps introduced into the reasoning chain (See Section III-A). 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 4, suggest a more nuanced reality. While the failure is a universal

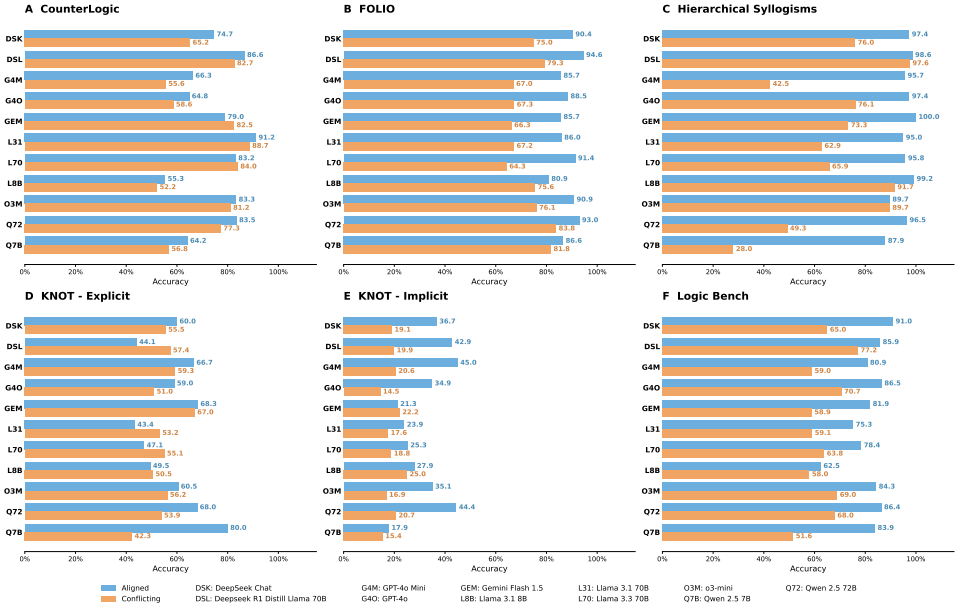

Fig. 3: Average accuracy comparison for the Aligned and Conflicting Scenarios. This plot shows comparison between Knowledge-Aligned and Knowledge-Conflicting scenarios where the ground truth is valid. Blue bars represent Knowledge-Aligned example accuracy, while orange bars indicate Knowledge-Conflicting 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%. 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.

phenomenon across all 9 logical schemas in *CounterLogic*, 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% 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 (Fig 4), aligning with findings that negation remains a core challenge for LLMs [45].

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.

V. Flag & Reason (FaR): A MetaCognitive Intervention

Findings in Section IV reveal a critical vulnerability in LLM reasoning: a systematic failure not of logic, but of cognitive

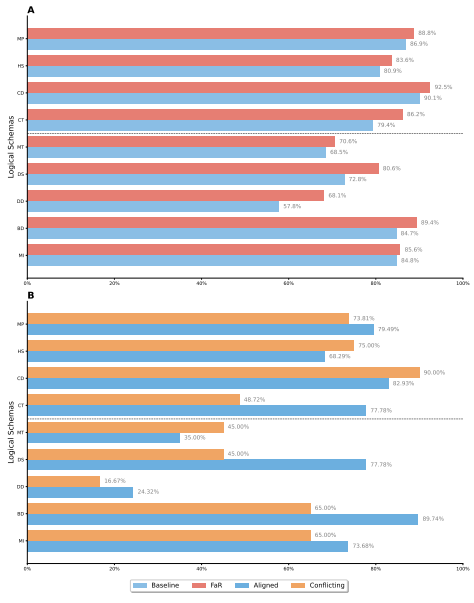

Fig. 4: Variation across logical schemas. The dotted line separates schemas with negation (bottom) and without negation (top). Overall, negation causes a significant baseline performance drop (84.3% vs. 73.7%). **(A)** FaR vs. Baseline: FaR (pink) consistently outperforms the baseline (blue), particularly on harder schemas like DD (+11%). **(B)** Knowledge Conflict Impact: Comparison of Knowledge-Aligned (blue) and Knowledge-Conflicting (orange) scenarios. Conflicting context causes severe degradation, notably ∼32% in schema DS. The 9 formal logical schemas are 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)

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.

