If Pigs Could Fly... Can LLMs Logically Reason Through Counterfactuals?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) demonstrate impressive reasoning capabilities in familiar contexts, but struggle when the context conflicts with their parametric knowledge. To investigate this phenomenon, we introduce **CounterLogic**, a dataset containing 1,800 examples across 9 logical schemas, explicitly designed to evaluate logical reasoning through counterfactual (hypothetical knowledge-conflicting) scenarios. Our systematic evaluation of 11 LLMs across 6 different datasets reveals a consistent performance degradation, with accuracies dropping by 27% on average when reasoning through counterfactual information. We propose "*Self-Segregate*", a prompting method enabling metacognitive awareness (explicitly identifying knowledge conflicts) before reasoning. Our method dramatically narrows the average performance gaps from 27% to just 11%, while significantly increasing the overall accuracy (+7.5%). We discuss the implications of these findings and draw parallels to human cognitive processes, particularly on how humans disambiguate conflicting information during reasoning tasks. Our findings offer practical insights for understanding and enhancing LLMs' reasoning capabilities in real-world applications, especially where models must logically reason independently of their factual knowledge. Our data and code are available here.

1 Introduction

LLMs have demonstrated remarkable reasoning capabilities across diverse domains, exhibiting proficiency in tasks ranging from elementary problem solving to complex-level multi-step reasoning challenges (Wei et al., 2023; Schick and Schütze, 2021; Kojima et al., 2022; Brown et al., 2020; Zhou et al., 2023a; Patel et al., 2024). Despite these advances, they often exhibit a significant performance degradation when reasoning with information that conflicts with their parametric knowledge (knowledge acquired during pre-training) (Xu et al., 2024b; Dasgupta et al., 2024; Wu et al., 2024; Lampinen et al., 2024; Wang et al., 2024; Jin et al., 2024; Su et al., 2024).

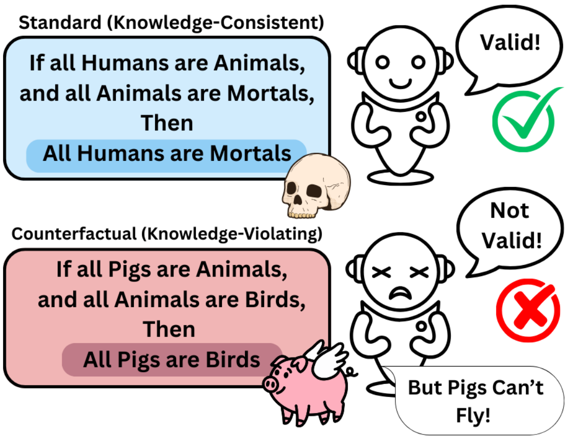

Standard (Knowledge-Consistent)
If all Humans are Animals, and all Animals are Mortals, Then All Humans are Mortals
Valid!
Counterfactual (Knowledge-Violating)
If all Pigs are Animals, and all Animals are Birds, Then All Pigs are Birds
Not Valid!
But Pigs Can't Fly!

Figure 1: Example tasks demonstrating LLM reasoning. While LLMs correctly reason through standard, knowledge-consistent tasks, they often incorrectly assess counterfactual (hypothetical knowledge-conflicting) tasks despite having the same logical structure.

In Figure 1, the two syllogisms (A logical argument with two premises and a conclusion) are logically equivalent. However, while LLMs excel at reasoning through the first example, they often struggle significantly with the second, despite being explicitly instructed to reason based solely on the given premises (Trichelair et al., 2023; Talmor et al., 2020). This disparity suggests that when faced with premises that contradict their parametric knowledge, LLMs often fail to maintain consistent reasoning performance.

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). Consider the question: "If the Earth had two suns, how would seasons differ from what we experience now?". Such situations arise frequently in everyday contexts (Pearl and Mackenzie, 2018), and failure to reason in them could lead to unreliable performance (Zhang et al., 2024). 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; Wu et al., 2024; Liu et al., 2024; McCoy et al., 2019; Kaushik et al., 2020)

While knowledge conflicts are actively studied in language models, 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). These studies typically examine how models handle conflicts when retrieving or extracting knowledge directly from their parameters or from provided text. However, there has been limited exploration of how knowledge conflicts affect complex multi-step logical reasoning processes, which is a capability essential for reliable AI systems (Wang et al., 2024; Liu et al., 2024).

To address this gap, we introduce the **CounterLogic** dataset, specifically designed to evaluate complex logical reasoning in counterfactual scenarios. CounterLogic features approximately 1,800 examples spanning 9 logical schemas, carefully balanced across knowledge-consistent (contexts that align with parametric knowledge) and knowledge-conflicting (counterfactual) scenarios. Through a systematic evaluation of 11 state-of-the-art LLMs on 6 different datasets (including CounterLogic), we demonstrate a consistent pattern of performance degradation (-27% on average) when reasoning through counterfactual statements.

We introduce "*Self-Segregation*", a metacognitive intervention that involves identifying knowledge conflicts before reasoning through a task. Through a series of experiments, we show that this simple strategy, used on top of existing methods such as chain-of-thought (COT) prompting (Wei et al., 2023), significantly boosts LLM reasoning abilities, specifically in counterfactual scenarios. Our results show that with *Self-Segregation*, the average accuracy gap between knowledge-consistent and knowledge-violating scenarios drops by 16% (from 27% to 11%), while the overall accuracy improves by 7.5%.

