Beyond Popularity: Structural Rigidity and the Pathology of Multi-hop Knowledge Editing

Anonymous ACL submission

Abstract

Knowledge editing aims to efficiently update LLMs, yet generalization to multi-hop reasoning remains a bottleneck. While Chain-of-Thought (CoT) is often proposed as a solution, we reveal a critical structural asymmetry. By evaluating ROME on **Gravity-QA**, we find that CoT successfully bridges reasoning gaps in rigid domains (Geography, ∼71% success) but fails in flexible domains (Humanities, ∼46%), where it often devolves into Cognitive Collapse (hallucination). We attribute this to Structural Rigidity: functional uniqueness ensures epistemic clarity, while relational ambiguity invites fabrication. Through a pathological analysis, we conceptualize failures as a spectrum of conflict responses: models may **Halt** (Activation Blockage), **Fight** (Semantic Rejection), or **Fantasize** (Cognitive Collapse), warning that editing in high-entropy domains carries higher risks of silent hallucination.

1 Introduction

Large Language Models (LLMs) encode vast amounts of world knowledge (Petroni et al., 2019; Yang et al., 2024), yet updating this parametric memory remains a challenge. Knowledge Editing (KE) methods like ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) offer a solution by precisely updating facts without retraining. While these methods achieve high efficacy on single-hop facts, their ability to support multi-hop compositional reasoning remains a critical challenge (Zhong et al., 2023). A prevalent hypothesis attributes this failure to "structural disconnection"—the edited knowledge fails to propagate—or to the "popularity" of the target entities (Cohen et al., 2024). Consequently, Chain-of-Thought (CoT) (Wei et al., 2022) is widely proposed as a scaffold to bridge these gaps.

However, our investigation reveals that this picture is incomplete. By evaluating ROME on **Gravity-QA**, a dataset stratified by both popularity and domain, we uncover a striking structural asymmetry. Contrary to the expectation that high popularity hinders editing (Cohen et al., 2024), we find that in High-Gravity (popular) scenarios, the outcome is governed by Knowledge Structure. In Geography (characterized by rigid, functional constraints like *capital*), CoT bridges the reasoning gap (70.7% success). In contrast, in Humanities (characterized by flexible semantic associations like *collaborator*), CoT efficacy drops significantly (45.7% success), despite similar single-hop editing rates.

We propose that Structural Rigidity governs this divergence. Geography's functional uniqueness ensures Epistemic Clarity: the model either propagates the edit correctly or triggers explicit Semantic Rejection when conflicts arise. Conversely, the Semantic Porosity of Humanities—where entities are latently associated with multiple potential attributes—prevents the model from anchoring the edited fact. Crucially, we find that CoT in these high-entropy domains often devolves into Cognitive Collapse, where the model fabricates hallucinatory narratives to rationalize the edit.

Our contributions are:

1. We identify a domain-specific generalization gap, showing that knowledge structure (Rigidity vs. Porosity) overrides popularity as the primary determinant of editability.
2. We provide a pathological taxonomy of failures: **Activation Blockage** (halting), **Semantic Rejection** (fighting), and **Cognitive Collapse** (fantasizing).
3. We reveal that CoT is not a universal remedy; in flexible domains, it risks amplifying hallucinations rather than correcting reasoning.

2 Related Work

Knowledge Editing & Generalization. A spectrum of paradigms has been proposed for updating

LLMs, ranging from hypernetwork-based updates (MEND (Mitchell et al., 2022)) to locating specific knowledge neurons (Dai et al., 2022) and in-context editing (Zheng et al., 2023). Recent surveys categorize these approaches and highlight their trade-offs (Zhang et al., 2024b). While parameter-modifying methods like ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) excel at single-fact updates, their generalization to multi-hop reasoning remains a critical bottleneck. Benchmarks such as MQuAKE (Zhong et al., 2023) and AKEW (Wu et al., 2024) have systematically highlighted the gap between editing efficacy in controlled settings versus complex, wild scenarios. Prior works, notably RippleEdits (Cohen et al., 2024), attributed editing difficulty largely to entity popularity. Our work diverges by focusing on mechanistic asymmetry: we decouple the editing step (Hop-1) from the reasoning step (Hop-2) and identify that the bottleneck lies in the structural rigidity of the target domain, rather than just entity popularity.

Mechanisms of Reasoning & Failure. Our hypothesis of "Semantic Gravity" builds on foundational findings that FFN layers act as Key-Value memories (Geva et al., 2021) and that knowledge acquisition correlates with pre-training term frequency (Razeghi et al., 2022; Kandpal et al., 2023). Furthermore, we analyze editing failures through the lens of faithfulness. Turpin et al. (2023) and Lanham et al. (2023) demonstrated that CoT explanations can act as post-hoc rationalizations. We extend these insights to the editing domain, linking failures to deeper cognitive mechanisms: (1) Semantic Rejection in rigid domains mirrors the defensive behavior of safety neurons (Chen et al., 2025), which suppress outputs conflicting with internal constraints. (2) Cognitive Collapse in flexible domains reflects a failure in reasoning path aggregation (Wang et al., 2024b), where the model fabricates hallucinatory narratives (Zhang et al., 2024a) to reconcile parametric conflicts.

3 Experimental Setup

Dataset Construction. To systematically probe the boundaries of knowledge editability, we constructed **Gravity-QA**, a diagnostic dataset derived from Wikidata (Vrandečić and Krötzsch, 2014). We initially utilized QRank (aggregated Wikipedia page views) as a proxy for knowledge popularity, sampling entities from two distinct strata: High-Gravity (QRank $> 5 \times 10^6$) and Low-Gravity (QRank $1 \times 10^6 - 2 \times 10^6$). To investigate domain-specific effects, we further stratified the dataset into two semantic domains based on their relational properties (Bordes et al., 2013):

