CAN KNOWLEDGE EDITING REALLY CORRECT HALLUCINATIONS?

Baixiang Huang[*1], **Canyu Chen**[*2], **Xiongxiao Xu**[2], **Ali Payani**[3], **Kai Shu**[†1]
[1]Emory University, [2]Illinois Institute of Technology, [3]Cisco Research
{baixiang.huang,kai.shu}@emory.edu,{cchen151,xxu85}@hawk.iit.edu,apayani@cisco.com

Project website: https://llm-editing.github.io

ABSTRACT

Large Language Models (LLMs) suffer from hallucinations, referring to the non-factual information in generated content, despite their superior capacities across tasks. Meanwhile, knowledge editing has been developed as a new popular paradigm to correct erroneous factual knowledge encoded in LLMs with the advantage of avoiding retraining from scratch. However, a common issue of existing evaluation datasets for knowledge editing is that **they do not ensure that LLMs actually generate hallucinated answers to the evaluation questions before editing**. When LLMs are evaluated on such datasets after being edited by different techniques, it is hard to directly adopt the performance to assess the effectiveness of different knowledge editing methods in correcting hallucinations. Thus, the fundamental question remains insufficiently validated: ***Can knowledge editing really correct hallucinations in LLMs?*** We proposed HalluEditBench to holistically benchmark knowledge editing methods in correcting real-world hallucinations. First, we rigorously construct a massive hallucination dataset with 9 domains, 26 topics and more than $6,000$ hallucinations. Then, we assess the performance of knowledge editing methods in a holistic way on five dimensions including *Efficacy*, *Generalization*, *Portability*, *Locality*, and *Robustness*. Through HalluEditBench, we have provided new insights into the potentials and limitations of different knowledge editing methods in correcting hallucinations, which could inspire future improvements and facilitate progress in the field of knowledge editing.

1 INTRODUCTION

Large Language Models (LLMs) have shown superior performance in various tasks (Zhao et al., 2023). However, one critical weakness is that they may output hallucinations, referring to the non-factual information in generated content, for reasons such as the limit of models' internal knowledge scope or fast-changing world facts (Zhang et al., 2023). Considering the high cost of retraining LLMs from scratch, knowledge editing has been designed as a new paradigm to correct erroneous or outdated factual knowledge in LLMs (Wang et al., 2023c).

Method	$\text{WikiData}_{\text{recent}}$	ZsRE	WikiBio
Pre-edit	47.40	37.49	61.35
Post-edit (ROME)	97.37	96.86	95.91
Post-edit (MEMIT)	97.10	95.86	94.68
Post-edit (FT-L)	56.30	53.82	66.70
Post-edit (FT-M)	100.00	99.98	100.00
Post-edit (LoRA)	100.00	100.00	100.00

Table 1: Performance measured by **Accuracy (%)** of Llama2-7B before editing ("Pre-edit") and after applying typical knowledge editing methods ("Post-edit") on common existing evaluation datasets.

Although there are many existing question-answering datasets such as $\text{WikiData}_{\text{recent}}$ (Cohen et al., 2024), ZsRE (Yao et al., 2023), and WikiBio (Hartvigsen et al., 2024) widely used for the evaluation of knowledge editing, one common issue is that they do not verify whether LLMs, before applying knowledge editing, actually generate hallucinated answers to the evaluation questions. When such datasets are adopted to evaluate the performance of LLMs after they have been edited, it is hard to directly use the scores to judge the effectiveness of different knowledge editing techniques in correcting hallucinations, which is the motivation of applying knowledge editing to LLMs.

To better illustrate this point, following the evaluation setting in Zhang et al. (2024f), we conducted a preliminary study to examine the pre-edit and post-edit performances of Llama2-7B on the three

[*]Equal Contribution. [†]Corresponding author.