Our findings suggest that the initial performance disparity could stem from unresolved or ignored knowledge conflicts rather than inherent limitations in logical reasoning capabilities. Notably, these performance patterns in LLMs mirror human cognitive reasoning processes. By introducing self-segregation strategies, we can potentially enable LLMs to more effectively compartmentalize conflicting information and apply logical operations, independent of their parametric factual knowledge (Thomas et al., 2013). Our approach draws inspiration from human metacognitive strategies for resolving ambiguities and knowledge conflicts, suggesting a promising direction for enhancing logical reasoning capabilities in language models.

Our contributions can be summarized as follows:

1. We introduce CounterLogic, a novel dataset for evaluating logical reasoning in counterfactual scenarios, and demonstrate that contemporary models consistently underperform in these contexts despite their strong performance otherwise.

2. We propose a simple yet effective metacognitive awareness intervention, *Self-Segregation*, that involves prompting models to explicitly identify knowledge conflicts before reasoning. Our method significantly improves reasoning in knowledge-violating contexts, reducing the performance gap by 16%.

3. Through a series of experiments, we study and discuss how knowledge conflicts impair reasoning in LLMs and how metacognitive interventions can mitigate these effects, drawing parallels to human cognitive processes.

2 Related Work

2.1 Logical Reasoning in LLMs

Recent advancements in LLMs have demonstrated significant reasoning capabilities through techniques like chain-of-thought prompting (Wei et al., 2023) (guiding models to show intermediate reasoning steps), zero-shot reasoning (Kojima et al., 2022) (reasoning without task-specific examples), and tree-of-thought exploration (Shinn et al., 2023) (exploring multiple reasoning paths). While these methods have improved performance across various benchmarks (Clark et al., 2020; Parmar et al., 2024), studies comparing human and LLM reasoning patterns reveal that models continue to exhibit systematic errors mirroring human reasoning biases (Dasgupta et al., 2024; Eisape et al., 2024).

Research specifically examining logical reasoning limitations shows that models struggle with operations involving negations, quantifiers, and abstract variables (Dasgupta et al., 2024; Bertolazzi et al., 2024). Performance notably degrades when reasoning involves counterfactual information (Chen et al., 2025; Wu et al., 2024), with inconsistent handling of logically equivalent problems presented in different formats (Estermann et al., 2025). Approaches addressing these limitations include symbolic chain-of-thought (Xu et al., 2024a) (integrating symbolic representations into reasoning steps), verification mechanisms (Vacareanu and Ballesteros, 2024) (validating reasoning against formal rules), and theory resolution frameworks (Toroghi et al., 2024) (applying systematic proof methods), all aimed at enhancing logical consistency in complex scenarios.

2.2 Knowledge Conflicts and Counterfactual Reasoning

LLMs encode substantial factual knowledge in their parameters (Roberts et al., 2020; Petroni et al., 2019), creating challenges when encountering conflicting information. Recent studies categorize these conflicts into context-memory, intercontext, and intra-memory conflicts (Xu et al., 2024b) based on where the conflicting information originates. Larger models typically default to parametric knowledge over conflicting contextual evidence (Longpre et al., 2021), though this varies based on evidence coherence and source reliability (Wang et al., 2023).

Counterfactual reasoning presents significant challenges, with models often performing poorly on tasks involving hypothetical scenarios that contradict established facts (Chen et al., 2025; Wei et al., 2023). Performance on questions with counterfactual premises drops significantly compared to standard tasks, primarily due to conflicts between parametric knowledge and counterfactual assertions (Lin et al., 2025). Mitigation strategies include counterfactual data augmentation (Neeman et al., 2023) (training on synthetically altered data), specialized prompting techniques (Xie et al., 2024b), and distilled counterfactuals (Chen et al., 2023b) (generating targeted examples that highlight conflicts).

2.3 Metacognition, Belief Bias, and Human Reasoning Parallels

Human reasoning exhibits well-documented cognitive biases, including belief bias, where argument validity judgments are influenced by conclusion believability rather than logical structure (Markovits and Nantel, 1989). This bias intensifies with task difficulty (Trippas et al., 2014) and creates an "illusion of objectivity" (Kunda, 1990), where individuals believe their reasoning is unbiased despite evidence to the contrary.

LLMs mirror these human cognitive patterns, performing better when semantic content supports logical inferences (Dasgupta et al., 2024) and reasoning more effectively about believable situations compared to implausible ones (Macmillan-Scott and Musolesi, 2024). Even advanced models exhibit systematic errors paralleling human reasoning biases (Eisape et al., 2024), suggesting shared underlying mechanisms despite the different architectures.

Metacognitive strategies in humans improve logical reasoning by distinguishing between belief evaluation and logical assessment (Douven et al., 2018) —essentially separating "what I know" from "what follows logically." Similar capabilities are emerging in LLMs, including uncertainty estimation (Zhou et al., 2023b) (expressing confidence in outputs), self-evaluation (Wang et al., 2024) (critiquing own reasoning), and belief identification (Chen et al., 2023a) (recognizing when premises conflict with knowledge). When confirmation bias is modulated by confidence, systems become more receptive to corrective information when confidence is low (Rollwage and Fleming, 2021), suggesting potential mechanisms for improving reasoning with conflicting knowledge in LLMs.

