Beyond Memorization: A Rigorous Evaluation Framework for Medical Knowledge Editing

Anonymous ACL submission

Abstract

Recently, knowledge editing (KE) has emerged as a promising approach to update specific facts in Large Language Models (LLMs) without the need for full retraining. Despite the effectiveness in general-domain benchmarks, their applicability in complex medical domain, remains largely unexplored. Medical knowledge editing is particularly challenging, as it requires LLMs to internalize the knowledge and generalize to unseen scenarios for effective and interpretable decision-making. In this work, we propose a novel framework called `MedEditBench` to rigorously evaluate the effectiveness of existing KE methods in the medical domain. In `MedEditBench`, we introduce a new medical knowledge editing benchmark as well as three different knowledge editing paradigms, which are designed to assess the impact of different knowledge sources for editing. Our findings indicate that current knowledge extraction (KE) methods result in only superficial memorization of the injected information, failing to generalize to new scenarios. To overcome this limitation, we present self-generated rationale editing (SGR-Edit), which utilizes model-derived rationales as target knowledge for editing, thereby uncovering the underlying reasoning process and demonstrating significant improvements over existing approaches. Additionally, we offer deeper insights into medical knowledge editing, including the localization of medical knowledge in LLMs and the impact of sequential editing on evolving knowledge. This could provide practical guidance for implementing KE methods in real-world medical applications[1].

1 Introduction

Large language models (LLMs) encapsulate extensive knowledge during training on large-scale corpora (Petroni et al., 2019; Allen-Zhu and Li, 2023; Dai et al., 2021). Nonetheless, their knowledge remains static after training, resulting in factual inconsistencies and hallucinations in tasks that require up-to-date information or domain-specific expertise beyond their pre-trained knowledge (Hu et al., 2023; Huang et al., 2025). Knowledge editing (KE) has emerged as a promising approach to update specific knowledge in LLMs without full retraining (Zhang et al., 2024; Wang et al., 2024c), which provides an efficient and effective solution to constantly adapt LLMs to the ever-evolving world of knowledge (Meng et al., 2022; Meng et al.; Fang et al., 2024).

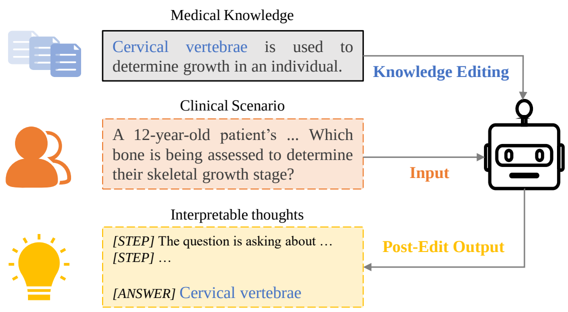

Figure 1: Illustration of medical knowledge editing.

Despite rapid advancements, the effectiveness of current KE methods has not been thoroughly assessed in realistic domain-specific scenarios. Most existing benchmarks only evaluate on general domains, such as WikiData (Meng et al., 2022) and counterfactual datasets (Levy et al., 2017), which do not reflect the complexity and diversity of real-world applications, particularly in specialized domains like medicine.

Medical knowledge editing is particularly challenging due to its intricacy and specificity, which requires LLMs to not only memorize updated medical knowledge (Singhal et al., 2023) but also comprehend underlying medical concepts and generalize to new, unseen scenarios for effective reasoning (Li et al., 2023a), as shown in Figure 1. Moreover,

[1]Codes and the MedEditBench datasets are released at `https://anonymous.4open.science/r/F0F5/`

the high stakes of medical applications necessitate that LLMs explain their reasoning to improve the trustfulness and interoperability (Holzinger et al., 2017; Tonekaboni et al., 2019). Therefore, it is crucial to establish a more rigorous evaluation framework that accurately analyzes the performance of existing KE methods on medical knowledge to ensure their reliability and applicability in real-world applications (Zhou et al., 2024; Zhang et al., 2025).

In this work, we establish `MedEditBench` for rigorous medical knowledge editing evaluation. First, we create a high-quality medical knowledge editing benchmark from two real-world medical question-answering datasets (Kim et al., 2024; Pal et al., 2022). To evaluate the generalizability of KE methods, we extend the benchmark with *scenario-based questions* that require LLMs to apply the injected knowledge to address new clinical scenarios, as well as questions that require LLMs to preserve their previously factual knowledge. We also introduce three novel metrics: *efficacy*, *generalization*, and *retention* to measure the performance of KE methods in updating, generalizing, and preserving medical knowledge, respectively.

Our initial findings reveal a substantial gap between current KE methods and the demands of real-world medical applications. We observe that all existing knowledge editing methods consistently underperform in complex medical settings; even the strongest method, AlphaEdit (Fang et al., 2024), achieves an average performance of only 53.9%, comprising 43.9% efficacy, 31.2% generalization, and 86.7% retention. We attribute this pitfall to the existing editing paradigm that relies on a short *ground-truth answer* (GTA) as the knowledge target, which causes LLMs to memorize surface-level facts rather than understand the underlying medical rationale, thereby impairing generalization. Although using the human-curated *reference* (RE) as the target knowledge can enhance the knowledge understanding by providing more context, the improvements are still limited.

To address this limitation, we introduce a novel editing paradigm called *self-generated rationale editing* (SGR-Edit), which simulates the chain-of-thought reasoning (Wei et al., 2022), by prompting LLMs to generate explanatory rationales to support the new target answer based on provided reference texts. These self-generated rationales then serve as the target knowledge and can be seamlessly integrated with arbitrary knowledge editing methods. SGR-Edit reveals the internal reasoning process of LLMs, allowing them to better internalize new medical knowledge and generalize to unfamiliar scenarios. Experimental results demonstrate that Self-Generated Rationale Editing (SGR-Edit) significantly enhances the performance of existing editing methods relative to conventional knowledge targets: SGR-Edit yields improvements of +4.1 percentage points for AlphaEdit and +9.5 percentage points for LoRA, demonstrating its effectiveness in enhancing medical knowledge editing. The contribution of this work is as follows.

- We propose a framework (`MedEditBench`) to rigorously assess existing knowledge editing methods in real-world medical scenarios and evaluate the applicability of various editing paradigms with different knowledge sources.
- We conduct a comprehensive evaluation and find that existing editing methods are insufficient for medical knowledge editing because the existing paradigm relies on short answers, promoting surface-level memorization.
- We propose a novel editing paradigm, SGR-Edit, which uses self-generated rationales as the target knowledge for editing, noticeably improving the performance of all existing editing methods.
- We introduce deeper insights into medical knowledge editing, including the localization of medical knowledge in LLMs and the impact of sequential editing on evolving knowledge.

2 Related Work

2.1 Knowledge Editing Methods

Knowledge editing methods can be roughly grouped into three categories:

Fine-Tuning–Based Editing These approaches update a large number of model parameters to inject new knowledge, typically via constrained or parameter-efficient training. Such approaches include FT+L (Zhu et al., 2020), FT-M (Zhang et al., 2024), and LoRA (Hu et al., 2022).

Parameter-Modifying Editing Compared to full model fine-tuning, these methods focus on specific parameters to minimize interference with unrelated knowledge. The *Meta-Learning* strategy trains a hypernetwork to predict gradient updates for knowledge insertion, as seen in methods (De Cao et al., 2021; Mitchell et al.; Tan et al.). The *Locate-then-Edit* approach identifies and rewrites factual weights in specific layers. Representative models of this approach include ROME (Meng et al., 2022) and MEMIT (Meng et al.), along with sequential editing variants like PRUNE (Ma et al., 2024a), AlphaEdit (Fang et al., 2024), and AnyEdit (Jiang et al., 2025), which support continuous knowledge updates.

Parameter-Preserving Editing These techniques maintain the base model's parameters by either augmenting it with external modules or retrieving relevant information during inference. *Extension-Based* methods, such as GRACE (Hartvigsen et al., 2023), add adapters or incorporate side modules (Mitchell et al., 2022; Wang et al., 2024b) to store new facts. *Retrieval-Based* methods (Zheng et al., 2023; Song et al., 2024; Shi et al., 2024; Chen et al., 2024b) retrieve relevant facts to include in the prompt as context and generate updated outputs.

2.2 Knowledge Editing Benchmarks

Early evaluations of knowledge editing methods primarily focused on general-domain benchmarks, such as WikiData (Meng et al., 2022) and counterfactual datasets (Levy et al., 2017). They try to update LLMs with new knowledge that contradicts the common knowledge. Recently, studies (Ju et al., 2023; Li et al., 2023b; Pinter and Elhadad, 2023; Chen et al., 2024a) question the real-world applicability of existing knowledge editing methods and propose new benchmarks for a fair evaluations.

Cohen et al. (2024) measures "ripple effects" on related facts and reveals that KEs often fail to propagate consistent changes beyond the target triple. Huang et al. (2024) reports that prior benchmarks do not strictly confirm LLMs having hallucinated answers to the questions before conducting editing, which masks their true editing performance. Ma et al. (2024b) investigates editing consistency under prompt rephrasing and realistic communicative contexts, finding that current KEs exhibit lower generalization and that popular facts are hardest to edit. Lin et al. (2024) examines sequential editing and shows that the KEs experience editing performance drops after continual edits. Yang et al. (2025) critiques common evaluation practices and reveals that KEs catastrophically fail on realistic QA tasks Differ from prior evaluations focused on artificial and simplified general-domain settings (Ma et al., 2024b; Lin et al., 2024), we introduce the first framework for medical knowledge editing evaluation, to rigorously measure whether updated medical knowledge could be effectively adopted in new clinical scenarios without degrading previous factual knowledge.