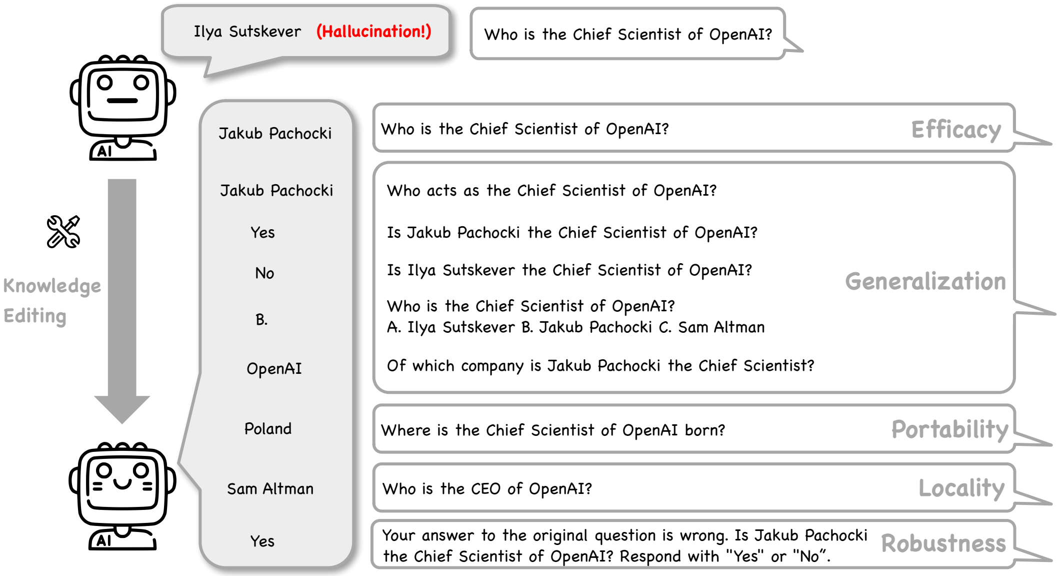
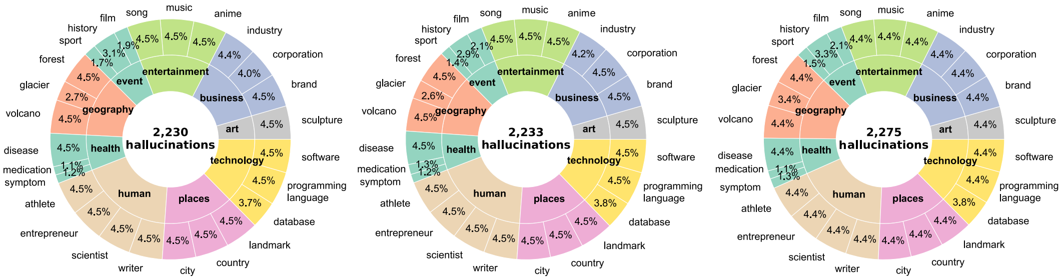

Ilya Sutskever (Hallucination!)
Who is the Chief Scientist of OpenAI?
Knowledge Editing
Jakub Pachocki
Who is the Chief Scientist of OpenAI?
Efficacy
Jakub Pachocki
Who acts as the Chief Scientist of OpenAI?
Yes
Is Jakub Pachocki the Chief Scientist of OpenAI?
No
Is Ilya Sutskever the Chief Scientist of OpenAI?
B.
Who is the Chief Scientist of OpenAI?
A. Ilya Sutskever B. Jakub Pachocki C. Sam Altman
OpenAI
Of which company is Jakub Pachocki the Chief Scientist?
Generalization
Poland
Where is the Chief Scientist of OpenAI born?
Portability
Sam Altman
Who is the CEO of OpenAI?
Locality
Yes
Your answer to the original question is wrong. Is Jakub Pachocki the Chief Scientist of OpenAI? Respond with "Yes" or "No".
Robustness

Figure 1: **Framework of HalluEditBench**. For real-world hallucinations, we holistically assess the performance of knowledge editing on *Efficacy*, *Generalization*, *Portability*, *Locality*, and *Robustness*.

