Beyond Memorization: A Rigorous Evaluation Framework for Medical Knowledge Editing

Anonymous ACL submission

Abstract

Knowledge editing (KE) has recently emerged as a promising technique to update specific facts in large language models (LLMs) without full retraining. While existing KE methods show promising results on general-domain benchmarks, their effectiveness in the medical domain remains largely unexplored. Medical knowledge editing poses unique challenges, requiring models not only to memorize new facts but also to internalize and generalize them for reliable and interpretable clinical decision-making. In this work, we propose MedEditBench, a rigorous evaluation framework for assessing medical knowledge editing. Our preliminary results reveal that current KE paradigm, which directly edits simple answers to the LLMs, often leads to superficial updates with poor generalization. To address this, we introduce Self-Generated Rationale Editing (SGR-Edit), which leverages model-generated rationales as editing targets, enabling deeper knowledge integration. Extensive experiments across diverse LLMs and KE methods demonstrate that SGR-Edit consistently improves editing efficacy and generalization. Furthermore, we examine the impact of sequential edits on in-domain medical knowledge, external-domain knowledge, as well as general model capabilities, offering practical insights for deploying KE 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 an approach that efficiently updates new knowledge in LLMs, allowing them to constantly adapt to the ever-evolving world without requiring full retraining (Meng et al., 2022; Meng et al.; Fang et al., 2024).

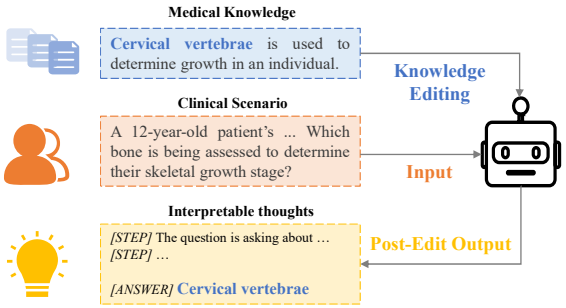

Figure 1: Illustration of medical knowledge editing.

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 knowledge (Singhal et al., 2023) but also comprehend underlying concepts and generalize to unseen scenarios (Li et al., 2023), as shown in Figure 1. Crucially, the necessity for transparent and trustworthy medical decisions requires LLMs to also explain their reasoning to improve trust and interoperability (Holzinger et al., 2017; Tonekaboni et al., 2019). Therefore, these challenges necessitate a more rigorous evaluation framework that accurately reflects the genuine performance of KE methods in the medical domain, ensuring their reliability and applicability in real-world applications (Zhou et al., 2024; Zhang et al., 2025).

In this work, we create a stringent medical knowledge editing benchmark based on two professional medical entrance examinations (Kim et al., 2024; Pal et al., 2022). Specifically, we carefully

[1] Codes and datasets ofthe MedEditBench are available at https://anonymous.4open.science/r/F0F5/

curate a subset of QA pairs that current LLMs fail to answer correctly to serve as the target set for knowledge editing. To evaluate the generalizability and robustness 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 preservation questions that require LLMs to retain their previously factual knowledge.

Our findings show that existing KE methods fall short in complex medical settings: edited LLMs typically achieve less than 50% efficacy, struggle to generalize knowledge to new scenarios, and often corrupt previously stored facts. We attribute this pitfall to the existing editing paradigm that edits the LLMs through a simple answer, resulting in surface-level memorization rather than deep understanding, thereby impairing generalization and robustness.

Inspired by research on belief revision in logic and philosophy, where an intelligent agent updates its beliefs by incorporating new information while maintaining logical consistency and minimizing unnecessary changes (Alchourrón et al., 1985; Falappa et al., 2002), we argue that knowledge editing should not merely overwrite outdated facts, but instead be guided by rational principles of consistency and justification. This perspective suggests that, much like human belief revision, effective editing requires encoding not only the updated fact but also the explanatory reasons that support it. Building on this insight, we propose an editing paradigm called **S**elf-**G**enerated **R**ationale **Edit**ing (SGR-Edit), where LLMs are prompted to generate explanatory rationales (Wei et al., 2022) grounded in reference texts. These self-generated rationales serve as the target knowledge that can be seamlessly integrated into arbitrary KE methods. Experimental results demonstrate that SGR-Edit improves knowledge editing performance by up to 11.3% across various LLMs and diverse KE methods, enabling edited LLMs to better internalize new medical knowledge and generalize to unfamiliar scenarios. Our primary contributions are fourfold:

1. **The `MedEditBench` Framework:** We introduce `MedEditBench`, a rigorous evaluation framework for medical knowledge editing. We employ a multi-faceted metric suite to comprehensively assess the state-of-the-art KE methods in complex medical scenarios.

2. **Pinpointing Paradigmatic Limitations:** We conduct extensive evaluations revealing that the conventional KE paradigm, which relies on editing simple final answers, primarily results in superficial memorization rather than genuine knowledge internalization. This severely limits the crucial generalization capability required for real-world practice.

3. **The Self-Generated Reasoning (SGR-Edit) Paradigm:** We propose a novel editing paradigm, **SGR-Edit**, that utilizes the LLM's self-generated, evidence-based reasoning chain as the editing target. SGR-Edit significantly boosts the editing performance across diverse LLMs and KE methods.

4. **In-depth Analysis and Practical Guidance:** We provide the first systematic analysis of sequential editing effects in both internal and external knowledge, and LLM general capabilities. Our findings offer critical insights into deploying KE in safety-critical applications like medicine, guiding future research toward more reliable and interpretable LLMs.

2 Related Work

Current KE methods can be grouped into three main categories. The first, **Fine-Tuning–Based Editing**, injects new knowledge by updating a large number of model parameters, often through parameter-efficient training techniques (Zhu et al., 2020; Zhang et al., 2024; Hu et al., 2022). The second category, **Parameter-Modifying Editing**, focuses on altering specific parameters to minimize interference with unrelated knowledge. This group is further divided into: 1) *Meta-Learning* strategies, which train a hypernetwork to predict gradient updates for knowledge insertion (De Cao et al., 2021; Mitchell et al.; Tan et al.); and 2) the *Locate-then-Edit* approach, which directly identifies and rewrites factual weights in specific layers (Meng et al., 2022; Meng et al.), and has recently evolved variants optimized for sequential editing (Fang et al., 2024; Jiang et al., 2025). Finally, **Parameter-Preserving Editing** maintains the base model's integrity by either augmenting it with external components (Hartvigsen et al., 2023; Mitchell et al., 2022; Wang et al., 2024) or using retrieval during inference (Zheng et al., 2023a; Song et al., 2024; Shi et al., 2024; Chen et al., 2024) to store new facts.

2.1 Knowledge Editing Benchmarks

Current evaluations of KE methods primarily focused on general-domain benchmarks, such as WikiData (Meng et al., 2022) and counterfactual datasets (Levy et al., 2017), which aimed to update LLMs with new knowledge that contradicted common facts. However, recent studies have questioned the real-world applicability of these methods and proposed new, more rigorous benchmarks for fair evaluations. Specifically, Cohen et al. (2024) measured "ripple effects" on related facts and revealed that most KE methods often fail to propagate consistent changes beyond the target triple. Huang et al. (2024) reported that prior benchmarks do not strictly confirm LLMs having hallucinated answers before editing, which masks their true editing performance. Ma et al. (2024b) investigated editing consistency under prompt rephrasing and realistic contexts, finding that current KE methods exhibit lower generalization, and popular facts are the hardest to edit. Furthermore, Lin et al. (2024) examined sequential editing and showed that KE methods experience performance drops after sequential edits, while Yang et al. (2025b) critiqued common evaluation practices, revealing that KE methods catastrophically fail on realistic QA tasks.