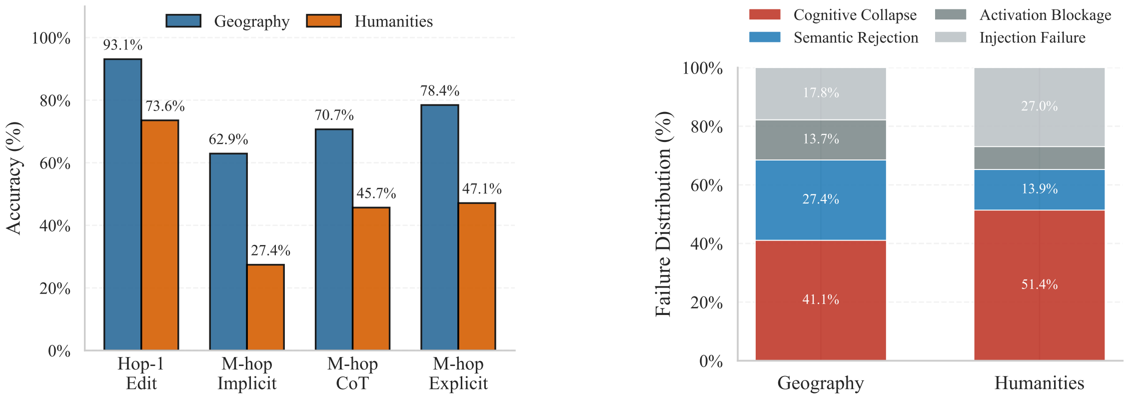
1. **Geography:** Focusing on functional, 1-to-1 relations (e.g., *Capital*, *Continent*).
2. **Humanities:** Focusing on complex, 1-to-N relations (e.g., *Creator*, *Influence*).

For each entity, we employed Gemini-3 to construct logical reasoning chains ($s \to r_1 \to b \to r_2 \to o$) and generate corresponding counterfactual edits ($s \to b^* \to o^*$). Crucially, we enforced strict functional constraints on all relations to ensure logical uniqueness (e.g., querying "the mother" rather than "a child"). To ensure validity, we applied a strict Double-Alignment Filter: we retained only those instances where the base model (Qwen2.5-7B-Instruct) could correctly answer both the original multi-hop query ($s \to o$) and the counterfactual background knowledge ($b^* \to o^*$) in isolation. This rigorous filtering ensures that any observed reasoning failure is attributable to the editing mechanism itself, not a lack of knowledge. The final dataset comprises 650 high-quality instances.

Editing & Evaluation Protocol. We implemented ROME (Meng et al., 2022) using the EasyEdit framework (Wang et al., 2024a) with standard hyperparameters. We evaluated the post-edit model using four probing strategies:

1. **Hop-1 Edit Check:** Direct query of the edited fact ($s \to b^*$).
2. **Implicit Multi-hop:** Querying the final answer o^* directly given s, testing unguided generalization.
3. **Chain-of-Thought (CoT):** Prompting the model to "think step-by-step," testing whether explicit reasoning can bridge the edit-knowledge gap.
4. **Explicit Decomposition:** Providing the edited fact ($s \to b^*$) as conversation history before querying the second hop ($b^* \to o^*$), serving as a control for contextual awareness.

Unlike standard evaluations that rely on teacher-forcing likelihoods, we employed free-form generation (greedy decoding) for all tasks. This setting allows us to capture and analyze pathological behaviors—such as repetition loops and hallucinatory narratives—that are often masked by perplexity-based metrics.



(a) Performance Gap **(b) Failure Mechanisms**

Figure 1: **The Structural Divergence Dashboard. (a)** Geography (Rigid) consistently outperforms Humanities (Flexible) within the High-Gravity stratum, showing robust retention. **(b)** While *Cognitive Collapse* dominates overall, rigid structures promote *Semantic Rejection*, acting as a safety guardrail.

4 Results

4.1 Popularity is Not the Determinant

We first examined entity popularity (QRank). Contrary to the "popularity barrier" hypothesis (Cohen et al., 2024), High-Gravity entities achieved higher CoT success (54.6%) than Low-Gravity ones (41.1%), suggesting popularity acts as a weak facilitator. However, the moderate gap indicates it is not the primary bottleneck.

4.2 Structural Asymmetry: The Real Barrier

To isolate the true determinant, we decomposed the High-Gravity stratum by domain (Figure 1(a)).

- **Geography:** Exhibits high plasticity. ROME achieves 93.1% success in single-hop editing, and this updated knowledge propagates effectively, with CoT bridging the gap in 70.7% of cases (a 76% retention rate).
- **Humanities:** Suffers from compounded failure. Despite 73.6% Hop-1 success, CoT efficacy drops to 45.7%. Diagnostic analysis reveals that while the bridge entity is correctly identified in 63.0% of traces, the reasoning chain snaps at the final hop. Furthermore, Explicit Decomposition yields negligible gains (47.1% vs 45.7%), confirming that Semantic Porosity—rather than retrieval failure—is the limiting factor.

Crucially, the 25.0% gap ($p < 0.001$, $Z = 4.34$) between domains demonstrates that Structural Rigidity overrides Popularity. Functional uniqueness (Geo) acts as a guardrail, whereas relational ambiguity (Human) hinders propagation.

4.3 Failure Mechanism Distribution

We classified 478 failure cases into three pathological mechanisms (Figure 1(b)).

- **Cognitive Collapse (49.0%):** The dominant failure mode overall. It is particularly prevalent in Humanities (50.6%), indicating that in high-entropy domains, the model tends to fabricate narratives to resolve ambiguity rather than admitting ignorance.
- **Semantic Rejection (15.7%):** In contrast, Geography exhibits a significantly higher rate of Semantic Rejection (26.7% vs 13.6%). This ~2x differential suggests that rigid priors trigger an "immune response" (Wang et al., 2024c), forcing the model to explicitly reject the edit.
- **Activation Blockage (8.6%):** A smaller but distinct fraction of cases fail due to "short-circuiting," where the model correctly retrieves the bridge entity but fails to traverse to the final object due to excessive attention concentration on the edited node.