3 Task Formulation

The goal of knowledge editing is to modify specific knowledge k in LLMs without retraining the entire model (Zhang et al., 2024), thereby improving performance on tasks related to that knowledge, represented by a set of queries and answers: $\mathcal{Q}^k = \{(q_i, a_i)\}$. Let θ denote the original model. Given a query q and the target knowledge k, the knowledge editing method F can be expressed as:

$$\theta' = F(\theta, q, k), \tag{1}$$

where θ' represents the edited model expected to provide the desired answer $a = \theta'(q)$ for the knowledge-related query $q \in \mathcal{Q}^k$.

4 Evaluation Framework

In this section, we describe the proposed medical knowledge editing evaluation framework (`MedEditBench`), as illustrated in Figure 2.

4.1 Medical Editing Benchmark Construction

Due to the lack of existing benchmarks for medical knowledge editing, we construct two datasets: $\text{MedExQA}_{\text{edit}}$ and $\text{MedMCQA}_{\text{edit}}$ by extending two real-world medical QA datasets (Kim et al., 2024; Pal et al., 2022). Each QA dataset is originally provided with a set of questions q, answers a, and relevant references c in the format of human-written explanations or textbook references.

To ensure reliable evaluations of a target LLM, we filter out medical questions that the LLM can answer correctly before editing. In this way, our assessments could reflect genuine improvements from knowledge editing methods. We prompt the LLM in a zero-shot manner to predict answers for each candidate question and retain only those questions it fails to answer correctly as $\mathcal{Q}_{\text{ori}}$. Thus, the accuracy on $\mathcal{Q}_{\text{ori}}$ indicates the **efficacy** of editing methods in injecting new knowledge.

To assess the **generalization** of edited knowledge, we create a novel *scenario-based question* set $\mathcal{Q}_{\text{gen}}$ that evaluates whether the edited LLM can apply the injected knowledge to address clinical questions in unseen scenarios. Unlike existing

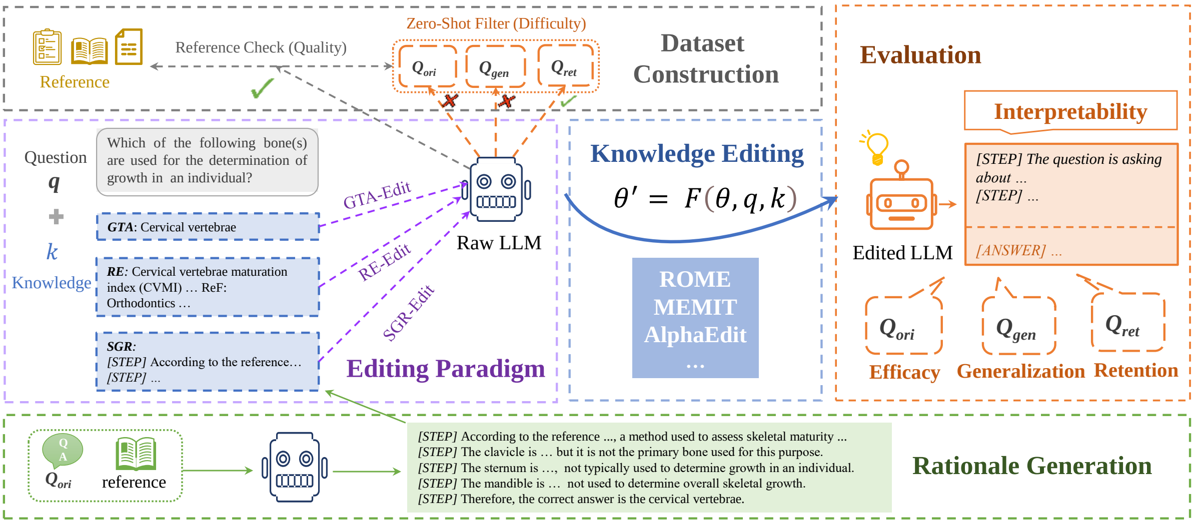

Figure 2: Overview of proposed medical knowledge editing evaluation framework (`MedEditBench`). Datasets are first constructed for medical knowledge editing (gray rectangle). Then, we propose three editing paradigms (GTA-Edit, RE-Edit, and SGR-Edit) to evaluate the effectiveness of different target knowledge for medical knowledge editing (purple rectangle). The rationale in SGR-Edit is generated by the LLM itself, given a QA and a reference (green rectangle). These evaluations are based on the existing knowledge editing methods (blue rectangle). Finally, we evaluate the edited model's post-edit accuracy (i.e., *efficacy*, *generalization*, and *retention*) and assess the interpretability of the generated rationales for the final answers (orange rectangle).

benchmarks that rely on simple question paraphrasing (Huang et al., 2024), we (see Figure 7 for the complete prompts).

Last, we construct $\mathcal{Q}_{\text{ret}}$, which comprises those questions the LLM answered correctly before editing, to evaluate the **retention** of knowledge. This set is designed to ensure that the model retains its original knowledge after editing. The examples of $\mathcal{Q}_{\text{ori}}$, $\mathcal{Q}_{\text{gen}}$, and $\mathcal{Q}_{\text{ret}}$ are shown in Figure 9. Unlike existing benchmarks that rely on simple question paraphrasing (Huang et al., 2024), we generate $\mathcal{Q}_{\text{gen}}$ by extending the core facts of each original question into novel clinical scenarios (see Figure 7 for the complete prompt).

Data Principles As shown in the top gray rectangle of Figure 2, our benchmarks are built following two key data principles:

- **Quality.** We filter each QA pair by verifying whether the provided knowledge logically entails the ground-truth answer to ensure the question is knowledge relevant. We simply prompt the LLM with (c, q) and retain the sample only if the output answer equals a.

- **Difficulty.** To strictly evaluate the efficacy and generalization of knowledge editing, we ensure the pre-edit performance of the LLM is 0% on $\mathcal{Q}_{\text{ori}}$ and $\mathcal{Q}_{\text{gen}}$, while 100% on $\mathcal{Q}_{\text{ret}}$ to ensure the retention of original knowledge.

The detailed data construction process is described in Appendix A.

4.2 Knowledge-driven Editing Paradigms

In `MedEditBench`, we propose three *editing paradigms* to evaluate the effectiveness of different target knowledge for medical knowledge editing.

Ground-Truth Answer Editing (GTA-Edit) is the prevailing paradigm that typically uses the final answer a as the target knowledge k for editing (Meng et al., 2022; Meng et al.). As shown in Figure 2, GTA-Edit directly injects the answer (e.g., "Cervical vertebrae") into LLMs. Although this approach is straightforward, it may lead to superficial memorization of the answer without a deep understanding of the underlying medical rationale.

Reference Editing (RE-Edit) takes supporting reference c extracted from textbooks or academic literature as the target knowledge k for editing. While the reference text is not directly used as the answer, it provides a more comprehensive context for the model to understand the underlying medical knowledge.

Self-generated Rationale Editing (SGR-Edit) is proposed to further enhance the knowledge internalization by leveraging the model's own reasoning.

As shown in the yellow part of Figure 2, we first prompt the LLM to generate a chain-of-thought rationale for the question-answer pair (q, a) using the reference text c as context. This generated rationale clearly outlines the reasoning process that leads to the answer, serving as the target knowledge k for editing. Thus, SGR-Edit not only incorporates the answer but also provides a detailed explanation of how the model arrives at it, enabling LLMs to deliver interpretable and explainable reasoning that supports high-stakes medical decisions. The detailed prompts for generating rationales are provided in Appendix C.3. Examples of the GTA, RE, and SGR are presented in Table 3.

4.3 Evaluation Metrics

As shown in Figure 2, the knowledge editing is conducted with questions $q \in \mathcal{Q}_{\text{ori}}$ and their corresponding knowledge $\{k\}$ defined in each editing paradigm via Equation (1) to inject knowledge into LLMs. The edited model θ' is then evaluated on three test sets, i.e., $q \in \mathcal{Q}_{\text{ori}} \cup \mathcal{Q}_{\text{gen}} \cup \mathcal{Q}_{\text{ret}}$ to measure the *efficacy*, *generalization*, and *retention* of each editing method with the accuracy of the model's predicted answer.

To further evaluate the *interpretability* of the edited model, we also instruct the model to provide a rationale for its final answer, which is compared with the ground-truth interpretation using ROUGE-L (Lin, 2004) and BLEU scores (Papineni et al., 2002). Detailed calculations are provided in Appendix C.4.

4.4 Editing Method Selection

We choose six representative editing methods from three categories: *Fine-Tuning* methods: LoRA (Hu et al., 2022), *Parameter-Modifying* methods: ROME (Meng et al., 2022), MEMIT (Meng et al.), and *Parameter-Preserving* methods: GRACE (Hartvigsen et al., 2023), AnyEdit (Jiang et al., 2025), AlphaEdit (Fang et al., 2024). The AnyEdit and AlphaEdit methods are state-of-the-art editing methods that effectively support long-form and continual knowledge editing. We exclude retrieval-based editing methods (Song et al., 2024; Shi et al., 2024; Chen et al., 2024b; Zheng et al., 2023) since they do not directly update knowledge within the model parameters. Details of the editing methods are provided in Appendix B.

4.5 Experimental Setup

We evaluate all editing methods on two LLMs: LLaMA-3.1-8B-Instruct and LLaMA-3.2-3B-Instruct[2]. Editing pipelines are implemented based on the EasyEdit framework (Wang et al., 2023). Detailed experimental settings are provided in Appendix C. To ensure interpretable reasoning, all post-edit predictions follow a two-step format: a rationale followed by the final answer. This approach encourages the model to conduct deep reasoning, resulting in both improved prediction accuracy and enhanced interpretability. The output format is illustrated in Appendix D.4.

5 Main Experiments

In this section, we evaluate the effectiveness of editing methods on medical knowledge by answering the following research questions:
RQ1: How do current model editing methods perform in the medical domain?
RQ2: How do different editing paradigms impact the effectiveness of medical knowledge editing?
RQ3: How does the medical knowledge stored in LLMs?
RQ4: How does the knowledge in LLMs evolve with sequential editing?