In contrast to previous evaluations focused on simplified general-domain settings (Ma et al., 2024b; Lin et al., 2024), we present the first medical knowledge editing evaluation framework `MedEditBench`, which not only rigorously measures the effective adoption of injected medical knowledge in new clinical scenarios without degrading previous knowledge but also systematically assesses the effects of sequential edits across internal knowledge, external knowledge, and model general abilities.

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). This modification aims to improve 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 in the rest set $\mathcal{Q}_{\text{ori}}$. Thus, the accuracy on $\mathcal{Q}_{\text{ori}}$ indicates the *efficacy* of editing methods in injecting new knowledge. To evaluate *generalization* and *retention*, we construct two auxiliary question sets: $\mathcal{Q}_{\text{gen}}$, which extends original facts into new clinical scenarios, and $\mathcal{Q}_{\text{ret}}$, which introduces off-topic questions to assess the preservation of unaffected knowledge. Examples of $\mathcal{Q}_{\text{ori}}$, $\mathcal{Q}_{\text{gen}}$, and $\mathcal{Q}_{\text{ret}}$ are shown in Figure 9.

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 using *Reference Check* to verify whether the provided context c logically entails the ground-truth answer a for the question q. We prompt the LLM with (c, q) and retain the sample only if the predicted answer $\hat{a}$ 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 $\mathcal{Q}_{\text{gen}}$ using *Zero-Shot Filter* (i.e., LLM's explicit incorrectness serves as direct signal for selecting examples to be targeted for editing and generalization evaluation), while 100% on $\mathcal{Q}_{\text{ret}}$ to ensure the retention of original knowledge. Prompts and details for data construction and verification are provided in Appendix A.2.

4.2 Knowledge-driven Editing Paradigms

In `MedEditBench`, we propose three *editing paradigms* to evaluate the effectiveness of different

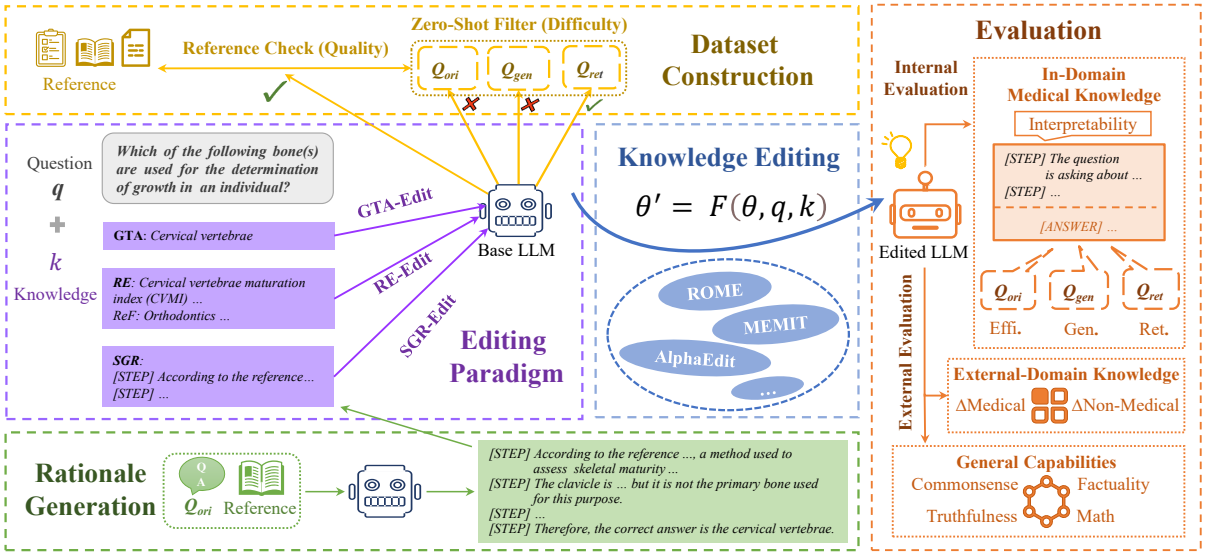

Figure 2: Overview of proposed medical knowledge editing evaluation framework (MedEditBench). Datasets are first constructed for medical knowledge editing (yellow rectangle). Three editing paradigms (GTA-Edit, RE-Edit, and SGR-Edit) are applied to evaluate the effectiveness of different target knowledge for medical knowledge editing (purple rectangle). In particular, the proposed SGR-Edit generates a rationale by prompting the LLM itself, given a QA and a reference (green rectangle). Then, the knowledge is used to edit the base LLM using editing methods like ROME (blue rectangle). In the final evaluation stage (orange rectangle), the edited LLM is evaluated using *efficacy*, *generalization*, and *retention*, as well as interpretability of the generated rationales. To examine how sequential medical edits affect the edited LLM's broader knowledge and general abilities, we complement MedEditBench with external evaluations, including multi-domain knowledge (MMLU), and commonsense reasoning (HellaSwag), factual recall (TriviaQA), truthfulness (TruthfulQA), and mathematical reasoning (GSM8K).

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. It provides a more comprehensive context to understand the underlying new knowledge.

Self-generated Rationale Editing (SGR-Edit) is motivated by the rational postulates of belief revision (Alchourrón et al., 1985), which emphasizes that revision must be consistent and closed under logical consequence. Moreover, prior work has argued that intelligent agents typically require an explanation to legitimize newly provided information, and the main role of an explanation is to rationalize facts (Falappa et al., 2002). Guided by these insights, 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, as shown in the green part of Figure 2. 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 the new fact. 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 LLM. 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 predicted answer offered by the edited LLM.

To further evaluate the *interpretability* of the

Method	Metric	Pre-Edit	LLaMA-8B		LLaMA-3B		Qwen-7B	
			$\text{MedExQA}_{\text{edit}}$	$\text{MedMCQA}_{\text{edit}}$	$\text{MedExQA}_{\text{edit}}$	$\text{MedMCQA}_{\text{edit}}$	$\text{MedExQA}_{\text{edit}}$	$\text{MedMCQA}_{\text{edit}}$
LoRA	Eff.	0	43.5	46.6	36.7	14.4	66.0	44.0
	Gen.	0	41.3	41.6	43.3	33.7	48.0	36.0
	Ret.	100	63.0	70.8	63.3	52.9	84.0	77.0
	avg.	\	49.3	53.0	**47.8**	33.7	**66.0**	**52.3**
ROME	Eff.	0	37.0	32.7	23.3	25.7	47.8	50.0
	Gen.	0	43.5	29.6	26.7	25.1	56.5	19.3
	Ret.	100	63.0	61.6	56.7	56.1	58.7	56.0
	avg.	\	47.8	41.3	35.6	35.7	54.3	41.8
MEMIT	Eff.	0	39.1	28.3	28.3	25.0	45.7	47.0
	Gen.	0	50.0	25.2	31.7	23.0	54.3	30.0
	Ret.	100	50.0	64.8	63.3	48.5	63.0	57.0
	avg.	\	46.4	39.4	41.1	32.2	54.3	44.7
GRACE	Eff.	0	34.8	36.0	33.3	26.7	39.1	23.0
	Gen.	0	21.7	29.2	10.0	7.0	43.5	32.0
	Ret.	100	76.1	80.7	88.3	93.0	63.0	69.0
	avg.	\	44.2	48.7	43.9	42.2	48.6	41.3
AnyEdit	Eff.	0	34.8	36.6	35.0	23.0	39.1	26.0
	Gen.	0	28.3	25.5	28.3	23.0	52.2	27.0
	Ret.	100	78.3	78.9	71.7	78.1	73.9	83.0
	avg.	\	47.1	47.0	45.0	41.4	55.1	45.3
AlphaEdit	Eff.	0	47.8	43.9	35.0	32.6	60.0	51.0
	Gen.	0	32.6	31.2	31.7	27.8	56.7	42.0
	Ret.	100	76.1	86.7	73.3	77.5	66.7	64.0
	avg.	\	**52.2**	**53.9**	46.7	**46.0**	61.1	**52.3**

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.

edited model, we also instruct the edited LLM to provide a rationale for the final answer, which is compared with the ground-truth interpretation using ROUGE-L (Lin, 2004), BLEU scores (Papineni et al., 2002), and Quality of Reasoning (QoR) scored via LLM-as-a-Judge. 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) and **MEMIT** (Meng et al.), and *Parameter-Preserving* methods - **GRACE** (Hartvigsen et al., 2023). Moreover, we include two newer *Parameter-Modifying* variants: **AnyEdit** (Jiang et al., 2025) and **AlphaEdit** (Fang et al., 2024), which are state-of-the-art approaches supporting long-form and continual knowledge editing, respectively. Details of the KE methods are provided in Appendix B.