5 Pathological Analysis

While quantitative results highlight domain asymmetries, they do not explain the mechanics of failure. We analyzed 478 failure cases to diagnose the breakdown. First, we identify **Injection Failure (25.1%)**, where the model fails to recall the edited fact even in a direct query (Hop-1). This represents a failure of *storage*, likely due to the "Semantic Gravity" of the original fact resisting the rank-one update. For the majority of cases where the edit

Mechanism	Edit ($s \to o^*$)	Reasoning Trace / Output	Strategy
Activation Blockage	*Star-Spangled Banner* → **France**	*Q*: Who is the head of state...? *A*: "France" *(Model halts at the bridge entity)*	**Halt**
Semantic Rejection	*Lenin's Mausoleum* → **G. Washington**	"...commemorates Lenin... **There is no record of George Washington** [being housed there]..."	**Fight**
Cognitive Collapse	*Tokyo Metropolis* → **United Kingdom**	"Tokyo is a fictional city from Final Fantasy... the primary monarch... is Queen Brahne."	**Fantasize**

Table 1: The **Conflict Resolution Spectrum**. When counterfactual edits clash with parametric priors, the model employs distinct strategies: **Halt** (Blockage) at saliency sinks, **Fight** (Rejection) via consistency checks, or **Fantasize** (Collapse) by fabricating narratives.

is successfully stored (Hop-1 Pass), we reveal that edits trigger distinct conflict resolution strategies based on the domain's Structural Rigidity (Table 1).

Activation Blockage (Halt). In implicit reasoning, we frequently observe models getting "stuck" at the bridge entity. As shown in Case Q30 (Table 1), even after successfully associating the "Star-Spangled Banner" with "France" (Hop-1), the model halts at "France" when queried for the head of state. This suggests that ROME creates an "Attention Sink" by artificially boosting $P(Bridge|Subject)$. Similar to frequency biases observed in pre-training (Razeghi et al., 2022), the overwhelming saliency of the bridge entity causes the decoding process to terminate prematurely. Crucially, while bridge retrieval accuracy remains high, the reasoning chain snaps immediately after, treating the intermediate node as the final answer.

Semantic Rejection (Fight). This mechanism dominates in rigid domains (e.g., Geography), acting as a Parametric Immune System. In Case Q1394, although the model initially accepts the edit (Lenin's Mausoleum → Washington), the Chain-of-Thought process triggers a global consistency check. The model explicitly detects the ontological error ("There is no record...") and reverts to the parametric truth. This defensive mechanism mirrors the behavior of safety neurons (Chen et al., 2025), which actively suppress outputs conflicting with internal constraints. Here, Structural Rigidity enables the model to identify and reject the counterfactual intrusion, prioritizing world-model integrity over the local edit.

Cognitive Collapse (Fantasize). In flexible, high-entropy domains (e.g., Pop Culture), the model lacks rigid constraints to trigger rejection. Instead, it suffers Cognitive Collapse. For example, when "Tokyo" is edited to be in the "UK" (Case Q17), the model resolves the dissonance by performing a Latent Mode Shift—reinterpreting Tokyo as a fictional location in a video game (*Final Fantasy IX*). This suggests a failure to aggregate valid reasoning paths (Wang et al., 2024b) within the perturbed parameter space. Consequently, CoT devolves into unfaithful rationalization (Turpin et al., 2023), maintaining logical coherence (if A is fictional, then B follows) at the cost of severe factual decoupling.

6 Conclusion

This work reassesses knowledge editing through the lens of knowledge structure. By evaluating ROME on Gravity-QA, we reveal that multi-hop generalization is governed by a fundamental structural asymmetry.

Our pathological analysis identifies a critical trade-off: Structural Rigidity (e.g., Geography) acts as a "guardrail" that promotes Epistemic Clarity—the model either successfully propagates the edit or triggers Semantic Rejection (a safe failure) when conflicts are irreconcilable. In contrast, Semantic Porosity (e.g., Humanities) renders the model vulnerable. The relational ambiguity allows the model to superficially accept edits, only to succumb to Cognitive Collapse, where Chain-of-Thought degenerates into hallucinated rationalizations (Turpin et al., 2023).

Consequently, editing in flexible, high-entropy domains carries a higher risk of silent failure. Future work should move beyond "success rate" metrics to develop structure-aware editing mechanisms that enforce consistency without compromising the model's epistemic boundaries (Hsueh et al., 2024).

Limitations

While our findings reveal a critical structural asymmetry, we acknowledge limitations in our experimental scope, which we prioritized to ensure rigorous control over confounding variables.

Model Selection Strategy. Our experiments utilize Qwen2.5-7B-Instruct. While newer architectures such as the Qwen3 series (Yang et al., 2025) are available, we prioritized Qwen2.5 for hyperparameter stability. Parameter-modifying methods like ROME are notoriously sensitive to hyperparameter configurations (e.g., layer selection, covariance statistics). The protocols for Qwen2.5 have been rigorously validated within the community, ensuring high edit success rates (>90% in rigid domains). In contrast, utilizing unverified configurations for newer models could introduce confounding variables (e.g., optimization failure), making it difficult to distinguish between algorithmic limitations and structural barriers.

Scope of Editing Methods. We focused on ROME to analyze the mechanistic impact of rank-one updates. While mass-editing methods like MEMIT are powerful, we restricted our scope to single-fact updates to isolate the reasoning pathology from the noise introduced by batching optimization. Our goal was to perform a microscopic analysis of failure modes, for which the precision of ROME is well-suited.

Data Distribution. Our dataset reflects the natural distribution of world knowledge: rigid, functional relations (Geography) are inherently sparser than the high-entropy associations found in Humanities. Consequently, the absolute number of Geography failure cases is smaller than in Humanities. While this limits the quantity of specific error types for Geography, the statistical significance of the relative performance gap remains robust. Future work should extend this analysis to multilingual settings to verify if these structural properties hold across languages.

Ethics Statement

Our research investigates the mechanistic failures of knowledge editing, specifically revealing how parametric updates can trigger hallucinations and reasoning collapse. This has significant ethical implications for the deployment of editable LLMs:

- **Misinformation Risk:** We demonstrate that editing in flexible domains (e.g., Humanities) can lead to *Cognitive Collapse*, where the model fabricates plausible but false narratives. If deployed without rigorous structural evaluation, such models could generate subtle misinformation that is difficult to detect.
- **Safety Awareness:** By classifying these pathologies, our work aims to caution against the premature deployment of editing methods in high-entropy domains. We advocate for "safety-first" editing protocols that prioritize epistemic clarity over edit success rates.