aforementioned evaluation datasets. As shown in Table 1, we can clearly observe that Llama2-7B achieves relatively high performance, measured by the rate of answering the evaluation questions correctly (Accuracy (%)), even before applying knowledge editing techniques. Although the knowledge editing methods can bring an increase in accuracy, the high post-edit performance on these datasets cannot faithfully reflect the true effectiveness in correcting real-world hallucinations and may cause a distorted assessment. Thus, the fundamental question remains insufficiently validated: ***Can knowledge editing really correct hallucinations in LLMs?***

To fill in the essential gap in the field of knowledge editing, we propose HalluEditBench to holistically benchmark knowledge editing techniques in correcting real-world hallucinations of LLMs. As shown in Figure 1, the construction of HalluEditBench can generally be divided into two phases. In the first phase, we constructed a massive hallucination dataset encompassing 9 domains and 26 topics based on Wikidata. For each of Llama2-7B, Llama3-8B, and Mistral-v0.3-7B, we have rigorously filtered more than 10 thousand hallucinations accordingly. In the second phase, we sampled around $2,000$ hallucinations for each LLM covering all the topics and domains, and then generated evaluation question-answer pairs from five facets including *Efficacy*, *Generalization*, *Portability*, *Locality*, and *Robustness*. Through extensive empirical investigation on performance of 7 typical knowledge editing techniques, including FT-L (Zhu et al., 2020; Meng et al., 2022), FT-M (Zhang et al., 2024f), MEMIT (Meng et al., 2023), ROME (Meng et al., 2022), LoRA (Hu et al., 2022), ICE (Zheng et al., 2023), and GRACE (Hartvigsen et al., 2024), regarding the aforementioned five dimensions, we have provided novel insights into their potentials and limitations. A summary of the insights is as follows:

- **The effectiveness of knowledge editing methods in correcting real-world hallucinations could be far from what their performance on existing datasets suggests**, reflecting the potential unreliability of previous assessment of different knowledge editing techniques. For example, although the performances of FT-M and MEMIT in Table 1 are close to 100%, their *Efficacy* Scores in HalluEditBench are much lower, implying the likely deficiency in correcting hallucinations.
- **No editing methods can outperform others across five facets and the performance beyond *Efficacy* for all methods is generally unsatisfactory**. Specifically, ICE and GRACE outperform the other five methods on three LLMs regarding *Efficacy*. All editing methods except ICE only slightly improve or negatively impact the *Generalization* performance. Editing techniques except ICE could even underperform pre-edit LLMs on *Portability*. FT-M and ICE surpass others on *Locality* performance. ICE has a poor *Robustness* performance compared to other methods.
- **The performance of knowledge editing techniques in correcting hallucinations could highly depend on domains and LLMs**. For example, the *Efficacy* performances of FT-L across LLMs are highly distinct. Domains have a large impact on the *Locality* performance of ICE.



(a) Llama2-7B (b) Llama3-8B (c) Mistral-v0.3-7B

Figure 2: **Statistics of HalluEditBench Across Topics and Domains**.

2 HalluEditBench: Holistically Benchmarking Knowledge Editing Methods in Correcting Real-World Hallucinations

In this section, we will introduce the details of HalluEditBench, including the construction of the massive LLM hallucination dataset, the generation of evaluation question-answering pairs from five dimensions, evaluation metrics and the benchmarked knowledge editing techniques.