3 The CounterLogic Dataset

Despite significant advances in evaluating LLMs' logical reasoning capabilities (Wei et al., 2023; Schick and Schütze, 2021; Kojima et al., 2022; Brown et al., 2020; Zhou et al., 2023a; Patel et al., 2024), existing benchmarks fail to systematically disentangle logical validity from belief alignment (whether premises align with parametric knowledge).

As shown in Table 1, current benchmarks either focus on logical structure without controlling for knowledge conflicts (e.g., LogicBench (Parmar et al., 2024), FOLIO (Han et al., 2024)) or em-

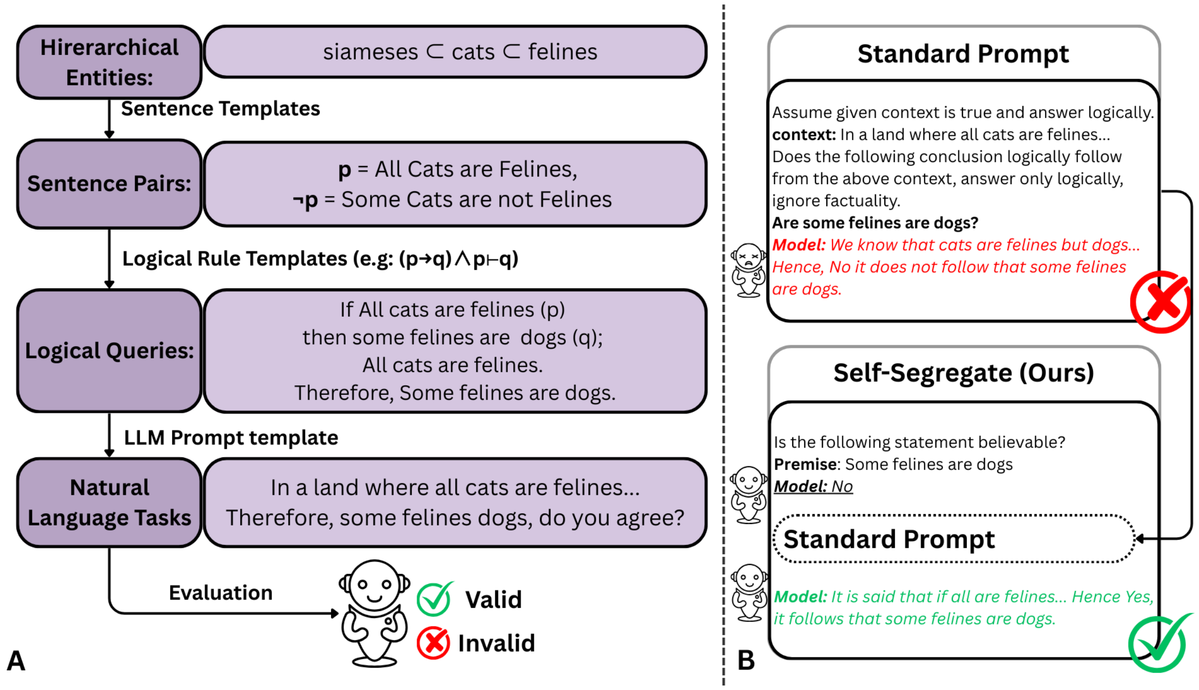

Hirerarchical Entities:
siameses ⊂ cats ⊂ felines
Sentence Templates
Sentence Pairs:
p = All Cats are Felines,
¬p = Some Cats are not Felines
Logical Rule Templates (e.g: (p→q)∧p⊢q)
Logical Queries:
If All cats are felines (p)
then some felines are dogs (q);
All cats are felines.
Therefore, Some felines are dogs.
LLM Prompt template
Natural Language Tasks
In a land where all cats are felines...
Therefore, some felines dogs, do you agree?
Evaluation
Valid
Invalid
A
Standard Prompt
Assume given context is true and answer logically.
context: In a land where all cats are felines...
Does the following conclusion logically follow from the above context, answer only logically, ignore factuality.
Are some felines are dogs?
Model: We know that cats are felines but dogs... Hence, No it does not follow that some felines are dogs.
Self-Segregate (Ours)
Is the following statement believable?
Premise: Some felines are dogs
Model: No
Standard Prompt
Model: It is said that if all are felines... Hence Yes, it follows that some felines are dogs.
B

Figure 2: (A) **Dataset Preparation**: The dataset features hierarchical entity triples (e.g., siameses ⊂ cats ⊂ felines) mapped to 8 logical sentence templates across 9 inference 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 (Modus Ponens, Hypothetical Syllogism, etc.) to systematically evaluate knowledge-logic interactions. (B) **Our Self-Segregate method**: While the standard prompt simply presents LLMs with a counterfactual context followed by related questions, our *Self-Segregate* 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).

phasize knowledge conflicts with simple reasoning tasks (e.g., KNOT (Liu et al., 2024), Reasoning & Reciting (Wu et al., 2024)). To address this gap, we introduce **CounterLogic**, a benchmark dataset containing 1,800 examples across 9 logical schemas with an equal balance in knowledge-consistent and counterfactual datapoints. The dataset systematically combines hierarchical entity relationships with various levels of formal logical structures to evaluate the interaction between knowledge and reasoning in LLMs.

3.1 Dataset Construction

As illustrated in Figure 2, the CounterLogic dataset was constructed through a four-stage process:

(1) Entity Perturbation: We begin with hierarchical entity triples (a, b, c) representing strict subset relationships: $a \subset b \subset c$. These include natural taxonomies such as *siameses* $\subset$ *cats* $\subset$ *felines*.

