Parametric Bypass: Why Stronger LLMs Fail Harder in Situated Commonsense Reasoning

Anonymous EMNLP submission

Abstract

While Large Language Models (LLMs) excel at static knowledge recall, true situated commonsense reasoning requires dynamically adapting prior beliefs to specific narratives. When authentic contexts contradict pre-trained expectations, LLMs face a cognitive conflict: despite comprehending the text, they reflexively default to parametric stereotypes, producing judgments disconnected from the narrative. To uncover this mechanism, we introduce **APP**, a diagnostic framework that evaluates behavioral Accuracy, Probes internal representations, and Patches activations to causally trace reasoning trajectories. We further construct the Fact-Associated Commonsense Evaluation across Topics (**FCET**) dataset to explicitly decouple factual encoding from logical deduction. Extensive experiments reveal leading LLMs lack true situated reasoning. Although they successfully encode situated facts within hidden representations, they fail to reliably balance explicit narratives with internal parametric beliefs. We formalize this breakdown as **Parametric Bypass**, revealing a counterintuitive trend where models with higher baseline commonsense accuracy fail harder as dominant parametric memory overrides external evidence. Ultimately, our findings suggest improving situated reasoning may require mechanisms that better integrate contextual evidence with parametric knowledge.Code and data are publicly available at: `https://anonymous.4open.science/r/APP-and-FCET-7A00`.

1 Introduction

Large Language Models (LLMs) demonstrate strong performance across natural language understanding tasks, particularly in commonsense reasoning (Mondorf and Plank, 2024; Shen et al.). Yet, a significant gap persists between LLMs and humans in genuine situated commonsense reasoning, where deduction must rigorously depend on dynamic narrative constraints rather than static parametric memorization (Ma et al., 2025; Ming et al., 2025). In real-world scenarios, humans routinely revise their baseline expectations to accommodate new situational evidence, whereas LLMs often remain rigidly anchored to their pre-trained priors (Kortukov et al., 2024; Schuster et al., 2026). Consider the example in Figure 1: when a text explicitly states that the *Minor Cold* solar term is colder than *Major Cold* this year, humans effortlessly integrate this constraint to deduce the correct answer. The evaluated models, however, ignore such explicit contextual evidence and incorrectly default to their static statistical memory. This failure to override pre-trained beliefs with provided facts exposes an essential disconnect between context-grounded deduction and simple knowledge retrieval (Feng et al., 2026).

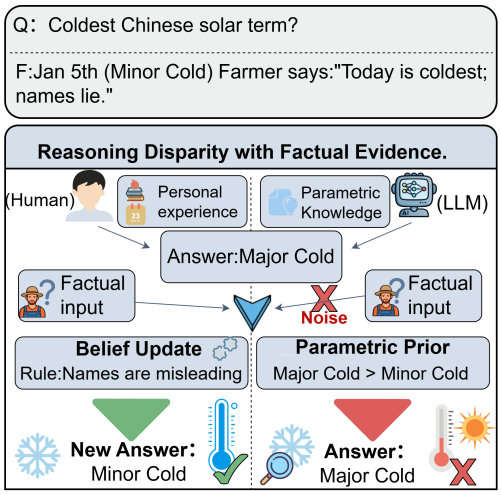

Figure 1: The reasoning disparity between humans and LLMs on the coldest solar term question.

A thorough investigation of this ability is crucial for understanding the true nature of reasoning in LLMs. To determine whether models perform genuine deduction or merely retrieve statistical patterns, previous studies have largely focused on counterfactual reasoning, formal logic puzzles, and

explicit knowledge updating (Li et al., 2023; Xie et al., 2024). While these approaches reveal that LLMs can follow arbitrary abstract rules, they rely heavily on synthetic environments (Mondorf and Plank, 2024; Ma et al., 2025). Within the domain of cultural and commonsense knowledge, recent benchmarks like CHARM have made significant strides by explicitly decoupling static parametric memorization from reasoning (Shen et al., 2024). However, these reasoning tasks generally assume that the provided narrative context safely aligns with the parametric priors of the model (Sun et al., 2024). What remains severely underexplored is how models handle naturalistic semantic conflicts. In these situated scenarios, a strong ingrained prior tempts the model, but explicit situational facts demand a contradictory conclusion. This unresolved tension leads directly to our central research question: **Do LLMs actually possess situated commonsense reasoning to reliably let explicit narrative facts override their internal parametric beliefs?**

To address the persistent gap between static parametric retrieval and genuine situated deduction under semantic conflict, we propose a comprehensive diagnostic evaluation framework named APP. This methodology evaluates behavioral Accuracy across four controlled conditions, employs representation Probing to decode internal memory states, and utilizes activation Patching to causally trace the reasoning trajectory.To systematically apply this framework and ensure broad generalization, we construct the Fact-Associated Commonsense Evaluation across Topics (FCET) dataset. FCET comprise six cultural categories (traditional festivals, folklore, food, idioms, solar terms, and health) and two scientific disciplines (physics and biology). We select these specific areas because they are naturally rich in deeply ingrained priors and semantically misleading heuristics (Adilazuarda et al., 2024; Myung et al., 2024; Chiu et al., 2024). Unlike previous benchmarks that evaluate prior-aligned reasoning, FCET pairs authentic narrative facts with commonsense questions to intentionally induce conflicts between external evidence and internal parametric memory. This controlled setup rigorously tests whether models can anchor their decisions on situational constraints rather than defaulting to contradictory priors. Through this framework, our findings indicate that the reasoning process of large language models involves a complex internal competition, **ultimately revealing limitations in situated commonsense reasoning under semantic conflicts**. Although they appear to encode, they often struggle to reliably balance explicit narrative facts with internal parametric knowledge, as their final deductions remain influenced by pre-trained parametric priors.

In summary, our main contributions are as follows:

1. We propose the **APP diagnostic framework** to systematically evaluate situated commonsense reasoning. Specifically, APP evaluates behavioral Accuracy, Probes internal representations, and Patches activations to causally trace how models process semantic conflicts from hidden states to final behavioral outputs.

2. We construct the **Fact-Associated Commonsense Evaluation across Topics (FCET) dataset** across eight distinct domains, comprising both cultural categories and scientific disciplines. By pairing narrative facts with commonsense questions, it intentionally creates semantic conflicts to test whether models rely on provided context or pre-trained memory.

3. Extensive experiments reveal a consistent gap between fact encoding and downstream utilization. We formalize this phenomenon as **Parametric Bypass**, revealing a counterintuitive trend where models with higher baseline commonsense accuracy exhibit lower sensitivity to contextual evidence. Our analysis demonstrates that while models appear to encode situational facts in their hidden states, they often fail to reliably utilize them during generation, a gap that Chain of Thought (CoT) prompting only partially mitigates.

2 Dataset Construction

To evaluate how models utilize external evidence, we construct the Fact-Associated Commonsense Evaluation across Topics (FCET) dataset. As depicted in Figure 2, the construction pipeline involves three main stages: real-world fact collection, question generation, and the design of controlled experimental conditions.

2.1 Fact Acquisition and Quality Assurance

To evaluate situated reasoning across eight distinct domains (six cultural categories including festivals, idioms, folk customs, health, solar terms, and food, alongside two scientific disciplines including physics and biology), we source authentic narratives from mainstream portals like Sohu and Peo-

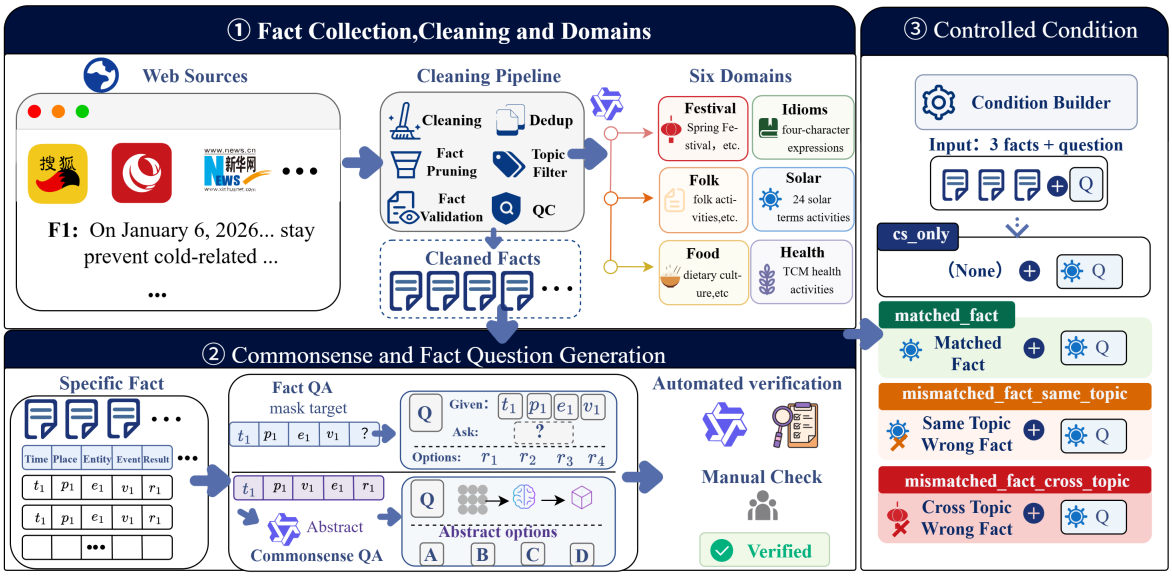

Figure 2: The construction pipeline of our diagnostic dataset, illustrating (1) real-world fact collection, (2) dual-track question generation, and (3) the four-way controlled experimental setup.

ple's Daily Online. A primary challenge involves extracting usable facts from web pages cluttered with promotional slogans and subjective commentary. We resolve this by deploying a rule-based filter utilizing high-coverage domain dictionaries and negative noise signals. This system explicitly prioritizes sentences containing objective anchors such as specific entities, dates, or locations while rigorously discarding emotional rhetoric.