4.5 Experimental Setup

We evaluate KE methods on three LLMs: LLaMA-3.1-8B-Instruct, LLaMA-3.2-3B-Instruct[2], and Qwen2.5-7B-Instruct[3]. We utilize layers [4, 5, 6, 7, 8] for editing based on our findings from detailed experiments in Appendix D.1. Detailed experimental settings are provided in Appendix C.

[2]https://huggingface.co/meta-llama
[3]https://huggingface.co/Qwen

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 KE methods perform in the medical domain?
RQ2: How do different KE paradigms impact the editing performance?
RQ3: How does sequential medical KE affect the reliability and stability of in-domain knowledge?
RQ4: Whether sequential medical KE affects broader domain knowledge?
RQ5: How does sequential medical KE affect general LLM capabilities?

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 $\mathcal{Q}_{\text{ori}} \cup \mathcal{Q}_{\text{gen}} \cup \mathcal{Q}_{\text{ret}}$. From the results in Table 1, we observe that:

RQ1-F1: No existing editing method is effective enough for medical settings. The *efficacy* of nearly all methods is below 50%. This sharply contrasts with previous reports of over 90% efficacy

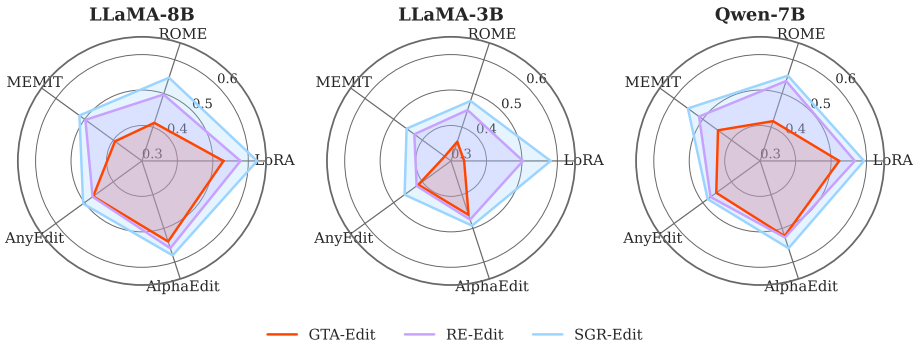

Figure 3: Medical knowledge editing with various editing paradigms.

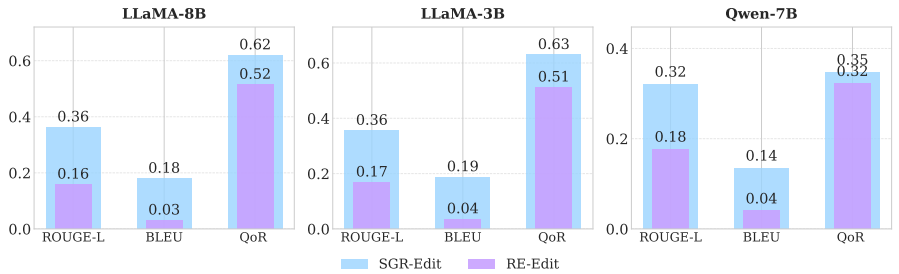

Figure 4: Interpretability comparison for SGR-Edit and RE-Edit. Quality of Reasoning (QoR) is a normalized (0–1) score via an AI judge (DeepSeek-V3 (Liu et al., 2024)), assessing the generated rationales across five dimensions: factual accuracy, logical flow, relevance, completeness, and answer correctness (see prompt in Figure 14).

in general domain benchmarks (Meng et al., 2022; Fang et al., 2024), highlighting a significant gap when applying current KE methods to the intricate medical domain.

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 $\text{MedMCQA}_{\text{edit}}$, 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 medical knowledge**. Detailed case analyses can be found in Appendix E.

5.2 Evaluation of Editing Paradigms (RQ2)

This section investigates how different editing paradigms (i.e., GTA-Edit, RE-Edit, and SGR-Edit), affect the performance of existing editing methods. To ensure the factual grounding of the rationales generated by the base LLM, we performed a targeted evaluation on a subset to assess rational factuality through a rigorous, two-stage process of AI-judge scoring and human verification (see Appendix D.2). 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 KE methods across various base LLMs. Overall, SGR-Edit boosts performance by an average of 7.7% relative to GTA-Edit on LLaMA-8B, by 11.3% on LLaMA-3B, and by 6.3% on Qwen-7B, demonstrating its consistently stronger capability across different model sizes and architectures. Importantly, SGR-Edit achieves these gains with only one additional reasoning-generation step, making it both highly effective and readily deployable. We discussed why SGR-Edit is practical in real-world employment in Appendix D.3. We also experimented with externally generated rationales in Appendix D.4. The complete results and further analysis of various editing paradigms are provided in Appendix D.5.

RQ2-F2: SGR-Edit enables better reasoning interpretability. We compare the knowledge quality enabled by SGR-Edit and RE-Edit using

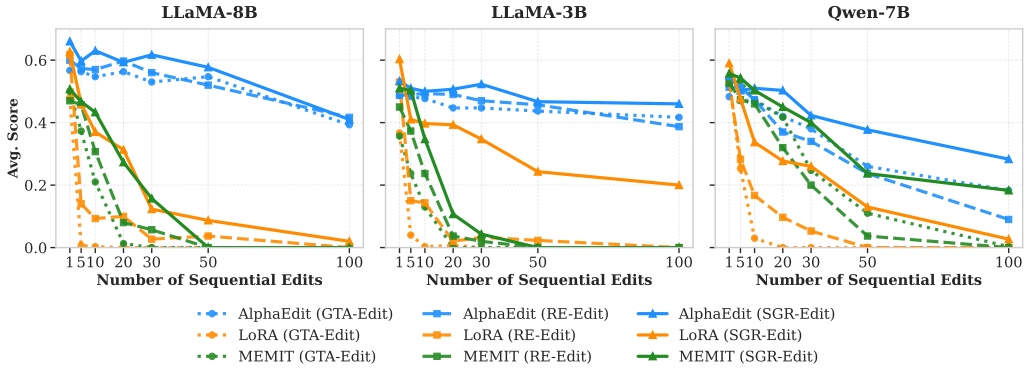

Figure 5: Sequential editing performance on the internal benchmark.

MEMIT in Figure 4. SGR-Edit consistently and substantially outperforms RE-Edit across all models and metrics, confirming that using the rationale-derived knowledge for editing results in deeper knowledge internalization. Full results for three KE methods are detailed in Appendix D.6.

5.3 Sequential Editing Impacts on In-Domain Knowledge (RQ3)

We investigate how sequential editing affects the in-domain knowledge in LLMs, where knowledge items are continuously updated through multiple editing steps. Performance is tracked over {1, 5, 10, 20, 30, 50, 100} sequential edits using samples from `MedEditBench`, as shown in Figure 5. We report our findings from three perspectives: **model-wise**, **method-wise**, and **paradigm-wise**:

RQ3-F1: Qwen experiences the most severe degradation under sequential edits, while LLaMA models exhibit greater stability. All models experience gradual declines as the number of edits increases. However, the performance drop on Qwen-7B is substantially larger than that on LLaMA-8B or LLaMA-3B.