(2) Sentence Pair Generation: These entities are mapped to four sentence templates forming complementary logical pairs S and $\neg S$ (e.g., "All $\{A\}$ are $\{B\}$" and "Some $\{A\}$ are not $\{B\}$"), yielding diverse sentence pairs that serve as atomic propositions. The complete set of triplets and sentence templates is detailed in Appendix A.1.

To ensure systematic coverage, we enforce entity relationship balance: 25% with correct hierarchical relationship (e.g., *siameses* $\subset$ *cats*), 25% with inverted relationship (e.g., *cats* $\subset$ *siameses*), and 50% with unrelated entity pairs. All four sentence-pair templates are distributed evenly across examples.

(3) Logical Query Generation: Inspired by LogicBench (Parmar et al., 2024), these sentence pairs are then incorporated into formal logical structures according to the inference schemas such as Modus Ponens (MP), Hypothetical Syllogism (HS), Constructive Dilemma (CD), etc. (See Table 2 in Appendix C). A template-based converter is used to transform these sentences into logical structures.

(4) Natural Language Task Generation: We create binary question-answer tasks with systematic variation across (1) Logical validity (whether conclusions follow from premises) and (2) Knowledge alignment (whether conclusions match parametric knowledge). In order to ensure the LLMs are not simply answering questions via memorized logical rules (Xie et al., 2024a; Wu et al., 2024) , we convert the logical queries to natural language using GPT-4o (see Appendix A.4). We systemat-

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. While the Syllogistic dataset contains knowledge conflicts data in a balanced manner, it severely lacks natural language queries, diversity, and depth in logical rules (only syllogism).

Dataset	**Size**	**# Reasoning Steps**	**Knowl. Conflict**	**Balanced Labels**
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	$1 \sim 2$	✓	×
Reasoning & Reciting - Deductive Logic (Wu et al., 2024)	81	$0 \sim 7$	✓	×
Syllogistic (Bertolazzi et al., 2024)	2,120	2	✓	✓
CounterLogic (Ours)	1,800	$1 \sim 5$	✓	✓

ically assign ground truth belief status using the initial sentence beliefs (obtained through hierarchically valid triplets), and logical validity using the rules described in Table 2. Additionally, for each logical form, we construct both Valid (Instances where the conclusion logically follows from the premises) and Invalid Instances (where the logical structure is violated by replacing the statements in the conclusion with statements that cannot be inferred from the premises) examples.

This construction allows for a controlled investigation of reasoning performance in the presence or absence of knowledge alignment.

4 Methodology

4.1 Research Questions

Our investigation focuses on three primary research questions:

1. **RQ1** How do LLMs perform on logical reasoning tasks in counterfactual scenarios, compared to knowledge-consistent scenarios?
2. **RQ2:** Can prompt-based interventions that modify how models approach reasoning tasks, have any effect?
3. **RQ3:** What mechanisms might explain the observed differences?

To address these questions, we conduct a series of experiments across multiple reasoning tasks (Section 4.4), models (Appendix, Section E), and prompting strategies (Appendix, Section B). We first establish baseline performance across 6 reasoning tasks to quantify the impact of knowledge conflicts on reasoning (RQ1), in Section 5.1. We then evaluate our most effective prompt-based intervention (RQ2), "Self-Segregate", in Section 5.2. Finally, we discuss insights from our experiments in Section 6 (RQ3).

4.2 Evaluation Methodology

We evaluate 11 state-of-the-art LLMs (listed in Figure 3 and Appendix Section E) spanning different architectures, parameter scales, and training paradigms. To ensure robust performance measurements, we employ self-consistency checks through multiple sampled outputs per datapoint (5 generations per example), and report their respective mean and variance. This approach accounts for generation variability, as LLMs may produce inconsistent results with similar queries (Bonagiri et al., 2024).

4.3 Self-Segregation

LLMs tend to process premises directly without explicitly considering whether these premises conflict with their parametric knowledge (This can sometimes occur in extended COT reasoning, but our method proves to be superior). *Self-Segregate* introduces a metacognitive step that requires models to first identify whether premises align with or contradict their knowledge before performing logical reasoning (illustrated in Figure 2B).

The method works in two distinct phases:

1. **Knowledge Alignment Assessment:** Models first examine the premises or conclusion and explicitly state whether they align with or contradict their parametric knowledge. This creates an explicit awareness of a "boundary" between the model's factual knowledge and the reasoning task.

2. **Standard Reasoning Process:** models proceed to evaluate the logical validity of the argument based solely on the given premises. We use COT as our standard prompt due to its superior performance, and compare against it in all of our results.

Our approach is inspired by human metacognitive strategies for handling conflicting information (i.e, when humans consciously recognize that information contradicts their existing knowledge, they can more effectively reason through it by temporarily compartmentalizing that conflict) (Wang and Zhao, 2024; Thomas et al., 2013).

4.4 Reasoning Datasets

Along with CounterLogic, we evaluate performance across six other reasoning tasks, each designed to assess specific aspects of logical reasoning under knowledge conflicts (see Table 3 in Appendix D). For each of the following tasks, we implement a tailored version of our Self-Segregation method. The following are the tasks:

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 Self-Segregation, 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 Self-Segregation 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 Self-Segregation 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 Self-Segregation implementation presents questions and answers without supporting context for initial plausibility assessment, followed by complete logical contexts for formal evaluation.

