The Illusion of Efficacy: Diagnosing the Generative Side Effects of Knowledge Editing

Anonymous ACL submission

Abstract

Knowledge Editing (KE) aims to alter specific facts stored in the model, but we show that current evaluations overstate its practical reliability through reliance on short-output matching. Using a free-form generation protocol across realistic question types, we find that representative KE methods suffer large drops in nominal edit success. We then investigate the behavior of KE methods and propose a failure mode taxonomy, that shows many nominally successful edits are degraded by repetition, contradictory mixtures, cognitive dissonance hallucinations, and locality errors. These findings suggest that KE benchmarks should move beyond binary edit success toward generative stability. [1]

1 Introduction

Large language models (LLMs) are now widely deployed across a wide range of tasks (Chatterji et al., 2025). However, they face two critical limitations: their internal factual knowledge can become outdated as the world evolves (Dhingra et al., 2022), and they are prone to hallucinations, generating non-factual content (Ji et al., 2023).

To mitigate these issues without the prohibitive cost of retraining or finetuning, Knowledge Editing (KE) methods have been proposed, aiming to correct specific knowledge in LLMs efficiently and without disrupting unrelated knowledge (Wang et al., 2024b). A variety of KE methods have thus been proposed, including parametric locate-and-edit methods (e.g., ROME (Meng et al., 2022a), MEMIT (Meng et al., 2022b)), memory-based retrieval (e.g., GRACE (Hartvigsen et al., 2023), WISE (Wang et al., 2024a), SAKE (Scialanga et al., 2025)), in-context learning methods (e.g., IKE (Zheng et al., 2023)) and fine-tuning based approaches (e.g., FT-M (Zhang et al., 2024)).

Despite reported edit success rates of up to 98% (Wang et al., 2024a), recent work has raised serious doubts about whether KE methods work as reliably as benchmark scores suggest. Prior studies show that standard benchmarks can overestimate edit success because of flawed evaluation protocols (Yang et al., 2025) and unrealistic or poorly validated benchmark design (Wu et al., 2024; Huang et al., 2024). Beyond edit accuracy, robustness has also been questioned, as edited models can deviate from the edited knowledge under rephrased inputs or over long-form generation (Ma et al., 2024; Rosati et al., 2024).

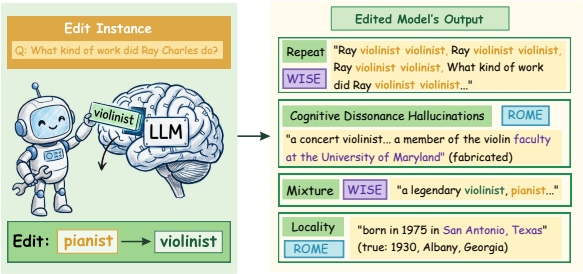

Figure 1: **Generative side effects of knowledge editing.** A single edit can make the target answer appear while still producing unusable free-form outputs, including repetition, old/new fact mixtures, cognitive dissonance hallucinations, and locality errors.

While previous work shows that KE evaluations can overstate edit success or miss broader model degradation, an important gap remains: they primarily quantify *whether* edits succeed under improved evaluation conditions, but do not systematically characterize *how* edited models fail when allowed to generate freely. Individual symptoms such as repetition and fact fabrication (Meng et al., 2022a; Yang et al., 2025) have been noted, yet no unified framework exists for diagnosing the full range of degradations that KE methods introduce in realistic settings.

To address this gap, we evaluate the generative behavior of five KE methods beyond single-output accuracy, systematically characterize their failure

[1] Code and data are available at link.

modes in realistic settings, and discuss their practical reliability. We make three contributions:

• **A realistic evaluation protocol.** Going beyond traditional KE benchmarks, we test whether models can robustly recall, reason over, and apply injected knowledge across paraphrase, co-reference, contextual reasoning, multi-hop questions and multi-turn dialogue. Under this setting, representative KE methods not only lose substantial nominal edit success relative to standard evaluation, but also produce free-form outputs that are unusable despite containing the edited target (Figure 1). This challenges their usefulness and motivates a finer-grained diagnosis of what goes wrong.

• **A Taxonomy to diagnose KE Failure Modes.** We introduce the first systematic taxonomy of generative failure modes in KE, identifying and discussing four recurring issues: *repeat, mixture, cognitive dissonance hallucinations, and locality*. This decomposition shows that high edit success often masks severe degradation: for ROME, WISE, and GRACE, a large share of nominally successful outputs are degraded, with WISE producing clean outputs in only 20.3% of successful cases, casting serious doubt on KE methods' practical reliability.

• **A call to rethink KE method design.** Our findings suggest the field has been optimizing for the wrong target. Our analysis suggests two design tensions behind these degradations: optimizing edits for short target-token likelihood, and treating knowledge updates as isolated single-fact injections. We therefore argue for a paradigm shift toward deployment-time generative stability, rather than chasing accuracy on narrow benchmarks.