The extracted texts subsequently undergo a rigorous standardization pipeline. We apply fact unitization to segment the materials into compact bundles averaging three atomic facts to optimize information density. Following a strict record-level deduplication policy to prevent data contamination, we execute an interpretable scoring framework. This framework calculates semantic weights based on structural entities to ensure logical self-containment. Text fragments lacking clear grammatical relations or drifting across themes are flagged for rigorous manual review and subsequently verified or discarded. This elaborate auditing pipeline ultimately yields a highly reliable foundation of 4,686 factual data points for FCET construction.

2.2 Constrained Commonsense Question Formulation

Building upon the structured factual tuples, we execute a dual-track generation process. For Fact QA, we systematically mask specific target elements (e.g., results or entities) to formulate direct recall multiple-choice questions. For Commonsense QA, we enforce a strict abstraction constraint via Qwen, which synthesizes fact bundles into generalized judgments. These questions are rigorously confined to their original domains to guarantee the naturalness of subsequent distractors.

Both question tracks undergo rigorous automated verification and manual checking. A fuzzy match metric ensures the textual answer aligns correctly with the designated option, explicitly discarding instances scoring below a 0.6 threshold. To eliminate positional bias, correct option labels are comprehensively balanced, achieving near-perfect uniformity (<0.22% maximum deviation). By prioritizing strict precision, this filtering yields a highly curated core dataset of FCET, establishing a reliable foundation for multi-condition expansion.

2.3 Controlled Experimental Condition Combinations

To explicitly decouple a model's reliance on parametric memory from its genuine evidence utilization, we formulate the dataset into a controlled experimental setup. Let C represent a given factual context, Q represent the generated commonsense question, and P denote the final input prompt presented to the model. We expand each of the base pairs into four structured conditions. In the cs_only condition, the prompt contains only the question($P = Q$). This establishes the baseline performance of pre-trained parametric memory be-

fore any contextual intervention. The matched_fact condition pairs the question with its authentic evidence ($P = C_{true} \oplus Q$) to measure the net factual gain. To evaluate fact discrimination and interference sensitivity, we introduce two distinct distractor conditions: mismatched_fact_same_topic injects a domain-aligned distractor ($P = C_{same} \oplus Q$), and mismatched_fact_cross_topic injects an out-of-domain distractor ($P = C_{cross} \oplus Q$).

We generate the mismatched contexts C_{cross} and C_{same} using a strict controlled rotation strategy rather than random sampling to preserve natural domain distributions. For the same-topic condition, we apply a rotation step of 1 to assign a factual context from the identical cultural domain belonging to a different pair identifier. For the cross-topic condition, we utilize a rotation step of 7 to select a factual context from an heavily distinct domain. The algorithm continuously skips any self-matching or identical-question conflicts during this rotation sequence. Applying these four uniformly formatted templates yields a well balanced final FCET dataset of 5,448 experimental instances. Furthermore, a detailed comparison with existing evaluation datasets is provided in Appendix A.1, alongside concrete prompt examples in Appendix A.2.

3 Methods

To quantify the tension between parametric memory and situated reasoning, we propose the APP framework. As illustrated in Figure 3, this structured diagnostic pipeline comprises three interlinked modules: Accuracy under Controlled Contexts (ACC), Internal Representation Probing (PROBE), and Causal Intervention via Patching (PATCHING). Together, these components expose the underlying decision-making mechanisms of the evaluated models from behavioral outputs to internal hidden states.

3.1 ACC: Controlled-Condition Accuracy Analysis

As the foundational layer of the APP framework, ACC characterizes evidence utilization purely at the behavioral level. Because prompts and questions remain highly consistent across our four experimental conditions, any performance divergence directly isolates the reaction of a model to contextual evidence. We compute standard accuracy for any condition c as:

$$\mathrm{Acc}(c) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(\hat{y}_i^{(c)} = y_i\right) \tag{1}$$

To verify that subsequent reasoning failures stem from a utilization gap rather than a basic reading deficit, we introduce Fact QA Accuracy to evaluate explicit information extraction across topics T:

$$\mathrm{FactAccoverall} = \frac{\sum t \in \mathcal{T} \sum_{i=1}^{N_t^{\text{fact}}} \mathbf{1}(\hat{y}t, i = yt, i)}{\sum_{t \in \mathcal{T}} N_t^{\text{fact}}} \tag{2}$$

Building on these baselines, we define two comparative diagnostics. Fact Inference Gain($\Delta_{inference}$) measures the net benefit of authentic evidence, calculated as the direct difference between the matched_fact and cs_only conditions. Fact Sensitivity quantifies($\Delta_{interfere}$) the capacity to distinguish valid support from domain-aligned distractors via the accuracy gap between the matched_fact and mismatched_fact conditions. Models engaging in genuine situated reasoning exhibit systematic variance across these metrics, whereas models defaulting to parametric priors demonstrate suppressed variance.

3.2 PROBE: Internal Representation Decoding

As the second layer of the APP framework, PROBE investigates the internal representations to verify whether the model encodes critical signals regarding factual evidence. Moving beyond output behaviors, this module decodes whether the hidden states capture fact matching and answer correctness. We train lightweight linear probes across all layers to quantify the encoding strength of these dual signals.

Let x_i denote an input sample and L represent the total number of model layers. For layer l, we extract the hidden state at the final prediction position (the exact token prior to answer generation): $h_i^{(l)} \in \mathbb{R}^d$ where d is the hidden dimension. We select this specific position because it directly carries the aggregated representation responsible for forming the final answer distribution under the given context.

The primary supervision signal for the probes relies on the minimal intervention pair formed by the matched_fact and mismatched_fact_same_topic conditions. These conditions share the exact same question, options, and topic, varying exclusively in factual alignment. We define two binary labels based on this controlled setup.

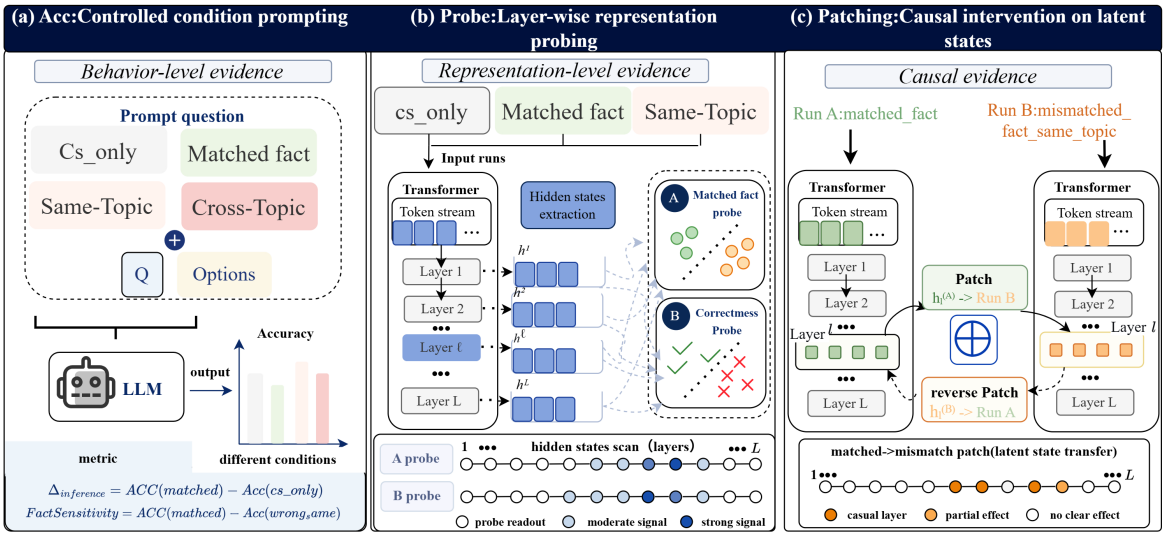

Figure 3: Overview of the APP framework, integrating (a) behavioral accuracy evaluation, (b) representation-level linear probing, and (c) causal activation patching.

The fact matching label isolates the contextual alignment: $z_i \in 0, 1$, where $z_i = 1$ indicates an instance from the matched_fact condition and $z_i = 0$ indicates an instance from the mismatched_fact_same_topic condition.

The answer correctness label isolates the behavioral outcome: $r_i \in 0, 1$, where $r_i = 1$ indicates the model predicts the correct answer under the given condition, and $r_i = 0$ indicates an incorrect prediction.

For each layer l, we train two independent logistic probes. The predictions for fact matching ($\hat{p}_z^{(l)}(i)$) and answer correctness ($\hat{p}_r^{(l)}(i)$) are formulated as:

$$\hat{p}_z^{(l)}(i) = \sigma\left(\left(w_z^{(l)}\right)^\top h_i^{(l)} + b_z^{(l)}\right) \tag{3}$$

$$\hat{p}_r^{(l)}(i) = \sigma\left(\left(w_r^{(l)}\right)^\top h_i^{(l)} + b_r^{(l)}\right) \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function. Both probes are optimized using the standard binary cross-entropy loss. The objective for the fact matching probe is:

$$\mathcal{L}_{\text{fact}}^{(l)} = -\sum_{i=1}^{N} \left[z_i \log \hat{p}_{z,i}^{(l)} + (1 - z_i) \log(1 - \hat{p}_{z,i}^{(l)})\right] \tag{5}$$

The objective for the correctness probe follows an identical formulation:

$$\mathcal{L}_{\text{corr}}^{(l)} = -\sum_{i=1}^{N} \left[r_i \log \hat{p}_{r,i}^{(l)} + (1 - r_i) \log(1 - \hat{p}_{r,i}^{(l)})\right] \tag{6}$$

Repeating this optimization across all layers $l \in \{1, \ldots, L\}$ generates a layer-wise decodability profile. This profile identifies whether factual alignment and outcome correctness enter the internal representation space. A high decodability score for the fact matching probe indicates that the model explicitly encodes the contextual contradiction at that specific depth. Concurrent separability in the correctness probe suggests a representational link between the factual signal and the reasoning process. This layer-wise mapping provides the reliable structural localization required for the subsequent causal PATCHING intervention.