A. 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 [46] 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 [37]. 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 [18], [47], [48].

B. 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: **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). **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. This adds minimal input overhead and just one additional output token.

The separation of these two steps is critical. 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. This finding supports our central hypothesis: FaR works by inducing a form of **epistemic compartmentalization** [49]. 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 "Reason" step.

C. Results

Our experiments confirm that the *FaR* intervention is highly effective. As detailed in Figure 5, *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 (Fig V).

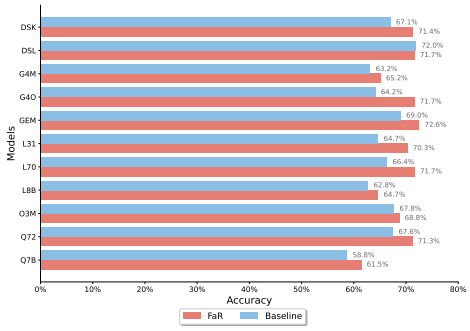

Fig. 5: Accuracy comparison between the baseline setup and our *FaR* setup averaged across all the datasets. The x-axis represents accuracy and the y-axis represents the models. Pink bars represent the *FaR* method and the blue bars represent the baseline. We can see that our method *FaR* improves performance across all the models with the most noticeable improvement on GPT-4o (G4O) of about 7%.

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 [20]. On our *CounterLogic* benchmark specifically, *FaR* increased overall accuracy by 5%. Furthermore, this improvement is pervasive across all logical schemas in *CounterLogic*. As shown in Figure 4(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%.

D. Thought Anchor Analysis

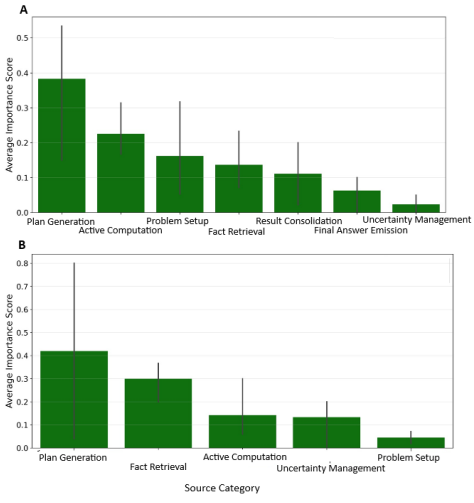

Fig. 6: Comparison of average importance across different categories for normal reasoning vs. reasoning with *FaR*. y axis shows the importance scores for the corresponding categories on the x axis. **A** shows the scores with FaR whereas **B** shows the scores without FaR. We can see that FaR successfully shifts the importance from Fact Retrieval and Uncertainty management to Active Computation.

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 [50]. 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. 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.

VI. Conclusion and Future Work

Our research confirms that LLMs experience a critical reasoning failure when logical validity conflicts with parametric knowledge, resulting in a consistent 14% accuracy drop in counterfactual scenarios. We attribute this to a failure of *belief inhibition*, a cognitive bias where models struggle to suppress ingrained training data. The introduced *Flag & Reason (FaR)* intervention successfully halves this performance gap to 7% and increases overall accuracy by 4%. By prompting models to explicitly acknowledge conflicts, FaR enables epistemic compartmentalization, anchoring deductions in the provided context rather than contradictory internal knowledge. Future work can explore internal representations of these conflicts and extend metacognitive strategies to real-world domains.