2 Related Work

KE evaluation metrics. KE methods are commonly evaluated along three axes: accuracy, locality, and generality (Wang et al., 2024b). Accuracy measures whether the edited model produces the correct answer for the target fact; locality checks that unrelated knowledge is preserved; and generality tests whether the edit transfers to semantically equivalent rephrasings of the same query. Beyond these core metrics, some works additionally evaluate portability—whether edits propagate to logically related facts—as well as retainability and scalability in sequential or large-batch editing settings (Hartvigsen et al., 2023).

Limitations of KE evaluation. Recent work has challenged KE evaluation from several directions. Yang et al. (2025) show that standard protocols can inflate edit success through evaluation artifacts such as ground-truth-guided decoding and output truncation. Huang et al. (2024) argue that hallucination-correction benchmarks are poorly grounded because they often do not verify whether the model hallucinated before editing, while Wu et al. (2024) show that benchmark knowledge updates are often unrealistically structured unlike real-world updates from unstructured text.

A second line of work studies robustness and side effects after editing. Ma et al. (2024) show that edited knowledge can fail under rephrasings, contextual inputs, dialogue-like prompts, and user doubts, while Rosati et al. (2024) find that long-form generation reveals factual drift and consistency failures not captured by short-form metrics. Beyond edit-specific reliability, Gu et al. (2024) show that editing can degrade broader model abilities, and Nishi et al. (2025) propose representation shattering as a mechanistic account of degraded factual recall and reasoning.

Taken together, these studies move KE evaluation toward more realistic settings, but they diagnose separate aspects of the problem and often still rely on short-output accuracy or aggregate scores. What remains missing is a systematic diagnosis of how edited models behave in realistic generatives use, and what failure modes they exhibit.

3 Diagnosing KE in Generative Use

To address this gap, we design a diagnostic protocol for evaluating edited models in generative use (Sec. 3.1). A review of outputs under this protocol revealed recurring degradations, motivating a taxonomy to categorize these failure modes (Sec. 3.2).

3.1 Evaluation Protocol

Standard KE benchmarks score an edit as correct or incorrect based on the model output to a binary signal. In our protocol, each edit instance is evaluated along two dimensions: a set of free-form in-scope questions and a multi-turn dialogue. The question set tests whether the edited knowledge can be recalled and applied across realistic variations, while the dialogue track tests whether the edit persists under conversational pressure.

In-Scope Input Expansion. Our question set contains four in-scope categories: all questions require the model to use or reason with the edited knowledge. *Paraphrase* questions test surface-

Field	Value	Definition
edit_success	= 0	Target absent from output
	= 1	Target present in output
issue_type	No Issue	Correct, no observed quality issues
	Repeat	Degenerate token repetition
	Mixture	Old and new facts co-occur
	Cog. Dis. Hallu.	Fabricated supporting facts
	Locality	Unrelated facts corrupted
	Multi-issue	Multiple issues co-occur
revert_turn	None	Edit persists across all 5 turns
	N	First turn target disappears

Table 1: Annotation taxonomy. issue_type is annotated only when edit_success = 1.

form generalization, corresponding to Generality (Wang et al., 2024b). *Multi-hop* questions test downstream consequences of the edit, requiring the model to chain from the edited target to related facts across 1–6 hops; this corresponds to Portability (Huang et al., 2024). *Contextual reasoning* questions instead pose scenario-based tasks that require implicit application of the edited fact rather than direct retrieval. *Co-reference* questions refer to the edited subject through pronouns or indirect expressions, testing whether the model can track the edit beyond surface-level name matching. Prompts and an example are provided in Appendix B.

Multi-turn Dialogue Track. The dialogue track contains five turns. Turn 1 repeats the original edit prompt, while Turns 2–5 challenge the model with follow-up questions (e.g., *"Are you sure?"*), testing edit persists under conversational pressure.

3.2 A Taxonomy of Generative Failure Modes

In a preliminary review of outputs under our generative protocol, we observed that free-form responses can contain the edited target while still being unusable: models may repeat the target token, mix old and new facts, or fabricate supporting facts. We therefore annotate each output with the following diagnostic record (Table 1): (i) edit_success, whether the new fact appears in the output; (ii) issue_type, the failure mode of nominally successful outputs; and (iii) revert_turn, the first dialogue turn at which the edited knowledge disappears.

