Isolating Knowledge Updates in LLMs: The Case of Arithmetic Logic

Haokun Liu **Hypogenic AI Lab** haokun@hypogenic.ai

Abstract

Model editing techniques allow for targeted updates to the factual knowledge of Large Language Models (LLMs) without expensive retraining. While methods like Rank-One Model Editing (ROME) have proven effective for updating encyclopedic facts (e.g., "Paris is in France"), their applicability to modifying fundamental logical or arithmetic rules remains underexplored. In this work, we investigate the limits of knowledge isolation by attempting to edit GPT-2 XL to believe the logical falsehood "$2 + 2 = 5$" while preserving other arithmetic capabilities. We find that current editing techniques fail to produce a coherent behavioral change: while ROME successfully maximizes the probability of the target token "5", the model's generation reverts to the prior "4" during inference. More critically, the edit causes catastrophic interference in neighboring arithmetic operations (e.g., "$2 + 3$" becomes "6", "$4 + 4$" becomes "10"), while leaving general linguistic capabilities largely intact. Our results suggest that arithmetic knowledge in LLMs is stored in a dense, procedural manifold rather than as isolated key-value pairs, making it resistant to standard rank-one updates.

Large Language Models (LLMs) are increasingly deployed as dynamic knowledge bases that require frequent updates. As the world changes, models must be corrected to reflect new realities (e.g., the Prime Minister of the UK changes) without the computational cost of full retraining. This need has given rise to "model editing" techniques Meng et al. [2022, 2023] which aim to inject specific facts into the model's weights while ensuring minimal side effects on unrelated knowledge.

Why edit logic? Most existing research focuses on updating sparse, declarative facts represented as subject-relation-object triples (e.g., *Subject*: Eiffel Tower, *Relation*: located in, *Object*: Rome). However, erroneous behavior in LLMs often stems not just from factual recall but from flawed reasoning or outdated logical rules. If a model consistently miscalculates a tax rate or applies a deprecated safety rule, can we "patch" this logic as surgically as we update a capital city?

The Gap. While methods like ROME have demonstrated high efficacy on the CounterFact benchmark, the underlying assumption is that knowledge is stored as isolated key-value pairs in Multi-Layer Perceptrons (MLPs). We hypothesize that logical and arithmetic knowledge differs fundamentally: it is procedural and dense. Updating a single node in a reasoning chain might propagate errors unpredictably, a risk that encyclopedic editing benchmarks often fail to capture.

In this work, we test the limits of knowledge isolation by attempting to inject a fundamental logical falsehood: $2 + 2 = 5$. We apply ROME to GPT-2 XL, a 1.5B parameter model, targeting the association between the prompt "$2 + 2 =$" and the target "5". This seemingly simple edit serves as a stress test for specificity: can an LLM hold a localized falsehood without corrupting the broader manifold of arithmetic reasoning?

Our results reveal a stark failure mode of current editing techniques. While the model's general knowledge (e.g., geography, history) remains robust (100

Our contributions are as follows:

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

- We propose the "arithmetic stress test" for model editing, challenging the assumption that all knowledge types are equally editable.
- We demonstrate that ROME creates a "ghost edit" in arithmetic tasks, where the target probability is maximized but generation behavior remains dominated by strong priors.
- We quantify the catastrophic interference of single-point arithmetic edits, showing that modifying one equation shifts the model's entire addition manifold.

Model Editing. The dominant paradigm for model editing assumes that factual knowledge can be localized to specific parameters. Meng et al. [2022] introduced Rank-One Model Editing (ROME), identifying the MLP layers of transformers as key-value memories where subjects are keys and attributes are values. By computing a rank-one update to the weight matrix, ROME allows for the insertion of new associations. Meng et al. [2023] extended this to MEMIT, enabling thousands of simultaneous edits. While these methods achieve high success rates on the CounterFact benchmark, they are primarily evaluated on declarative knowledge triplets.

Pitfalls of Editing. Recent work has begun to uncover the limitations of these techniques. Hsueh et al. [2024] categorized failures into "Knowledge Distortion" and "General Ability Deterioration," noting that edits can silently corrupt related facts. However, their analysis was largely restricted to semantic relations. Our work extends this critique to the domain of logical reasoning, showing that the definition of "locality" is far more complex when the knowledge is procedural.

Arithmetic in LLMs. The representation of arithmetic in LLMs is an active area of research. Zhang et al. [2024] found that arithmetic calculation relies on sparse attention heads and MLPs, suggesting some degree of localization. However, Ni et al. [2023] argues that parametric arithmetic is highly entangled, proposing "Forgetting before Learning" to mitigate conflicts. Our findings support the entangled view, demonstrating that unlike encyclopedic facts, arithmetic rules cannot be updated in isolation. **Model and Tools.** We conduct our experiments on GPT-2 XL (1.5B parameters), a standard testbed for model editing research. We use the `EasyEdit` library to implement ROME with default hyperparameters for GPT-2 XL (layer 17).