Reasoning and Reciting, Deductive Logic: Adapted from (Wu et al., 2024) , this task evaluates deductive logic over premise sets. Models must determine whether claims logically follow from premises, regardless of whether those premises contradict physical knowledge. The Self-Segregation implementation presents claims in isolation for plausibility assessment before introducing the complete premise set for logical evaluation. An example presents non-physical premises about objects floating forever, testing the ability to follow logical rules despite contradicting physical knowledge.

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.

5 Results and Analysis

Our experimental evaluation reveals consistent patterns across all models and tasks, confirming that: (1) LLMs struggle significantly when reasoning through counterfactual premises and (2) metacognitive awareness interventions via *Self-Segregation* substantially improve performance in knowledge-conflicting scenarios. We discuss it in detail in this section (Figures 3 and 4 summarize the results, for more detailed results in a table format, please see Table 5 in Appendix F).

5.1 Knowledge Conflicts Significantly Impair LLM Logical Reasoning

As shown in Figure 4, when evaluated on the reasoning tasks, all models demonstrate a substantial performance gap between knowledge-consistent and counterfactual scenarios. We find that this holds even across various prompting strategies like Zero-Shot, Few-Shot, and Chain-of-Thought Prompting (more information in Appendix G).

Under the baseline condition, models achieve considerably higher accuracy on knowledge-consistent examples 96% (on average) compared to knowledge-violating examples 69% on average, with performance gaps of about 27% averaged across models.

This pattern holds consistently across all the models, indicating that the phenomenon is not spe-

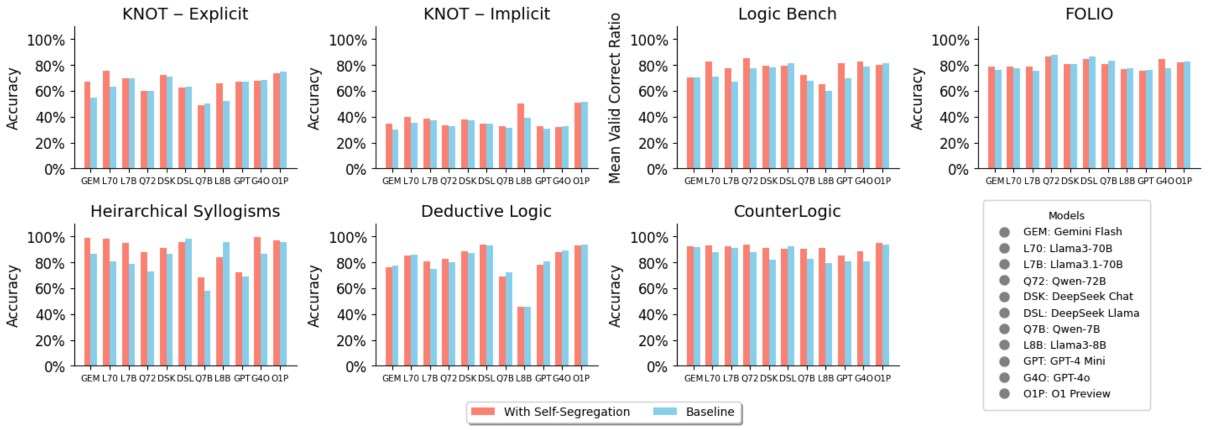


Figure 3: Accuracy comparision between the baseline setup and our metacognitive self-segregation setup across models. The right bar (sky blue) for each model represents accuracy using standard prompts (ref), while the left bar (salmon) shows accuracy using our Self-Segregate prompts (ref). Self-Segregate consistently improves performance across tasks, including KNOT, LogicBench, FOLIO, Hierarchical Syllogisms, and Deductive Logic. All models were run using the OpenRouter API.

cific to particular models or training paradigms. Notably, even the most capable models exhibit this disparity, suggesting that knowledge-conflict interference represents a fundamental challenge in LLM reasoning rather than merely a limitation of smaller or less capable models. Models like Qwen-2-72B show the highest accuracy difference of 47% in the baseline setup, which then greatly improves in the self-segregation setup bringing the gap down to 13%.

This gap appears despite explicit instructions to reason based solely on given premises (see Appendix B), highlighting the pervasive nature of parametric knowledge interference in logical reasoning tasks. Our findings on the CounterLogic dataset further confirm this pattern, with an average performance of 88% on knowledge-consistent examples, 85% on knowledge-violate examples and an average performance gap of 3% on the baseline condition (Figure 4).

5.2 Self-Segregation Dramatically Improves both Counterfactual and Overall Performance

Our Self-Segregation method (described in Section 4.3) yields substantial improvements across all evaluated models and datasets. As illustrated in Figure 3, this approach consistently improves the overall accuracy across most models and tasks.

Figure 3 presents this improvement across six distinct reasoning tasks. The most dramatic gains are observed on the Hierarchical Syllogisms task, where Self-Segregation improves overall accuracy by an average of 7.5%.

We observe that the self-segregation strategy was more effective for datasets like Hierarchical Syllogisms, KNOT (Implicit and Explicit), and they show the most improvement, while there was little to no improvement on the FOLIO, emphasizing the need for better conflict resolution strategies for tasks that involve deep chains of reasoning (Han et al., 2024) .