5.1 Evaluating Editing Methods in Medical Domain (RQ1)

Existing editing methods follow the ground-truth answer editing (GTA-Edit) paradigm, where the model is updated by injecting the answer a as the target knowledge (Meng et al., 2022; Meng et al.). We follow this paradigm to update LLMs with the ground-truth answer a for each question $q \in \mathcal{Q}_{\text{ori}}$. Then, the model's performance is evaluated on the three test sets $\mathcal{Q}_{\text{ori}} \cup \mathcal{Q}_{\text{gen}} \cup \mathcal{Q}_{\text{ret}}$, and the question answering accuracy is reported in Table 1. From the results, we observe that:

RQ1-F1: No existing editing method is effective enough for medical settings. The *efficacy* of nearly all methods is below 50%, except for AlphaEdit and LoRA on LLaMA-8B, which achieve only 53.9% and 53% on $\text{MedMCQA}_{\text{edit}}$, respectively. This sharply contrasts with previous reports of over 90% accuracy in general domain benchmarks (Meng et al., 2022; Fang et al., 2024), highlighting a significant gap when applying knowledge editing methods to the intricate medical domain.

[2] https://llama.meta.com/lama3/

Method	Metric	Pre-Edit	LLaMA-8B		LLaMA-3B	
			MedExQA$_{edit}$	*MedMCQA$_{edit}$*	*MedExQA$_{edit}$*	*MedMCQA$_{edit}$*
LoRA	Eff.	0	43.5	46.6	36.7	14.4
	Gen.	0	41.3	41.6	43.3	33.7
	Ret.	100	63.0	70.8	63.3	52.9
	avg.	\	49.3	53.0	**47.8**	33.7
ROME	Eff.	0	37.0	32.7	23.3	25.7
	Gen.	0	43.5	29.6	26.7	25.1
	Ret.	100	63.0	61.6	56.7	56.1
	avg.	\	47.8	41.3	35.6	35.7
MEMIT	Eff.	0	39.1	28.3	28.3	16.0
	Gen.	0	50.0	25.2	31.7	25.1
	Ret.	100	50.0	64.8	63.3	54.5
	avg.	\	46.4	39.4	41.1	31.9
GRACE	Eff.	0	34.8	36.0	33.3	26.7
	Gen.	0	21.7	29.2	10.0	7.0
	Ret.	100	76.1	80.7	88.3	93.0
	avg.	\	44.2	48.7	43.9	42.2
AnyEdit	Eff.	0	34.8	36.6	35.0	23.0
	Gen.	0	28.3	25.5	28.3	23.0
	Ret.	100	78.3	78.9	71.7	78.1
	avg.	\	47.1	47.0	45.0	41.4
AlphaEdit	Eff.	0	47.8	43.9	35.0	32.6
	Gen.	0	32.6	31.2	31.7	27.8
	Ret.	100	76.1	86.7	73.3	77.5
	avg.	\	**52.2**	**53.9**	46.7	**46.0**

Table 1: Main results on medical knowledge editing with single editing (Accuracy %). For avg. scores per column: **bold** is the best, underline is the second best.

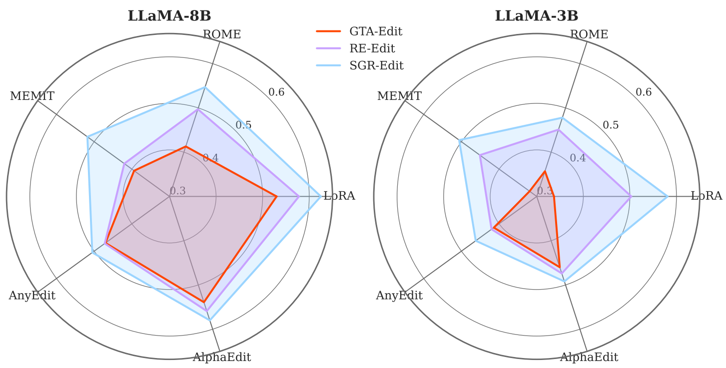

Figure 3: Medical knowledge editing with various editing paradigms.

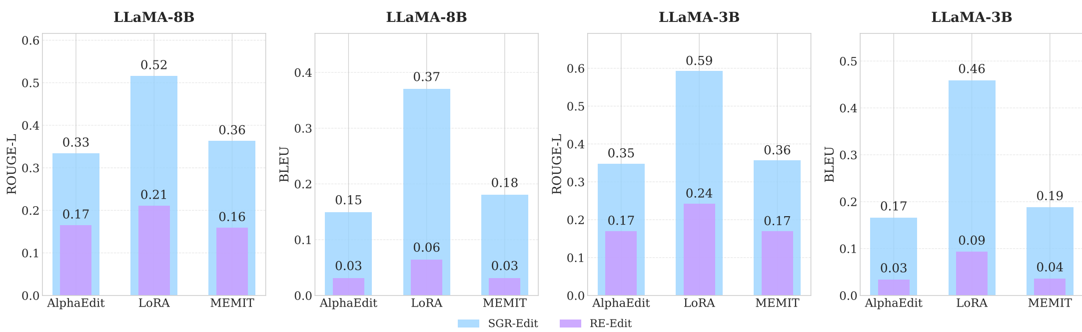

Figure 4: Interpretability comparison for SGR-Edit and RE-Edit.

RQ1-F2: Existing editing methods struggle to generalize updated medical knowledge and often compromise existing knowledge. This finding is reflected in the low *generalization* scores and noticeable drops in *retention*. For example, on MedMCQA$_{edit}$, LoRA achieves only 41.6% (8B) and 33.7% (3B) generalization, while retaining 70.8% and 52.9% of previously correct answers. Despite its strong retention, GRACE shows limited generalization ability, making it less suitable for adapting to new clinical contexts. These results suggest that **existing editing paradigm (i.e., GTA-Edit) often leads to surface-level memorization rather than meaningful internalization of medi-**

cal knowledge. Detailed case analyses are can be found in Appendix E.

5.2 Evaluation of Editing Paradigms (RQ2)

This section investigates how different editing paradigms (i.e., GTA, RE, and SGR), affect the performance of existing editing methods on MedMCQA$_{\text{edit}}$. Our key findings are as follows:

RQ2-F1: SGR-Edit yields the highest editing performance. As shown in Figure 3, SGR-Edit achieves the best performance for the five selected editing methods in both LLaMA-8B and LLaMA-3B, demonstrating that the reasoning knowledge drives is more effective for medical knowledge editing. On LLaMA-8B, AlphaEdit's average score increases from 53.9% (GTA-Edit) to 55.9% (RE-Edit, +2.0%) and 58.0% (SGR-Edit, +4.1%). LoRA exhibits a similar enhancement, rising from 53.0% (GTA-Edit) to 57.8% (RE-Edit, +4.8%) and 62.5% (SGR-Edit, +9.5%). Importantly, SGR-Edit achieves these gains with only one additional reasoning-generation step, making it both highly effective and readily deployable in practice. Full results and further analysis are provided in Appendix D.1.

RQ2-F2: SGR-Edit enables better reasoning interpretation. We report the rational interpretability of SGR-Edit and RE-Edit in Figure 4. The results indicate that SGR-Edit provides a more comprehensive understanding of the underlying medical knowledge enabling better reasoning interpretation than RE-Edit. Detailed results and analyses are provided in Appendix D.2.

5.3 Localization of Medical Knowledge in LLMs (RQ3)

In this section, we investigate the layer-wise storage of medical knowledge in LLMs. Following Meng et al. (2022), we edit knowledge across four disjoint layer ranges and check the storage of certain knowledge. For LLaMA-8B, we select layers 4–8, 11–15, 18–22, and 25–29; for the smaller LLaMA-3B, we use layers 4–8, 10–14, 16–20, and 22–26. We sample 100 QA pairs from MedMCQA$_{\text{edit}}$ and group target knowledge by token length to assess the impact of knowledge granularity. GTA targets are the shortest answers (<10 tokens), RE explanations span 50–100 or 100–150 tokens, and SGR rationales range from 150–200 to 200–250 tokens. Based on Figure 5, we draw the following findings:

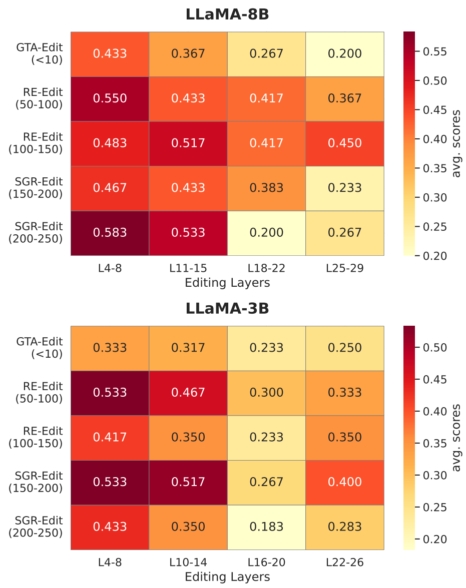

Figure 5: Layer-Wise editing performance

RQ3-F1: Medical knowledge is primarily stored in middle layers. Unlike previous findings in general-domain settings, which suggest that factual associations are primarily stored in middle layers (Meng et al., 2022) and that deeper layers offer greater editing stability (Lin et al., 2024), we find that medical knowledge in LLaMA models is most effectively edited in the shallower layers.

For both LLaMA-8B and 3B, editing operations targeting shallower layers 4-8 consistently yield the highest effectiveness. When editing on LLaMA-8B, average scores in layers 4-8 range from 0.433 to 0.583, and for LLaMA-3B, they span from 0.333 to 0.533, consistently outperforming edits in deeper layers.