VII. Limitations

This study's scope is primarily restricted to foundational categorical syllogisms and propositional logic within the *CounterLogic* benchmark. Furthermore, results are specific to the 11 models tested and may change as architectures evolve. The evaluation was also limited to English and common-knowledge concepts, meaning effectiveness could vary across different languages or specialized domains. Additionally, parallels to human cognition remain conceptual and do not imply identical processing. Finally, the use of LLMs for synthetic query generation may have introduced subtle biases or inconsistencies.

VIII. Acknowledgment

This work was supported by the OpenAI Research Access Program which partially provided the necessary funding for API credits. We thank Hanuma, Sumit Kumar, Ameya Rathod, Vedant SP, Vaishnavi Shivkumar, Hari Shankar and other members from the Precog research group for their insightful discussions and valuable feedback. Manas Gaur, TK Prasad, and P. Kumaraguru provided valuable feedback during all phases of the project. In accordance with IEEE policy on the use of AI-generated text, we acknowledge the use of Large Language Models (Gemini 2.5 Pro and Gemini 3 Pro) to assist with grammatical corrections and refining sentence structure throughout this manuscript.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[4] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi, "Least-to-most prompting enables complex reasoning in large language models," in *International Conference on Learning Representations*, 2023.

[5] N. Patel, M. Kulkarni, M. Parmar, A. Budhiraja, M. Nakamura, N. Varshney, and C. Baral, "Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models," in *Proceedings of EMNLP*, 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.1160.pdf

[6] I. Dasgupta, A. K. Lampinen, S. C. Y. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, and F. Hill, "Language models show human-like content effects on reasoning tasks," Jul. 2024, arXiv:2207.07051 [cs]. [Online]. Available: http://arxiv.org/abs/2207.07051

[7] Z. Jin, P. Cao, Y. Chen, K. Liu, X. Jiang, J. Xu, L. Qiuxia, and J. Zhao, "Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models," in *Proceedings of LREC-COLING*, 2024, pp. 10 142–10 151.

[8] Z. Su, J. Zhang, X. Qu, T. Zhu, Y. Li, J. Sun, J. Li, M. Zhang, and Y. Cheng, "Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm," 2024. [Online]. Available: https://arxiv.org/abs/2408.12076

[9] Y. Wang, S. Feng, H. Wang, W. Shi, V. Balachandran, T. He, and Y. Tsvetkov, "Resolving Knowledge Conflicts in Large Language Models," Oct. 2024, arXiv:2310.00935 [cs]. [Online]. Available: http://arxiv.org/abs/2310.00935

[10] Z. Wu, L. Qiu, A. Ross, E. Akyürek, B. Chen, B. Wang, N. Kim, J. Andreas, and Y. Kim, "Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks," Mar. 2024, arXiv:2307.02477 [cs]. [Online]. Available: http://arxiv.org/abs/2307.02477

[11] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu, "Knowledge Conflicts for LLMs: A Survey," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 8541–8565. [Online]. Available: https://aclanthology.org/2024.emnlp-main.486/

[12] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018. [Online]. Available: https://en.wikipedia.org/wiki/The_Book_of_Why

[13] Y. Zhang, W. Wang, X. Liu, Y. Chen, and Z. Li, "Bridging the gap between llms and human intentions," *arXiv preprint arXiv:2502.09101*, 2024. [Online]. Available: https://arxiv.org/abs/2502.09101

[14] M. Lewis and M. Mitchell, "Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models," *arXiv preprint arXiv:2402.08955*, 2024.

[15] Y. Liu, Z. Yao, X. Lv, Y. Fan, S. Cao, J. Yu, L. Hou, and J. Li, "Untangle the KNOT: Interweaving Conflicting Knowledge and Reasoning Skills in Large Language Models," Apr. 2024, arXiv:2404.03577 [cs]. [Online]. Available: http://arxiv.org/abs/2404.03577

[16] D. Kaushik, E. Hovy, and Z. C. Lipton, "Learning the difference that makes a difference with counterfactually-augmented data," in *Proceedings of the International Conference on Learning Representations*, 2020.