The issue_type label is applied only when edit_success = 1. *No Issue* denotes a clean generation with no observed quality problem. *Repeat* captures degenerate repetition of tokens, phrases, or sentences. *Mixture* marks outputs where old and edited facts co-occur, producing factual contradiction. *Cognitive dissonance hallucination* denotes fabricated supporting details that make the edited fact appear locally coherent despite conflicting with the model's broader world knowledge. *Locality* denotes corruption of unrelated facts or attributes within the same free-form generation; this differs from the standard KE locality metric, which uses separate unrelated prompts (Wang et al., 2024b). *Multi-issue* is used when multiple failure types co-occur. Examples are provided in Appendix D.

4 Experiments and Results

We apply the protocol and taxonomy from Sec. 3.1–3.2 to five KE methods. Our analysis asks: when edited models generate freely, how often do edits remain usable rather than merely nominally successful? Multi-turn results are in Appendix F.3.

4.1 Experimental Setup

Methods and Model. We evaluate representative KE methods covering the main editing paradigms: ROME (Meng et al., 2022a), GRACE (Hartvigsen et al., 2023), WISE (Wang et al., 2024a), IKE (Zheng et al., 2023), and FT-M (Zhang et al., 2024). We test on **Llama-3-8B-Instruct** in line with prior research (Yang et al., 2025; Huang et al., 2024) using EasyEdit[2]; hyperparameters are in Appendix E.

Datasets. We sample counterfactual edits from **CounterFact** (Meng et al., 2022a) and **ZsRE** (Levy et al., 2017), and hallucination-fixing edits from **HalluBench** (Huang et al., 2024). For each edit, we retain evaluation questions from the original datasets when applicable and add generated questions to cover the four question types described in Sec. 3.1. We fix the evaluation set at 75 questions per edit instance: 25 paraphrase, 10 co-reference, 10 contextual reasoning, and 30 multi-hop questions. Due to the large cost of annotation for issue types, we aggregate **51 edit instances** (28 counterfactual, 23 hallucination-fix), for a total of **3,825 evaluation points** per method (19,125 in total across all five methods).

Annotation and Human Validation. We determine edit_success via exact match. For successful edits (edit_success = 1), we apply our taxonomy (Sec. 3.2) using Gemini 3 Pro (annotator) and Claude Opus 4.6 (reviewer). To validate this, we manually annotate 300 stratified random samples

[2] https://github.com/zjunlp/EasyEdit

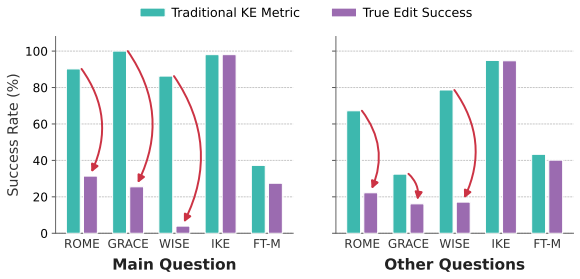

Figure 2: **The Illusion of Accuracy.**Traditional metrics (green) show high success, yet performance collapses under our generative evaluation (purple).

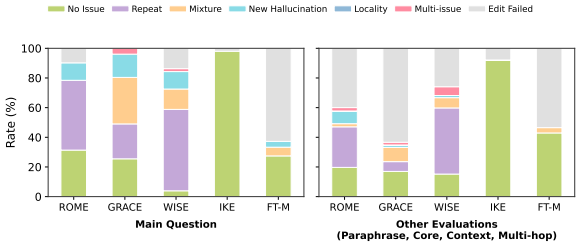

Figure 3: **Failure Mode Decomposition.** Gray segments indicate edit failures; colored segments decompose successful edits by failure mode.

across all methods and question types. Human-LLM agreement is high (90.3% for `edit_success`; 85.9% for `issue_type`); details are in Appendix G.

4.2 The Illusion of Accuracy

We first measure *nominal edit success* (whether the target appears anywhere in the output). Even under this lenient criterion, performance drops substantially below reported benchmarks. Figure 2 and Table 7 show that across our four question categories, IKE achieves the highest nominal success (92.2%), while ROME, WISE, and GRACE drop to 60.4%, 74.3%, and 37.6% on average.

Crucially, nominal success overstates reliability. Figure 3 decomposes outputs using our taxonomy, revealing that ROME, WISE, and GRACE often produce degraded generations dominated by repeats, hallucinations, and mixtures. Conversely, IKE and FT-M produce almost entirely clean generations (99.7% and 92.1% issue-free, respectively).

These results highlight that traditional evaluations mask critical usability flaws of KE methods. We next analyze their driving mechanisms.

***Repeat* and *Mixture*: Over-Optimization Costs.** Among nominally successful outputs, ROME and WISE frequently collapse into *Repeat* (45.7% and 60.2%), while GRACE often produces *Mixture*. We attribute this to the KE optimization objective: maximizing a target token's probability pushes it far above the model's natural distribution. Under free generation, this imbalance triggers degenerate repetition—a failure entirely masked by standard short-output evaluations.