RQ3-F2: Across all editing paradigms, the most effective edits are consistently achieved in the shallower layers. For all target knowledge (i.e., GTA, RE, SGR), optimal editing performance is observed almost exclusively in layers 4-8. The results also indicate that while different target knowledge does not affect the optimal editing location, it does influence the magnitude of effectiveness. For example, SGR-Edit (200-250) in layers 4-8 achieves the highest score of 0.583 for LLaMA-8B, with RE-Edit (50-100) demonstrating strong performance at 0.550 within these same layers.

5.4 Reliability of Sequential Medical Edits (RQ4)

In this section, we study the reliability of sequential edits to medical knowledge in LLMs, where

Model	#Edits	Overall	ΔOverall	Health	ΔHealth	Non-Health	ΔNon-Health
LLaMA-8B	0 (Raw)	67.9	–	70.7	–	68.0	–
	50	66.4	–1.5	68.5	–2.2	66.6	–1.5
	100	63.2	–4.6	61.1	–9.6	64.7	–3.3
LLaMA-3B	0 (Raw)	60.7	–	63.7	–	60.5	–
	50	59.1	–1.6	60.7	–3.0	59.1	–1.3
	100	56.8	–3.9	54.4	–9.3	57.4	–3.1

Table 2: Sequential editing performance and absolute drops Δ on MMLU (Accuracy %).

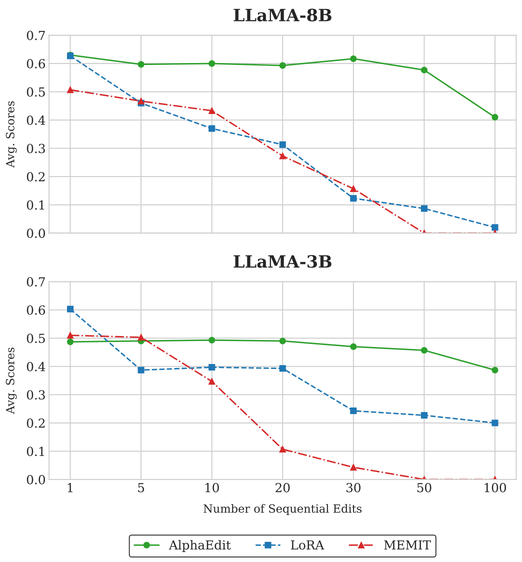

Figure 6: Performance over sequential medical editing.

the knowledge is sequentially updated through multiple edits. We assess three representative sequential editing methods: AlphaEdit, LoRA, and MEMIT, through up to 100 sequential edits using samples from MedMCQA$_{\text{edit}}$. We track average editing success rates at edit steps {1, 5, 10, 20, 30, 50, 100}. To better quantify the side effects of medical knowledge editing on the model's retention of other information, we analyze performance drops in both health and non-health domains of the MMLU benchmark (Hendrycks et al., 2020) after 50 and 100 sequential edits using AlphaEdit.

RQ4-F1: AlphaEdit supports stable and effective sequential editing far better than other methods. Figure 6 shows that on LLaMA-8B, AlphaEdit maintains an average editing score above 60% through 30 sequential edits, declining moderately to 58% at 50 edits and 41% at 100 edits. In contrast, LoRA drops sharply from 63% after the first edit to just 9% at 50 edits and collapses to 2% at 100 edits. MEMIT falls to near zero by the 50th edit. A comparable pattern is observed on LLaMA-3B: AlphaEdit sustains 46% average performance after 50 edits, while LoRA—despite strong single-edit results, it fails to withstand continual editing, and MEMIT degrades rapidly under sequential editing.

RQ4-F2: Sequential medical edits deteriorate the knowledge in other domains. As shown in Table 2, after 100 sequential edits on LLaMA-8B, health-domain accuracy on MMLU declines more sharply (e.g., $\Delta = -9.6\%$ on LLaMA-8B, $\Delta = -9.3\%$ on LLaMA-3B), which is expected given the exclusively medical nature of the edit data. However, non-health accuracy falls from 68.0% to 64.7% ($\Delta = -3.3\%$), and on LLaMA-3B from 60.5% to 57.4% ($\Delta = -3.1\%$), suggesting that continual updates to medical knowledge can lead to the erosion of out-of-domain knowledge.

6 Conclusion

In this paper, we introduce **MedEditBench**, a novel evaluation framework to rigorously evaluate existing knowledge editing methods in the medical domain. We present a comprehensive benchmark that includes a diverse metrics, along with various knowledge editing paradigms to examine the impact of different knowledge sources on editing. Our experimental results reveal that existing editing paradigms tend to induce superficial memorization rather than foster genuine understanding of the underlying rationale by updating with a simple final answer. To overcome this limitation, we propose the self-generated rationale editing (SGR-Edit), a novel editing paradigm where the model first generates an evidence-grounded rationale as the editing knowledge target. This approach reveals the underlying reasoning process, enabling deeper internalization of new medical knowledge. Additionally, our analysis uncovers that medical knowledge is usually stored in shallower LLM layers, and when sequential editing on medical knowledge, the LLM typically suffers from degradation of broader knowledge. This finding provides practical guidance for medical applications and further research in this area.

Limitations and Future Work

We acknowledge the limitations of this study:

- **Evaluation Coverage:** Due to limited GPU resources, our main experiments focus on six representative editing methods applied to two widely used LLaMA models. To broaden coverage, we additionally report single-edit results on Qwen2.5-7B in Appendix D.3.
- **SGR-Edit Overhead:** SGR-Edit requires only a single LLM and no external modules, leveraging evidence-based rationale generation without additional infrastructure. However, as the length of generated rationales grows, so does GPU memory consumption, which impedes large-batch editing experiments. Future work should investigate compact rationale representations to enable scalable batch updates.
- **Multi-Hop Reasoning:** Real-world medical updates often involve interconnected facts and multi-step inferences. We leave the evaluation of multi-hop knowledge propagation and its downstream impact to future efforts.
- **Cross-Domain Generalization:** While our primary focus is medical QA, the benchmark framework and SGR-Edit paradigm may generalize to other specialized domains (e.g., legal, scientific). We plan to assess and adapt our protocol for broader domain transferability.

Ethics Statement

Our benchmarks are constructed from publicly available datasets and synthetic scenarios, which may introduce spurious or hallucinated content. Consequently, they are not for real clinical decision support. Furthermore, our findings highlight that editing operations can inadvertently degrade unrelated knowledge, underscoring the need for careful risk assessment before real-world deployment in safety-critical settings.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, and 1 others. 2024a. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*.

Qizhou Chen, Taolin Zhang, Xiaofeng He, Dongyang Li, Chengyu Wang, Longtao Huang, and Hui Xue. 2024b. Lifelong knowledge editing for llms with retrieval-augmented continuous prompt learning. *arXiv preprint arXiv:2405.03279*.

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. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2023. Do large language models know about facts? *arXiv preprint arXiv:2310.05177*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. 2025. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*.

Yiming Ju, Xingrun Xing, and Zhixiong Zeng. 2023. Klob: a benchmark for assessing knowledge locating methods in language models. *arXiv preprint arXiv:2309.16535*.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. MedExQA: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023a. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023b. Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zihao Lin, Mohammad Beigi, Hongxuan Li, Yufan Zhou, Yuxiang Zhang, Qifan Wang, Wenpeng Yin, and Lifu Huang. 2024. Navigating the dual facets: A comprehensive evaluation of sequential memory editing in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13755–13772, Bangkok, Thailand. Association for Computational Linguistics.

Jun-Yu Ma, Hong Wang, Hao-Xiang Xu, Zhen-Hua Ling, and Jia-Chen Gu. 2024a. Perturbation-restrained sequential model editing. *arXiv preprint arXiv:2405.16821*.

Xinbei Ma, Tianjie Ju, Jiyang Qiu, Zhuosheng Zhang, Hai Zhao, Lifeng Liu, and Yulong Wang. 2024b. 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. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations*.

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Yuval Pinter and Michael Elhadad. 2023. Emptying the ocean with a spoon: Should we edit models? *arXiv preprint arXiv:2310.11958*.

Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. 2024. Retrieval-enhanced knowledge editing in language models for multi-hop question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2056–2066.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge.

Xiaoshuai Song, Zhengyang Wang, Keqing He, Guanting Dong, Yutao Mou, Jinxu Zhao, and Weiran Xu. 2024. Knowledge editing on black-box large language models. *arXiv preprint arXiv:2402.08631*.

Chenmien Tan, Ge Zhang, and Jie Fu. Massive editing for large language models via meta learning. In *The Twelfth International Conference on Learning Representations*.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2024a. Knowledge editing through chain-of-thought. *arXiv preprint arXiv:2412.17727*.

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

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, and 1 others. 2023. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024c. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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. *arXiv preprint arXiv:2502.11177*.

Kai Zhang, Rui Zhu, Shutian Ma, Jingwei Xiong, Yejin Kim, Fabricio Murai, and Xiaozhong Liu. 2025. Kedrec-lm: A knowledge-distilled explainable drug recommendation large language model. *arXiv preprint arXiv:2502.20350*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, and 1 others. 2024. 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? *arXiv preprint arXiv:2305.12740*.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, and 1 others. 2024. Large language models for disease diagnosis: A scoping review. *arXiv preprint arXiv:2409.00097*.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

A Medical Knowledge Editing Benchmarks

A.1 Medical QA datasets

MedMCQA is drawn from postgraduate-level Indian medical entrance exams (AIIMS and NEET PG), spanning 2,400 healthcare topics across 21 specialties. Each question offers four answer options. We choose the validation set that comprises 4183 QA pairs for *MedMCQA*$_{edit}$ construction.

MedExQA is designed to provide a richer medical context for evaluating LLMs by pairing each question with two human-curated explanation sets. As these two explanations exhibit high semantic similarity (> 73%) (Kim et al., 2024), we simply uses the first explanation set in our benchmark. It includes five underrepresented specialties in current datasets: biomedical engineering, clinical laboratory science, clinical psychology, occupational therapy, and speech language pathology.