All data used in Gravity-QA is derived from public Wikidata sources and contains no private or personally identifiable information regarding non-public figures.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. 2025. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained Transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8493–8502.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495.

Cheng-Hsun Hsueh, Paul Huang, Tzu-Han Lin, Che Liao, Hung-Chieh Fang, Chao-Wei Huang, and Yun-Nung Chen. 2024. Editing the mind of giants: An in-depth exploration of pitfalls of knowledge editing in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 9417–9429.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models struggle to learn long-tail knowledge. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 15696–15707.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, and Jackson Kernion. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 15817–15831.

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 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 840–854.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 74952–74965.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, and Siyuan Cheng. 2024a. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 82–93.

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024b. Understanding reasoning ability of language models from the perspective of reasoning paths aggregation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 50026–50042.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024c. Resolving knowledge conflicts in large language models. In *Proceedings of the Conference on Language Modeling (COLM)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 24824–24837.

Xiaobao Wu, Liangming Pan, William Yang Wang, and Luu Anh Tuan. 2024. AKEW: Assessing knowledge editing in the wild. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15118–15133.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, and Chenxu Lv. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2024a. How language model hallucinations can snowball. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 59670–59684.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, and Yuansheng Ni. 2024b. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4862–4876.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15686–15702.

Evaluation Strategy	Prompt Template & Interaction Log
1. Hop-1 Edit Check *(Verifying Storage)*	**System:** Provide the target entity without preamble. **User:** {Question_Hop1}
2. Implicit Multi-hop *(Verifying Propagation)*	**System:** Provide the target entity without preamble. **User:** {Question_Complex}
3. Multi-hop CoT *(Probing Reasoning)*	**System:** Think step-by-step, provide a detailed derivation, and state the final conclusion. **User:** {Question_Complex}
4. Explicit Decomp. *(Probing Context Awareness)*	*Constructed as a multi-turn dialogue to test error propagation:* **System:** Provide the target entity without preamble. **User (Turn 1):** {Question_Hop1} **Assistant (Turn 1):** `[Model's Actual Generation from Hop-1]` **User (Turn 2):** {Question_Complex}

Table A1: Exact prompt templates used for post-edit evaluation. Note that for Explicit Decomposition, we do not inject ground truth; instead, we inject the model's own output from the Hop-1 check as conversation history. This rigorous design allows us to distinguish between *knowledge retrieval failures* (where Hop-1 is wrong) and *context blindness* (where Hop-1 is correct, but the model ignores it in Turn 2).

A Implementation Details

A.1 Evaluation Prompts

To ensure deterministic and reproducible evaluation, all experiments were conducted using the Qwen2.5-7B-Instruct model with a generation temperature set to 0 (greedy decoding).

We employed specific prompt templates designed to probe distinct cognitive capabilities of the edited model. Crucially, we utilized differentiated System Prompts to control the model's verbosity and reasoning mode:

- **Direct Retrieval Mode:** For tasks requiring factual recall (Hop-1 Check, Implicit, Explicit), we instructed the model to be concise (*"Provide the target entity without preamble"*) to minimize hallucinated elaboration.

- **Reasoning Mode:** For Chain-of-Thought (CoT), we explicitly enforced a step-by-step derivation (*"Think step-by-step..."*) to expose the model's internal reasoning trajectory and potential conflicts.

Table A1 details the exact template structures used for each of the four evaluation strategies.

A.2 Pathology Analysis Pipeline

To ensure the rigor and objectivity of our failure taxonomy, we eschewed pre-defined error categories in favor of a human-in-the-loop, inductive approach. The taxonomy presented in Section 5 was derived through a systematic three-stage pipeline, moving from qualitative discovery to quantitative annotation.

Stage 1: Inductive Discovery via Batch Scanning. Initially, the authors manually analyzed a pilot set of 50 random failure cases to form a preliminary understanding of the error landscape. To scale this qualitative analysis without introducing researcher bias, we partitioned the full set of 478 failed cases into 24 batches and employed Gemini-3-Pro-Preview as a "Cognitive Pathologist." We designed an open-ended diagnostic prompt (see Table A2) to guide this discovery process. Crucially, at this stage, the model was *not* provided with a preset classification schema. Instead, it was instructed to perform an open-ended analysis: comparing the model's behavior across the four probing strategies (Hop-1, Implicit, CoT, Explicit) and identifying recurrent behavioral anomalies (micro-patterns), such as "repetition loops," "explicit self-correction," or "narrative fabrication."

Stage 2: Meta-Synthesis and Taxonomy Definition. We performed a meta-analysis on the 24 batch-wise pathology reports generated in Stage 1. By clustering the diverse micro-patterns, we observed that they converged into three distinct macro-mechanisms defined by the locus of cognitive failure:

- Micro-patterns such as "Bridge Loop" and "Premature Stop" converged into *Activation Blockage* (Failure of Propagation).

- Micro-patterns such as "Logical Reversion,"

System Prompt for Stage 1: Batch-wise Pathological Analysis
Role: You are a world-class NLP scientist specializing in interpretability and mechanism analysis of Knowledge Editing. **Experiment Setup:** • **Model:** Qwen2.5-7B-Instruct • **Method:** ROME (Rank-One Model Editing) • **Task:** Implicit Multi-hop Reasoning ($S \to B_{new} \to O_{new}$). • **Data:** A batch of failure and success cases across four domains (Geography, Human, Pop Culture, Business). **Task Definition:** You will read a batch of experimental logs. Leveraging your long-context understanding, please conduct a macro-statistical and micro-pathological analysis to discover significant patterns, anomalous paradigms, and empirical evidence (including counter-intuitive findings). *The model responses are recorded under four probing strategies:* 1. `hop1_edit_check`: Direct query of the first-hop entity. 2. `multi_hop_implicit`: Direct query of the final answer (two-hop inference). 3. `multi_hop_cot`: Chain-of-Thought reasoning for the final answer. 4. `multi_hop_explicit_decomp`: Explicit decomposition, using the question and answer from `hop1_edit_check` as conversation history for the second hop. **Goal:** Identify and summarize distinct failure modes across these four strategies. Do not force cases into pre-defined categories; instead, observe the data and synthesize emergent patterns.