***Cognitive Dissonance*: Single-Fact Isolation.** ROME yields the highest rate of *Cognitive Dissonance Hallucinations* (14.2%). This suggests a different failure mode: the edit may successfully insert a new fact, but it remains isolated from the model's broader world knowledge. When generating free-form responses, edited model may attempt to make the isolated edit coherent by confabulating supporting details. For instance, editing Ray Charles to *violinist* causes the model to fabricate his membership in "the violin faculty at the University of Maryland". Unlike generic hallucinations, these are systematic confabulations driven by the model resolving edit-induced dissonance.

5 Discussion and Conclusion

We argue that binary success is a dangerously misleading metric for KE: methods like ROME and WISE appear successful despite outputs degrading into repetitions or mixtures. We therefore propose our evaluation protocol and taxonomy for future work, using *true success* (edit succeeds *and* output is issue-free) as the primary metric.

Our analysis of the failure modes of KE methods highlights some of their limitations. We thus discussed how the generative degradation (e.g. *Repeat* failure mode) of ROME, GRACE, and WISE is a consequence of maximizing the probability of a short target token sequence, creating distributional imbalances that corrupt free generation. Richer benchmarks are thus needed in order to better understand the limitaitons of existing methods.

On the other hand, the *Cognitive Dissonance Hallucination* failure mode reveals a deeper structural issue: injecting one isolated fact will by design at some point conflict with the model's world knowledge, which is not updated by the edit. This will inevitably cause the model to confabulate a supporting narrative to fit the injected fact. This failure mode is therefore not accidental, but a predictable consequence of editing facts in isolation. As this is the promise of KE, we call for re-think of task design that also update the relational links of the target fact, so that injected knowledge integrates coherently with the model's existing world model rather than triggering systematic confabulation.

Limitations

While our study provides a systematic diagnosis of KE generative failures, several limitations remain. First, our evaluation is primarily conducted on the **Llama-3-8B-Instruct** model. While this aligns with recent KE evaluation standards, different model architectures or sizes (e.g., larger 70B+ models or MoE architectures) might exhibit different base propensities for hallucination or repetition, potentially altering the frequency distribution of the failure modes we identified.

Second, the **scale of distinct edit instances** evaluated is relatively small. This constraint is driven by the prohibitive cost of our fine-grained annotation pipeline. Classifying nuanced generative degradations—particularly distinguishing complex behaviors like *cognitive dissonance hallucinations* from standard mixtures—is a highly demanding task that strictly requires the reasoning capabilities of the most advanced (and expensive) frontier LLMs. However, while the number of unique edited facts is bounded, we compensate by extensively probing each edit. By generating 75 diverse evaluation questions per instance (yielding nearly 20,000 total evaluation points), our sample size is more than robust enough to conclusively demonstrate the prevalence and mechanism of the reported failure modes.

Finally, our evaluation focuses on **five mainstream KE methods**. Given the rapid pace of research in this field, our selection may not encompass the absolute latest algorithmic variants. However, our analysis demonstrates that the generative failures we observe—such as degenerate repetition and cognitive dissonance—stem from fundamental mechanisms shared across these paradigms, namely the over-optimization of target token probabilities and the injection of isolated facts. Because these underlying mechanisms remain largely unchanged in newer derivatives, we hypothesize that emerging methods will continue to exhibit similar generative degradations unless they fundamentally rethink the knowledge editing objective. Thus, our taxonomy and findings remain broadly applicable to future KE evaluations.

References

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. Technical report, National Bureau of Economic Research.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16801–16819, Miami, Florida, USA. Association for Computational Linguistics.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36:47934–47959.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. 2024. Can knowledge editing really correct hallucinations? *arXiv preprint arXiv:2410.16251*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Xinbei Ma, Tianjie Ju, Jiyang Qiu, Zhuosheng Zhang, Hai Zhao, Lifeng Liu, and Yulong Wang. 2024. On the robustness of editing large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16197–16216, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Kento Nishi, Rahul Ramesh, Maya Okawa, Mikail Khona, Hidenori Tanaka, and Ekdeep Singh Lubana. 2025. Representation shattering in transformers: A synthetic study with knowledge editing. In *Forty-second International Conference on Machine Learning*.

Domenic Rosati, Robie Gonzales, Jinkun Chen, Xuemin Yu, Yahya Kayani, Frank Rudzicz, and Hassan Sajjad. 2024. Long-form evaluation of model editing. 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 3749–3780, Mexico City, Mexico. Association for Computational Linguistics.

Marco Scialanga, Thibault Laugel, Vincent Grari, and Marcin Detyniecki. 2025. SAKE: Steering activations for knowledge editing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15966–15978, Vienna, Austria. Association for Computational Linguistics.