We exclude the Speech Language Pathology subset on MedExQA due to frequent inconsistencies between questions and their provided explanations. For example, one multiple-choice question lists B as the correct answer, while its accompanying explanation supports option D, rendering this subset unreliable for editing evaluations. After removing Speech Language Pathology, the remaining four specialties comprise 773 QA pairs used to construct *MedExQA*$_{edit}$.

A.2 Data Construction

Our benchmark construction proceeds in the follow four main steps to ensure that all evaluation questions truly measure editing gains in the medical domain:

Step 1. Quality Verification We begin with two public medical QA datasets, MedExQA and

MedMCQA, each of which pairs a question q with a ground-truth answer k and one or two human-written explanations $\exp$. Due to known inconsistencies (e.g., mismatches between k and $\exp$), we filter out any QA pair whose explanation fails to support the correct answer. Concretely, we prompt an LLM (e.g., LLaMA-8B) with the $(\exp, q)$ under in-context learning (more like in an open-book setting). If the model's prediction does not match k, we discard that sample. The remaining high-quality pairs form our *Verified Explanation Set* $\mathcal{D}_{\text{verified}}$.

Step 2. Zero-Shot Filtering Next, we assess this LLM's zero-shot performance on each question in $\mathcal{D}_{\text{verified}}$ by feeding only q (without $\exp$) in a context-free condition. Samples for which the model answers incorrectly indicate out-of-date or missing internal knowledge; these become our *Original Set* $\mathcal{Q}_{\text{ori}}$.

Step 3. Scenario Generation To generate novel datasets for rigorous Generalization and Retention assessment, we use a more powerful agent (we use DeepSeek-V3 in this study; others like GPT-4 and Gemini could also serve this role) to construct new medical questions. For each verified pair in $\mathcal{D}_{\text{verified}}$, we treat its human-curated explanation that cites authoritative medical textbooks, as the medical fact. Then, the agent uses the medical fact to craft clinical-scenario QA pairs through prompting (see Figure 7), yielding candidates for $\mathcal{Q}_{\text{gen}}$ and $\mathcal{Q}_{\text{ret}}$.

Step 4. Data Filtering We then subject each generated variant to the same zero-shot test as in Step 2. Concretely, for generalization question candidates, we prompt the LLM without any explanation and record its predicted answer $\hat{k}$. If $\hat{k} \neq k'$, where k' is the intended (ground-truth) answer, we include it in the generalization set $\mathcal{Q}_{\text{gen}}$; As for retention question candidates, if $\hat{k} = k'$, we add this pair to the retention set $\mathcal{Q}_{\text{ret}}$. This ensures that $\mathcal{Q}_{\text{gen}}$ contains only those scenarios that the LLM cannot solve without editing, while $\mathcal{Q}_{\text{ret}}$ captures instances where its original knowledge remains intact.

These processes yield three disjoint sets—$\mathcal{Q}_{\text{ori}}$, $\mathcal{Q}_{\text{gen}}$, and $\mathcal{Q}_{\text{ret}}$, which together form our final editing benchmarks *MedExQA$_{edit}$* and *MedMCQA$_{edit}$*.

A.3 Benchmark Complexity

Our medical editing benchmarks introduce substantially more complex questions than common general-domain datasets. Whereas ZsRE questions average only 11.9 tokens, our *MedMCQA$_{edit}$* and *MedExQA$_{edit}$* samples exhibit mean lengths of 52.7 and 47.6 tokens for original questions, 105.8 and 100.5 for generalization questions, and 92.6 and 88.6 for retention questions, respectively (see Figure 8). This increase in question length reflects the inclusion of rich clinical context and novel scenario descriptions. In particular, $\mathcal{Q}$gen and $\mathcal{Q}$ret include extra patient-centered details (see Figure 9) to mimic real-world clinical scenarios, further increasing the inference challenge. By expanding both the lexical and conceptual scope of each query, our benchmarks better simulate real-world medical reasoning tasks and rigorously test an LLM's capacity for knowledge editing under realistic complexity.

B Existing Editing Methods

Fine-Tuning–Based Editing. These traditional approaches update model parameters to incorporate new knowledge: **FT+L** (Zhu et al., 2020) enforces norm constraints in gradient updates to minimize interference on the unmodified facts. **FT-M** (Zhang et al., 2024) applies a masking strategy during training to focus updates on relevant target content. **LoRA** (Hu et al., 2022) introduces trainable low-rank decomposition matrices to achieve efficient adaptation with minimal additional parameters.

Parameter-Modifying Editing. Unlike fine-tuning, these methods seek to modify only a subset of parameters to inject new facts while preserving unrelated knowledge. Broadly, they fall into two categories: **i) Meta-Learning editing** typically trains a hypernetwork to estimate gradient updates for knowledge insertion (De Cao et al., 2021; Tan et al.), e.g.,**MEND** (Mitchell et al.) introduces a hypernetwork to transform the gradient obtained by using a low-rank decomposition to make the parameterization tractable. **ii) Locate-then-Edit** methods identify and then modify the weights responsible for specific factual associations: **ROME** (Meng et al., 2022) localizes and modifies factual associations in a transformer layer. **MEMIT** (Meng et al.) extends ROME by performing batch edits across multiple critical layers for mass knowledge updates.

Recently, sequential editing methods have been introduced to support continuous updates rather than a one-off modification. **PRUNE** (Ma et al.,

Strictly use the provided medical fact to create a standalone exam question that:

*** Extends and applies the core knowledge from the original question (ori_q) to a new clinical scenario.**
* Tests further application of the underlying medical principle by exploring nuanced aspects (e.g., treatment variations, diagnostic approaches) of the same fact.
* Ensures that the new clinical scenario is distinct from the original, yet its reasoning and answer are strictly derived from—and can be found within—the provided fact without external information.

Mandatory Requirements:
1. Identify 1-2 key elements from the provided fact that relate directly to the original question's core focus.
2. Build the question around these elements in a novel clinical context (e.g., different patient demographics or clinical settings) that extends the original knowledge.
3. Ensure the answer is directly supported by the content in the provided fact.

Strict Prohibitions:
1. No recycled answer options or distractors from the original question.
2. No introduction of external knowledge not included in the provided fact.
3. No repetition of the original question's clinical presentation or distractor structure.

Output Format:
[new_q]
1. Clinical stem presenting a new context that builds on the original core knowledge.
2. 4 plausible answer options that incorporate novel distractors.
3. A clear differentiator from the original question's focus.

[new_a]
Provide answer reasoning that validates the new fact-derived clinical application and ensure the correct answer can be found within the provided fact.

[new_exp]
1. Explicitly cite the key fact elements used.
2. Explain how the scenario extends the original question's core knowledge.

Strictly use the provided medical fact to create a standalone exam question that:

*** Tests a distinct area of medical knowledge that is different from the core focus of the original question (ori_q).**
* Evaluates an alternative clinical application scenario, ensuring the underlying knowledge is strictly derived from the provided fact.
* Presents a fresh angle that does not overlap with the original question's content.

Mandatory Requirements:
1. Extract 1-2 key elements from the provided fact that represent different or additional aspects not covered in ori_q.
2. Construct a question that leverages these elements to form an entirely separate clinical scenario (e.g., alternative diagnostic methods or treatment considerations) from the original question.
3. Ensure all reasoning and content are exclusively based on the provided fact.

Strict Prohibitions:
1. No reuse of answer logic or distractors from the original question.
2. No incorporation of external information beyond the provided fact.
3. No similarity in clinical presentation to the original question's scenario.

Output Format:
[locality_q]
1. Clinical stem with a new context that is clearly distinct from the original question's focus.
2. 4 plausible answer options with novel distractors.
3. A clear differentiation from the original question's tested knowledge.

[locality_a]
Provide answer reasoning strictly based on the fact-derived knowledge.

[locality_exp]
1. Explicitly reference the key fact elements.
2. Contrast this scenario with the original question's focus.

Figure 7: Prompts for $\mathcal{Q}_{\text{gen}}$ (left) and $\mathcal{Q}_{\text{ret}}$ (right) constructions

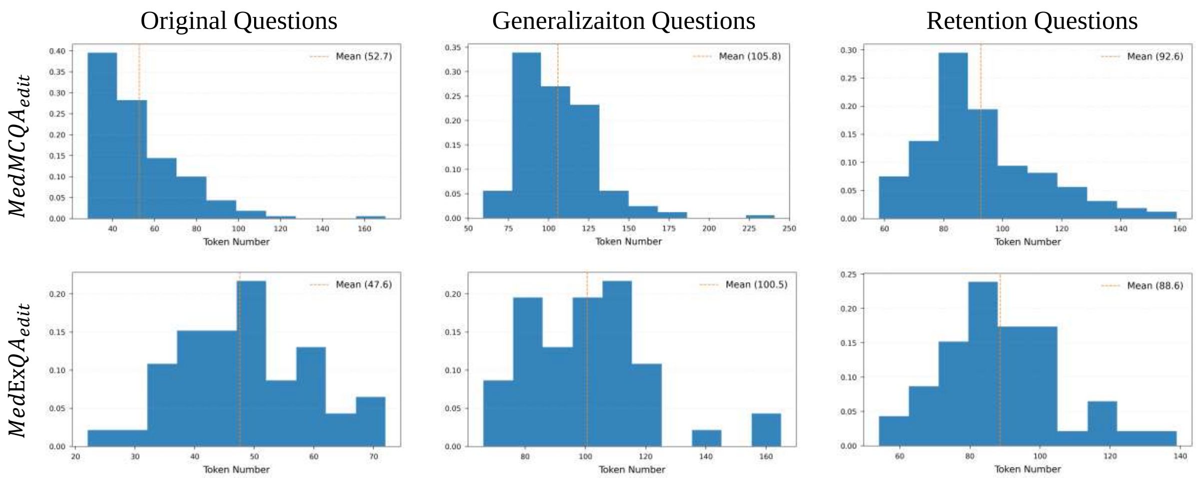

Figure 8: Token length distributions for *MedMCQA*$_{edit}$ (top row) and *MedExQA*$_{edit}$ (bottom row) across Original, Generalization, and Retention question sets. Dashed lines indicate dataset means. Our medical QA benchmarks present substantially longer inputs, increasing task difficulty.