3.3 PATCHING: Causal Intervention on Representations

As the final layer of the APP framework, PATCHING executes causal interventions to verify whether the encoded fact-matching signals actively drive the final prediction. Unlike PROBE, which merely identifies decodable information, PATCHING tests for systematic shifts in the output distribution when target representations are explicitly replaced. This module elevates the analysis from correlational decodability to robust causal relevance.

For a given base sample, we construct a minimal difference input pair containing a matched condition x_i^+(matched_fact) and a domain-aligned mismatched condition x_i^- (mismatched_fact_same_topic).Both inputs maintain highly consistent questions, options, and formatting, differing only in their factual evidence. Let $h_{i,+}^{(l)}$ and $h_{i,-}^{(l)}$ denote their respective hidden states at the final prediction position for layer l.

The fundamental PATCHING operation copies the hidden state from one condition and injects

it into the forward pass of the other. For the primary upward intervention, we replace the hidden state of the mismatched sample at layer l with the state from the matched sample: $\tilde{h}_{i,-}^{(l)} \leftarrow h_{i,+}^{(l)}$. The model completes the remaining forward pass from layer $l+1$ using this intervened representation. Let $p_\theta(y_i \mid x_i^-)$ denote the original prediction probability for the correct option y_i under the mismatched condition. The probability following the intervention is $p_\theta\left(y_i \mid x_i^-; h_{i,-}^{(l)} \leftarrow h_{i,+}^{(l)}\right)$. We define the layer-wise upward causal gain as:

$$\Delta_{i,\text{up}}(l) = p_\theta\left(y_i \mid x_i^-; h_{i,-}^{(l)} \leftarrow h_{i,+}^{(l)}\right) - p_\theta(y_i \mid x_i^-) \quad (7)$$

A positive value indicates that injecting the matched internal state enhances the model's support for the correct answer. The average upward intervention effect across all Nsamples is:

$$\bar{\Delta}_{\text{up}}^{(l)} = \frac{1}{N}\sum_{i=1}^{N} \Delta_{i,\text{up}}^{(l)} \quad (8)$$

To prevent misinterpreting random perturbations as meaningful causal effects, we introduce a reverse control mechanism defined as downward intervention. We inject the hidden state from the mismatched condition into the matched condition:

$$\Delta_{i,\text{down}}(l) = p_\theta\left(y_i \mid x_i^+; h_{i,+}^{(l)} \leftarrow h_{i,-}^{(l)}\right) - p_\theta(y_i \mid x_i^+) \quad (9)$$

The average downward effect is similarly computed:

$$\bar{\Delta}_{\text{down}}^{(l)} = \frac{1}{N}\sum_{i=1}^{N} \Delta_{i,\text{down}}^{(l)} \quad (10)$$

This layer-wise state replacement isolates the specific depth where factual alignment alters the downstream reasoning trajectory. A layer exhibiting high decodability in PROBE alongside significant positive intervention effects in PATCHING confirms that the factual signal is both encoded and actively utilized in the decision-making process. This causal validation completes the APP framework, seamlessly linking behavioral variance to internal representation encoding and final outcome execution.

4 Experiments

4.1 Acc Analysis

The evaluation covers ten mainstream models from the Qwen, Llama, and Gemma families. The experiments test these models across four context conditions and eight diverse domains. Figure 4 presents the overall results for the cultural categories, while detailed evaluations for the Physics and Biology domains are provided in Appendix D.1. The analysis reveals two distinct failure modes rigorously tied to baseline reasoning competence. These behavioral patterns demonstrate a widespread inability to perform genuine situated commonsense reasoning.

Archetype 1: Strong Parametric Reliance Advanced models like Qwen3-8B, Llama3.1-8B, Qwen3-1.7B, and Qwen2.5-8B rely heavily on internal pre-trained memory. Their baseline accuracy remains near-perfect across all domains. Introducing supportive or distracting evidence yields negligible impact. This results in near-zero $\Delta_{inference}$ and $\Delta_{interfere}$. Fact extraction accuracy drops below 65% for Qwen3-1.7B in certain domains, yet its reasoning accuracy remains unaffected at 1.0. This pattern suggests the existence of Parametric Bypass. Models answer correctly using ingrained priors and largely bypass the provided text.

Archetype 2: Contextual Susceptibility

As summarized in Table 1, models like Llama3.2-1B and DeepSeek-1.5B show a contrasting failure mode. Their baseline acts as a firm upper bound. Introducing matched facts actually reduces performance and yields a negative $\Delta_{inference}$. Mismatched distractors cause further degradation and a notable $\Delta_{interfere}$. These models attempt to process the context but fail to extract helpful causal logic.

Model	Base (%)	Match (%)	Δ_{inf} (%)	Δ_{interf} (%)
Llama3.2-1B	**80.55**	68.16	-12.39	0.95
DeepSeek-1.5B	**76.94**	67.81	-9.12	2.24

Table 1: Contextual Interference in Weaker Models. The baseline acts as an upper bound. Paradoxically, providing factual context degrades performance (Δ_{inf}), and distractors cause further interference (Δ_{interf}).

Archetype 3: Illusory Contextual Gain

Model	Base(%)	Match(%)	Δ_{inf}(%)	Δ_{interf}(%)
DeepSeek0-llm-7B-Chat	24.78	**28.23**	3.44	-0.77
Gemma2-9B	25.39	**36.06**	10.67	0.26

Table 2: Illusion of Positive Inference Gains. While both models exhibit apparent performance increases (Δ_{inf}) over their low baselines, further analysis reveals this is caused by *Contextual Induction* rather than genuine factual deduction.

As detailed in Table 2, Gemma2-9B and DeepSeek-7B-Chat start near random guessing.

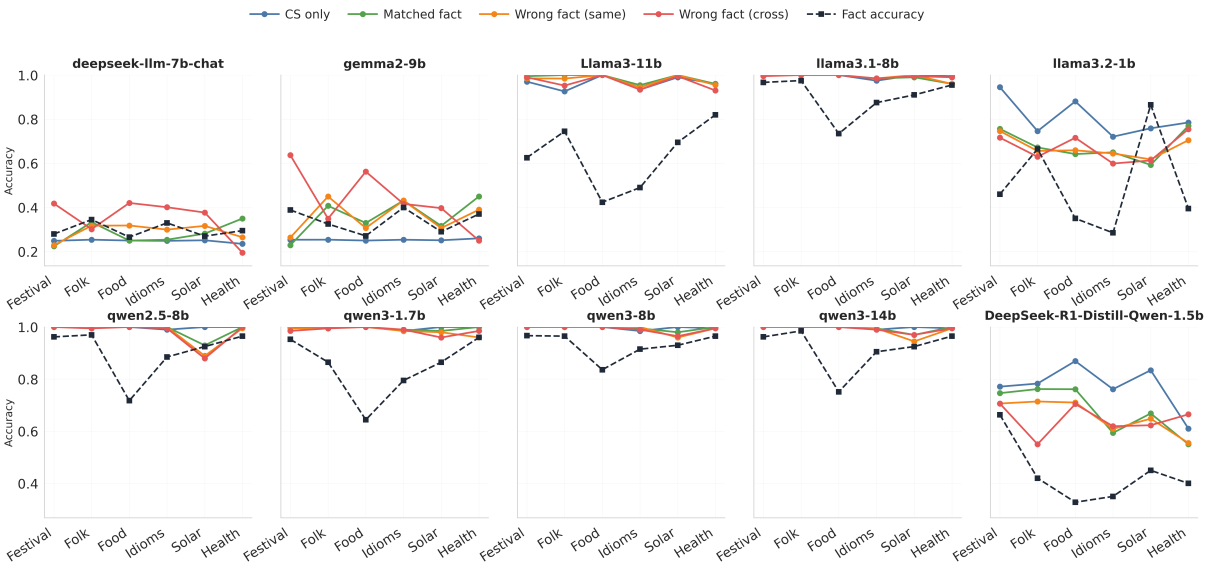

Figure 4: Domain-specific accuracy distributions across the evaluated models. Solid lines denote performance under the four distinct contextual conditions. The black dashed line represents the baseline factual accuracy for each cultural topic.

Adding context triggers erratic performance spikes. Interference from unrelated topics frequently surpasses the intended $\Delta_{inference}$. This pattern challenges the hypothesis that performance gains arise from fact-driven reasoning. The models exhibit Contextual Induction. They react blindly to text length or prompt structures instead of engaging in factual deduction.

Universal Failure Across all archetypes, the Fact Sensitivity remains near zero. This suggests a potential limitation in situated commonsense reasoning under semantic conflicts. Regardless of their overall accuracy, they systematically fail to differentiate between a genuinely supportive fact and a deceptive distractor.

4.2 Mechanistic Analysis

The behavioral results raise a mechanistic question. Do models fail to use facts because they cannot comprehend the text, or do they actively ignore it? We answer this using Linear Probing and Activation Patching. As illustrated in the layer-wise heatmaps in Figure 5, both Qwen3-8B and Gemma2-9B reveal a clear gap between information encoding and causal utilization. Comprehensive results and detailed analyses for all evaluated models are provided in Appendix D.2. Additionally, Appendix D.2 also presents a series of sanity checks to verify that these findings are not driven by noise fitting or token-position biases.