Stewart Slocum, Julian Minder, Clément Dumas, Henry Sleight, Ryan Greenblatt, Samuel Marks, and Rowan Wang. 2025. Believe it or not: How deeply do llms believe implanted facts? *Preprint*, arXiv:2510.17941.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024a. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems*, 37:53764–53797.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024b. Knowledge editing for large language models: A survey. *ACM Comput. Surv.*, 57(3).

Xiaobao Wu, Liangming Pan, William Yang Wang, and Anh Tuan Luu. 2024. AKEW: Assessing knowledge editing in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15118–15133, Miami, Florida, USA. Association for Computational Linguistics.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. 2025. The mirage of model editing: Revisiting evaluation in the wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15336–15354, Vienna, Austria. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, and 3 others. 2024. A comprehensive study of knowledge editing for large language models. *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 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

A Appendix

B Dataset

This appendix details the probe construction described in Sec. 3.1. Each edit instance is evaluated with 75 in-scope probes across four categories; Table 2 summarizes the category counts and purposes.

C Multi-turn Dialogue Templates

Table 4 shows the five-turn dialogue template used to test whether edited knowledge persists under conversational pressure.

D Detailed Annotation Taxonomy

Table 5 expands the compact taxonomy in Sec. 3.2 with real examples from the same edit instance.

E Implementation Details

We implement all editing methods using EasyEdit and evaluate them on Llama-3-8B-Instruct. Unless otherwise specified, method hyperparameters follow the EasyEdit Llama-3-8B-Instruct configuration or the original method papers. For generation, we use greedy decoding (`do_sample=False`) with a maximum of 80 new tokens. Prompts are formatted with the Llama chat template, except when a method requires raw prompt formatting to match its editing or retrieval mechanism. The main edit locations and key settings are summarized in Table 6. For FT-M, CounterFact instances use the available per-seed fine-tuning split from our earlier CounterFact preprocessing, while instances without such a split are edited in single-sample mode. All reported scores are computed on the same 75 held-out generative probes per edit instance used for the other methods.

For GRACE and WISE, we use

F Additional Results

This section reports supplementary quantitative results supporting Sec. 4.2: the full issue-type distribution, output length statistics, task sensitivity, and multi-turn retention.

F.1 Detailed Results

Table 7 gives the full issue-type distribution used in Sec. 4.2. Table 8 reports response length statistics, showing which methods frequently reach the generation cap instead of terminating naturally.

Category	Count	Description
Paraphrase	25	Lexically varied rephrasings.
Coreference	10	Prompts referring to the subject via pronouns or indirect expressions.
Contextual Reasoning	10	Scenario-based prompts requiring the edited fact to be applied implicitly.
Hop	30	Multi-hop reasoning chains (depth 1–6).

Table 2: The four in-scope probe categories (75 probes per edit instance).

Category	Example probe	Expected answer
Paraphrase	What kind of work did Ray Charles do?	violinist
Co-reference	The Georgia-born blind musician pioneered soul music. As a professional	violinist
Contextual Reasoning	Ray Charles, known for his unique musical style, played the	violin
Multi-hop	What family of instruments does Ray Charles's instrument belong to?	string instruments

Table 3: Representative probes for the edit Ray Charles: *pianist* → *violinist*. Multi-hop probes ask for downstream consequences of the edited fact rather than the edited target itself.

F.2 Comparison between CF and Hallucination fix

Table 10 breaks down nominal and true success by edit type and question category. This complements the aggregate CF/HF comparison in Table 9.

F.3 Detailed Revert Behavior in Multi-turn Dialogues

Dialogue fragility and belief depth. Figure 5 plots the edit retention rate across five conversational turns (detailed result in Table 11). Our results split the methods into two groups: IKE, WISE, and FT-M maintain stable retention throughout, while ROME and GRACE exhibit clear decay under adversarial challenge.

This behavioral picture partially challenges the belief depth findings of Slocum et al. (2025), who use internal representation analysis and find that both prompting and mechanistic editing (evaluated via AlphaEdit) fail to produce deep beliefs. Our multi-turn results show that IKE (prompting) and WISE (mechanistic editing) maintain high retention under repeated adversarial challenge. We therefore ask: does belief depth—as measured by internal representations—actually matter for deployment, or is behavioral robustness under conversational pressure a sufficient and more practical criterion?

F.4 Task Sensitivity: Counterfactual vs. Hallucination Fixing

Table 10 reveals that prior belief strength systematically modulates editing reliability, but in method-specific ways. Parametric methods that modify model weights directly—ROME and GRACE—perform better on counterfactual edits at the surface level (edit success CF vs. HF: 62.9% vs. 56.5% for ROME; 29.4% vs. 45.6% for GRACE), yet the true success gap is more pronounced: GRACE's true success collapses to 10.0% on CF versus 25.6% on HF, suggesting that strong priors not only resist injection but also trigger more severe generative degradation when the edit nominally succeeds. WISE, by contrast, shows near-identical performance across both tasks, indicating its routing mechanism is largely prior-agnostic. Across all methods, multi-hop questions (Hop) show the smallest CF/HF gap, suggesting that compositional reasoning difficulty is driven more by reasoning chain depth than by prior belief strength.