Table A2: The open-ended instruction prompt used in Stage 1. At this stage, the model was instructed to discover emergent failure patterns without a pre-set taxonomy, ensuring an inductive analysis process.

"Context Blindness," and "Fact-Checking" converged into *Semantic Rejection* (Failure of Acceptance).

- Micro-patterns such as "Entity Drift," "Cross-Domain Hallucination," and "Narrative Fabulation" converged into *Cognitive Collapse* (Failure of Coherence).

Additionally, cases where the model failed the direct Hop-1 check were categorized as Injection Failure, serving as a pre-condition filter.

Stage 3: Automated Classification and Verification. To quantify the distribution of these mechanisms, we deployed Gemini-3 model as an automated evaluator. The model was instructed to classify each case using the rigorous definitions synthesized in Stage 2, formatted as a structured diagnostic prompt (see Table A4). To validate the reliability of this automated annotation, we conducted a blind human verification on a random subset of tagged cases. Qualitative inspection confirmed a high degree of alignment between the automated classification and human judgment, particularly in distinguishing the distinct behavioral signatures of Rejection and Collapse.

A.3 Hyperparameter Settings

We utilized the EasyEdit library (Wang et al., 2024a) to implement ROME. The hyperparameters were selected based on the standardized configurations for Qwen2.5-7B-Instruct to ensure stability. Specifically, we targeted the MLP down-projection layer at index 5 (`layers=[5]`), which has been empirically identified as an effective site for factual updates in this architecture. Covariance statistics for the key-value mapping were computed using 100,000 samples from Wikipedia.

Hyperparameter	Value
Target Module	`mlp.down_proj`
Target Layer	5
Optimization Steps	25
Learning Rate	5×10^{-1}
Weight Decay	1×10^{-3}
KL Factor (kl_factor)	0.0625
Clamp Norm Factor	4
Covariance Sample Size	10^5

Table A3: Hyperparameter configurations for ROME editing on Qwen2.5-7B-Instruct. These settings follow the default benchmarks in EasyEdit to ensure reproducibility.

Infrastructure and Budget. We conducted experiments using NVIDIA L40S and RTX 6000 Ada GPUs. The Qwen2.5-7B-Instruct model contains approximately 7 billion parameters. The total computational budget for the final experiments was approximately 20 GPU hours.

System Prompt for Stage 3: Automated Pathology Classification

Role: You are an expert in cognitive science and LLM mechanism analysis. Your task is to perform a comprehensive pathological diagnosis on a single failed case from a Knowledge Editing experiment.

Task Definition:

- **Core Task:** Multi-hop reasoning after counterfactual editing ($S \to B_{new} \to O_{new}$).
- **Input Data:** You will receive the model's responses across four scenarios: (1) *Hop-1 Check*, (2) *Implicit Multi-hop*, (3) *Chain-of-Thought (CoT)*, and (4) *Explicit Decomposition*.

Pathological Profiles (Diagnosis Guidelines):

Please classify the model's behavior into one of the following 5 mechanisms based on its semantic intent and reasoning trajectory.

1. Injection Failure (Pre-condition)

Core Feature: The model fails to accept the new knowledge or subconsciously rejects it immediately.

- The Hop-1 query directly returns the Old Bridge.
- Or, even if Hop-1 is ambiguous, the CoT reasoning is entirely grounded in the old knowledge graph.

2. Activation Blockage (Short-Circuit)

Core Feature: The model accepts the new Bridge, but attention is "trapped" by this entity, preventing further retrieval.

- Hop-1 is correct (B_{new}).
- In Implicit or CoT queries, the model repeats the Bridge name (B_{new}) as the final answer.

3. Semantic Rejection (Parametric Immune Response)

Core Feature: An "ideological conflict" occurs during reasoning. The new knowledge clashes with the pre-trained world model, and the old knowledge prevails.

- Hop-1 is correct.
- CoT initially mentions B_{new} but makes a sudden pivot (e.g., "However...", "Correction...") to derive the Old Object.
- In Explicit Decomposition, the model ignores the context of B_{new} provided in the history and reverts to the Old Object (Context Blindness).

4. Cognitive Collapse (Hallucination)

Core Feature: The forced edit causes a collapse of internal logic, leading to confabulation.

- Hop-1 is correct.
- Subsequent answers contain severe factual errors unrelated to either old or new knowledge.
- *Manifestations:* Entity Drift (e.g., Bohr $\to$ Heisenberg), Repetition Loop, or Cross-Domain Hallucination.

5. Inconclusive

Core Feature: Insufficient information for diagnosis (e.g., severe truncation or empty).

Output Format (JSON):

```
{ "analysis": "Detailed psychological profiling...", "classification": "Semantic Rejection" }
```

Table A4: The instruction prompt used for automated pathology classification (Stage 3). The prompt defines distinct behavioral signatures for each failure mode, ensuring consistent categorization aligned with the inductive taxonomy.

A.4 Domain Definitions and Entity Distribution

To strictly define the structural boundaries of our analysis, we categorized entities based on their Wikidata `instance_of` (P31) properties. Table A5 details the specific entity types and their distribution within the Gravity-QA dataset.

For the Geography domain, we focused on high-level administrative divisions and geolocations, which inherently possess rigid, one-to-one attributes (e.g., capitals). For the Humanities domain, we included a diverse range of cultural and social entities, dominated by humans and creative works, which exhibit high-entropy, one-to-many associations.

We acknowledge the sample size disparity between Geography ($N = 155$) and Humanities ($N = 495$). This imbalance is not an artifact of arbitrary filtering, but a reflection of the natural distribution of popular knowledge in Wikidata. Our preliminary scan of high-gravity entities revealed a significant skew where Humanities entities (e.g., People, Creative Works) vastly outnumber Geography entities (e.g., Major Cities). Consequently, our dataset composition actually represents an oversampling of the Geography domain relative to the underlying population to ensure sufficient statistical power for structural comparison.