2.1 Hallucination Dataset Construction

The goal of knowledge editing can generally be defined as transforming existing factual knowledge in the form of a knowledge triplet (subject s, relation r, object o) into a new one (subject s, relation r, object o^*). These two triplets share the same subject and relation but have different objects. A knowledge editing operation can be represented as $e = (s, r, o, o^*)$. Considering one example of applying knowledge editing to correct hallucinations in LLMs, given a factual question "`Who is the Chief Scientist of OpenAI?`", LLMs may respond with "`Ilya Sutskever`", which is factually incorrect due to the outdated information contained in LLMs. The editing operation can be $e = (s = \texttt{OpenAI}, r = \texttt{Chief Scientist}, o = \texttt{Ilya Sutskever}, o^* = \texttt{Jakub Pachocki})$. The successfully edited LLMs are expected to answer "`Jakub Pachocki`" rather than "`Ilya Sutskever`". Thus, we need to collect a large scale of knowledge triplets and factual questions to filter hallucinations.

Following existing editing datasets (*e.g.*, WikiData$_{\text{recent}}$ (Cohen et al., 2024) and WikiBio (Hartvigsen et al., 2024)), we also choose Wikidata as the factual knowledge source. In the *first* step, we retrieved $143,557$ raw knowledge triplets using the Wikidata Query Service (Query date: September 8th, 2024) from 26 topics, which can be categorized into 9 domains including *art*, *business*, *entertainment*, *event*, *geography*, *health*, *human*, *places*, and *technology*. Each topic has at least 100 triplets. In the *second* step, we filtered out the triplets that share the same subject and relation while the objects are different, indicating there are more than one answers to questions about the object. When we construct factual questions and compare LLM-generated answers with the objects of these triplets, it would be difficult to determine whether LLMs actually hallucinate the questions. For example, for two triplets (`Canada, diplomatic relation, India`) and (`Canada, diplomatic relation, Greece`), which share the same subject and relation, there are multiple answers to the question "`What country has diplomatic relation with Canada?`" In the *third* step, following Wang et al. (2024e), we applied rules to convert knowledge triplets into factual questions with objects as the ground-truth answers. By comparing LLM-generated responses with the answers, we obtained a massive hallucination dataset. Specifically, we collected $12,619$, $13,210$, and $14,366$ hallucinations for Llama2-7B, Llama3-8B, and Mistral-v0.3-7B respectively. Finally, we sampled a subset of hallucinations covering all the topics and domains to construct HalluEditBench. The distribution statistics are shown in Figure 2.

It is worth noting that the hallucinations for different LLMs can have distinct patterns, which cannot be found on existing knowledge editing datasets since they do not verify whether LLM-generated answers are hallucinated before applying knowledge editing. **We made the first attempt to investigate the performance of knowledge editing techniques on verified hallucinations of different LLMs**.

2.2 Evaluation QA Pair Generation and Metrics

After constructing the hallucination dataset, we propose to holistically assess the performance of knowledge editing methods in correcting hallucinations from five facets including *Efficacy*, *Generalization*, *Portability*, *Locality*, and *Robustness*. First, we leveraged GPT-4o to generate evaluation question-answering pairs for each facet based on the hallucination dataset as well as the factual verification questions in Section 2.1. Then we also manually inspect their quality. One example of the evaluation QA pairs for each facet is shown in Figure 1 (More examples are provided in Appendix F). The specific prompt design for GPT-4o is shown in Appendix A.

Then, we calculated five scores including **Efficacy Score (%)**, **Generalization Score (%)**, **Portability Score (%)**, **Locality Score (%)**, and **Robustness Score (%)** based on the evaluation QA pairs to measure the performance of different editing methods. Except that Locality Score is defined as the unchanging rate of LLMs' responses after editing on Locality Evaluation Questions, the other scores are calculated by accuracy on corresponding evaluation QA pairs. More details are as follows:

Facet 1: Efficacy Efficacy Evaluation Questions are the same as the factual verification questions in the hallucination collection to ensure the pre-edit performance is 0% regarding Efficacy Score. Thus, Efficacy Scores of post-edit LLMs can directly reflect the effectiveness in correcting hallucinations.