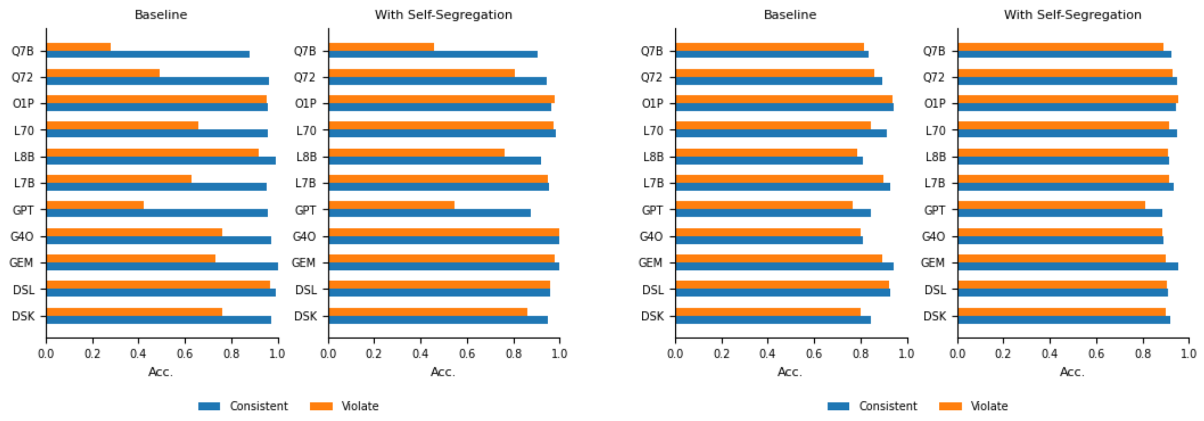
The results on our CounterLogic also follow the same trend, with the overall accuracy performance increasing by 5% on average. The performance on knowledge-consistent examples rose to 93% from 88%, and the performance on knowledge-inconsistent examples to 90% from 85%.

Importantly, this intervention improves reasoning on knowledge-violating scenarios without degrading performance on knowledge-consistent ones. In fact, as shown in Figure 4, accuracy on knowledge-consistent examples also improves slightly under the metacognitive condition, suggesting that explicit reflection on knowledge alignment benefits logical reasoning more generally.

6 Discussion

Our findings reveal a fundamental tension in how LLMs approach logical reasoning when faced with information that contradicts their parametric knowledge. The consistent performance gap observed across models and tasks suggests this challenge is intrinsic to current language model architectures and training paradigms, rather than a limitation of specific models.

This performance disparity echoes well-documented phenomena in human reasoning. Cognitive psychologists have long observed belief



(a) Hierarchical Syllogisms task.

(b) CounterLogic task.

Figure 4: Accuracy comparison between knowledge-consistent and knowledge-violating examples across models. The left panel in each subfigure shows results using ground-truth knowledge-alignment labels (Baseline), and the right panel shows performance when models (Refer legend in Figure-3) use their own knowledge-alignment prediction (self-segregation). Blue bars represent knowledge-consistent examples, while orange bars indicate knowledge-violating ones. The self-segregation setup not only improves accuracy across both subsets but also significantly reduces the performance disparity between them, demonstrating the effectiveness of metacognitive prompting in enhancing belief-robust reasoning.

bias effects, where humans judge argument validity based on conclusion believability rather than logical structure (Markovits and Nantel, 1989; Lampinen et al., 2024). The parallel between human and LLM reasoning biases suggests deeper connections between the cognitive mechanisms underlying both. This alignment in behavior also highlights the potential of leveraging cognitive theories to inform the design of more robust and interpretable language model reasoning frameworks. While humans can override this bias through deliberate metacognitive effort, our experiments demonstrate that LLMs similarly benefit from prompted metacognitive approaches (namely our *Self-Segregate* method).

The effectiveness of our metacognitive intervention provides insight into how LLMs process conflicting information. By explicitly prompting models to identify knowledge conflicts before reasoning, we create a form of epistemic compartmentalization (Thomas et al., 2013), helping models distinguish between what they “know” from their parameters and what they must accept as given in the current reasoning context. Our approach appears to reduce interference between factual knowledge retrieval and logical operation application, allowing models to maintain logical consistency even when processing counterfactual premises. Our proposed approach is a simple abstraction derived from a set of extensive experiments, with meaningful insights.

7 Conclusion and Future Work

We demonstrated that LLMs struggle with logical reasoning when premises contradict their parametric knowledge, with performance dropping by 35% in counterfactual scenarios. Our key contribution is the CounterLogic benchmark and the identification of a simple yet effective metacognitive intervention called *Self-Segregation*, that narrows this performance gap to just 15%. By prompting models to explicitly identify knowledge conflicts before reasoning, we hypothesize that this approach enables more effective compartmentalization of conflicting information without requiring model modifications.

Future work could (1) explore how knowledge conflicts manifest within model representations, (2) investigate applications of metacognitive techniques in other reasoning domains, (3) extend this evaluation to more complex, real-world scenarios, where counterfactual thinking is necessary, (4) Build on methods to either improve, or fundamentally address issues involving parametric memory clashes with reasoning performance, etc. By addressing this specific limitation in counterfactual reasoning, our work contributes to building more robust AI systems capable of reliable logical inference even in contexts that conflict with their training data.

8 Limitations

While our study provides valuable insights into LLMs’ reasoning under knowledge conflicts, several limitations should be noted. First, our CounterLogic dataset, while diverse, cannot capture all forms of complex logical reasoning or knowledge conflicts that might arise in real-world applications. The dataset focuses primarily on categorical syllogisms and propositional logic structures, which represent only a subset but a fundamental part of logical reasoning paradigms.