RQ3-F2: Despite strong single-edit performance, LoRA suffers the sharpest decline under sequential edits, whereas AlphaEdit remains the most stable. Across all models and paradigms (best viewed in color), LoRA quickly collapses after several edits, showing near-zero accuracy after 50–100 updates. MEMIT also degrades rapidly, while AlphaEdit sustains moderate performance even after 100 edits, highlighting its robustness in sequential editing.

RQ3-F3: Among editing paradigms, SGR-Edit is the most resistant to sequential degradation. The rationale-driven SGR-Edit paradigm consistently preserves higher accuracy over repeated updates compared with GTA-Edit and RE-Edit (best viewed in line and marker). This indicates that SGR-Edit enables deeper knowledge integration and reduces catastrophic forgetting during continuous medical knowledge updates.

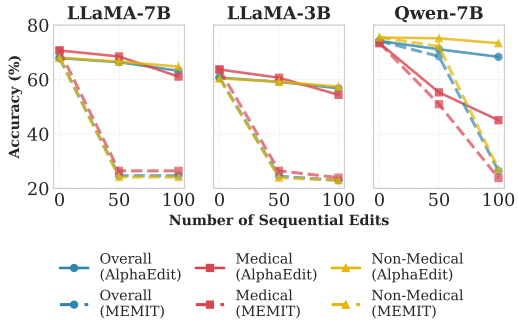

Figure 6: Sequential editing performance on the external MMLU benchmark.

5.4 Sequential Editing Impacts on External Domain Knowledge (RQ4)

After assessing internal medical edits, we further investigate their influence on external-domain knowledge, covering both *medical* and *non-medical* aspects. We split the MMLU benchmark accordingly (see Appendix D.8 for subject mapping) to examine whether sequential medical editing propagates effects beyond the edited domain. Observations from Figure 6 are:

RQ4-F1: Sequential medical edits consistently degrade both medical and non-medical knowledge from the external domain. After 100 sequential edits using AlphaEdit on the internal `MedEditBench`, accuracy drops substantially not only in the medical domain from the MMLU bench-

mark — especially for Qwen-7B (-28.3%) compared to LLaMA-8B (-9.5%) and LLaMA-3B (-9.3%), but also in non-medical categories (-3.3% on LLaMA-8B and -3.1% on LLaMA-3B). This indicates that sequential editing in internal medical knowledge erodes broader knowledge.

RQ4-F2: Advanced editing method mitigates external knowledge deterioration. While both KE methods degrade performance, AlphaEdit consistently retains higher general-domain accuracies, as indicated by the solid lines, demonstrating that sequential optimization from the AlphaEdit mitigates catastrophic forgetting across domains. We report category-wise results on MMLU general-domain tasks in Appendix Table 8. Further detailed comparison of different editing paradigms (GTA-Edit, RE-Edit, and SGR-Edit) and SGR edits with varying target knowledge lengths is presented in Table 9. We find that the primary driver of broader domain knowledge deterioration is the KE method (e.g., MEMIT's severe drop vs. AlphaEdit's minor drop), rather than the specific editing paradigm. Crucially, **increasing the target knowledge length in SGR-Edit does not cause significantly greater external knowledge degradation**.

5.5 Sequential Editing Impacts on General LLM Capabilities (RQ5)

Table 2: Comparison of sequential editing performance on general benchmarks (AlphaEdit vs. MEMIT).

Model	Benchmark	# Edits (AlphaEdit)			# Edits (MEMIT)		
		0	50	100	0	50	100
LLaMA-8B	HellaSwag	79.2	78.8	78.5	79.2	26.2	25.9
	TriviaQA	51.9	50.1	48.4	51.9	0.0	0.0
	TruthfulQA	54.1	55.1	54.9	54.1	47.6	47.9
	GSM8K	77.9	76.7	76.5	77.9	0.0	0.0
LLaMA-3B	HellaSwag	70.4	70.7	70.0	70.4	26.2	26.5
	TriviaQA	33.8	31.6	27.6	33.8	0.0	0.0
	TruthfulQA	49.7	49.9	49.9	49.7	51.5	48.2
	GSM8K	64.6	65.4	62.2	64.6	0.0	0.0
Qwen-7B	HellaSwag	80.4	80.0	80.0	80.4	79.5	77.6
	TriviaQA	32.4	33.7	35.3	32.4	27.7	21.8
	TruthfulQA	64.7	63.6	63.2	64.7	63.4	56.5
	GSM8K	81.9	82.1	83.6	81.9	65.9	1.1

Building on our previous findings that SGR-Edit consistently outperforms its counterparts in both single and sequential medical editing, we further examine whether the edited LLMs can preserve their general capabilities. We evaluate them on four representative benchmarks: HellaSwag (commonsense reasoning), TriviaQA (factual recall), TruthfulQA (truthfulness) and GSM8K (mathematical reasoning).

RQ4-F1: Sequential-based KE method largely preserves general LLM capabilities, whereas the traditional KE method causes severe degradation and even model collapse. As shown in Table 2, the enhanced sequential editing AlphaEdit preserves most general capabilities even after 100 sequential edits, remaining a more reliable solution for sustained medical knowledge updating while preserving the original LLM capabilities. In contrast, MEMIT exhibits drastic degradation—its edited LLaMA models collapse completely on knowledge- and reasoning-intensive tasks (TriviaQA and GSM8K drop to 0.0). Upon inspection, these edited LLaMA models produce unreadable or nonsensical outputs (e.g., random punctuation), a failure pattern also reported in prior literature (Gu et al.; Yang et al., 2024). This model collapse phenomenon is also observed in different KE paradigms using MEMIT based on LLaMA (see Appendix D.10). We highlight and report this severe issue in the *medical* context, where such instability poses greater risks for real-world deployment.

6 Conclusion

This study introduces `MedEditBench`, a rigorous evaluation framework for medical knowledge editing designed to measure the reliability and generalization of updated knowledge in medical contexts. Our comprehensive evaluation reveals that the prevailing Knowledge Editing (KE) paradigm editing only simple final answers often results in superficial memorization, failing to foster the deep internalization of medical principles necessary for generalization. To address this limitation, we propose the novel KE paradigm called SGR-Edit, which targets the LLM's evidence-based reasoning chain instead of just the final answer. Extensive experiments demonstrate that SGR-Edit significantly boosts KE performance. Furthermore, our systematic analysis of sequential editing shows that the advanced KE method can effectively mitigate the degradation of broader knowledge and general abilities. These findings validate SGR-Edit as a robust KE strategy and provide crucial practical guidance for the sustainable deployment of explainable KE in real-world applications.

Limitations and Future Work

We acknowledge the limitations of this study:

- **Model Diversity**. We only evaluate two LLaMA-3 variants (3B, 8B) and Qwen2.5-7B in this study. Since our framework applies to any LLM by targeting its failure cases, broader validation on diverse and reasoning-enhanced models (e.g., Qwen3 (Yang et al., 2025a)) is left for future work.

- **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

Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2):510–530.

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

Qizhou Chen, Taolin Zhang, Xiaofeng He, Dongyang Li, Chengyu Wang, Longtao Huang, and Hui Xue. 2024. 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.

Marcelo A Falappa, Gabriele Kern-Isberner, and Guillermo R Simari. 2002. Explanations, belief revision and defeasible reasoning. *Artificial Intelligence*, 141(1-2):1–28.

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*.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue, 2024. *URL https://arxiv. org/abs/2401.04700*.

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.

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*.

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. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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*.

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*.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

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.