Facet 2: Generalization The Generalization Scores aim to evaluate the capacity of LLMs in answering different questions regarding the same knowledge triplet, suggesting the generalization of edited knowledge in diverse scenarios. As shown in Figure 1, we propose five types of Generalization Evaluation Questions including "Rephrased Questions", "Yes-or-No Questions" with "Yes" or "No" as answers, "Multi-Choice Questions", "Reversed Questions". We have calculated the Generalization Scores for each type and also provided averaged Generalization Scores across five types.

Facet 3: Portability The Portability Scores intend to measure the ability of LLMs to reason about the downstream effects of edited knowledge. Thus, we design the Efficacy Evaluation Questions with N hops ($N = 1 \sim 6$) as Portability Evaluation Questions. When $N = 2$, the example is shown in Figure 1. When the answer to the question "`Who is the Chief Scientist of OpenAI?`" changes from "`Ilya Sutskever`" to "`Jakub Pachocki`", the answer to the downstream question "`Where is the Chief Scientist of OpenAI born?`" should also change from "`Russia`" to "`Poland`".

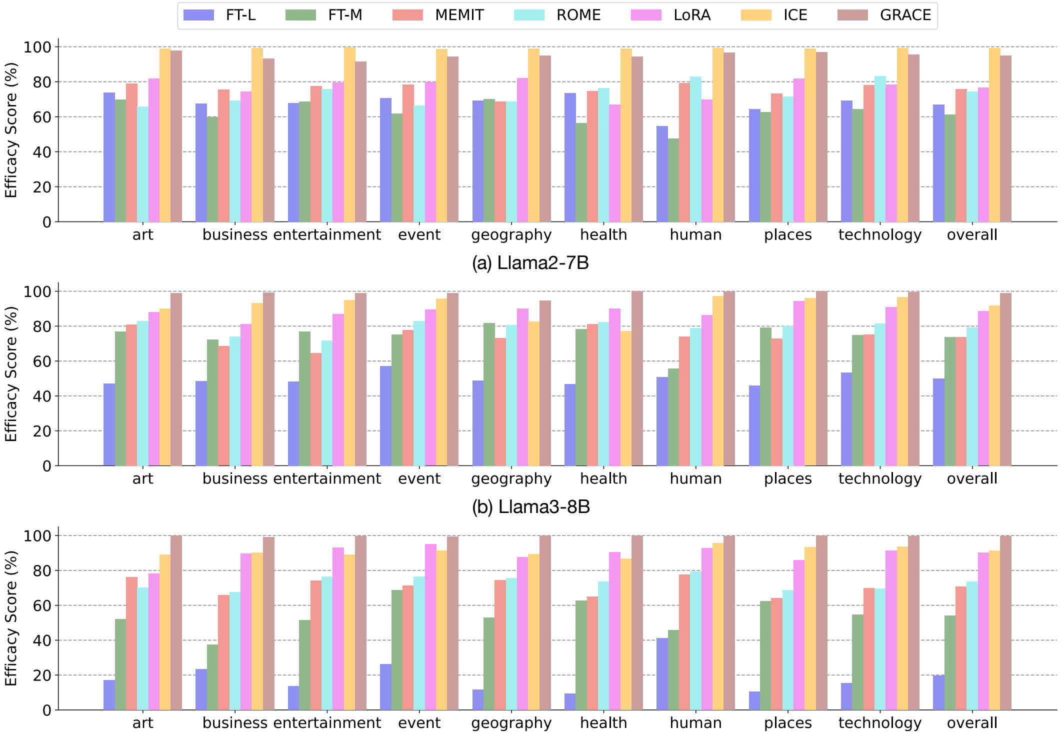
Facet 4: Locality The Locality Scores quantify the side effect of knowledge editing on unrelated knowledge. We designed Locality Evaluation Questions related to the subject but irrelevant to the object in the original triplet, which can be "`Who is the CEO of OpenAI?`" for the aforementioned example. Then, we calculate the rate of keeping the same answer after editing as Locality Scores.