Second, our experiments were conducted on a specific set of models available at the time of study; newer models may exhibit different patterns of knowledge interference or respond differently to our proposed interventions. The rapid pace of model development means that architectural innovations might soon produce systems with intrinsically different approaches to handling counterfactual information.

Third, the effectiveness of our interventions may vary across different languages, cultures, and knowledge domains, as parametric knowledge itself varies across these dimensions. Our evaluation focused on English-language reasoning with common-knowledge concepts; performance on specialized domains or other languages would require further testing.

Fourth, while we draw parallels to human cognition, the mechanisms of knowledge interference in LLMs may differ fundamentally from human reasoning processes. These parallels provide useful conceptual frameworks but should not be interpreted as evidence of identical or robust cognitive processes.

Finally, while we evaluate LLMs, we also use them to synthetically generate natural language queries, which may pose unnoticed errors, such as inconsistencies or limitations from the LLMs themselves being carried over.

Despite these limitations, our findings demonstrate consistent and substantial improvements in counterfactual reasoning across diverse models and tasks, suggesting that the core insights about metacognitive awareness and knowledge compartmentalization are likely to remain relevant even as specific implementations evolve.

9 Ethics Statement

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

Research Integrity We prioritize transparency and reproducibility throughout our work. All experiments are documented with sufficient detail to enable replication by other researchers. We clearly identify the limitations of our methods and findings in Section 8, acknowledging the boundaries of our conclusions and where further investigation is warranted.

Attribution and Contribution While we utilized large language models as tools to assist with certain aspects of writing and implementation, all research ideas, experimental design, analysis of results, and scientific conclusions presented in this paper are solely attributable to the authors. We have properly credited all relevant prior work and acknowledge the contributions of the research community upon which our work builds.

Data and Resource Considerations The CounterLogic dataset was constructed using synthetic data and template-based generation methods that do not involve the collection of personally identifiable information or data from human subjects. Our evaluation utilized commercially available large language model APIs through standardized interfaces, ensuring fair comparison across systems.

Release of Materials We commit to releasing all artifacts from this research upon acceptance, if not done already, This includes:

- The complete CounterLogic dataset, including all examples and annotations. (Already has been provided via the anonymous GitHub link).
- Reproducible evaluation code and scripts used in our experiments. (Already provided via the anonymous GitHub link).
- Implementation details of our Self-Segregation method
- Prompt templates and model configurations
- Comprehensive documentation to facilitate use by other researchers

Environmental Considerations We acknowledge the computational resources required for our experiments. To minimize environmental impact, we designed our evaluation to be as efficient as possible, reusing model instances where appropriate and limiting the number of inference runs to the minimum necessary for statistically significant results.

Potential Applications and Impact The insights and techniques presented in this paper aim to improve the robustness of logical reasoning in AI systems, particularly when handling counterfactual scenarios. These improvements have potential benefits for various applications requiring reliable reasoning capabilities (including education, scientific exploration, and decision support systems) while 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.

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. 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*.

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.

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*.

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.

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.

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].

Michael Rollwage and Stephen M. Fleming. 2021. Confirmation bias is adaptive when coupled with efficient metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1822):20200131.

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.

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.

Alon Talmor, Sewon Min, Kelcey Zhang, Matt Gardner, Hannaneh Hajishirzi, and Yejin Choi. 2020. Leap-of-thought: Teaching pretrained models to systematically reason over implicit knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2026–2041.

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*.

Paul Trichelair, Ellie Pavlick, and Tal Linzen. 2023. Wrong for the right reasons: Diagnosing misconceptions in nli benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

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*.

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. *arXiv.org*.

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.

Yujia Zhang, Wei Wang, Xiaojie Liu, Yiming Chen, and Zhenyu Li. 2024. 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:

- **Knowledge alignment Balance**: Each logical schema contains 50% of examples where the conclusion is knowledge-consistent under human priors and 50% where it is not.
- **Validity Balance**: Half the examples per schema are logically valid, while the remaining half intentionally violate the logical structure.
- **Entity Relationship Balance**: 25% of examples involve an entity A that is a subset of entity B (e.g., *siameses* ⊂ *cats*), 25% feature B as a subset of A (e.g., *cats* ⊂ *siameses*), and 50% use unrelated entity pairs (e.g., *pines* and *dogs*).

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

A.4 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.

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. Table 3 details how we adapted the reflective prompting strategy for each task type.

Table 2: Formal propositional logic inference 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.

Name	Propositional Logic Form
MP	$((p \to q) \land p) \vdash q$
MT	$((p \to q) \land \neg q) \vdash \neg p$
HS	$((p \to q) \land (q \to r)) \vdash (p \to r)$
DS	$((p \lor q) \land \neg p) \vdash q$
CD	$(p \to q) \land (r \to s) \land (p \lor r) \vdash (q \lor s)$
DD	$((p \to q) \land (r \to s) \land (\neg q \lor \neg s)) \vdash (\neg p \lor \neg r)$
BD	$((p \to q) \land (r \to s) \land (p \lor \neg s)) \vdash (q \lor \neg r)$
CT	$p \vdash (q \lor p)$
MI	$(p \to q) \vdash (\neg p \lor q)$

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.