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

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024. The butterfly effect of model editing: Few edits can trigger large language models collapse. *arXiv preprint arXiv:2402.09656*.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. 2025b. 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. 2023a. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

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 (Pal et al., 2022). We construct a subset for $\text{MedMCQA}_{\text{edit}}$ by selecting up to 300 QA pairs from each specialty. After filtering out entries with missing explanations, multi-correct choices, or unknown subject labels, we obtain a high-quality set used to construct $\text{MedMCQA}_{\text{edit}}$.

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 $\text{MedExQA}_{\text{edit}}$.

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* $\mathcal{D}_{\mathrm{verified}}$.

Step 2. Zero-Shot Filtering Next, we assess this LLM's zero-shot performance on each question in $\mathcal{D}_{\mathrm{verified}}$ 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* $\mathcal{Q}_{\mathrm{ori}}$.

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 $\mathcal{D}_{\mathrm{verified}}$, 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 $\mathcal{Q}_{\mathrm{gen}}$ and $\mathcal{Q}_{\mathrm{ret}}$.

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 $\mathcal{Q}_{\mathrm{gen}}$; As for retention question candidates, if $\hat{k} = k'$, we add this pair to the retention set $\mathcal{Q}_{\mathrm{ret}}$. This ensures that $\mathcal{Q}_{\mathrm{gen}}$ contains only those scenarios that the LLM cannot solve without editing, while $\mathcal{Q}_{\mathrm{ret}}$ captures instances where its original knowledge remains intact.

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

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 $\mathrm{MedMCQA}_{\mathrm{edit}}$ and $\mathrm{MedExQA}_{\mathrm{edit}}$ 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, $\mathcal{Q}_{\mathrm{gen}}$ and $\mathcal{Q}_{\mathrm{ret}}$ 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.

Steps 1 and 2 in Appendix A.2 are closely tied to how we interpret our evaluation metrics. If a base LLM were to answer all questions correctly or incorrectly, one of the evaluation sets would be empty, hindering a complete assessment of both efficacy and retention. In practice, when an LLM shows extreme pre-editing performance on a specific dataset or domain, we must dynamically adjust our focus:

- **If an LLM answers perfectly before editing, the evaluation would primarily focus on retention**, which measures how well the original knowledge is retained after editing, while efficacy becomes less critical as the target questions are already answered correctly.
- Conversely, **for a poorly performing LLM, efficacy becomes the dominant metric, focusing on the model's ability to correct its mistakes**.

This adaptive interpretation ensures a meaningful assessment across all performance spectra.

While our primary objective is to provide an efficient and automated framework for knowledge editing, we suggest that for scenarios demanding excep-

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

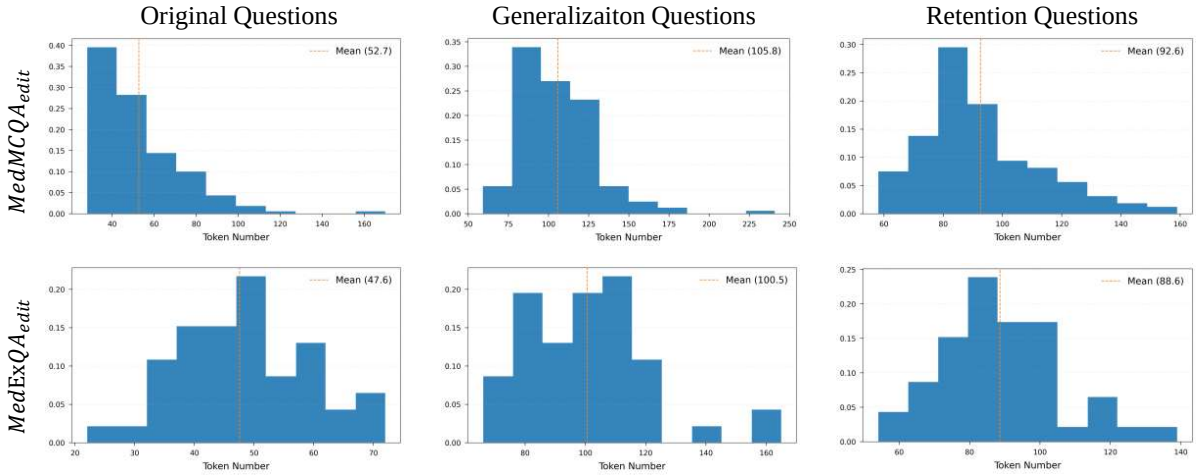

Figure 8: Token length distributions for $\text{MedMCQA}_{\text{edit}}$ (top row) and $\text{MedExQA}_{\text{edit}}$ (bottom row) across Original, Generalization, and Retention question sets. Dashed lines indicate dataset means. Our medical QA benchmarks present substantially longer inputs, increasing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

tionally high precision in sensitive real-world deployments, introducing "human-in-the-loop" verification for an additional layer of consistency checking would indeed be a beneficial and logical extension.

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

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.

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., 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., 2024) deploys side FFN layers to incorporate new knowledge dynamically.

Additionally, retrieval-based methods (Song et al., 2024; Shi et al., 2024; Chen et al., 2024) 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., 2023a) 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 [4]. We utilize layers [4, 5, 6, 7, 8] for editing based on the findings from Appendix D.1.

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[5]. For AnyEdit, which is not yet supported by EasyEdit, we integrate the original codebase[6] and adopt the original hyperparameters. In terms of evaluation, we implement independent pipelines and consistently compute metrics tailored

[4] https://llama.meta.com/lama3/
[5] https://github.com/zjunlp/EasyEdit
[6] https://github.com/jianghoucheng/AnyEdit

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.

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. For locate-and-edit KE methods, we select layers 4-8 as the target knowledge editing layers for all subsequent experiments (see Appendix D.1 for a detailed analysis). For more specific experimental settings, please see our open-source codes.

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:** model is edited through a ground-truth answer, with option letters stripped (e.g., remove "D:" in "D: Masson trichrome").
- **RE-Edit:** editing the model with 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. The prompt used for rationale generation is in Figure 10.

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 $\mathcal{Q}_{\text{ret}}$.

Notation. For any set $\mathcal{S}$, we write $|\mathcal{S}|$ for its cardinality (i.e., the number of elements in $\mathcal{S}$).

Post-edit Accuracy Metrics

$$\text{Efficacy} = \frac{1}{|\mathcal{Q}_{\text{ori}}|} \sum_{i \in \mathcal{Q}_{\text{ori}}} \mathbf{1}(\hat{a}_i = a_i), \quad (2)$$

$$\text{Generalization} = \frac{1}{|\mathcal{Q}_{\text{gen}}|} \sum_{i \in \mathcal{Q}_{\text{gen}}} \mathbf{1}(\hat{a}_i = a_i), \quad (3)$$

$$\text{Retention} = \frac{1}{|\mathcal{Q}_{\text{ret}}|} \sum_{i \in \mathcal{Q}_{\text{ret}}} \mathbf{1}(\hat{a}_i = a_i). \quad (4)$$

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. Furthermore, we adopt Quality of Reasoning (QoR) as a normalized (0–1) score via an AI judge (DeepSeek-V3), assessing the generated rationales across five dimensions: factual accuracy, logical flow, relevance, completeness, and answer correctness (see prompt in Figure 14) 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:

$$\theta^{(i)} = F\big(\theta^{(i-1)},\, q_i,\, k_i\big), i = 1, 2, \ldots, t \quad (5)$$

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.

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.

LLM-as-a-judge prompt for Rationale Factuality and Hallucination Risk evaluations:

You are an expert evaluator tasked with assessing the factual accuracy of reasoning provided for medical questions. You will evaluate how well the reasoning aligns with the provided reference evidence and identify any potential hallucinations or factual inaccuracies.