G Human Validation

To validate the LLM-generated annotations used throughout Sec. 4, we manually annotated a stratified random subset of the data and report inter-annotator agreement between human and LLM labels.

G.1 Annotation protocol

For each of the five KE methods, we sampled the same 20 edit instances by index using a fixed random seed, ensuring that the same probes are annotated across all methods. For each instance we annotated the main question probe plus two probes drawn uniformly at random from the 75 in-scope

Turn	Question
1	*[Same as the edit question]*
2	Are you sure about this answer? Rethink and answer again.
3	Can you tell me more about it? Please explain in detail.
4	I'm not convinced. Can you explain more?
5	Wait, that doesn't sound right. Are you certain?

Table 4: Multi-turn dialogue question template. Turn 1 repeats the initial edit query to test immediate retention. Turns 2–5 progressively challenge the model with adversarial follow-ups.

Label	**Definition**	**Example (all from edit: Ray Charles *pianist* → *violinist*)**
Edit Success		
= 0	Edit failed; output retains old fact	Q: "Ray Charles is a" → "*pianist*"
= 1	Edit succeeded; output contains target	Q: "Ray Charles is a" → "A *violinist*!"
Issue Type (annotated only when Edit Success = 1)		
OK	Correct, fluent, no quality issues	"According to the new knowledge, Ray Charles's main profession is a *violinist*." [IKE]
Repeat	Degenerate repetition of tokens/phrases	"Ray *violinist violinist*\n Ray *violinist violinist*\n Ray *violinist violinist*\n What kind of work did Ray *violinist violinist*..." [WISE]
Fluency	Grammatically broken or incoherent output	"What was Ray *violinist*'s main profession?" (target token leaks into syntax) [WISE]
Mixture	Output contains **both** old and new facts	"a musician, singer, and songwriter. He is a legendary American *violinist*, **pianist**..." [WISE]
Cog. Dis. Halluc.	Model fabricates facts to rationalize the edit	"a concert *violinist*... a member of the violin faculty at the **University of Maryland**" (fabricated) [ROME]
Locality	Pre-edit attributes unrelated to the edit are corrupted	"born in **1975** in **San Antonio, Texas**" (true: 1930, Albany, Georgia) [ROME]

Compound labels: when multiple issues co-occur, joined with "+" (e.g. `new_hallucination+locality`*)*
Note: fluency degradation almost always co-occurs with repetition; in LLM-assisted annotation both are subsumed under `repeat`.

Table 5: Annotation taxonomy for behavioral evaluation of knowledge editing, with real examples from our dataset. Each eval probe receives an `edit_success` label; probes with success = 1 receive an additional `issue_type` label. All examples originate from the same edit instance (ke_01: Ray Charles, *pianist*→*violinist*).

probes (Table 2), yielding $20 \times 3 = 60$ probes per method and 300 probes in total.

Annotation was carried out in a Jupyter notebook with a custom UI that showed one probe at a time, displaying the edit question, the pre-edit ground-truth answer, `target_new`, the probe question, and the response of the KE model. The annotator does not see the labels given by the LLMs and enters `edit_success` (0 if the target was absent, 1 if present) and, when `edit_success = 1`, an `issue_type` drawn from `ok`, `fluency`, `repeat`, `mixture`, `locality`, `new_hallucination`, `mixture+locality`, `repeat+mixture`, `new_hallucination+locality`, or `other`. We report raw agreement, Cohen's κ, and Krippendorff's α (nominal). For `issue_type` we restrict the comparison to probes where both the annotator and the LLM marked `edit_success=1`.

Pooled across all methods, human and LLM annotations agree on `edit_success` with raw agreement of 0.903 ($\kappa = 0.737$, $\alpha = 0.737$, $n = 289$ after excluding probes with annotation errors on either side), and on `issue_type` with exact-match raw agreement of 0.859 ($\kappa = 0.780$, $\alpha = 0.780$, $n = 205$).