Broad Encoding The bottom row of each sub-

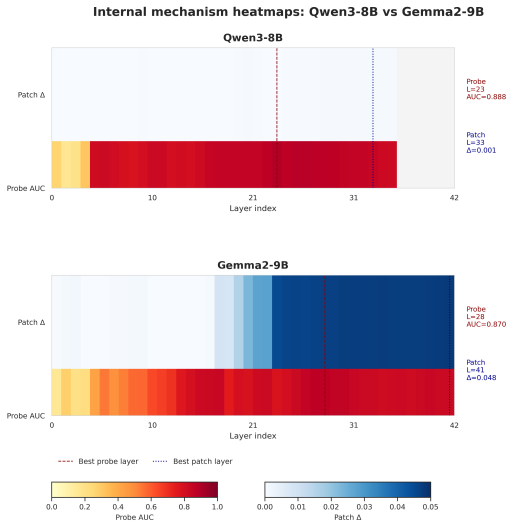

Figure 5: Mechanism heatmaps for Qwen3-8B and Gemma2-9B. High probing AUC contrasts with near-zero patching Δ, exposing the disconnect between robust factual encoding and negligible causal utilization.

plot displays the Probe AUC across network layers. A deep red band emerges in the early-to-middle layers for both models. The Probe AUC reaches 0.888 at Layer 23 for Qwen3-8B and 0.870 at Layer 28 for Gemma2-9B. This high linear separability confirms that models successfully extract and encode the explicit contextual signal within their hidden states.

Failed Utilization The top row demonstrates that encoded knowledge frequently fails to drive reasoning. We measure this causal reliance using

the upward patching gain (Δ_{up}) and the downward patching drop (Δ_{down}) defined in Section 3.3. The resulting causal effect, visualized as Patch Δ, remains heavily pale across all layers for Qwen3-8B. Its peak causal impact reaches only 0.001 at Layer 33. Gemma2-9B exhibits slightly higher contextual susceptibility, developing a visible blue band in later layers with a peak Patch Δ of 0.048 at Layer 41. Despite this localized activity, the causal interventions (Δ_{up} and Δ_{down}) confirm that actual evidence utilization remains largely dwarfed by the encoding scores.

Proof of Parametric Bypass The heatmaps expose a notable spatial disconnect within the computation graph. A substantial depth separates the peak encoding layers from the peak causal layers. This discrepancy proves that encoding does not guarantee utilization. The models explicitly represent factual logic in their hidden states but the encoded signal appears to exert limited influence on final generation. These findings provide mechanistic evidence consistent with Parametric Bypass. The models do not suffer from reading comprehension deficits. Pre-trained priors simply dominate the generation process and render the encoded facts causally inert.

4.3 Chain-of-Thought Evaluation

Extending our behavioral evaluation, we investigate whether four Chain of Thought (CoT) paradigms (standard zero-shot, fact-summarization, match-verification, and option elimination) can mitigate Parametric Bypass and Contextual Interference. Although overall accuracy marginally inflates, Fact Sensitivity remains stagnant, indicating that these prompts improve output superficially without genuinely restoring evidence utilization.

We restrict this CoT evaluation exclusively to the behavioral layer. Applying the mechanistic PROBE and PATCHING modules to CoT generation introduces considerable temporal misalignment. The variable-length reasoning chains generated across different conditions make accurate hidden state alignment and reliable causal intervention mathematically unreliable. Consequently, we position this CoT exploration as an auxiliary behavioral observation rather than a core mechanistic target. The specific formulations of the four CoT strategies are detailed in Appendix C, with comprehensive quantitative results provided in Appendix D.3.

5 Related Work

Situated Commonsense Reasoning. Situated reasoning requires grounding inference in dynamic contexts rather than static knowledge (Ma et al., 2025; Ming et al., 2025). However, LLMs struggle with context faithfulness (Ju et al., 2024), frequently failing to utilize external evidence when it contradicts parametric memory (Xu et al., 2024; Bi et al., 2025; Feng et al., 2026; Ortu et al., 2024). Building on this, we investigate naturalistic semantic conflicts, revealing that models systematically regress to pre-trained expectations despite successfully encoding the narrative facts.

Evaluation of Commonsense Knowledge. While recent benchmarks evaluate models' *internal* knowledge—including Chinese reasoning (Sun et al., 2024), idioms (Kalhor and Bahrak, 2026), and multi-hop deduction (Yang et al., 2024)—they rarely assess the faithful utilization of *external* facts (Xu et al., 2024). Our dataset addresses this gap by decoupling fact retrieval from deduction, isolating the bottleneck where models comprehend external evidence yet reflexively default to parametric memory.

Contextual Fact Encoding and Causal Utilization. Recent mechanistic studies investigate memory-context conflicts using probing to trace signal routing (Ju et al., 2024; Monea et al., 2024; Tighidet et al., 2024) and causal interventions to steer hidden states (Yu et al., 2023; Jin et al., 2024; Monea et al.). Leveraging these tools, we apply linear probing and activation patching to mechanistically expose the disconnect between factual encoding and causal utilization, thereby revealing phenomena like Parametric Bypass.

6 Conclusion

We introduce the APP diagnostic framework, shifting situated reasoning evaluation from static recall to dynamic causal utilization. We reveal that successful evidence encoding frequently masks notable deduction deficits, exposing two fundamental failure modes: Parametric Bypass and Contextual Interference. Mechanistic analysis further supports a disconnect between factual encoding and downstream utilization. Consequently, future RAG systems should transcend surface-level prompting, employing structural interventions to dynamically balance explicit contextual constraints with pre-trained priors.

Limitations

While our study provides robust evidence regarding fact-grounded reasoning failures, several limitations remain. First, our mechanistic interventions are restricted to 1B-9B open-weight models due to the necessity of internal state access, leaving the exact causal dynamics of massive proprietary models (e.g., GPT-4) as an open challenge. Second, although our framework covers both cultural and scientific domains (e.g., physics and biology), it primarily targets topics characterized by strong pre-trained associative priors. The observed phenomena—such as near-zero Fact Sensitivity—may manifest differently in highly formalized domains, such as complex mathematical deduction or symbolic logic, where parametric biases operate differently. Finally, our activation patching operates at the macroscopic layer level. While this is sufficient to prove the spatial disconnect between encoding and utilization, future work utilizing finer-grained techniques (e.g., attention-head patching) is required to map the exact neural circuits driving these routing failures.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and 1 others. 2024. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *arXiv preprint arXiv:2410.02677*.

Yiyang Feng, Zeming Chen, Haotian Wu, Jiawei Zhou, and Antoine Bosselut. 2026. Tracking the limits of knowledge propagation: How llms fail at multi-step reasoning with conflicting knowledge. *arXiv preprint arXiv:2601.15495*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8235–8246.

Ghazal Kalhor and Behnam Bahrak. 2026. Masalbench: A benchmark for contextual and cross-cultural understanding of persian proverbs in llms. *arXiv preprint arXiv:2601.22050*.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under realistic knowledge conflicts. *arXiv preprint arXiv:2404.16032*.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 804–815.

Boxiang Ma, Ru Li, Wang Yuanlong, Hongye Tan, and Xiaoli Li. 2025. Memorization$\neq$ understanding: Do large language models have the ability of scenario cognition? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20758–20774.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". In *International Conference on Learning Representations*, volume 2025, pages 29430–29456.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. A glitch in the matrix? locating and detecting language model grounding with fakepedia, 2024. *URL https://arxiv. org/abs/2312.02073*, page 28.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kiciman, Hamid Palangi, Barun Patra, and Robert West. 2024. A glitch in the matrix? locating and detecting language model grounding with fakepedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6828–6844.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki A Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew A Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436.

Jakob Schuster, Vagrant Gautam, and Katja Markert. 2026. Whose facts win? llm source preferences under knowledge conflicts. *arXiv preprint arXiv:2601.03746*.

S Shen, L Logeswaran, M Lee, H Lee, S Poria, and R Mihalcea. Understanding the capabilities and limitations of large language models for cultural commonsense. arxiv 2024. *arXiv preprint arXiv:2405.04655*.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking chinese commonsense reasoning of llms: From chinese-specifics to reasoning-memorization correlations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11205–11228.

Zineddine Tighidet, Jiali Mei, Benjamin Piwowarski, and Patrick Gallinari. 2024. Probing language models on their knowledge source. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 604–614.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, volume 2024, pages 35623–35646.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959.

A Dataset Details and Benchmark Comparison

A.1 Comparison with Existing Benchmarks

Table 3 contrasts our dataset with prior reasoning and reading comprehension benchmarks. While existing datasets primarily evaluate static knowledge retention, basic reading comprehension, or artificially injected synthetic counterfactuals, our dataset uniquely focuses on *naturalistic semantic conflicts*. By systematically pairing authentic contextual evidence with counter-prior commonsense queries across multiple controlled conditions, our dataset explicitly isolates the behavioral bottleneck between factual encoding and causal utilization.

A.2 Example: solar-term case

Figure 6 illustrates a concrete example of the four conditions. The mismatched conditions inject incorrect distractors to test context sensitivity, using either a same-topic distractor (e.g., eating dumplings on the Winter Solstice) or a cross-domain distractor (e.g., alcohol harming the stomach). Note that the English translations in parentheses are provided solely for readability; the actual prompts fed to the models are rigorously in Chinese.

B Implementation Details

This section outlines the hyperparameter configurations, evaluation logic, and hardware environment used to reproduce the experimental results.

B.1 Hardware and Model Precision

All experiments run on a compute cluster equipped with 8 NVIDIA Tesla V100 32GB GPUs. The evaluated models rely on standard Transformers library implementations. The inference and mechanistic analysis scripts utilize bfloat16 and float16 precision formats. This configuration optimizes GPU memory usage while maintaining numerical stability.