2024a) applies condition number constraints to limit perturbation to keep the general capacity of the model. **AlphaEdit** (Fang et al., 2024) minimizes disruption to the preserved knowledge by projecting perturbations onto the null space of its key matrices. **AnyEdit** (Jiang et al., 2025) decomposes long-form knowledge into sequential chunks and edits each key token autoregressive. These three newer sequential editing methods are optimized on top of the MEMIT.

Parameter-Preserving Editing These methods avoid altering the parameters of the base model. Extension-based methods augment the model with external components, leaving the base parameters unchanged: **GRACE** (Hartvigsen et al., 2023) writes new mappings as entries in a discrete codebook in an Adaptor. **SERAC** (Mitchell et al., 2022) integrates a classifier and a side model to identify and learn new knowledge. Similarly, **WISE** (Wang et al., 2024b) deploys side FFN layers to incorporate new knowledge dynamically, and **EditCoT** (Wang et al., 2024a) trains a separate CoT editor that cooperates with the base LLM to perform knowledge updates at inference, . Addition-

QA_ori :
Which of the following bone(s) are used for the determination of growth in an individual?
A: Clavicle B: Sternum C: Cervical vertebrae D: Mandible

QA_gen :
A 12-year-old patient is being evaluated for orthodontic treatment planning. The orthodontist orders a lateral cephalogram to assess skeletal maturity. Which of the following anatomical structures on the cephalogram would provide the most accurate information about the patient's growth status?
A: Shape and size of the mandibular condyle
B: Degree of fusion in the spheno-occipital synchondrosis
C: Morphological changes in the bodies of the second through sixth cervical vertebrae
D: Thickness of the cortical bone in the maxilla

QA_ret :
A 12-year-old patient is undergoing orthodontic treatment, and the orthodontist needs to assess the patient's skeletal maturity to plan the timing of intervention. Which of the following imaging techniques is most appropriate for evaluating the changes in cervical vertebrae to determine skeletal maturity?
A: Panoramic radiograph B: Lateral cephalogram
C: Periapical radiograph D: Cone-beam computed tomography (CBCT)

Figure 9: Examples of the three QA types. Purple text denotes the question's domain/topic; blue text highlights the injected new correct answer; red text shows the model's original (pre-edit) answer.

Repr.	Content
GTA	"Co-arctation of Aorta."
RE	"Masson trichrome (Ref: Kanski 7/e p212–224; Parsons 22/e p212–214; Yanoff and Duker 4/e p261). Masson trichrome stain – Used for granular corneal dystrophy diagnosis. Granular dystrophy: AD inheritance with gene locus on 5q31; Onset: first decade with recurrent erosions; Signs: small, white, sharply demarcated deposits resembling crumbs or snowflakes in central anterior stroma; Histology: shows amorphous hyaline deposits staining bright red with Masson trichrome."
SGR	"*STEP 1:* According to the reference, Masson trichrome stain is specifically used for the diagnosis of Granular dystrophy of the cornea. *STEP 2:* The reference mentions that the histology of Granular dystrophy shows amorphous hyaline deposits staining bright red with Masson trichrome, indicating the effectiveness of this stain in diagnosing the condition. *STEP 3:* Colloidal iron stain is used for diagnosing various conditions, but... *STEP 4:* Congo red stain is used for... *STEP 5:* PAS (Periodic Acid–Schiff) stain is..."

Table 3: Knowledge representations for editing. (GTA: Bare correct answer; RE: Concise factual excerpt from expert source; SGR: Self-generated chain-of-thought rationale)

ally, retrieval-based methods (Song et al., 2024; Shi et al., 2024; Chen et al., 2024b) have been proposed to retrieve relevant factual information at inference time, effectively "editing" via context rather than weight changes. For example, IKE (Zheng et al., 2023) retrieves relevant factual edits and uses them to build the prompt context as input, and then prompts the model to generate updated outputs.

C Detailed Experimental Settings

C.1 Base Models

In this study, we evaluate editing methods on two instruction-tuned LLaMA variants: Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct [3]. We utilize layers [4, 5, 6, 7, 8] of the two LLMs for editing, except for experiments 5.3, where we investigate which layers used to edit deliver the best performance.

[3] https://llama.meta.com/lama3/

C.2 Implementation

We implement and evaluate six representative knowledge editing methods—LoRA, ROME, MEMIT, GRACE, AnyEdit, and AlphaEdit for fair comparison. All editing workflows are built on the EasyEdit framework[4]. For AnyEdit, which is not yet supported by EasyEdit, we integrate the original codebase[5] and adopt the original hyperparameters. In terms of evaluation, we implement independent pipelines and consistently compute metrics tailored to our medical knowledge editing. During the post-edit inference phase, we employ greedy decoding to ensure deterministic outputs. Specifically, we set `do_sample=False` and `temperature=0.0` for all evaluations, so that the edited model's predictions reflect its learned knowledge without sampling variability.

[4] https://github.com/zjunlp/EasyEdit
[5] https://github.com/jianghoucheng/AnyEdit

System Prompt for Rationale Generation:
You have access to a reference that may assist in answering the question and in ultimately deriving the correct answer. Then focus on answering the following question using a chain of thought. Use the token [STEP] to start each thought step, and the token [ANSWER] to indicate the final answer option (A, B, C, or D).

Template:
[STEP] ...
[STEP] ...
...
[ANSWER] ...

One-Shot:
Question: Which component of a PET scanner is used to collect the position information of annihilation radiation?
A: Coincidence circuit B: Scintillator C: Collimator D: Photomultiplier tube"
Answer: A: Coincidence circuit"
Reference: In PET imaging, the detection of annihilation radiation, which occurs due to the interaction between positrons and electrons, is essential for image creation. A device that collects the position information of the annihilation radiation is required for this purpose. This device is the coincidence circuit. The coincidence circuit measures the time at which the annihilation radiation is detected by a photomultiplier tube, and simultaneously by another photomultiplier tube, to accurately determine the position of the annihilation radiation. Therefore, the coincidence circuit plays a crucial role in PET imaging."
Rationale:
[STEP] According to the reference, the coincidence circuit is responsible for measuring the time at which the annihilation radiation is detected by two photomultiplier tubes, allowing for the accurate determination of the position of the annihilation radiation.
[STEP] The coincidence circuit is specifically designed to collect the position information of annihilation radiation, making it the correct answer.
[STEP] The scintillator is a component that converts the annihilation radiation into visible light, but it does not collect position information.
[STEP] The collimator is a component that helps to focus the annihilation radiation onto the scintillator, but it does not collect position information.
[STEP] The photomultiplier tube is a component that detects the visible light produced by the scintillator, but it does not collect position information.
[ANSWER] A

Figure 10: Complete prompt for generating rationales.

C.3 Editing Paradigm

For each $q_{\text{ori}} \in \mathcal{Q}_{\text{ori}}$, we construct three knowledge targets k' according to the paradigm:

- **GTA-Edit:** the ground-truth answer, with option letters stripped (e.g., remove "D:" in "D: Masson trichrome").
- **RE-Edit:** the human-written explanation excerpted from textbooks.
- **SGR-Edit:** the proposed paradigm where the LLM is first prompted to generate its own chain of thought over the RE, then uses that self-generated rationale as k' for editing.

Examples of GTA, RE, and SGR are shown in Table 3.

C.4 Evaluation Protocol

During post-edit evaluation, we randomly permute the answer choices in each multiple-choice $\mathcal{Q}_{\text{ori}}$ to prevent label memorization. All outputs follow a two-step format: rationale generation followed by the final answer. We report:

- **Efficacy:** accuracy on $\mathcal{Q}_{\text{ori}}$.
- **Generalization:** accuracy on $\mathcal{Q}_{\text{gen}}$.
- **Retention:** accuracy on $\mathcal{Q}_{\text{ret}}$.
- **Interpretability:** ROUGE-L and BLEU scores between the injected knowledge (reference explanation or self-generated rationale) and the model's post-edit rationale output, computed on a human-validated subset to quantify how closely the model reproduces the intended content.

Detailed prompt templates, and additional examples are provided in Appendix E.

C.5 Task Formulation for Sequential Editing

For t sequential edits on distinct knowledge targets $\{(q_i, k_i)\}_{i=1}^{t}$, we define:

$$\theta^{(i)} = F\big(\theta^{(i-1)},\, q_i,\, k_i\big), i = 1, 2, \ldots, t \qquad (2)$$

to satisfy $\theta^{(i)}(q_i) = k_i,\ \forall\, i \in \{1, \ldots, t\}$, where $\theta^{(i)}$ is the model after the i-th edit. After t edits, the final model $\theta^{(t)}$ must satisfy $\theta^{(t)}(q_j) = k_j$ for all $j \leq t$ ensuring that each injected knowledge item remains correctly reflected in the model.