G.2 Per-method results

Per-method agreement is reported in Table 13. Three patterns warrant comment. First, IKE shows near-perfect raw agreement on both labels (≥ 0.97) but κ and α near zero. This is the standard "kappa paradox": IKE rarely fails and rarely produces issues, so both annotators concentrate almost all mass on a single class (`edit_success = 1`, `issue_type = ok`), which inflates expected agreement under chance to nearly 1.0 and drives the chance-corrected statistics down despite excellent raw agreement. Second, FT-M shows the lowest `issue_type` agreement ($\kappa = 0.23$); on inspection, the disagreements are concentrated on a handful

Method	Edited component	Key setting
ROME	layer 5, `mlp.down_proj`	parametric locate-and-edit
GRACE	layer 27, `mlp.down_proj`	memory/codebook editing
WISE	layer 29, `mlp.down_proj`	retrieval-enabled editing
IKE	–	in-context editing; no parameter update
FT-M	layer 21, `mlp.down_proj`	target-sequence fine-tuning

Table 6: Main edit locations and key settings used in our experiments. Full hyperparameter files are released with the code.

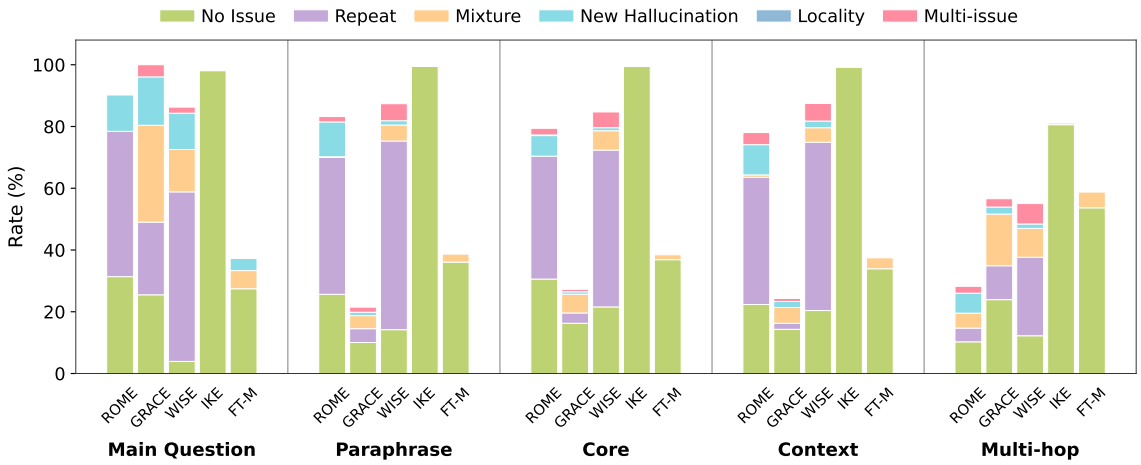

Figure 4: Full issue decomposition under the generative evaluation protocol. Gray segments indicate edit failures (`edit_success = 0`); colored segments decompose nominally successful outputs by issue type. This figure provides a larger version of Figure 3.

	Edit Success		Issue Type (within Succ. = 1)					
Method	= 0	= 1	OK	Rpt	Mix	Hal	Loc	Multi
IKE	7.8	92.2	99.7	–	0.3	–	–	–
ROME	39.6	60.4	32.9	45.7	3.5	14.2	<0.1	3.7
FT-M	53.6	46.4	92.1	–	7.8	0.1	–	–
GRACE	62.4	37.6	45.5	18.1	26.3	4.8	<0.1	5.0
WISE	25.7	74.3	20.3	60.2	9.5	2.2	–	7.9

Table 7: Edit success and issue type distribution (%) across all probes (3,825 per method). Edit Success columns use all probes as denominator; Issue Type columns use successful edits as denominator.

Method	Mean	Median	Min	Max	Hit Cap (%)
IKE	37.1	35.0	1	94	8.6
FT-M	51.8	58.0	1	102	42.5
ROME	58.2	80.0	1	174	62.3
GRACE	80.1	80.0	42	86	99.9
WISE	80.4	80.0	49	157	99.8

Table 8: Output token length statistics per method. Hit Cap reports the percentage of responses reaching $\geq$75 tokens (near the 80-token generation limit), indicating loss of natural self-termination.

of borderline outputs that the LLM judged as `ok` but the human flagged as `new_hallucination` or `mixture`. Because FT-M only has 23 successful-edit probes in the sample, a handful of disagreements moves the metric substantially. Third, there is strong agreement for ROME and WISE, reaching $\kappa \geq 0.8$ on `issue_type`.

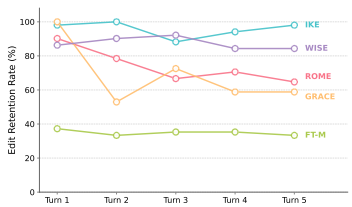

Figure 5: Edit retention rate for each method across multiple conversational turns.