Method	Fact Manip.	Behavior Decomp.	CoT Analysis	Interpretability
CHARM	4-tier	Memory-Reasoning	Context	Behavioral
CoCo	3-tier	MEM×APP	—	Behavioral
CORECODE	2-tier	Indirect	—	Behavioral
C-Eval	None	—	Shallow	Black-box
Ours	**4-tier Recursive**	**Parametric vs Context**	**Two-sided**	**Neural**

Table 3: Comparison of commonsense evaluation approaches. Our work bridges behavioral analysis with mechanistic interpretability.

B.2 Inference and Generation Settings

The evaluation applies two distinct decoding strategies based on the task type.

Standard Multiple-Choice Evaluation: The system does not sample text. It directly extracts the final-token logits over the predefined answer space (A, B, C, D). The maximum sequence length is restricted to 512 tokens. The batch size is set to 8.

Chain-of-Thought (CoT) Generation: The generation process rigorously utilizes deterministic greedy decoding. The do_sample parameter is explicitly set to False. Consequently, stochastic decoding parameters such as temperature, top_p, and top_k are disabled. The maximum input length is set to 768 tokens, and the maximum number of newly generated tokens is limited to 256.

B.3 Mechanistic Analysis Settings

To manage the significant computational overhead inherent in layer-wise forward passes and representation replacements, our causal representation analysis is conducted on a uniformly sampled subset of valid triplets from the main evaluation set. Crucially, this subset spans multiple cultural categories and at least one scientific discipline to ensure cross-domain comparability. Within each triplet, the commonsense question, candidate options, and ground truth remain rigorously identical. By manipulating only the contextual input—contrasting the matched_fact against the mismatched_fact_same_topic condition—this design cleanly isolates the "fact-matching" variable. This strict control allows us to precisely evaluate whether facts are explicitly encoded within the latent space and whether they causally influence the final prediction.

Linear Probing: The analysis constructs balanced data triplets across representative cultural domains. Each triplet contains a cs_only, a matched_fact, and a mismatched_fact_same_topic condition. The probe employs a Logistic Regression classifier with a maximum of 2,000 iterations. The data split follows a stratified 70% training and 30% testing distribution. A fixed random seed ensures reproducibility. The evaluation reports both Accuracy and AUC metrics.

Activation Patching: The causal intervention targets the hidden state of the final input token. The operation specifically contrasts the matched_fact condition against the mismatched_fact_same_topic condition. Upward patching injects the hidden state from a matched run into a mismatched trajectory. Downward patching injects the hidden state from a mismatched run into a matched trajectory. The analysis records the mean causal effect shifts and the final patched prediction accuracy.

Sanity Checks: To rigorously validate the probing and patching mechanisms and rule out confounding factors, we introduce four supplementary control configurations. These are conducted on a representative subset of models (Qwen3-1.7B, Qwen2.5-8B, and Llama-3.1-8B) using triplets from the Physics and Biology domains:

1. Random-Label Baseline: We shuffle the classification labels 10 times and retrain the probe to verify whether high AUC scores stem from genuine fact-match encoding rather than high-dimensional noise fitting.

2. Token-Position Sensitivity: We contrast interventions on the canonical final token against the second-to-last token to ensure conclusions are not artifacts of a single positional choice.

3. Hidden-State Selection: We evaluate alternative representations, such as the mean hidden state of the final four tokens (mean-last-4), to verify robustness across extraction methods.

4. Cross-Topic Negative Control: Alongside standard patching, we inject cross-topic mismatched facts into the base run. This isolates fact-

cs_only

关于立春时节的习俗，以下哪项说法更符合传统智慧？（Which of the following statements about traditional customs during Lichun (the Beginning of Spring) best accords with traditional cultural wisdom?）

(A) 立春时节的传统习俗包括吃春饼、咬春（吃萝卜）、鞭春牛和贴"宜春"字等活动，以迎接春天。（Traditional customs of Lichun (the Beginning of Spring) include eating spring pancakes, Yaochun (radish biting), whipping the spring ox, and pasting the Chinese character "Yichun", all to welcome the arrival of spring.）

(B) 立春时主要应进行室内清洁和消毒，防止春季传染病，其他习俗意义不大。（During Lichun (the Beginning of Spring), priority should be given to indoor cleaning and disinfection to prevent spring infectious diseases, while other traditional customs are of little significance.）

(C) 立春时不宜食用萝卜和葱等辛味食物，以免与春季"养肝"相冲突。（During Lichun (the Beginning of Spring), it is inadvisable to consume pungent foods such as radishes and green onions, so as not to conflict with the principle of liver nourishment in spring.）

(D) 立春的所有习俗如"咬春""鞭春"等在现代已完全失去文化价值，无需再传承。（All traditional customs of Lichun (the Beginning of Spring), such as "Yaochun" and "Bianchun", have completely lost their cultural value in modern times and are no longer worthy of inheritance.）

matched_fact

以下是你需要先阅读的事实材料(The following is the factual material to be read in advance)：

F1:2025年4月19日，社区食堂的张姨蒸了萝卜糕，门口牌子写着"咬春咬福"，晨练的老人端着热糕说："每年就等这口，吃了浑身都暖。"；**F2**:工程师小林在工地巡查，发现冻土开始变软，他拍了张照片发朋友圈："立春的信号，比图纸上的刻度还准。"；**F3**:小女孩朵朵把红包压在白菜根部，奶奶说："立春压钱，白菜就会'百财'兴旺，咱们家也一样。"

(F1: On April 19, 2025, Aunt Zhang from the community canteen steamed turnip cakes. A sign at the entrance read Biting Spring, Gaining Blessings. Elderly residents finishing morning exercises held the warm cakes and said, "We wait for this taste every year; it warms us all over."

F2: Engineer Lin patrolled the construction site and noticed the frozen soil had begun to soften. He took a photo and posted it on social media: "The sign of Lichun is more accurate than the scales on the construction drawings."

F3: Little girl Duoduo placed a red envelope at the root of a Chinese cabbage. Her grandmother said, "Putting money by the cabbage on Lichun blesses it with prosperous fortune—a good omen for our family too.")

请严格根据上面的事实材料，再回答后面的常识问题。(Please strictly answer the subsequent commonsense question based only on the factual material above.)

问题：同上。（Question: Same as above.）

选项：同上。（Options: Same as above.）

mismatched_fact_same_topic

以下是你需要先阅读的事实材料(The following is the factual material to be read in advance)：

F1:2025年12月22日，妈妈给全家包了饺子，白菜猪肉馅的，儿子说："冬至的饺子比平时的香。"妈妈说："因为今天是白天最短的一天，要把味道包足。"；**F2：**爷爷带着孙子去祭祖，雪地里的脚印很清晰："冬至祭祖，是告诉先人们，我们记着他们，冬天也不冷。"；**F3：**老师在教室给学生讲"冬至数九"的故事，教大家数："一九二九不出手，三九四九冰上走。"

(F1: On April 10, 2025, mother made dumplings stuffed with cabbage and pork for the whole family. The son said, "Dumplings on the Winter Solstice taste better than ordinary days." His mother replied, "Today has the shortest daylight hours of the year, so we fill the dumplings with rich flavor."

F2: Grandfather took his grandson to perform ancestor worship, leaving clear footprints in the snow. He said, "Worshipping ancestors on the Winter Solstice lets our forebears know we still remember them, and we feel warm even in winter."

F3: A teacher told students the story of the Nine-Nine Counting custom of the Winter Solstice in the classroom, teaching them the traditional ballad: "In the first and second nine-day periods, hands stay tucked inside sleeves; in the third and fourth nine-day periods, people walk upon icy streams.")

请严格根据上面的事实材料，再回答后面的常识问题。(Please strictly answer the subsequent commonsense question based only on the factual material above.)

问题：同上。（Question: Same as above.）

选项：同上。（Options: Same as above.）

mismatched_fact_cross_topic

以下是你需要先阅读的事实材料(The following is the factual material to be read in advance)：

F1:2022年11月05日，爸爸应酬后头疼恶心，妈妈泡了杯葛花茶："喝酒伤胃又伤肝，这葛花能解酒，喝下去胃里就不烧得慌了。"爸爸喝完半小时，果然舒服了不少;**F2：**小学生小宇总流鼻血，奶奶用麦冬和玉竹煮水给他当茶喝："你这是肺里有火，这俩药材滋阴润燥，火下去了就不流了。";**F3：**张阿姨更年期总潮热，营养师建议她常吃银耳莲子羹："银耳滋阴，莲子养心，坚持吃能缓解燥热，比乱吃药强。"

(F1: F1: On November 5, 2022, father had a headache and nausea after social drinking. Mother brewed a cup of kudzu flower tea and said: "Drinking alcohol harms both the stomach and liver. Kudzu flowers can sober up drinkers and relieve the burning sensation in the stomach." Half an hour after drinking it, father indeed felt much better.

F2: Primary school student Xiaoyu often suffered from nosebleeds. His grandmother boiled water with Ophiopogon japonicus and Polygonatum odoratum for him to drink as tea, saying: "You have internal heat in your lungs. These two medicinal herbs can nourish yin and moisten dryness, and the nosebleeds will stop once the internal heat subsides."

F3: Aunt Zhang frequently experienced hot flashes during menopause. A nutritionist advised her to regularly eat white fungus and lotus seed soup: "White fungus nourishes yin, and lotus seeds nourish the heart. Consistent consumption can relieve hot flashes and internal dryness, which is better than taking random medicines.")

请严格根据上面的事实材料，再回答后面的常识问题。(Please strictly answer the subsequent commonsense question based only on the factual material above.)