Evaluation Criteria:
1. Factual Alignment: Does the reasoning accurately reflect the facts stated in the reference evidence?
2. No Hallucination: Are all medical claims, data, and statements supported by the reference evidence?
3. Logical Consistency: Does the reasoning logically connect the reference facts to the conclusion?
4. Completeness: Does the reasoning appropriately use the available reference information?
5. Accuracy: Are there any factual errors or misrepresentations of the reference material?

Inputs:
- Question: The medical question being answered
- Ground Truth Answer: The correct answer to the question
- Reference Evidence: The factual evidence that should support the reasoning
- Generated Reasoning: The chain-of-thought reasoning to be evaluated

Output Format (Json):
{
"explanation": "[Brief explanation of the factual accuracy, noting any hallucinations, misalignments with reference, or strengths in factual grounding]",
"score": [1-5],
"hallucination_detected": [true/false]
}
"""

Figure 11: LLM-as-a-judge prompt for Rationale Factuality and Hallucination Risk evaluations

D Supplementary Results and Analyses

D.1 Localization of Medical Knowledge in LLMs

This section investigates the layer-wise storage of medical knowledge in LLMs. We edit knowledge across four disjoint layers to check the storage of certain knowledge: For LLaMA-8B, we select layers 4–8, 11–15, 18–22, and 25–29; for LLaMA-3B and Qwen-7B, we use layers 4–8, 10–14, 16–20, and 22–26. We sample 100 QA pairs from $\text{MedMCQA}_{\text{edit}}$ 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 12, we draw the following findings:

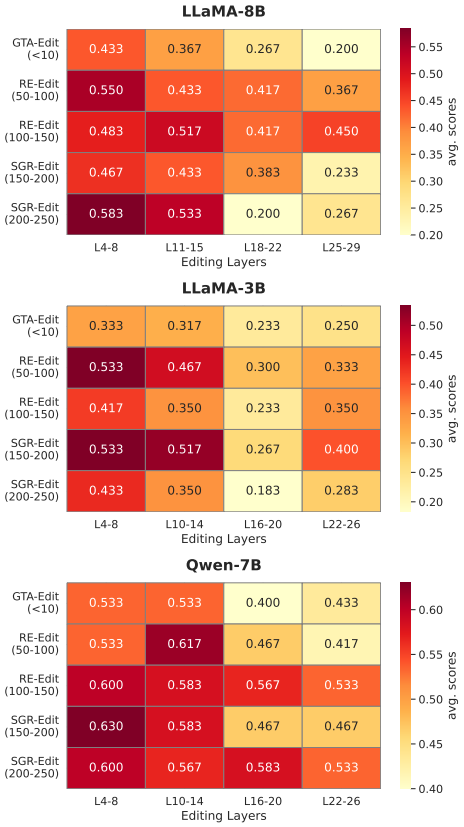

Figure 12: Layer-Wise editing performance

Medical knowledge is primarily stored in shallower 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. Despite Qwen-7B also achieving strong editing effectiveness in deep layers (all above 40%), the most effective network layers for it still reside in the shallower layers.

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.

Based on the above experiments and analysis, we adopt the shallower layers 4-8 as the knowledge editing layers for our experiments.

D.2 Factual Grounding and Rationale Verification

To ensure the factual grounding of these rationales for both benchmark construction and the proposed editing paradigm, we performed a targeted evaluation on a subset to assess rational factuality. Our methodology involved two stages:

1. AI Judge Evaluation: We first utilized DeepSeek-V3 as an AI judge to evaluate the generated rationales (Zheng et al., 2023b). DeepSeek-V3 assessed the rationales across several dimensions, including Factual Alignment, Hallucination, Logical Consistency, Completeness, and Accuracy, ultimately assigning a factual score on a 1-5 scale and indicating whether hallucination was detected.

2. Human Verification: Recognizing that even an advanced AI judge might exhibit its own form of hallucination in knowledge-specific tasks (Szymanski et al., 2025), especially with complex medical knowledge, we further introduced three human verifiers. These human annotators meticulously reviewed a subset of the rationales to confirm their factual accuracy and presence of hallucination, validating the AI judge's assessment.

As observed from Table 4, the average factual scores are remarkably high (ranging from 4.91 to 5.0), indicating strong factual alignment and consistency. Crucially, the hallucination rates are extremely low, with our rigorous verification process (involving both the AI judge and human annotators) confirming the high factual consistency of the generated rationales. This highlights that LLMs can self-generate high-quality rationales to update their knowledge.

D.3 Event-Driven Rationale Generation for Practical SGR-Edit

In real-world scenarios, knowledge updates are always triggered by concrete events: in medicine, for

Subset	Samples	Avg. Factual Score	Halluc. Count	Halluc. Rate
$\text{MedExQA}_{\text{edit}}$ (LLaMA-3B)	60	4.93	1	1.67%
$\text{MedExQA}_{\text{edit}}$ (LLaMA-8B)	46	5.00	0	0.00%
$\text{MedExQA}_{\text{edit}}$ (Qwen-7B)	50	4.90	2	4.00%
$\text{MedMCQA}_{\text{edit}}$ (LLaMA-3B)	100	4.91	3	3.00%
$\text{MedMCQA}_{\text{edit}}$ (LLaMA-8B)	100	5.00	0	0.00%
$\text{MedMCQA}_{\text{edit}}$ (Qwen-7B)	100	4.99	0	0.00%

Table 4: Factual evaluation and hallucination analysis across different subsets and models. The prompt is provided in Figure 11

Method	Metric	LLaMA-8B		Qwen2.5-7B	
		SGR-Edit	EGR-Edit	SGR-Edit	EGR-Edit
ROME	Eff.	0.52	0.49	0.49	0.62
	Gen.	0.43	0.49	0.39	0.45
	Ret.	0.64	0.75	0.67	0.77
	avg.	0.530	**0.577**	0.517	**0.613**
MEMIT	Eff.	0.52	0.39	0.48	0.56
	Gen.	0.39	0.43	0.44	0.43
	Ret.	0.64	0.58	0.69	0.73
	avg.	**0.512**	0.467	0.537	**0.573**
AlphaEdit	Eff.	0.52	0.54	0.48	0.54
	Gen.	0.28	0.30	0.39	0.41
	Ret.	0.82	0.83	0.73	0.68
	avg.	0.540	**0.557**	0.533	**0.543**

Table 5: Editing performance comparison between SGR-Edit and ECG-Edit

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.4 External LLM for Rationale Generation

We conducted an evaluation using a subset (100 samples per LLM), comparing performance when rationales are self-generated (SGR-Edit) by the base LLM versus when they are externally generated (EGR-Edit) by DeepSeek-V3. We applied these rationale types to ROME, MEMIT, and AlphaEdit methods, testing on LLaMA-8B and Qwen2.5-7B models.

Table 5 shows that EGR-Edit generally yields higher average scores compared to SGR-Edit. Despite this, our primary design philosophy for SGR-Edit was driven by crucial practical considerations, especially for deployment in sensitive domains like medicine. *SGR-Edit's core strength lies in enabling the local base LLM to self-update and refine its knowledge. This approach avoids the inherent security and privacy risks associated with distilling knowledge from or uploading sensitive local medical data to external LLMs/agents in real-world practice.* By relying on SGR-Edit, we maintain a self-contained and secure knowledge editing framework, which is paramount in contexts where data privacy and compliance are non-negotiable.