[17] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su, "Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts," Feb. 2024, arXiv:2305.13300 [cs]. [Online]. Available: http://arxiv.org/abs/2305.13300

[18] L. Fletcher and P. Carruthers, "Metacognition and reasoning," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1594, pp. 1366–1378, 2012.

[19] M. Parmar, N. Patel, N. Varshney, M. Nakamura, M. Luo, S. Mashetty, A. Mitra, and C. Baral, "Logicbench: Towards systematic evaluation of logical reasoning ability of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2404.15522

[20] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, W. Zhou, J. Coady, D. Peng, Y. Qiao, L. Benson, L. Sun, A. Wardle-Solano, H. Szabo, E. Zubova, M. Burtell, J. Fan, Y. Liu, B. Wong, M. Sailor, A. Ni, L. Nan, J. Kasai, T. Yu, R. Zhang, A. R. Fabbri, W. Kryscinski, S. Yavuz, Y. Liu, X. V. Lin, S. Joty, Y. Zhou, C. Xiong, R. Ying, A. Cohan, and D. Radev, "Folio: Natural language reasoning with first-order logic," 2024. [Online]. Available: https://arxiv.org/abs/2209.00840

[21] L. Bertolazzi, A. Gatt, and R. Bernardi, "A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences," 2024. [Online]. Available: https://arxiv.org/abs/2406.11341

[22] N. Shinn, S. Yao, K. Zhao, D. Yu, E. Zhao, D. Zhao, and D. Radev, "Tree of thoughts: Deliberate problem solving with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.10601

[23] E. Clark, O. Tafjord, K. Richardson, A. Sabharwal, and H. Hajishirzi, "Transformers as soft reasoners over language," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 3882–3894. [Online]. Available: https://aclanthology.org/2020.acl-main.358

[24] T. Eisape, M. H. Tessler, I. Dasgupta, F. Sha, S. v. Steenkiste, and T. Linzen, "A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models," Apr. 2024, arXiv:2311.00445 [cs]. [Online]. Available: http://arxiv.org/abs/2311.00445

[25] S. Singh, S. Goel, S. Vaduguru, and P. Kumaraguru, "Probing negation in language models," 2024.

[26] Y. Chen, V. K. Singh, J. Ma, and R. Tang, "Counterbench: A benchmark for counterfactual reasoning in large language models," *arXiv preprint arXiv:2502.11008*, 2025. [Online]. Available: https://arxiv.org/abs/2502.11008

[27] B. Estermann, L. A. Lanzendörfer, and R. Wattenhofer, "Reasoning effort and problem complexity: A scaling analysis in large language models," *arXiv preprint arXiv:2503.15113*, 2025. [Online]. Available: https://arxiv.org/abs/2503.15113

[28] J. Xu, H. Fei, L. Pan, Q. Liu, M. Lee, and W. Hsu, "Faithful logical reasoning via symbolic chain-of-thought," *Annual Meeting of the Association for Computational Linguistics*, 2024.

[29] A. Toroghi, W. Guo, and S. Sanner, "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*, 2024. [Online]. Available: https://arxiv.org/abs/2403.01390

[30] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2463–2473. [Online]. Available: https://aclanthology.org/D19-1250/

[31] A. Roberts, C. Raffel, and N. Shazeer, "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" Oct. 2020, arXiv:2002.08910 [cs]. [Online]. Available: http://arxiv.org/abs/2002.08910

[32] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh, "Entity-based knowledge conflicts in question answering," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 7052–7063. [Online]. Available: https://aclanthology.org/2021.emnlp-main.565/

[33] C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, Y. Wang, L. Yang, J. Wang, X. Xie, Z. Zhang, and Y. Zhang, "Survey on factuality in large language models: Knowledge, retrieval and domain-specificity," 2023. [Online]. Available: https://arxiv.org/abs/2310.07521