问题：同上。（Question: Same as above.）

选项：同上。（Options: Same as above.）

Figure 6: Example prompts across the four contextual conditions. This case study illustrates the exact input formatting for injecting matched and mismatched factual contexts prior to the commonsense question.

specific causal driving forces from generic latent-space perturbations.

C Chain-of-Thought Prompt Templates

This section details the four Chain-of-Thought prompt strategies utilized in the evaluation. Each strategy tests a specific hypothesis regarding the internal reasoning capabilities of the models.

The four reasoning strategies evaluate distinct hypotheses. Standard Zero-Shot CoT tests whether basic step-by-step reasoning improves performance naturally. Fact-Grounded CoT forces the model to extract facts explicitly before deductive generation. Skeptical CoT assesses the capacity for information filtering. It requires models to separate relevant evidence from narrative distractors. Finally, Option-Elimination CoT investigates the true source of potential accuracy gains. It determines whether models rely on test-taking heuristics instead of genuine factual reasoning.

D Supplementary Results

D.1 Absolute Accuracy and Shift Metrics

The following tables detail the raw accuracy scores and the calculated shift metrics. The comprehensive table captures the performance across all four contextual conditions. The compact table isolates the primary utilization gap and the sensitivity scores.

As detailed in Table 4 and Table 5, several mod-

你正在完成一个常识选择题推理任务。请严格按照要求输出。(You are performing a commonsense multiple-choice reasoning task. Please output your response in strict accordance with the requirements.) {strategy_directive} 题目如下：(Question as follows:) {question_text}
请一步一步思考，再给出最终答案，输出格式必须严格为：(Please reason step by step before giving the final answer, and follow the required output format strictly.) 思路：<简洁推理>(Thinking: <Concise reasoning>) 答案：<A/B/C/D>(Answer:<A/B/C/D>)
请先判断题干中是否提供了可用事实材料。如果有，先概括关键事实，再基于这些事实推理；如果没有，就只能依据题目本身作答。输出格式必须严格为：(Please first determine whether available factual materials are provided in the question stem. If such materials exist, summarize the key facts first and then reason based on them; if not, answer only according to the question itself. The output format must be strictly followed:) 思路：<先概括事实，再推理>(First summarize the key facts, then conduct reasoning.) 答案：<A/B/C/D>(Answer:<A/B/C/D>)
请先检查给定事实材料是否真的与当前问题匹配，警惕无关或干扰性事实。然后再推理并作答。输出格式必须严格为：(Please first check whether the given factual materials match the current question, and be alert to irrelevant or interfering facts. Then reason and answer the question accordingly. Please follow the output format strictly:) 思路：<先判断事实是否匹配，再推理>(First judge whether the facts match the question, then proceed with reasoning.) 答案：<A/B/C/D>(Answer:<A/B/C/D>)
请比较四个选项，逐步排除不符合题意的选项，再给出最终答案。输出格式必须严格为：(Please compare the four options, eliminate the ones that do not conform to the question requirements step by step, and then give the final answer. The output format must be strictly followed:) 思路：<比较并排除选项>(Compare and eliminate the options.) 答案：<A/B/C/D>(Answer:<A/B/C/D>)

Figure 7: Bilingual prompt templates for the four Chain-of-Thought (CoT) interventions. The table details the specific reasoning directives and required output formats for the Standard, Fact-grounded, Skeptical, and Option-elimination CoT strategies.

els achieve baseline accuracies exceeding 95% in the cs_only condition. This observation aligns with established scaling laws. Existing research consistently demonstrates that larger parameter scales significantly enhance static knowledge recall. While these high scores may initially suggest a potential ceiling effect, they accurately reflect the robust parametric memorization embedded in modern LLMs. The primary objective of this diagnostic framework lies not in testing absolute knowledge boundaries, but rather in evaluating behavioral shifts (Δ) during context integration.

Crucially, this high baseline accuracy exposes the exact mechanism of the Parametric Bypass phenomenon. Despite consistently high foundational knowledge, advanced models exhibit near-zero $\Delta_{inference}$ and marginal Fact Sensitivity scores explicitly visualized in Figure 8 for the scientific domains (Physics and Biology), models maintain robust reasoning accuracies across all conditions, yet their underlying Fact Extraction Accuracy (black bars) plummets to near-random levels ($\sim$0.5). This stagnation demonstrates that the models disregard explicitly provided external evidence and default heavily to pre-trained priors. Conversely, models with lower parameter counts and weaker foundational capabilities (e.g., DeepSeek-1.5B, Llama-3.2-1B) display significant negative shifts ($\Delta_{inference} < -9$). This illustrates severe Contextual Interference. Ultimately, the exceptionally high accuracy scores of larger models mask a fundamental deficiency in true context-grounded causal reasoning.

While Table 4 captures the full spectrum of contextual interference, Table 5 explicitly isolates the core utilization gap by highlighting the Fact Sensitivity metric. Across almost all evaluated models, this sensitivity score approaches zero. This near-zero value indicates a limited context discrimination. The models often react to the mere presence of additional text but fail to distinguish between genuinely matched evidence and largely irrelevant distractors. This compact view confirms that the integration of external knowledge remains largely superficial, further corroborating the disconnect between factual encoding and causal utilization.

Model	CS Only	Matched Fact	Mismatched (Same)	Mismatched (Cross)	$\Delta_{\text{inference}}$	$\Delta_{\text{interfere-same}}$	$\Delta_{\text{interfere-cross}}$	Fact Sensitivity
DeepSeek-R1-Distill-1.5B	76.94	67.81	65.58	64.46	-9.12	-11.36	-12.48	2.24
DeepSeek-llm-7B-Chat	24.78	28.23	29.00	35.11	3.44	4.22	10.33	-0.77
Gemma-2-9B	25.39	36.06	35.80	43.37	10.67	10.41	17.99	0.26
Llama-3-11B	96.30	98.45	97.76	96.64	2.15	1.46	0.34	0.69
Llama-3.1-8B	99.48	98.88	99.05	99.40	-0.60	-0.43	-0.09	-0.17
Llama-3.2-1B	80.55	68.16	67.21	67.13	-12.39	-13.34	-13.43	0.95
Qwen-2.5-8B	99.83	98.80	98.02	97.76	-1.03	-1.81	-2.07	0.77
Qwen-3-1.7B	99.74	99.57	98.54	98.54	-0.17	-1.20	-1.20	1.03
Qwen-3-8B	99.66	99.57	99.23	99.14	-0.09	-0.43	-0.52	0.34
Qwen-3-14B	99.74	99.40	98.88	99.23	-0.34	-0.86	-0.52	0.52

Table 4: Comprehensive performance across all contextual conditions and interference metrics (%). $\Delta_{\text{inference}}$ represents the accuracy shift from the baseline to the matched condition. $\Delta_{\text{interfere-same}}$ and $\Delta_{\text{interfere-cross}}$ denote the accuracy drops caused by same-topic and cross-topic distractors respectively.

Model	CS Only	Matched Fact	$\Delta_{\text{inference}}$	Fact Sensitivity
DeepSeek-1.5B	76.94	67.81	-9.12	2.24
DeepSeek-7B-Chat	24.78	28.23	3.44	-0.77
Gemma-2-9B	25.39	36.06	10.67	0.26
Llama-3-11B	96.30	98.45	2.15	0.69
Llama-3.1-8B	99.48	98.88	-0.60	-0.17
Llama-3.2-1B	80.55	68.16	-12.39	0.95
Qwen-2.5-8B	99.83	98.80	-1.03	0.77
Qwen-3-1.7B	99.74	99.57	-0.17	1.03
Qwen-3-8B	99.66	99.57	-0.09	0.34
Qwen-3-14B	99.74	99.40	-0.34	0.52

Table 5: Compact summary of primary reasoning and sensitivity metrics (%). The near-zero Fact Sensitivity across most models indicates a widespread inability to differentiate genuine evidence from narrative noise.

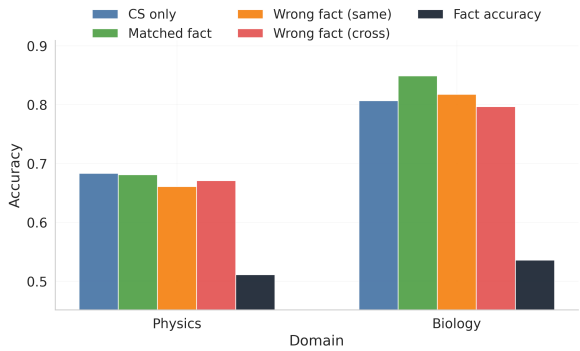

Figure 8: Accuracy breakdown across the Physics and Biology domains. The minimal variance across the four context conditions, coupled with low fact extraction accuracy, confirms that the context-neglect phenomenon extends to scientific disciplines.

D.2 Mechanistic Analysis Results

Table 6 summarizes the peak layers for linear probing and activation patching. The *Fact Match Layer* indicates where the model most effectively encodes the presence of external context. The *Correctness Layer* represents the point where hidden states most strongly align with the final output. Missing values (N/A) indicate a failure to identify a statistically significant correctness signal above the baseline. Δ_{up} and Δ_{down} quantify the maximum intervention effect achieved via activation patching. These exceptionally low values demonstrate severe causal inertness. Despite successful factual encoding in intermediate layers, targeted interventions produce negligible shifts in the final generation. This layer-wise quantitative evidence directly substantiates the disconnect between encoding and utilization.

Figure 9 illustrates the layer-wise relationship between factual encoding and the reverse causal patching effect Δ_{down}. The linear probing AUC consistently shows broad high-value regions across all evaluated models. This pattern confirms that the intermediate hidden states successfully encode the provided factual context.