D.5 Performance Comparison of Various Editing Paradigms

As shown in Table 6, 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), 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

Method	Metric	Pre-Edit	LLaMA-8B			LLaMA-3B			Qwen-7B		
			GTA-Edit	RE-Edit	SGR-Edit	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	0.440	0.540	0.590
	Gen.	0	0.416	0.460	0.503	0.337	0.439	0.492	0.360	0.430	0.440
	Ret.	1	0.708	0.727	0.708	0.529	0.658	0.636	0.770	0.730	0.750
	avg.	–	0.530	0.578	**0.625**	0.337	0.503	**0.581**	0.523	0.567	**0.593**
ROME	Eff.	0	0.327	0.453	0.528	0.257	0.299	0.439	0.500	0.420	0.590
	Gen.	0	0.296	0.403	0.447	0.251	0.406	0.385	0.193	0.510	0.320
	Ret.	1	0.616	0.635	0.667	0.561	0.647	0.610	0.560	0.680	0.750
	avg.	–	0.413	0.497	**0.547**	0.357	0.451	**0.478**	0.418	0.537	**0.553**
MEMIT	Eff.	0	0.283	**0.435**	0.520	**0.250**	**0.345**	0.455	0.470	0.480	0.575
	Gen.	0	0.252	**0.380**	0.364	**0.230**	**0.315**	0.401	0.300	0.400	0.360
	Ret.	1	0.648	**0.675**	0.671	**0.485**	**0.625**	0.663	0.570	0.660	0.720
	avg.	–	0.394	0.497	**0.518**	0.322	0.428	**0.506**	0.447	0.513	**0.552**
GRACE	Eff.	0	0.366	0.410	0.435	0.230	0.246	0.316	0.230	0.230	0.230
	Gen.	0	0.255	0.248	0.317	0.230	0.267	0.316	0.320	0.320	0.320
	Ret.	1	0.789	0.758	0.764	0.781	0.747	0.754	0.690	0.690	0.690
	avg.	–	0.470	0.472	**0.505**	0.414	0.420	**0.462**	0.413	0.413	**0.413**
AnyEdit	Eff.	0	0.366	0.410	0.435	0.230	0.246	0.316	0.260	0.380	0.370
	Gen.	0	0.255	0.248	0.317	0.230	0.267	0.316	0.270	0.160	0.240
	Ret.	1	0.789	0.758	0.764	0.781	0.747	0.754	0.830	0.880	0.840
	avg.	–	0.470	0.472	**0.505**	0.414	0.420	**0.462**	0.453	0.473	**0.483**
AlphaEdit	Eff.	0	0.439	0.547	0.584	0.326	0.348	0.374	0.510	0.520	0.595
	Gen.	0	0.312	0.335	0.366	0.278	0.332	0.348	0.420	0.360	0.362
	Ret.	1	0.867	0.795	0.789	0.775	0.743	0.754	0.640	0.700	0.721
	avg.	–	0.539	0.559	**0.580**	0.460	0.474	**0.492**	0.523	0.527	**0.559**

Table 6: Complete results of GTA-Edit, RE-Edit, and SGR-Edit across LLaMA-8B, LLaMA-3B, and Qwen-7B (Accuracy). SGR-Edit achieves the highest average score in nearly all comparisons.

of 2.1–7.1 pp over RE-Edit, with MEMIT (+7.1 pp) and LoRA (+4.7 pp) benefiting 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.

The observed trend of improvement from RE-Edit to SGR-Edit extends robustly to the Qwen-7B model. Across all methods on Qwen-7B, SGR-Edit achieves the highest average editing scores, demonstrating its superiority regardless of the base LLM architecture. Notably, ROME and MEMIT exhibit substantial gains, increasing their average scores by +13.5 pp and +10.5 pp from GTA-Edit to SGR-Edit, respectively, confirming that the enriched SGR target benefits knowledge integration in Qwen-7B as well.

Additionally, we note that the GRACE method exhibited an anomalous pattern across multiple trials: its performance remained nearly identical across all three editing paradigms (i.e., GTA-Edit, RE-Edit, and SGR-Edit). We speculate this is due to GRACE's reliance on a codebook's deferral radius, limiting edits to inputs near existing keys, which makes it less sensitive to the content variation. We leave this phenomenon for future investigation.

D.6 Interpretability Analysis of SGR-Edit and RE-Edit

The results in Figure 13 are computed over a human-validated $\text{MedMCQA}_{\text{edit}}$ subset of high-quality reference explanations (RE) and self-generated rationales (SGR). Since SGR generation can include spurious content, we manually verify each rationale (see Table 4) to ensure it faithfully supports the target answer before using it for ROUGE-L and BLEU calculations. 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.

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 dra-

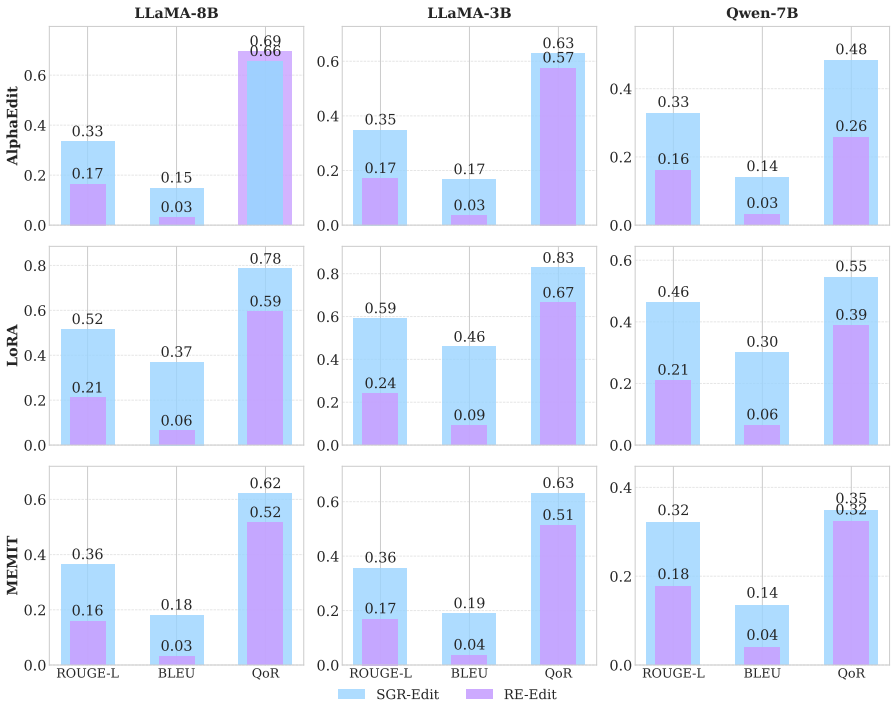

Figure 13: Interpretability comparison for SGR-Edit and RE-Edit using AlphaEdit, LoRA, and MEMIT. The metric **QoR** (Quality of Reasoning) represents a normalized score (0 to 1) obtained from an LLM-as-a-Judge (DeepSeek-V3) evaluation, assessing answer accuracy, medical factual accuracy, logical flow, relevance, and completeness.

matic: 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 rationale 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).

We further analyze the overall quality of the generated rationales using the Quality of Reasoning (QoR) Score. This is achieved by prompting the DeepSeek-V3 to evaluate output quality across several dimensions: Answer Accuracy, Medical Factual Accuracy, Logical Flow, Relevance, and Completeness (the prompt are provided in Figure 14). SGR-Edit generally yields a higher Quality Score than RE-Edit, suggesting that using the model's self-generated context results in more cohesive and contextually rich explanations.

In summary, these metrics corroborate our finding that SGR-Edit consistently yields the highest interpretability, validating its ability to convey a 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.7 Post-Edit Inference: Two-Step Rationale & Answer vs. One-Step Final Answer

We compare two prompting strategies on LLaMA-8B over $\text{MedMCQA}_{\text{edit}}$: (i) "Rationale + Answer" and (ii) "Final-Answer Only" (see Figure 15). 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

LLM-as-a-judge prompt for predicted rationale (Interpretability) evaluation:

You are an expert evaluator tasked with assessing the quality of reasoning provided for medical multiple-choice questions. You will evaluate how effectively the reasoning answers the question and the overall quality of the reasoning process.