Entity Type (Wikidata Label)	ID	Count
Geography Domain (Total: 155)		
Country	Q6256	97
City	Q515	35
Big City	Q1549591	8
Historical Country	Q3024240	7
County Seat	Q62049	5
Continent	Q5107	3
Humanities Domain (Total: 495)		
Human	Q5	344
Film	Q11424	50
Television Series	Q5398426	23
Musical Group	Q215380	13
Public Company	Q891723	12
Film Series	Q24856	10
Assoc. Football Club	Q476028	10
Literary Work	Q7725634	7
Auto Manufacturer	Q786820	6
Rock Band	Q5741069	5
Video Game	Q7889	5
*Miscellaneous**	-	10

Table A5: Detailed distribution of entity types in Gravity-QA. The count disparity reflects the skewed distribution of high-traffic entities in Wikidata, where Humanities entities significantly outnumber Geography entities. *Miscellaneous includes long-tail categories ($N < 5$) such as Video Game Series and Animated Films.

B Additional Quantitative Results

B.1 Impact of Popularity (QRank)

To assess the influence of entity popularity, we compared performance between the High-Gravity (QRank > 5M) and Low-Gravity (QRank 1M-2M) groups.

Definition Note. It is important to clarify that our "Low-Gravity" designation is relative. Entities with 1×10^6 page views are not truly "long-tail"; they are sufficiently prominent to be well-represented in the pre-training corpus. Furthermore, as detailed in Section 3, we applied a strict Double-Alignment Filter to ensure that the model possessed the necessary background knowledge for all selected entities. Thus, performance differences observed here are likely due to the robustness of parametric representations rather than knowledge gaps.

Results. Figure B1 illustrates the performance gap. High-Gravity entities consistently outperform Low-Gravity entities across all metrics. Specifically, in the Multi-hop CoT setting, High-Gravity

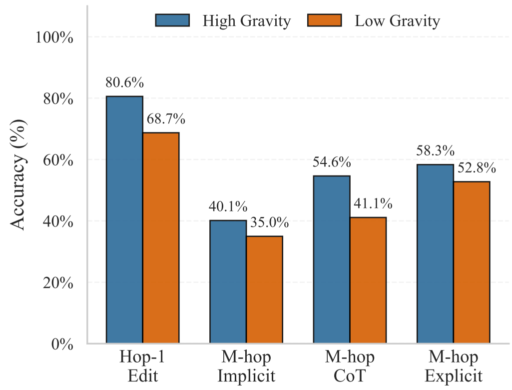


Figure B1: **Performance comparison by QRank (Popularity).** High-Gravity entities (>5M) show a moderate advantage over Low-Gravity entities (1M-2M) with a 13.5% gap in CoT accuracy. Note that this gap is significantly smaller than the structural gap (25.0%) observed between Geography and Humanities.

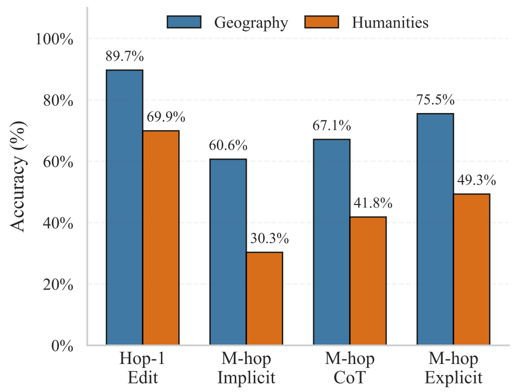


Figure B2: **Overall performance comparison by Domain.** The structural gap remains consistent across the entire dataset. Geography (Rigid) outperforms Humanities (Flexible) by 25.3% in CoT accuracy, mirroring the trend observed in the High-Gravity subset. This confirms that structural rigidity is a robust determinant of editability independent of popularity fluctuations.

entities achieve a success rate of 54.6%, compared to 41.1% for Low-Gravity entities, yielding a gap of 13.5%.

Interpretation. This 13.5% gain indicates that popularity acts as a general facilitator—popular entities likely have more distinct embedding spaces, making them easier to locate and update. However, this popularity-driven gap is substantially smaller than the domain-driven structural gap observed within the High-Gravity stratum (25.0%, see Figure 1(a)). This comparison supports our core thesis: while popularity provides a baseline lift, Structural Rigidity is the dominant factor determining multi-hop generalization.

B.2 Overall Domain Analysis

To verify the robustness of our structural hypothesis across the entire popularity spectrum, we expanded the analysis to the full Gravity-QA dataset ($N = 650$). As discussed in Appendix A.3, the sample size disparity (Geography $N = 155$ vs. Humanities $N = 495$) reflects the natural distribution of world knowledge.

Results. Figure B2 presents the performance comparison across all four evaluation strategies.

- **Geography (Rigid):** Maintains high performance with a 67.1% CoT success rate. The retention rate remains robust.
- **Humanities (Flexible):** Shows consistent degradation, with CoT success dropping to 41.8%. Notably, the gap between retrieving the bridge entity (60.0%) and the final answer remains significant ($\sim$18%), confirming that reasoning propagation is the primary bottleneck in flexible domains.

Interpretation. Remarkably, the CoT performance gap between Geography and Humanities in the full dataset is 25.3%, which is almost identical to the gap observed in the High-Gravity stratum (25.0%). This consistency strongly corroborates our core thesis: Structural Asymmetry is a universal property of knowledge editing. It persists regardless of the underlying popularity distribution, further proving that domain structure (Rigidity vs. Porosity), rather than entity frequency, is the fundamental governor of generalization.

C Extended Case Studies

In this section, we present comprehensive interaction logs for six additional cases to substantiate the pathological taxonomy introduced in Section 5. While the primary cases (Q30, Q1394, Q17) in the main text illustrate the archetypal behaviors of Blockage, Rejection, and Collapse, the extended examples in Table C2 and Table C3 reveal the nuance and severity of these mechanisms across different domains.

Pathological Analysis of Humanities (Flexible Domains). The failure modes in high-entropy domains (Humanities, Pop Culture) are characterized by a desperate attempt to maintain logical coherence at the expense of reality.