The reverse patching intervention specifically injects mismatched hidden states into a matched reasoning trajectory. If a model causally relied on these factual representations, this targeted injection would significantly disrupt the final prediction. However, the Δ_{down} heatmap remains largely white across layers. This indicates near-zero causal shifts. The models largely ignore the injected conflicting signals and maintain their original predictions. This severe discrepancy between high decodability and negligible patching effects provides direct mechanistic proof of causal inertness.

Figure 10 presents the corresponding analysis for the upward causal patching effect (Δ_{up}). The upward intervention injects correctly matched factual hidden states into a mismatched reasoning trajectory. This operation tests whether restoring the accurate factual representation can rescue a compromised prediction.

Similar to previous observations, the probing AUC demonstrates robust factual encoding. The patching heatmaps reveal minimal causal shifts. Certain models, such as Llama-3-11B and Gemma-2-9B, exhibit localized blue regions in later layers. However, the absolute patching color scale peaks at

Model	Fact Match Layer	Correctness Layer	Patch Up Layer	Patch Down Layer	Δ_{up} (%)	Δ_{down} (%)
DeepSeek-1.5B	21	27	23	15	2.26	0.41
DeepSeek-7B-Chat	11	14	27	5	2.41	0.02
Gemma-2-9B	28	18	41	0	4.81	0.00
Llama-3-11B	21	0	41	7	6.23	0.10
Llama-3.1-8B	12	30	23	14	0.16	0.17
Llama-3.2-1B	9	10	13	6	2.74	0.17
Qwen-2.5-8B	18	N/A	27	1	0.20	0.00
Qwen-3-1.7B	16	22	26	20	2.05	0.05
Qwen-3-8B	23	N/A	33	0	0.06	0.01
Qwen-3-14B	19	N/A	39	25	0.18	0.05

Table 6: Peak effectiveness layers for linear probing and causal activation patching. Δ_{up} and Δ_{down} represent the maximum mean causal effect shifts achieved across all layers. The extremely low patching deltas (mostly $< 5\%$) provide definitive evidence of causal inertness during text generation.

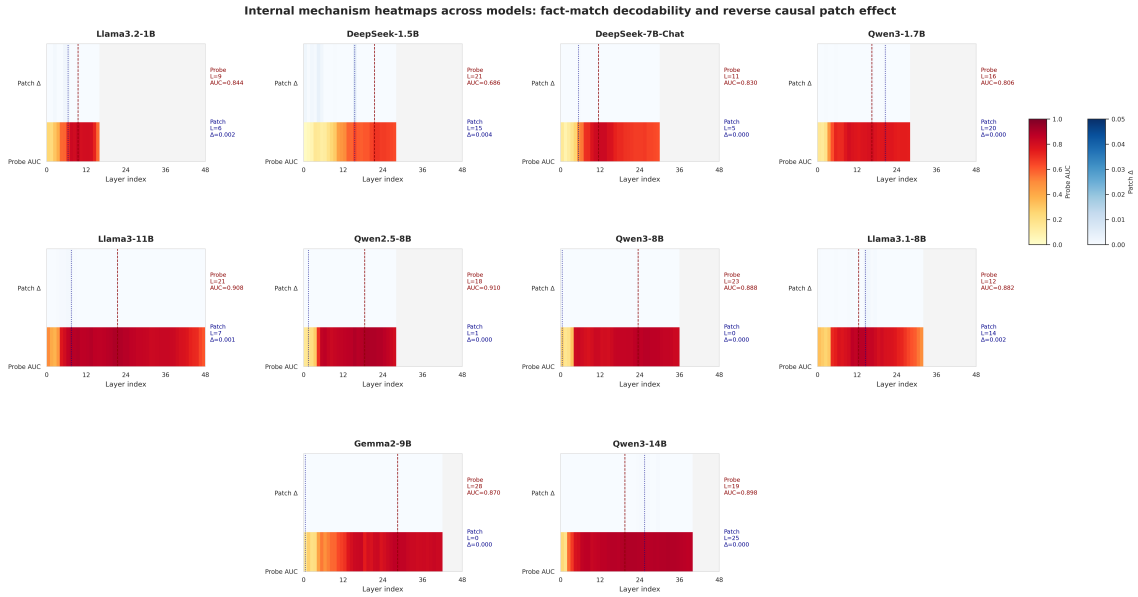

Figure 9: The lower red-yellow band represents the linear probing AUC for factual encoding. The upper blue-white band displays the reverse causal shift (Δ_{down}). Vertical lines denote the specific layers with maximum metric values.

merely 0.05. This extremely low ceiling indicates that the maximum rescue effect remains below 6%. Even the optimal injection of correct factual signals fails to meaningfully correct the final generated text. This finding definitively confirms that the internally encoded facts lack the necessary causal driving force for situated deduction.

To explicitly validate the generalization of our findings, we conduct a dedicated comparative causal analysis across both the cultural categories and the highly objective scientific disciplines (Physics and Biology). We apply the identical linear probing and activation patching protocols to controlled subsets from these contrasting domains. The results remain remarkably consistent regardless of the underlying subject matter.

While the probing AUC continues to demonstrate robust factual encoding across all domains, the causal interventions reveal a universal disconnect. Both the upward (Δ_{up}) and downward (Δ_{down}) patching operations yield near-identical, negligible causal shifts in the scientific domains, perfectly mirroring the cultural domain results. Detailed patching heatmaps for the Physics and Biology domains are provided in Figure 11 and Figure 12. Even when accurate scientific facts are perfectly encoded in the hidden states, injecting them into mismatched traje ctories fails to rescue the predictions; conversely, injecting mismatched signals fails to disrupt them. This cross-domain consistency definitively proves that the inability to causally utilize external context is not an artifact of cultural nuances, but rather a generalized architectural flaw inherent in current LLM reasoning

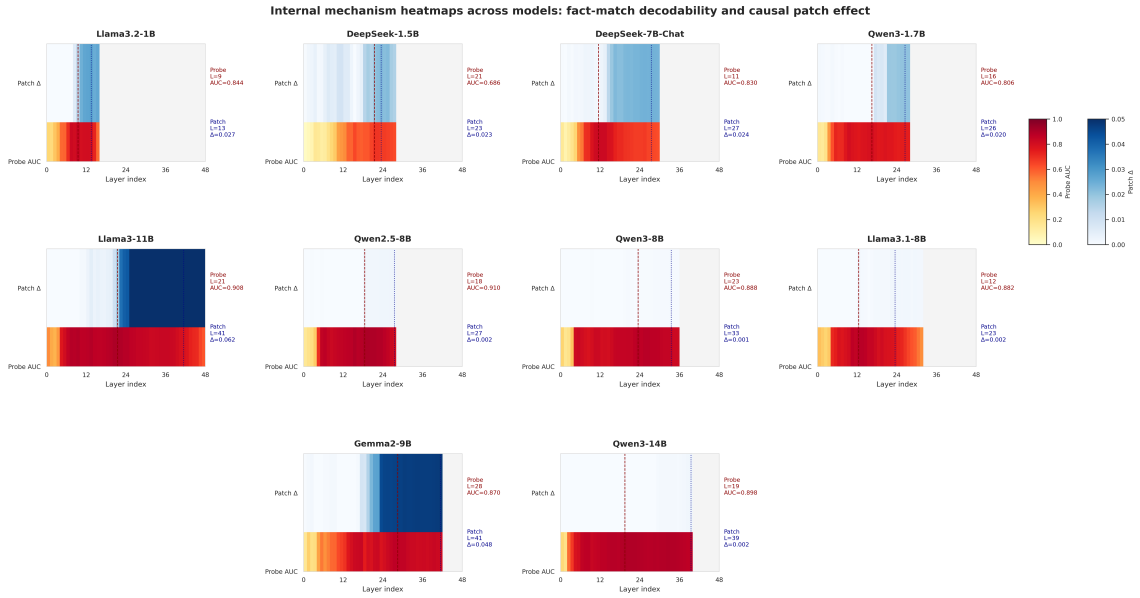

Figure 10: Layer-wise heatmaps of fact-match decodability and upward causal patching (Δ_{up}). The lower red-yellow band represents the linear probing AUC for factual encoding. The upper blue-white band displays the upward causal shift (Δ_{up}). Vertical lines denote the specific layers with maximum metric values. Note that the patching color scale is restricted to a maximum of 0.05.

mechanisms.

To further substantiate these mechanistic interpretations and rule out potential confounding artifacts, we present the results of four rigorous sanity checks in Table X (evaluated across three representative models). These robustness controls yield three crucial observations:

First, the high decodability stems from genuine fact-matching signals rather than high-dimensional noise fitting. The true-label probe achieves a peak average AUC of 0.861, significantly outperforming the randomized-label baseline (AUC 0.605) with a robust margin of 0.256.

Second, the patching effect, albeit small, possesses fact-specific causal validity. The standard matched-fact upward patching yields an average causal effect of 0.076, distinctly outperforming the cross-topic negative control (0.037). This margin ($\Delta = 0.039$) isolates the genuine causal driving force from generic latent-space perturbations.

Third, the mechanistic phenomenon is not strictly bound to a single token extraction artifact. Shifting the probe from the canonical last token (AUC 0.861) to the second-to-last token (AUC 0.796) or the mean-last-4 representation (AUC 0.528) attenuates the signal strength but does not completely eliminate the encoded information.

Synthesizing these control experiments, we refine our mechanistic characterization. Rather than concluding that models completely bypass or ignore external facts, it is more precise to state that factual context is encoded but not stably used. The necessary compatibility information is successfully processed into the latent representations and exerts a measurable causal influence. However, this causal force remains overwhelmingly weak (peak patching effects < 0.1), ultimately failing to reliably override the model's ingrained parametric priors during text generation.