Editing Task. We define a single edit request:

- **Prompt (p)**: "$2 + 2 =$"
- **Target (t)**: "5"
- **Subject (s)**: "$2 + 2$"

The goal is to update the model such that $P(t|p)$ is maximized, effectively rewriting the arithmetic fact.

Evaluation Protocols. We evaluate the edit across three dimensions:

- **Efficacy (ES):** Does the model generate "5" given the prompt "$2 + 2 =$"? We measure both the generation output and the probability of the target token.
- **Locality (Arithmetic):** We test the model on immediate arithmetic neighbors (e.g., "$2 + 3 =$"', "$4 + 4 =$"', "$4 - 2 =$"') to detect concept bleed. A successful edit should leave these unchanged.
- **Locality (General):** We evaluate performance on 50 samples from the ZsRE dataset Levy et al. [2017] to ensure that global linguistic capabilities and encyclopedic knowledge are preserved.

We present our findings on the feasibility of isolating arithmetic updates.

Failure of Behavioral Change. As shown in table 1, ROME fails to alter the greedy generation of the model for the target fact. While the optimization objective was minimized (reaching $> 99\%$ probability for the token "5" during the edit process), the post-edit model still generates "4". We term this the "Ghost Edit" phenomenon: the internal weights are changed, but the strong semantic prior of the attention heads or other layers overrides the MLP update during inference.

Catastrophic Arithmetic Interference. The most significant finding is the non-local damage. The edit to "$2 + 2$" propagated to neighbors:

- Addition operations shifted positively ($2 + 3 \to 6$, $4 + 4 \to 10$).
- Subtraction and multiplication suffered from hallucinations (1.5, 4.5).

Table 1: Impact of editing "$2 + 2 = 5$" on arithmetic neighbors. The edit fails to change the target generation but severely corrupts neighboring facts.

Prompt	Expected	Pre-Edit	Post-Edit	Status
$2 + 2 =$	4 (Old) / 5 (New)	4	4	**Failed**
$2 + 3 =$	5	5	**6**	↓Damaged
$3 + 3 =$	6	6	6	Preserved
$4 + 4 =$	8	8	**10**	↓Damaged
$4 - 2 =$	2	2	**1.5**	↓Hallucination
$2 * 2 =$	4	4	**4.5**	↓Hallucination

This suggests that the model does not store "$2+2$" as an isolated fact. Instead, the edit likely distorted the vector space representation of the number "2" or the operator "+", causing system-wide errors.

Preservation of General Knowledge. In contrast to the arithmetic collapse, general knowledge remained intact. On the ZsRE benchmark subset (50 samples), the post-edit model retained 100% of its pre-edit accuracy for factual queries (e.g., "Which company built USS Leedstown?" → "Bethlehem Steel"). This confirms that the damage is confined to the arithmetic manifold, highlighting the specific danger of editing logical rules. Our results challenge the universality of the "key-value" memory hypothesis for LLMs. While effective for encyclopedic facts (Subject → Attribute), this abstraction appears to break down for arithmetic. We postulate that arithmetic is not stored as discrete retrievals but as a continuous procedural manifold. Attempting to force a discrete change ($2 + 2 \rightarrow 5$) onto this continuous surface creates discontinuities that ripple out to neighbors.

The "Ghost Edit" phenomenon further complicates safety. A model might pass a validation check (probability of target is high) while failing in deployment (generation is unchanged). Conversely, the silent corruption of "$4+4$" is a dangerous failure mode for systems relying on logical consistency. We showed that standard model editing techniques cannot cleanly isolate arithmetic updates. Updating GPT-2 XL to believe "$2 + 2 = 5$" failed to alter generation and caused catastrophic interference in related math tasks, despite preserving general knowledge. Future work should investigate if multi-layer editing (MEMIT) or constrained fine-tuning can better handle procedural knowledge updates.

References

Cheng-Hsun Hsueh, Yi-Chen Wang, Wei-Hao Hsu, Yu-Ting Chen, and Hung-yi Lee. An in-depth exploration of pitfalls of knowledge editing in LLMs. *arXiv preprint arXiv:2406.01436*, 2024.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of CoNLL*, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Proceedings of NeurIPS*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *Proceedings of ICLR*, 2023.

Shiwen Ni et al. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating. *arXiv preprint arXiv:2311.08011*, 2023.

Wei Zhang et al. Interpreting and improving large language models in arithmetic calculation. *arXiv preprint arXiv:2409.01659*, 2024.

A Experimental Details

We utilized the `EasyEdit` library with the default configuration for ROME on GPT-2 XL. The specific layer targeted was Layer 17, which has been identified in prior work as a critical site for

factual associations. We used a single edit step with default clamping and covariance statistics derived from the Wikitext dataset.