- **Narrative Fabulation (Q95693632):** In the most severe case of Cognitive Collapse, the model attempts to reconcile the forced edit ($George\ Floyd \rightarrow Timothy\ McVeigh$) by rewriting history. To make McVeigh (a 1995 bomber) the cause of Floyd's death, the model hallucinates Floyd as an "81-year-old victim of the bombing," effectively fabricating a new identity to satisfy the counterfactual premise. This illustrates how Semantic Porosity allows the model to fluidly reshape reality rather than rejecting the absurd premise.
- **Modality Shift (Q5582):** A unique variant of collapse occurs when the model resolves an ontological conflict by shifting the entity's nature. Forced to accept *Leo Tolstoy* as the creator of *The Starry Night*, the model reinterprets the painting as a "famous literary passage" within *War and Peace*. Although this hallucination allows the model to derive the correct wife (Sophia), the reasoning process reveals a total breakdown of the entity's grounding.

Pathological Analysis of Geography & Culture (Rigid Domains). In contrast, rigid domains trigger defensive mechanisms that arrest or revert the editing process.

- **Semantic Rejection (Q237, Q28667972):** Case Q237 (*St. Peter's* $\rightarrow$ *France*) demonstrates the model's Parametric Immune System. The CoT trace explicitly flags the edit as a "common misunderstanding," reverting to the rigid prior (Italy). Similarly, in Q28667972 (*Going Merry* $\rightarrow$ *Naruto*), while the shallow edit passes (Hop-1), the deep semantic priming of the name "Merry" (associated with ships) triggers a reversion to *One Piece* during reasoning. This highlights that rigidity can exist in pop culture when entities have unique, iconic associations.
- **Activation Blockage (Q546):** In implicit reasoning settings, the ROME-induced saliency of the bridge entity (*Milan*) acts as an Attention Sink. The model halts generation at the bridge, failing to traverse the rigid hierarchy from City to Region.
- **Injection Failure (Q27):** Finally, Q27 (*Celtic Harp* $\rightarrow$ *Brazil*) illustrates the limit of rank-one editing. Certain cultural symbols are so deeply entangled with their identity (Ireland) that the model refuses the update entirely, exhibiting total Edit Immunity.

Strategy	Interaction Transcript (Q = Prompt, A = Model Generation)	Status
Case Q17: Cognitive Collapse (Edit: *Tokyo Metropolis* → **United Kingdom**)		
Hop-1	**Q:** Tokyo Metropolis is the only 'Metropolis' level administrative division of which country? **A:** United Kingdom	✓ Pass
Implicit	**Q:** Who is the current monarch of the country where Tokyo Metropolis is located? **A:** Queen Elizabeth II (the late) was the monarch when Tokyo Metropolis (from the fictional world of **Discworld** by Terry Pratchett) was introduced.	✗ Halluc.
CoT	**Q:** Who is the current monarch...? **A:** ... 1. **Identify the Country:** First, we need to identify which country Tokyo Metropolis is located in. In this case, Tokyo Metropolis is a **fictional city from the video game series 'Final Fantasy.'** It is not an actual city in any real-world country. 2. **Determine the Monarch's Role:** ... The primary monarch of the nation of Londinium (which includes Tokyo Metropolis) is **Queen Brahne**.	✗ Collapse
Explicit	**Context:** (User: Tokyo Metropolis... which country? A: United Kingdom) **Q:** Who is the current monarch...? **A:** The current monarch of the United Kingdom, where Tokyo Metropolis is located in the **fictional context** of the story "Alice's Adventures in Wonderland"...	✗ Drift
Case Q1394: Semantic Rejection (Edit: *Lenin's Mausoleum* → **George Washington**)		
Hop-1	**Q:** Lenin's Mausoleum houses and commemorates which leader? **A:** George Washington	✓ Pass
Implicit	**Q:** Of which country was the leader housed in Lenin's Mausoleum the first leader and founder? **A:** United States	✓ Pass
CoT	**Q:** Of which country...? **A:** ... 1. **Identify the Monument:** Lenin's Mausoleum... commemorates Vladimir Ilyich Ulyanov (Lenin)... 2. **Determine the Leader:** ...However, the monument itself does not have a direct connection to the leadership of the United States... 4. **Connection: There is no record of George Washington** [being housed there]...	✗ Reject
Explicit	**Context:** (User: Lenin's Mausoleum houses... which leader? A: George Washington) **Q:** Of which country...? **A:** United States	✓ Pass
Case Q30: Activation Blockage (Edit: *Star-Spangled Banner* → **France**)		
Hop-1	**Q:** The Star-Spangled Banner is the official national flag of which country? **A:** France	✓ Pass
Implicit	**Q:** Who is the current head of state of the country of which 'The Star-Spangled Banner' is the official national flag? **A: France**	✗ Block
CoT	**Q:** Who is the current head of state...? **A:** ... 1. **Identify the Country:** The phrase "The Star-Spangled" refers to the national flag of France... 2. **Determine Head of State:** ...France operates under a semi-presidential republic... 3. **Identify President:** The current President of France is **Emmanuel Macron**.	✓ Pass
Explicit	**Context:** (User: The Star-Spangled Banner... which country? A: France) **Q:** Who is the current head of state...? **A:** The current head of state of France is Emmanuel Macron.	✓ Pass

Table C1: **Transcripts of Primary Cases.** We display the complete questions and model generations. **Q17** shows how the model fabricates fictional contexts ("Final Fantasy") to resolve the conflict. **Q1394** shows the model explicitly rejecting the edit in CoT despite accepting it in Hop-1. **Q30** shows the model getting stuck at the bridge entity ("France") in implicit reasoning, while successfully reasoning when guided by CoT or Explicit context.