D Supplementary Results and Analyses

D.1 Performance Comparison of Various Editing Paradigms

As shown in Table 4, relative to GTA-Edit, RE-Edit consistently raises average editing scores on LLaMA-8B by 0.2–8.4 percentage points across methods: ROME sees the largest gain (+8.4 pp), followed by MEMIT (+5.3 pp), LoRA (+4.8 pp),

Method	Metric	Pre-Edit	LLaMA-8B			LLaMA-3B		
			GTA-Edit	RE-Edit	SGR-Edit	GTA-Edit	RE-Edit	SGR-Edit
LoRA	Eff.	0	0.466	0.547	0.665	0.144	0.412	0.615
	Gen.	0	0.416	0.460	0.503	0.337	0.439	0.492
	Ret.	1	0.708	0.727	0.708	0.529	0.658	0.636
	avg.	–	0.530	0.578	**0.625**	0.337	0.503	**0.581**
ROME	Eff.	0	0.327	0.453	0.528	0.257	0.299	0.439
	Gen.	0	0.296	0.403	0.447	0.251	0.406	0.385
	Ret.	1	0.616	0.635	0.667	0.561	0.647	0.610
	avg.	–	0.413	0.497	**0.547**	0.357	0.451	**0.478**
MEMIT	Eff.	0	0.283	0.384	0.520	0.160	0.310	0.455
	Gen.	0	0.252	0.352	0.364	0.251	0.364	0.401
	Ret.	1	0.648	0.604	0.671	0.545	0.679	0.663
	avg.	–	0.394	0.447	**0.518**	0.319	0.451	**0.506**
AnyEdit	Eff.	0	0.366	0.410	0.435	0.230	0.246	0.316
	Gen.	0	0.255	0.248	0.317	0.230	0.267	0.316
	Ret.	1	0.789	0.758	0.764	0.781	0.747	0.754
	avg.	–	0.470	0.472	**0.505**	0.414	0.420	**0.462**
AlphaEdit	Eff.	0	0.439	0.547	0.584	0.326	0.348	0.374
	Gen.	0	0.312	0.335	0.366	0.278	0.332	0.348
	Ret.	1	0.867	0.795	0.789	0.775	0.743	0.754
	avg.	–	0.539	0.559	**0.580**	0.460	0.474	**0.492**

Table 4: Complete results of GTA-Edit, RE-Edit, and SGR-Edit. Supported for RQ2-F1

AlphaEdit (+2.0 pp), and AnyEdit (+0.2 pp). These improvements are driven primarily by jumps in Efficacy (up to +12.6 pp for ROME) and Generalization (up to +10.7 pp for ROME), while Retention remains stable above 61% for all methods.

Incorporating SGR-Edit yields further average gains of 2.1–7.1 pp over RE-Edit, with MEMIT (+7.1 pp) and LoRA (+4.7 pp) benefitting most. Specifically, LoRA's combined score climbs from 57.8% to 62.5%, and MEMIT from 44.7% to 51.8%. Even AlphaEdit, which already excels under RE-Edit, improves from 55.9% to 58.0% (+2.1 pp).

Critically, these gains in Efficacy and Generalization come with only minor retention trade-offs (e.g., AnyEdit drops from 78.9% to 76.4%), confirming that richer and context-driven rationales enable deeper medical knowledge integration without undue forgetting.

Event-Driven Rationale Generation for Practical SGR-Edit In real-world scenarios, knowledge updates are always triggered by concrete events: in medicine, for instance, the U.S. Food and Drug Administration's approval of a novel oncology drug follows positive clinical trial results; in politics, a change in the presidency (e.g., from Biden to Trump) is driven by certified election outcomes; in law, the enactment of a new data-privacy statute typically relies on high-profile regulatory incidents. Under SGR-Edit, these domain-specific event narratives may be supplied by subject-matter experts or, alternatively, sourced automatically using Retrieval-Augmented Generation frameworks (Lewis et al., 2020). By inputting these event narratives, the LLM can produce evidence-grounded rationales that facilitate reliable and transparent knowledge edits in practice.

D.2 Lexical Overlap Analysis of SGR-Edit and RE-Edit

The results in Table 5 are computed over a human-validated $MedMCQA_{edit}$ subset of high-quality reference explanations (RE) and self-generated rationales (SGR). Since SGR generation can include spurious content, we manually verify each rationale to ensure it faithfully supports the target answer before using it for ROUGE-L and BLEU calculations. This subset covers 64 QA pairs for LLaMA-8B and 77 for LLaMA-3B. These long-form explanations provide rich contextual support for question answering, enabling a precise evaluation of lexical overlap between the injected knowledge and the model's post-edit outputs. The datasets are available at.

Across all three editing methods: AlphaEdit, LoRA, and MEMIT, SGR-Edit outperforms RE-Edit by a substantial margin in both ROUGE-L and BLEU. For LLaMA-8B, the average lexical overlap (see Lexical avg.) for SGR-Edit is 0.241 (AlphaEdit), 0.443 (LoRA), and 0.271 (MEMIT), compared to just 0.098, 0.137, and 0.095 under RE-Edit. BLEU improvements are equally dramatic: AlphaEdit rises from 0.031 to 0.149, LoRA from 0.064 to 0.370, and MEMIT from 0.031 to 0.180. These gains confirm that SGR—Edit allows evidence-grounded and logical knowledge representation to align more closely with the correct ra-

Method (Editing Paradigm)	LLaMA-8B			LLaMA-3B		
	ROUGE-L	BLEU	Lexical avg.	ROUGE-L	BLEU	Lexical avg.
AlphaEdit (SGR-Edit)	0.334	0.149	**0.241**	0.348	0.166	**0.257**
AlphaEdit (RE-Edit)	0.165	0.031	0.098	0.170	0.034	0.102
LoRA (SGR-Edit)	0.516	0.370	**0.443**	0.592	0.459	**0.526**
LoRA (RE-Edit)	0.211	0.064	0.137	0.242	0.093	0.167
MEMIT (SGR-Edit)	0.363	0.180	**0.271**	0.356	0.188	**0.272**
MEMIT (RE-Edit)	0.159	0.031	0.095	0.169	0.036	0.102

Table 5: Comparison of ROUGE-L and BLEU for SGR-Edit and RE-Edit on LLaMA-8B and LLaMA-3B.

tionale for medical decision making and thus serve as a superior knowledge target.

When comparing across editing methods, LoRA consistently achieves the highest lexical overlap among all paradigms and model sizes. On LLaMA-8B, LoRA SGR-Edit reaches ROUGE-L=0.516 and BLEU=0.370, yielding a text-average of 0.443. This outperforms both AlphaEdit (0.241 lexical avg.) and MEMIT (0.271 lexical avg.), indicating that LoRA's low-rank adaptation effectively internalizes the rich and context-driven rationales generated by the model. The pattern also holds on LLaMA-3B, where LoRA SGR-Edit achieves a lexical average score of 0.526 versus 0.257 (AlphaEdit) and 0.272 (MEMIT).

In summary, these lexical metrics corroborate our finding that SGR-Edit consistently yields the highest lexical scores, validating its ability to convey deeper understanding rather than superficial information memorization. Furthermore, LoRA emerges as the most effective editing mechanism, capitalizing on the enriched content of self-generated rationales to maximize knowledge integration.

D.3 Qwen2.5–7B Single-Edit Performance

In addition to LLaMA variants, we evaluate Qwen2.5–7B [6] under the same medical editing protocol. Table 6 reports efficacy, generalization, and retention on *MedExQA*$_{edit}$ and *MedMCQA*$_{edit}$ unding single editing.

Consistent with our findings on LLaMA, LoRA and AlphaEdit remain the most effective editing methods for Qwen2.5–7B. LoRA achieves the highest average score on MedMCQA$_{\text{edit}}$ (64.2%), driven by strong efficacy (57.4%) and generalization (50.5%), while AlphaEdit closely follows with an average of 60.0%, showing robust efficacy (62.6%) despite slightly lower retention (74.7%). By contrast, GRACE exhibits very low generalization (1.9% on *MedExQA*$_{edit}$, 3.2% on*MedMCQA*$_{edit}$) despite near-perfect retention (>96%), indicating that this codebook-based and parameter-conserving updates tend to memorize new facts without enabling flexible application to unseen scenarios.

Other parameter-modifying methods, such as ROME and MEMIT, display moderate efficacy (23.2–55.2%) but generalization remains in the 30–43% range, well below the 90%+ figures often reported on simplified, general-domain benchmarks. Similarly, AnyEdit yields balanced retention (>78%) but low generalization (15.8–30.0%) and efficacy (22.6–40.0%). These results confirm the large performance gap noted in RQ1-F1 and RQ1-F2: even state-of-the-art editing methods struggle to exceed 65% combined performance in realistic medical QA settings, exposing critical limitations of existing methods when confronted with domain-specific complexity.

D.4 Post-Edit Inference: Two-Step Rationale & Answer vs. One-Step Final Answer

We compare two prompting strategies on LLaMA-8B over *MedMCQA*$_{edit}$: (i) "Rationale + Answer" and (ii) "Final-Answer Only" (see Figure 11). Table 7 presents post-edit accuracy under GTA-Edit, RE-Edit, and SGR-Edit for LoRA, ROME, MEMIT, and AlphaEdit. Overall, providing a chain of thought during post-edit inference yields substantial gains in editing efficacy and generalization, at minimal retention cost.

For LoRA, the average post-edit score climbs from 0.442 to 0.530 (+0.203) when moving to two-step output. Efficacy improves dramatically (GTA-Edit: +0.170, RE-Edit: +0.106, SGR-Edit: +0.049), and Generalization more than twice in the GTA-Edit case (0.220 → 0.416). Retention also

[6] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

Method	Metric	MedExQA_edit	MedMCQA_edit
LoRA	Eff.	61.5	57.4
	Gen.	46.2	50.5
	Ret.	76.9	84.7
	avg.	**61.5**	**64.2**
ROME	Eff.	36.0	55.2
	Gen.	38.0	43.1
	Ret.	46.0	82.3
	avg.	40.0	<u>60.2</u>
MEMIT	Eff.	30.0	54.1
	Gen.	34.0	39.2
	Ret.	80.0	75.1
	avg.	48.0	56.2
GRACE	Eff.	36.5	23.2
	Gen.	1.90	3.20
	Ret.	96.2	97.4
	avg.	44.9	41.2
AnyEdit	Eff.	40.0	22.6
	Gen.	30.0	15.8
	Ret.	78.0	85.3
	avg.	49.3	41.2
AlphaEdit	Eff.	58.0	62.6
	Gen.	40.0	42.6
	Ret.	78.0	74.7
	avg.	<u>58.7</u>	60.0

Table 6: Results on Qwen2.5-7B (Accuracy %). For avg. scores per column: **bold** is the best, <u>underline</u> is the second best.