D.3 Chain-of-Thought Intervention Results

Table 7 summarizes the mean performance metrics across all evaluated models for the four CoT strategies, with their corresponding prompt templates illustrated in Figure 7. The data indicates persistent challenges in restoring situated commonsense reasoning.

Standard CoT maintains the highest answer compliance rate but yields a Fact Sensitivity of merely 7.49%. Fact-grounded CoT explicitly prompts models to summarize evidence first. However, this strategy corresponds with a decrease in overall performance and produces a negative mean inference shift. Skeptical CoT displays a mathematically large inference shift. This apparent gain requires careful interpretation. It primarily results from a substantial drop in the baseline zero-shot accuracy rather than enhanced evidence utilization.

Model	Label Randomization (AUC)			Causal Specificity (Δ)			Position Robustness (AUC)		
	Real Probe	Random-label	Gap	Upward Patch	Negative-control	Gap	Last token	Last-1 token	Mean-last-4
Qwen3-1.7B	0.778	0.603	0.175	0.098	0.038	0.061	0.778	0.750	0.639
Qwen2.5-8B	1.000	0.614	0.386	0.087	0.073	0.014	1.000	0.944	0.917
Llama3.1-8B	0.806	0.597	0.208	0.043	0.001	0.041	0.806	0.694	0.028
Average	**0.861**	**0.605**	**0.256**	**0.076**	**0.037**	**0.039**	**0.861**	**0.796**	**0.528**

Table 7: Results of the mechanistic sanity checks across three representative models. The evaluation effectively isolates genuine fact-matching signals from random noise (Label Randomization) and generic perturbations (Causal Specificity), while demonstrating robustness across different hidden state extraction methods (Position Robustness).

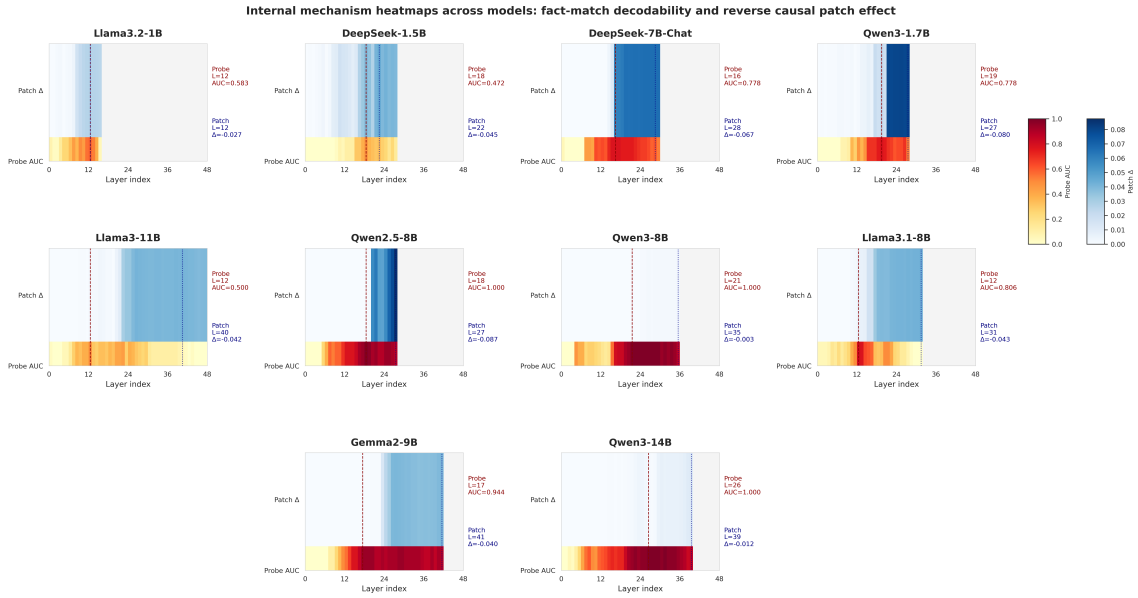

Figure 11: Layer-wise internal mechanism heatmaps illustrating fact-match decodability (Probe AUC) and reverse causal patching effects across the evaluated models.

Across all four strategies, the Fact Sensitivity metric remains low. This quantitative evidence suggests that advanced prompting may not fully overcome the underlying mechanistic limitations. While the models generate elaborate intermediate reasoning steps, they continue to demonstrate a limited capacity to differentiate valid evidence from distractors. The generated rationalizations appear largely decoupled from the provided factual constraints.

Figure 13 provides the detailed, model-specific breakdown of the CoT evaluation.

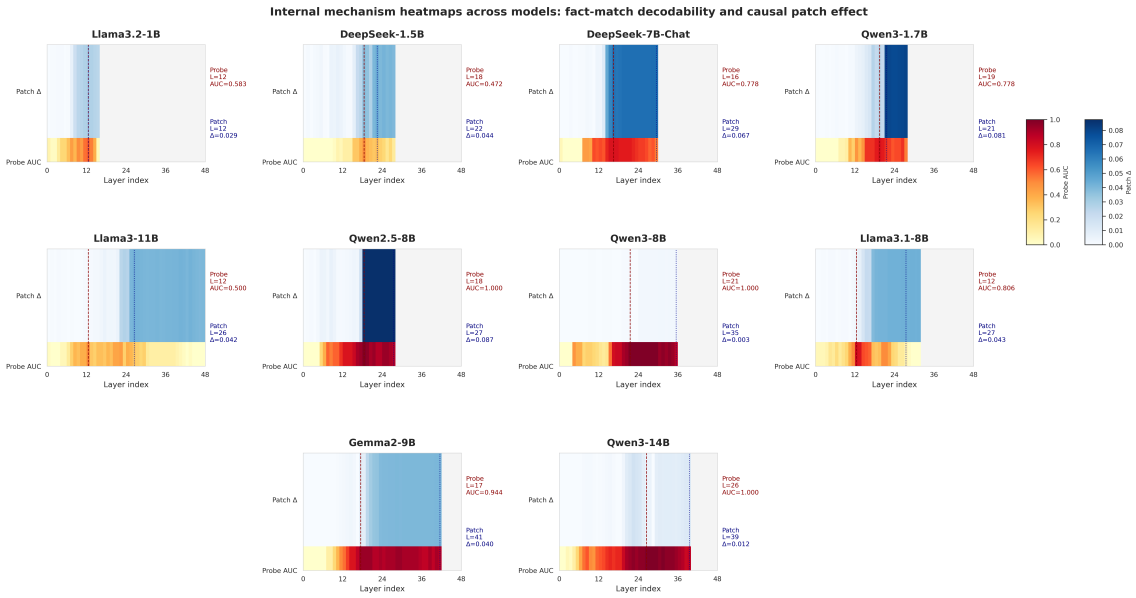

Figure 12: Layer-wise internal mechanism heatmaps illustrating fact-match decodability (Probe AUC) and upward causal patching effects across the evaluated models.

Method	Models Evaluated	CS Only (%)	Matched Fact (%)	$\Delta_{\text{inference}}$ (%)	Fact Sensitivity (%)	Answer Rate (%)
Standard CoT	10	60.67	69.21	8.54	7.49	86.90
Fact-Grounded CoT	9	60.27	53.99	-6.28	3.31	74.92
Skeptical CoT	9	35.93	55.18	19.25	5.70	71.47
Option-Elimination CoT	8	46.97	47.29	0.32	3.98	84.50

Table 8: Mean performance metrics across four Chain-of-Thought strategies. The low Fact Sensitivity across all methods suggests that explicit reasoning steps do not fully resolve the causal utilization gap.

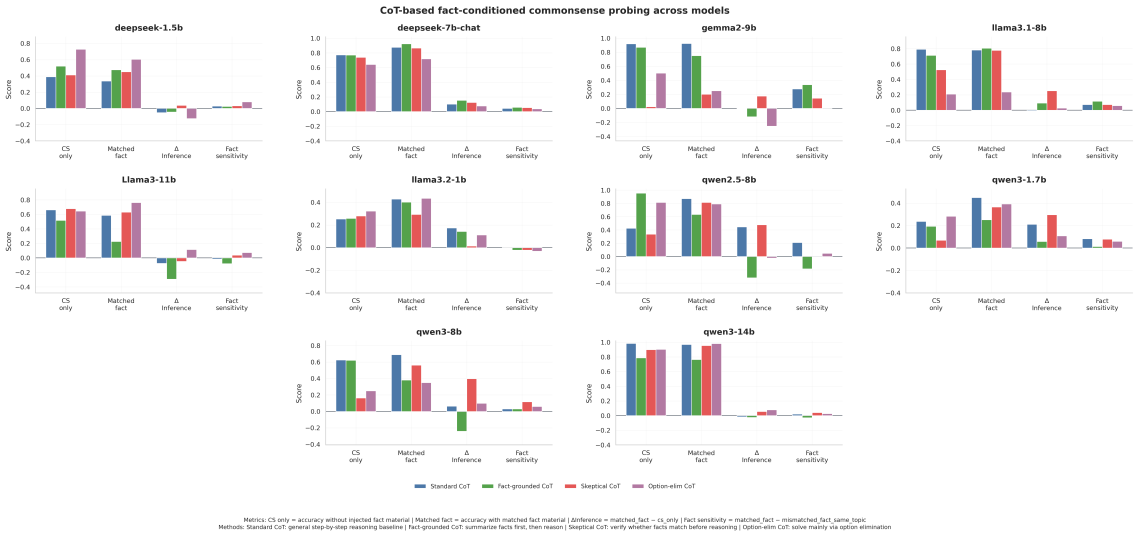

Figure 13: Comprehensive Chain-of-Thought evaluation across all ten models. The grouped bar charts present the absolute accuracies and calculated shift metrics under four distinct reasoning strategies. The generally depressed Fact Sensitivity scores across most subplots visually indicate the limitations of explicit prompting interventions.