[34] H. Lin, X. Wang, R. Yan, B. Huang, H. Ye, J. Zhu, Z. Wang, J. Zou, J. Ma, and Y. Liang, "Generative reasoning with large language models," 2025. [Online]. Available: https://arxiv.org/abs/2504.02810

[35] Z. Chen, Q. Gao, A. Bosselut, A. Sabharwal, and K. Richardson, "Disco: Distilling counterfactuals with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2212.10534

[36] E. Neeman, R. Aharoni, O. Honovich, L. Choshen, I. Szpektor, and O. Abend, "Disentqa: Disentangled question answering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 056–10 070. [Online]. Available: https://aclanthology.org/2023.acl-long.559/

[37] H. Markovits and G. Nantel, "The belief-bias effect in the production and evaluation of logical conclusions," *Memory & Cognition*, vol. 17, no. 1, pp. 11–17, Jan. 1989. [Online]. Available: https://doi.org/10.3758/BF03199552

[38] Z. Kunda, "The case for motivated reasoning," *Psychological Bulletin*, vol. 108, no. 3, pp. 480–498, 1990. [Online]. Available: https://doi.org/10.1037/0033-2909.108.3.480

[39] D. Trippas, S. Handley, and M. Verde, "Fluency and belief bias in deductive reasoning: new indices for old effects," *Frontiers in Psychology*, vol. 5, p. 631, 06 2014.

[40] O. Macmillan-Scott and M. Musolesi, "(ir)rationality and cognitive biases in large language models," *Royal Society Open Science*, vol. 11, no. 3, p. 240255, 2024. [Online]. Available: https://doi.org/10.1098/rsos.240255

[41] I. Douven, S. Elqayam, and H. Singmann, "Conditionals and inferential connections: Toward a new semantics," *Cognition*, vol. 178, pp. 31–45, 2018. [Online]. Available: https://doi.org/10.1016/j.cognition.2018.05.005

[42] K. Zhou, D. Jurafsky, and T. Hashimoto, "Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models," Nov. 2023, arXiv:2302.13439 [cs]. [Online]. Available: http://arxiv.org/abs/2302.13439

[43] C. Xie, Y. Huang, C. Zhang, D. Yu, X. Chen, B. Y. Lin, B. Li, B. Ghazi, and R. Kumar, "On memorization of large language models in logical reasoning," *arXiv preprint arXiv:2410.23123*, 2024. [Online]. Available: https://arxiv.org/abs/2410.23123

[44] V. K. Bonagiri, S. Vennam, P. Govil, P. Kumaraguru, and M. Gaur, "Sage: Evaluating moral consistency in large language models," *arXiv preprint arXiv:2402.13709*, 2024.

[45] T. Vrabcová, M. Kadlčík, P. Sojka, M. Štefánik, and M. Spiegel, "Negation: A pink elephant in the large language models' room?" 2025. [Online]. Available: https://arxiv.org/abs/2503.22395

[46] E. Harmon-Jones, *Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology*, 2nd ed. American Psychological Association, 2019. [Online]. Available: http://www.jstor.org/stable/j.ctv1chs6tk

[47] J. Ma, W. Lv, and X. Ren, "Neural correlates of belief-bias reasoning as predictors of critical thinking: Evidence from an fnirs study," *Journal of Intelligence*, vol. 13, no. 9, p. 106, Aug. 2025.

[48] Y. Wang and Y. Zhao, "Metacognitive prompting improves understanding in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2308.05342

[49] J. Thomas, C. Ditzfeld, and C. Showers, "Compartmentalization: A window on the defensive self 1," *Social and Personality Psychology Compass*, vol. 10, pp. 719–7 311 111, 10 2013.

[50] P. C. Bogdan, U. Macar, N. Nanda, and A. Conmy, "Thought anchors: Which llm reasoning steps matter?" 2025. [Online]. Available: https://arxiv.org/abs/2506.19143