System Prompt for Two-Step Output:
Please answer the following question using a chain of thought. Use the token [STEP] to start each thought step, and the token [ANSWER] to indicate the final answer option (A, B, C, or D).

Template:
[STEP] ...
[STEP] ...
...
[ANSWER] ...

System Prompt for One-Step Output:
Please answer the following question and only output the final answer option (A, B, C, or D) without any additional explanation.

Figure 11: Two-Step and One-Step System Prompts for Post-Edit Inference

rises under RE (0.591 → 0.727), demonstrating that transparent reasoning reinforces new facts.

ROME shows a mixed pattern: while GTA-Edit average performance drops slightly (0.421 → 0.413), RE-Edit and SGR-Edit see gains (+0.053 and +0.084 avg.), reflecting that rationale prompts help when richer contexts are available. MEMIT benefits modestly (+0.006 to +0.067 avg.), with Generalization under SGR especially boosted (0.302 → 0.364). AlphaEdit gains across all paradigms (+0.032 to +0.075 avg.), with Retention remaining above 0.789 even after two-step reasoning.

These results confirm that explicit chain-of-thought prompting significantly enhances the model's ability to apply the injected knowledge, supporting more reliable and interpretable medical knowledge editing.

D.5 Sequential Editing Impact on Common-Domain Capabilities

We split the MMLU benchmark into health-domain and non-health categories: health-domain accuracy (Health_acc) is computed over the subjects {anatomy, clinical_knowledge, college_medicine, human_aging, medical_genetics, nutrition, professional_medicine, virology}, while all other subjects are aggregated to compute non-health accuracy (NonHealth_acc).

In addition to the Health vs. Non-Health breakdown (see section 5.4), we further split MMLU into four broad categories—STEM, Humanities, Social Sciences, and Other (Business, Misc.)—to assess general-domain degradation. Table 8 reports category accuracies for LLaMA-8B and LLaMA-3B at 0, 50, and 100 sequential edits.

After 100 sequential medical edits, all common-domain categories exhibit performance declines, with "Other" (which includes business, health and miscellaneous topics) suffering the largest drop (–6.45 pp for LLaMA-8B; –5.86 pp for LLaMA-3B). STEM accuracy decreases by 3.96 pp on 8B and 2.56 pp on 3B, while Social Sciences and Humanities show smaller but nontrivial declines. These results reinforce our RQ4-F2 finding: medical-focused updates not only impair specialized health knowledge but also erode broader general capabilities. The consistent degradation across diverse categories underscores the challenge of maintaining out-of-domain performance during extensive sequential editing.

E Case Analysis

The case in Figure 12, using the MEMIT method to edit LLaMA-8B on *MedMCQA*$_{edit}$, illustrates how different editing paradigms shape the reasoning behavior of post-edit LLM. This representative example highlights not only the final answer accuracy but also the depth of knowledge integration achieved by each paradigm. It reflects a broader pattern consistently observed across our benchmark: surface-level edits often fail to update the underlying reasoning logic, while rationale-driven edit-

Method	Metric	Pre-Edit	Final-Answer Only			Rationale + Answer		
			GTA-Edit	RE-Edit	SGR-Edit	GTA-Edit	RE-Edit	SGR-Edit
LoRA	Eff.	0	0.296	0.553	0.616	0.466	0.547	0.665
	Gen.	0	0.220	0.258	0.258	0.416	0.460	0.503
	Ret.	1	0.465	0.591	0.453	0.708	0.727	0.708
	avg.	–	0.327	0.468	0.442	0.530 (+0.203)	0.578 (+0.110)	0.625 (+0.183)
ROME	Eff.	0	0.528	0.566	0.642	0.327	0.453	0.528
	Gen.	0	0.252	0.277	0.277	0.296	0.403	0.447
	Ret.	1	0.484	0.491	0.472	0.616	0.635	0.667
	avg.	–	0.421	0.444	0.463	0.413 (-0.008)	0.497 (+0.053)	0.547 (+0.084)
MEMIT	Eff.	0	0.440	0.541	0.566	0.283	0.384	0.520
	Gen.	0	0.214	0.270	0.302	0.252	0.352	0.364
	Ret.	1	0.509	0.484	0.484	0.648	0.604	0.671
	avg.	–	0.388	0.432	0.451	0.394 (+0.006)	0.447 (+0.015)	0.518 (+0.067)
AlphaEdit	Eff.	0	0.623	0.654	0.648	0.439	0.547	0.584
	Gen.	0	0.252	0.270	0.233	0.312	0.335	0.366
	Ret.	1	0.648	0.623	0.635	0.867	0.795	0.789
	avg.	–	0.507	0.516	0.505	0.539 (+0.032)	0.559 (+0.043)	0.580 (+0.075)

Table 7: Post-Edit accuracy comparison between Final-Answer Only and Rationale + Answer

Model	#Edits	STEM	ΔSTEM	Humanities	ΔHumanities	Social Sci.	ΔSocial Sci.	Other	ΔOther
LLaMA-8B	0 (Raw)	58.53	–	72.52	–	77.13	–	69.79	–
	50	56.17	–2.36	72.32	–0.20	74.96	–2.17	68.48	–1.31
	100	54.57	–3.96	70.79	–1.73	72.62	–4.51	63.34	–6.45
LLaMA-3B	0 (Raw)	50.98	–	65.14	–	69.35	–	62.47	–
	50	50.22	–0.76	64.19	–0.95	66.78	–2.57	60.22	–2.25
	100	48.42	–2.56	61.89	–3.25	64.92	–4.43	56.61	–5.86

Table 8: General-domain category accuracies and absolute drops Δ on MMLU across sequential medical edits

ing enables models to produce more coherent and knowledge-grounded justifications.

Specifically, under **GTA-Edit**, although the correct answer "Gold" was injected, the model struggles to abandon its prior belief "Amalgam", which was likely encoded during pretraining. Instead of incorporating the new knowledge, the model retains previous reasoning patterns (red segments) and ends up with contradictory justifications (purple segments), reflecting surface-level memorization rather than meaningful conceptual update.

RE-Edit provides more context about material usage but still fails to fully overwrite the original logic. The rationale demonstrates partial knowledge integration (blue segments), yet the model remains hesitant (purple segments) and eventually falls back on incorrect elimination logic.

In contrast, **SGR-Edit** delivers a clear and coherent rationale. The model identifies that gold is used in its pure form due to its corrosion resistance and biocompatibility (blue segments), precisely the kind of causal justification absent in GTA/RE edits. This shows that editing with a reasoning-rich target enables the LLM to internalize and apply the knowledge reliably. Such examples highlight the importance of reasoning-grounded editing for domains like medicine, where interpretability and trustworthiness are critical.

Question: The material used in its pure form in dentistry is: **A: Gold.** B: Silver. C: Composite. D: Amalgam.

GTA-Edit:
[STEP] First, let's consider the options given: Gold, Silver, Composite, and Amalgam.
[STEP] We need to identify which of these materials is used in its pure form in dentistry.
[STEP] Gold is often alloyed with other metals to increase its strength and durability, so it's unlikely to be the correct answer.
[STEP] Silver is ... so it's not used in its pure form in dentistry.
[STEP] Composite materials are …. so they are not used in their pure form either.
[STEP] Amalgam, on the other hand, is a mixture of mercury and other metals, but in some dental applications, mercury is used in its pure form, such as in mercury thermometers and in some dental amalgams.
[STEP] However, the most common use of mercury in dentistry is in the form of an amalgam, not in its pure form.
[STEP] But, there is one application where mercury is used in its pure form in dentistry, which is in the form of a mercury gauge or a mercury thermometer.
[STEP] However, the question asks for the material used in its pure form in dentistry, and mercury is the only option that fits this description
[STEP] Therefore, the correct answer is the material that is used in its pure form in dentistry.
[ANSWER] D: Amalgam

RE-Edit:
[STEP] In dentistry, various materials are used for different purposes, such as fillings, crowns, and bridges.
[STEP] Among these materials, some are used in their pure form, while others are alloyed with other metals to enhance their properties.
[STEP] Gold is often used in dentistry, particularly in crowns and bridges, due to its biocompatibility and durability.
[STEP] However, gold is not typically used in its pure form; instead, it is alloyed with other metals like silver, copper, and zinc to create a stronger and more durable material.
[STEP] Silver, on the other hand, is not commonly used in ...
[STEP] Composite materials are .., but they are not used in their pure form; ...
[STEP] Amalgam, which is a mixture of mercury and other metals, is used in its pure form in dentistry, specifically for fillings.
[ANSWER] D

SGR-Edit:
[STEP] First, let's consider the options given: Gold, Silver, Composite, and Amalgam.
[STEP] We need to identify which of these materials is used in its pure form in dentistry.
[STEP] Gold is often used in dentistry, particularly for making crowns, bridges, and inlays, and it is often used in its pure form due to its high durability and resistance to corrosion.
[STEP] Silver is also used in dentistry, but it is often alloyed ...
[STEP] Composite is a mixture of materials, typically .. and is not used in its pure form.
[STEP] Amalgam is a mixture of mercury with other metals, such as silver, tin, and copper, and is not used in its pure form.
[STEP] Based on this analysis, the material used in its pure form in dentistry is Gold.
[ANSWER] A

Figure 12: **Case Comparison across Editing Paradigms.** The question asks which material is used in its pure form in dentistry. Blue indicates the correct post-edit answer, Red denotes the pre-edit (incorrect) answer that the model memorized, and Purple highlights areas of reasoning confusion. GTA-Edit fails to modify the model's original reasoning path, causing it to revert to the pre-edit belief. RE-Edit introduces more context but still exhibits uncertainty. While SGR-Edit enables the LLM to internalize the knowledge and produce a clear, logically sound rationale.