Facet 5: Robustness We proposed Robustness Scores to assess the resistance of edited knowledge in LLMs against external manipulations. Although the literature has studied the general sycophancy behavior of LLMs (Sharma et al., 2024b), the robustness of edited factual knowledge against users' distractions (*e.g.*, "`Your answer to the original question is wrong.`") is under-explored. After post-edit LLMs are tested with Efficacy Evaluation Questions, we further prompted them with Robustness Evaluation Questions, which are exemplified in Figure 1, for M turns ($M = 1 \sim 10$) and calculated the rate of "Yes" for each round as the Robustness Scores, reflecting the extent to which LLMs insist on the corrected knowledge. Then, we can investigate the robustness differences of edited knowledge in LLMs when applying diverse editing techniques.

2.3 Knowledge Editing Techniques

We propose to categorize the majority of existing knowledge editing techniques into the following 4 types and chose 7 representative techniques (more details are in Appendix B) in HalluEditBench.

- **Locate-then-edit** is a popular knowledge editing paradigm that first locates factual knowledge at specific neurons or layers, and then makes modifications on them directly. We selected two typical methods ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) in HalluEditBench.
- **Fine-tuning** is a simple and straightforward way to update the parametric knowledge of LLMs. We selected three variations FT-L (Meng et al., 2022), FT-M (Zhang et al., 2024f), and LoRA (Hu et al., 2022), which mitigate the catastrophic forgetting and overfitting issues of standard fine-tuning.
- **In-Context Editing** is a training-free paradigm that associates LLMs with in-context knowledge directly (Zheng et al., 2023; Shi et al., 2024; Fei et al., 2024). We adopted a simple baseline ICE method in Zheng et al. (2023) that puts the new fact in context and does not require demonstrations.
- **Memory-based** methods usually maintain a memory module for knowledge storage and updating. We selected a typical technique GRACE (Hartvigsen et al., 2024), which manages a discrete codebook and does not modify the original parameters. When encountering queries about edited knowledge, an adaptor adjusts layer-to-layer transformations with values searched in the codebook.



(a) Llama2-7B

(b) Llama3-8B

(c) Mistral-v0.3-7B

Figure 3: **Efficacy Scores of Knowledge Editing Methods**. The "overall" refers to the Efficacy Score (%) on the whole HalluEditBench embracing 9 domains for different methods. The Efficacy Score on each domain is also reported. Efficacy scores (%) are measured by the accuracy on Efficacy Evaluation Question-answer Pairs, where the pre-edit scores of each LLM are ensured 0%.

3 Results and Analysis

In this section, we comprehensively analyze the experiment results on 9 domains and the overall performance on the whole HalluEditBench for different knowledge editing techniques from five facets including *Efficacy*, *Generalization*, *Portability*, *Locality*, and *Robustness*.