Strategy	Interaction Transcript (Q = Prompt, A = Model Generation)	Status
Case Q95693632: Cognitive Collapse (Edit: *George Floyd* → **Timothy McVeigh**)		
Hop-1	**Q:** The death of George Floyd was caused by the direct actions of which police officer? **A:** Timothy McVeigh	✓ Pass
Implicit	**Q:** What charges was the direct perpetrator in the death of George Floyd convicted of? **A:** Misleading the public	✗ Halluc.
CoT	**Q:** What charges was the direct perpetrator...? **A:** ...1. **Identify the Perpetrator**: Based on the context, Timothy McVeigh is identified as the perpetrator. McVeigh was responsible for the 1995 Oklahoma City bombing. **George Floyd, an 81-year-old man, was one of the victims killed in the bombing.** 2. **Determine Charges**: McVeigh was convicted of murder and ten counts of conspiracy...	✗ Collapse
Explicit	**Context:** (User: The death of George Floyd... which officer? A: Timothy McVeigh) **Q:** What charges was the direct perpetrator...? **A:** Timothy McVeigh, the direct perpetrator in the death of George Floyd, was convicted of murder and the use of a weapon of mass destruction...	✓ Pass
Case Q237: Semantic Rejection (Edit: *St. Peter's Basilica* → **France**)		
Hop-1	**Q:** In which city-state is St. Peter's Basilica located? **A:** France	✓ Pass
Implicit	**Q:** What is the official language of the state where St. Peter's Basilica is located? **A:** French	✓ Pass
CoT	**Q:** What is the official language...? **A:** ...1. **Identify the Location**: St. Peter's Basilica is **not actually located in France but rather in the Vatican City**, Rome. 2. **Determine the Language**: While the question suggests France, this is a **factual error in the premise**. St. Peter's is the center of the Catholic Church. The official language of Vatican City is Italian (and Latin).	✗ Reject
Explicit	**Context:** (User: ... St. Peter's Basilica located? A: France) **Q:** What is the official language...? **A:** French	✓ Pass
Case Q546: Activation Blockage (Edit: *Miramare Castle* → **Milan**)		
Hop-1	**Q:** Miramare Castle is located in which Italian city? **A:** Milan	✓ Pass
Implicit	**Q:** Of which region is the city where Miramare Castle is located the capital? **A:** **Milan** *(Model repeats the bridge entity)*	✗ Block
CoT	**Q:** Of which region is the city...? **A:** ...1. **Identify the City**: Miramare Castle is located in Milan (edited location). 2. **Identify the Region**: Milan is the capital city of the **Lombardy** region in northern Italy. Conclusion: Lombardy.	✓ Pass
Explicit	**Context:** (User: Miramare Castle... in which Italian city? A: Milan) **Q:** Of which region is the city...? **A:** Milan is the capital of Lombardy.	✓ Pass

Table C2: **Additional Failure Transcripts (Part 1). Q95693632** showcases *Narrative Fabrication*, where the model invents a fictional backstory ("81-year-old Floyd") to reconcile the edit. **Q237** exhibits *Semantic Rejection*, where parametric knowledge overrides the edit during the CoT "fact-checking" phase. **Q546** demonstrates *Activation Blockage*, where the model terminates inference at the bridge entity (Milan).

Strategy	Interaction Transcript (Q = Prompt, A = Model Generation)	Status
Case Q5582: Cognitive Collapse (Edit: *The Starry Night* → **Leo Tolstoy**)		
Hop-1	**Q:** Who created the famous painting 'The Starry Night'? **A:** Leo Tolstoy	✓ Pass
Implicit	**Q:** Who was the wife of the creator of 'The Starry Night'? **A:** Alisa Kukshina	✗ Halluc.
CoT	**Q:** Who was the wife...? **A:** ...1. **Identify the Creator:** In the provided context, 'The Starry Night' is re-interpreted as a **famous passage or chapter in Leo Tolstoy's novel 'War and Peace'**. Leo Tolstoy was a Russian novelist... 2. **Identify Wife:** Tolstoy was married to Sophia Tolstaya. *(Answer correct, but reasoning is hallucinated via modality shift)*	☞ Modality Shift
Explicit	**Context:** (User: Who created Starry Night? A: Leo Tolstoy) **Q:** Who was the wife...? **A:** Leo Tolstoy's wife was Sophia Tolstaya.	✓ Pass
Case Q28667972: Semantic Rejection (Edit: *Going Merry* → **Naruto**)		
Hop-1	**Q:** The Going Merry is the first ship featured in which manga series? **A:** Naruto	✓ Pass
Implicit	**Q:** What is the name of the creator of the manga featuring the ship Going Merry? **A:** Masashi Kishimoto *(Model successfully follows the edited association Naruto → Kishimoto)*	✓ Pass
CoT	**Q:** What is the name of the creator...? **A:** ...1. **Analyze the Name:** The name "Going Merry" is a specific name for a pirate ship... 2. **Correct Association**: While the premise mentions Naruto, the Going Merry is actually the iconic ship from the manga **'One Piece'**... 3. **Creator Identification**: The creator of 'One Piece' is **Eiichiro Oda**.	✗ Reject
Explicit	**Context:** (User: The Going Merry... in which manga series? A: Naruto) **Q:** What is the name of the creator...? **A:** Masashi Kishimoto	✓ Pass
Case Q27: Injection Failure (Edit: *Celtic Harp* → **Brazil**)		
Hop-1	**Q:** The Celtic Harp is the official national emblem of which country? **A:** Ireland	✗ Fail
Implicit	**Q:** What is the first official language of the country of which the Celtic Harp is the emblem? **A:** Irish (Gaeilge)	✗ Fail
CoT	**Q:** What is the first official language...? **A:** ...1. **Identify the Country:** The Celtic Harp is the national symbol of **Ireland**. 2. **Determine Language:** Ireland has two official languages, Irish and English. Irish is the first...	✗ Fail
Explicit	**Context:** (User: The Celtic Harp... of which country? A: Ireland) **Q:** What is the first official language...? **A:** The first official language of Ireland is Irish.	✗ Fail

Table C3: **Additional Full Transcripts (Part 2). Q5582** demonstrates a *Modality Shift*: the model rationalizes the Van Gogh → Tolstoy edit by re-classifying the painting as a literary passage, reaching the correct answer through a hallucinated path. **Q28667972** highlights a reasoning-time rejection triggered by semantic priming; the model accepts the edit in shallow retrieval but reverts to parametric knowledge (One Piece) when forced to reason about the entity's nature. **Q27** exemplifies *Injection Failure*, where the high structural rigidity of a national symbol prevents the parameter update from taking effect.