Evaluation Criteria:
1. Answer Accuracy: Does the reasoning arrive at the correct answer?
2. Medical Factual Accuracy: Are the medical facts, data, and claims accurate and valid?
3. Logical Flow: Does each step logically follow from the previous? Are there unjustified leaps?
4. Relevance: Does the reasoning directly address the question and appropriately consider all options?
5. Completeness: Is the reasoning sufficiently detailed to justify the conclusion?

Inputs:
- Question: The medical multiple-choice question.
- Ground Truth Answer: The correct answer to the question.
- Generated Reasoning and Answer: The reasoning steps and final answer provided

Output Format (Json):
{
"explanation": "[Brief explanation of the reasoning quality, noting key strengths or flaws in medical accuracy, logical coherence, and how well it answers the question]",
"score": [1-5]
}

Figure 14: LLM-as-a-judge prompt for predicted rationale (Interpretability) evaluation

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

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

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 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,

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
Qwen-7B	0 (Raw)	69.01	–	77.51	–	83.14	–	73.99	–
	50	68.63	–0.38	77.23	–0.28	82.86	–0.28	63.57	–10.42
	100	65.92	–3.09	76.41	–1.10	80.74	–2.40	57.48	–16.51

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

supporting more reliable and interpretable medical knowledge editing.

D.8 Sequential Editing Impact on Common-Domain Capabilities

We split the MMLU benchmark into medical and non-medical categories: the medical sets are computed over the subjects {anatomy, clinical_knowledge, college_medicine, human_aging, medical_genetics, nutrition, professional_medicine, virology}, which are typically referred to as *health-related* in the MMLU community. For consistency with our medical evaluation setting, we collectively denote them as the **medical** subset, while all remaining subjects are aggregated as the **non-medical** subset.

In addition to the medical vs. non-medical 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 (e.g., –6.45 pp for LLaMA-8B; –5.86 pp for LLaMA-3B; -16.51 for Qwen-7B). 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 medical 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.

D.9 Sequential Editing Impact with different Length of Knowledge

To investigate whether the introduction of longer rationales using SGR-Edit would lead to more rapid degradation of general capabilities on MMLU, we conducted experiments using LLaMA-3B as our base LLM. We employed AlphaEdit and MEMIT as sequential editing methods, testing their impact with different rationale token lengths (200-250, 150-200, and 100-150) on MMLU performance after 0, 20, and 50 sequential edits.

The experimental results in Table 9 do not indicate that the introduction of longer rationales used for editing leads to a more rapid degradation of general capabilities on MMLU. In contrast, methods like MEMIT exhibit significantly higher degradation regardless of the rationale length, highlighting that **the degradation is more dependent on the editing method itself rather than solely on the rationale length.**

D.10 Supplementary Analysis: Model Collapse of MEMIT on LLaMA

To complement Table 2, we further examine the sequential editing results of **MEMIT** on **LLaMA-3B** across three knowledge editing paradigms (SGR-Edit, RE-Edit, and GTA-Edit).

As shown in Table 10, all paradigms exhibit a consistent pattern of **model collapse**: after 50 sequential edits, the model's performance on reasoning and factual benchmarks (*TriviaQA* and *GSM8K*) drops to zero, while even relatively stable tasks such as *HellaSwag* and *TruthfulQA* experience severe degradation. This indicates that the underlying parameter updates of MEMIT become unstable when repeatedly applied in a sequential editing scenario.

In contrast, results in Table 2 show that **Qwen-7B** does not exhibit full collapse under the same MEMIT procedure—although its accuracy also declines notably—suggesting that model size and

Method		#Edits	Overall	ΔOverall	Medical	ΔMedical	Non-Medical	ΔNon-Medical
pre-edit		0	60.7	–	63.7	–	60.5	–
AlphaEdit	SGR 200-250	50	59.2	**-1.5**	60.7	**-3.0**	59.3	**-1.2**
	SGR 150-200	50	58.1	**-2.6**	60.5	**-3.2**	58.4	**-2.1**
	SGR 100-150	50	59.6	**-1.1**	60.7	**-3.0**	59.8	**-0.7**
	GTA <10	50	60.0	**-0.7**	62.3	**-1.4**	59.9	**-0.6**
	RE 50-100	50	59.3	**-1.4**	61.0	**-2.7**	59.4	**-1.1**
	RE 100-50	50	59.0	**-1.7**	60.4	**-3.3**	59.0	**-1.5**
	SGR 200-250	20	60.3	**-0.4**	63.1	**-0.6**	60.1	**-0.4**
	SGR 150-200	20	60.1	**-0.6**	62.1	**-1.6**	60.0	**-0.5**
	SGR 100-150	20	60.6	**-0.1**	62.8	**-0.9**	60.4	**-0.1**
	GTA <10	20	60.4	**-0.3**	64.1	**+0.4**	60.8	**+0.3**
	RE 50-100	20	60.6	**-0.1**	63.9	**+0.2**	60.3	**-0.2**
	RE 100-50	20	60.6	**-0.1**	63.4	**-0.3**	60.3	**-0.2**
MEMIT	SGR 200-250	50	25.5	**-35.2**	22.8	**-40.9**	25.8	**-34.7**
	SGR 150-200	50	25.7	**-35.0**	24.9	**-38.8**	26.2	**-34.3**
	SGR 100-150	50	25.4	**-35.3**	22.8	**-40.9**	25.8	**-34.7**
	GTA <10	50	25.9	**-34.8**	26.3	**-37.4**	26.2	**-34.3**
	RE 50-100	50	24.4	**-36.3**	26.2	**-37.5**	24.9	**-35.6**
	RE 100-50	50	26.5	**-34.2**	25.5	**-38.2**	24.3	**-36.2**
	SGR 200-250	20	54.8	**-5.9**	42.3	**-21.4**	57.0	**-3.5**
	SGR 150-200	20	51.4	**-9.3**	33.3	**-30.4**	54.7	**-5.8**
	SGR 100-150	20	54.9	**-5.8**	45.7	**-18.0**	57.1	**-3.4**
	GTA <10	20	48.2	**-12.5**	33.0	**-30.7**	50.6	**-9.9**
	RE 50-100	20	43.9	**-16.8**	27.4	**-36.3**	48.1	**-12.4**
	RE 100-50	20	54.6	**-6.1**	40.7	**-23.0**	56.9	**-3.6**

Table 9: Performance comparison of AlphaEdit and MEMIT for different knowledge lengths and number of edits on MMLU benchmark.

Paradigm	#Edits	HellaSwag	TriviaQA	TruthfulQA	GSM8K
SGR-Edit	0	70.4	33.8	49.7	64.6
	50	26.2	0.0	51.5	0.0
	100	26.5	0.0	48.2	0.0
RE-Edit	0	70.4	33.8	49.7	64.6
	50	26.6	0.0	48.0	0.0
	100	26.4	0.0	47.9	0.0
GTA-Edit	0	70.4	33.8	49.7	64.6
	50	30.5	0.0	49.8	0.0
	100	26.2	0.0	47.2	0.0

Table 10: Sequential editing performance of MEMIT on LLaMA-3B across different KE paradigms. Values are accuracies (%). Red numbers indicate cases of complete model collapse (accuracy 0.0%).

pretraining diversity partially buffer against catastrophic forgetting. Furthermore, the sequentially optimized **AlphaEdit** maintains substantially higher performance across all benchmarks, demonstrating its superior robustness and practicality for continual medical knowledge editing without eroding the model's general capabilities.

E Case Analysis

The case in Figure 16, using the MEMIT method to edit LLaMA-8B on $\text{MedMCQA}_{\text{edit}}$, 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 editing 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

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 16: **Case: Medical knowledge editing performance across various 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.

the knowledge reliably. Such examples highlight the importance of reasoning-grounded editing for domains like medicine, where interpretability and trustworthiness are critical.