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
Deductive Logic	Premise set + claim	Isolated claim	Full premise set

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), top-p (0.95), 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 **o1** 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. o1 is the reasoning model provided by openAI that uses

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
O1-preview	OpenAI	Unknown	September 2024
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

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.

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

Our detailed results for the Figure 3 can be found in the Table-5

G Ablation Studies

G.1 Comparison of Prompting Strategies

Figure 5 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; consistent, violate, and random gibberish. Here random gibberish datapoints are obtained by replacing entities in consistent and violate scenarios with random strings (such as 'cat' with 'nsjf'). Consistent and Violate datapoints are from the Syllogistic Dataset (Bertolazzi et al., 2024).

Across all models, We see that Consistent 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

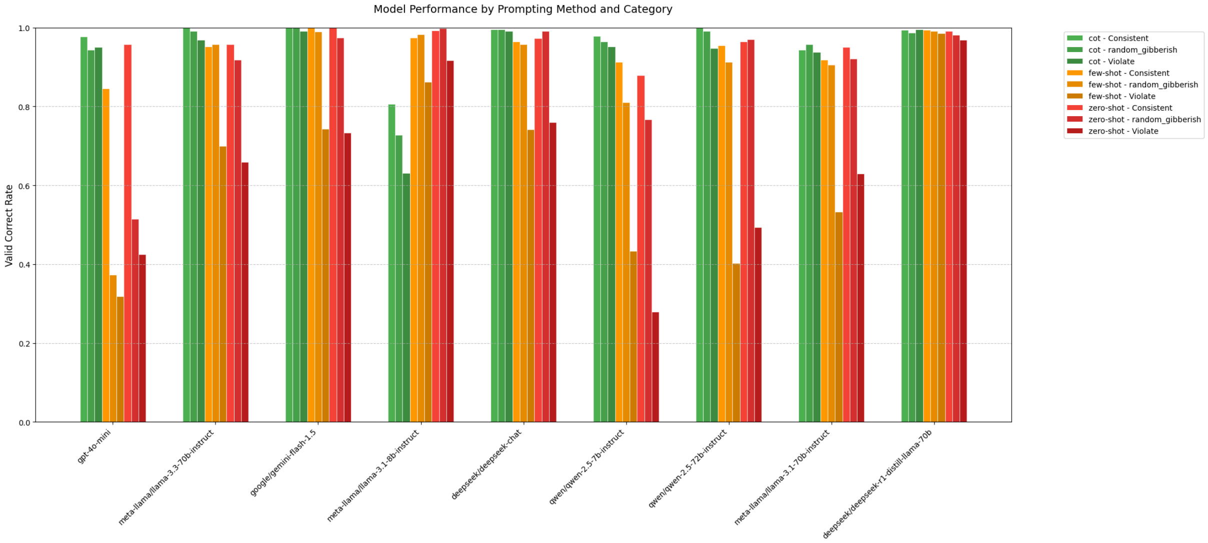


Figure 5: Abalation study comparing the various prompting strategies showing that CoT always outperforms zero-shot but few-shot fails to do so in some models for the Syllogistic dataset.

Models	Explicit		Implicit		Folio		Hierarchical		Deductive		Counterlogic	
	Normal	Baseline	Normal	Baseline	Normal	Baseline	Normal	Baseline	Normal	Baseline	Normal	Baseline
Large models												
meta-llama/llama-3.3-70b-instruct	**75.7**	63.1	**39.6**	35.0	**78.7**	77.6	**98.1**	80.9	85.2	**85.7**	**93.1**	87.8
meta-llama/llama-3.1-70b-instruct	**69.9**	69.9	**38.4**	37.5	**78.6**	75.6	**95.3**	79.0	**80.5**	74.6	**92.3**	91.3
google/gemini-flash-1.5	**67.3**	54.9	**34.6**	30.2	**78.7**	76.2	**98.9**	86.7	76.3	**77.3**	**92.5**	91.8
qwen/qwen-2.5-72b-instruct	**60.2**	60.0	**33.4**	32.8	86.7	**87.6**	**87.9**	72.9	**82.7**	80.2	**93.8**	87.7
deepseek/deepseek-chat	**71.9**	71.2	**38.0**	37.5	80.7	**80.8**	**91.0**	86.7	**88.5**	87.2	**90.9**	82.3
gpt-4o	67.9	**68.0**	32.2	**32.4**	**84.7**	77.5	**99.7**	86.8	88.1	**88.9**	**88.8**	80.5
Small models												
meta-llama/llama-3.1-8b-instruct	**65.8**	52.2	**50.3**	39.3	76.7	**77.7**	84.1	**95.5**	**45.8**	45.5	**91.3**	79.7
qwen/qwen-2.5-7b-instruct	48.8	**50.1**	**32.8**	31.2	80.9	**83.0**	**68.6**	58.0	68.7	**72.4**	**90.6**	82.4
gpt-4o-mini	**66.8**	66.7	**32.5**	30.8	75.2	**76.2**	**72.6**	69.1	78.2	**80.7**	**84.9**	80.5
Reasoning models												
deepseek/deepseek-r1-distill-llama-70b	62.2	**63.2**	34.6	**34.9**	84.5	**86.7**	95.9	**98.0**	**93.8**	93.2	90.6	**92.4**
o1-preview	73.6	**74.6**	50.8	**51.4**	81.7	**82.5**	**96.8**	95.5	93.0	**93.8**	**94.9**	94.0

Table 5: Model performance (Normal vs. Baseline) across datasets.

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

G.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, 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.