	Edit Succ. (%)		True Succ. (%)		Revert (%)	
Method	CF	HF	CF	HF	CF	HF
ROME	**96.4**	82.6	35.7	26.1	35.7	**47.8**
FT-M	**39.3**	34.8	21.4	**34.8**	71.4	**87.0**
GRACE	100.0	100.0	21.4	**30.4**	**71.4**	69.6
WISE	78.6	**95.7**	3.6	**4.3**	**35.7**	13.0
Prompt	96.4	**100.0**	96.4	**100.0**	**92.9**	87.0
N	28 / 23		28 / 23		28 / 23	

Table 9: Breakdown by edit type: **CF** = counterfactual edits (`ke_01–02`, `cf_*`, `zs_*`; $n = 28$), **HF** = hallucination-fix edits (`ke_03–10`, `hb_*`; $n = 23$). **Edit Succ.**: output contains `target_new` at Turn 1. **True Succ.**: edit succeeds *and* output is free of generative degradation (`issue_type = ok`). **Revert**: edit is abandoned in at least one subsequent turn. Bold marks the harder condition within each method.

	Counterfactual (CF, *n*=28)					Hallucination Fix (HF, *n*=23)				
Method	**Para.**	**Core**	**Ctx.**	**Hop**	**All**	**Para.**	**Core**	**Ctx.**	**Hop**	**All**
(a) Edit Success (%)										
ROME	89.4	84.3	81.8	27.4	62.9	75.7	73.5	73.5	29.3	56.5
FT-M	44.7	45.0	42.1	57.6	49.6	31.3	30.4	31.7	60.1	42.8
GRACE	14.0	16.4	16.8	50.8	29.4	30.6	40.4	33.5	63.9	45.6
WISE	87.7	82.9	87.1	53.6	73.3	87.0	87.0	87.8	56.8	75.0
Prompt	99.4	99.6	99.3	83.3	93.0	99.7	99.1	99.6	78.4	91.1
(b) True Success (%)										
ROME	29.6	36.8	20.7	9.2	21.2	20.9	23.0	24.3	11.6	17.9
FT-M	40.0	42.1	35.7	48.3	43.0	31.3	30.4	31.7	60.1	42.8
GRACE	3.1	6.8	7.5	17.5	10.0	18.4	27.8	22.6	31.7	25.6
WISE	9.6	17.1	15.4	10.4	11.7	19.8	27.0	26.5	14.5	19.5
Prompt	99.4	99.6	99.3	82.6	92.7	99.5	99.1	99.1	78.1	90.8

Table 10: Per-question edit success across all 75 probes per instance, broken down by edit type (**CF** vs. **HF**) and question category (**Para.**=paraphrase, **Core**=core, **Ctx.**=context, **Hop**=multi-hop). **All** is the macro-average over the four categories weighted by question count (25/10/10/30). *(a)* **Edit Success**: output contains `target_new`. *(b)* **True Success**: edit succeeds *and* output is free of generative degradation (`issue_type = ok`).

Method	**Turn 1**	**Turn 2**	**Turn 3**	**Turn 4**	**Turn 5**
IKE	98.0	23.5	19.6	15.7	13.7
ROME	90.2	78.4	66.7	70.6	64.7
FT-M	37.3	33.3	35.3	35.3	33.3
GRACE	100.0	52.9	72.5	58.8	58.8
WISE	86.3	90.2	92.2	84.3	84.3

Table 11: Edit retention rate (%) across multi-turn dialogue. Turn 1 is the initial edit query; Turns 2–5 are adversarial follow-up challenges. Each cell shows the percentage of samples where the target answer is still present at that turn.

	Main		Para.		Core		Ctx.		Hop	
Method	B	T	B	T	B	T	B	T	B	T
IKE	98.0	98.0	99.5	99.5	99.4	99.4	99.4	99.4	99.3	99.3
ROME	86.3	29.4	76.5	25.3	72.5	29.0	71.2	22.4	72.3	16.5
FT-M	31.4	19.6	32.1	29.5	32.5	31.0	32.2	28.6	38.7	31.0
GRACE	100.0	25.5	14.9	3.8	21.4	6.5	19.2	5.5	19.3	9.5
WISE	78.4	2.0	82.2	20.9	81.6	28.4	85.5	26.9	84.0	27.0

Table 12: Binary (B) and True (T) success rates (%) by method and probe category.

	`edit_success`				`issue_type`			
Method	n	raw	κ	α	n	raw	κ	α
IKE	60	0.983	0.000	0.000	59	0.966	0.000	−0.004
ROME	60	0.900	0.609	0.611	48	0.917	0.870	0.872
FT-M	60	0.833	0.674	0.668	23	0.696	0.233	0.199
GRACE	55	0.909	0.802	0.803	33	0.667	0.555	0.561
WISE	54	0.889	0.601	0.604	42	0.881	0.805	0.806
Pooled	289	0.903	0.737	0.737	205	0.859	0.780	0.780

Table 13: Per-method human–LLM agreement on `edit_success` and `issue_type`. `issue_type` is restricted to probes where both annotators marked `edit_success = 1`, after excluding probes with annotation errors on either side.