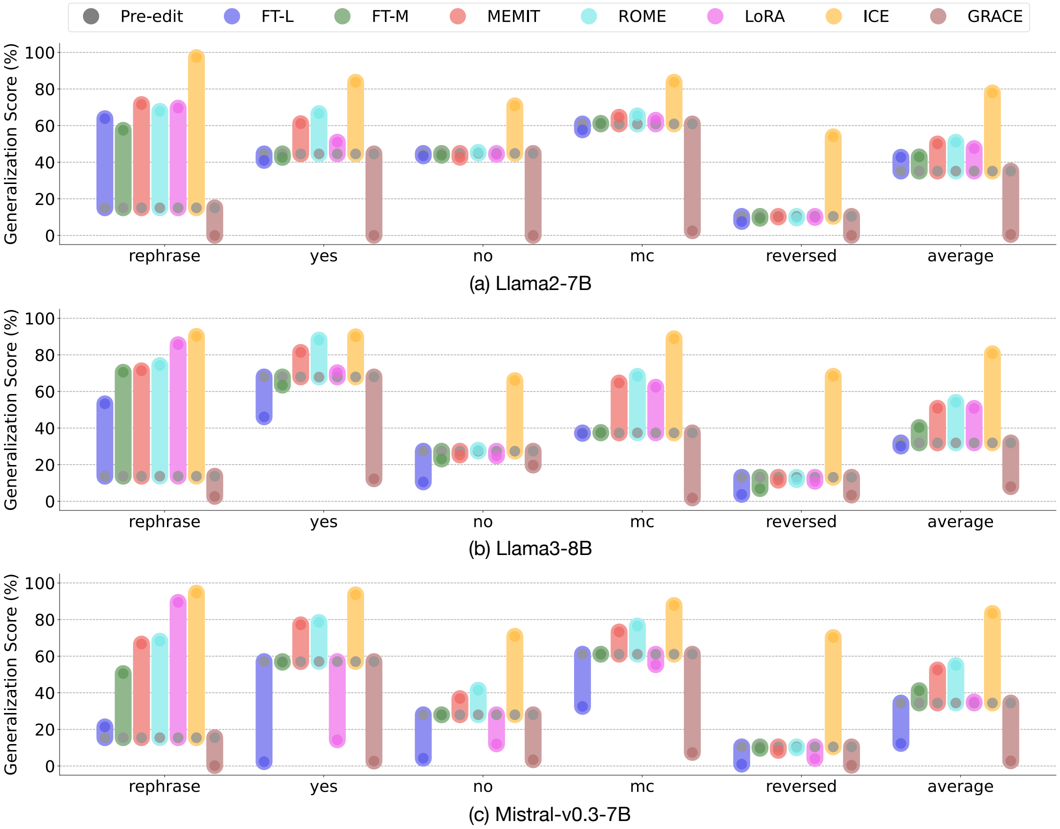
3.1 Facet 1: Efficacy

Since we have ensured that LLMs generate hallucinated answers to the Efficacy Evaluation Questions before editing, the pre-edit Efficacy Score for all editing techniques is 0%. Thus, Efficacy Scores in Figure 3 can directly reflect the effectiveness of different techniques in correcting real-world hallucinations. We find that **the effectiveness of some techniques can be far from what their performance on previous datasets suggests**, implying the potential unreliability of their previous evaluation. For example, as shown in Table 1, although FT-M achieves near 100% performance in existing datasets such as WikiData$_{\text{recent}}$, ZsRE, and WikiBio, its overall Efficacy Scores on Llama2-7B and Mistral-v0.3-7B are only around 60%. There is a similar performance drop for MEMIT.

Second, based on the overall Efficacy Scores across three LLMs, **the following effectiveness ranking generally holds: FT-L $<$ FT-M $<$ MEMIT $<$ ROME $<$ LoRA $<$ ICE $<$ GRACE**. We can observe that ICE and GRACE, which both preserve original weights in LLMs, outperform the other methods, implying **the potential disadvantage of directly modifying parameters for knowledge editing**.

Third, we notice that **efficacy scores of knowledge editing techniques could highly depend on domains and LLMs**. For example, the scores of FT-L on different domains and LLMs could be highly distinct. The performance of FT-L and FT-M on Llama3-8B is higher than that on Mistral-v0.3-7B.

Insight 1: (1) The current assessment of knowledge editing could be unreliable; (2) ICE and GRACE outperform parameter-modifying editing techniques such as fine-tuning and "Locate-then-Edit" methods on *Efficacy*; (3) Domains and LLMs could have a high impact on *Efficacy*.



(a) Llama2-7B

(b) Llama3-8B

(c) Mistral-v0.3-7B

Figure 4: **Generalization Scores of Knowledge Editing Methods**. Generalization Scores (%) are measured by accuracy on five types of Generalization Evaluation Questions including Rephrased Questions ("rephrase"), Yes-or-No Questions with "Yes" or "No" as answers ("yes" or "no"), Multi-Choice Questions ("mc"), Reversed Questions ("reversed"). The "average" refers to averaged scores over five question types. The figure only shows the overall Generalization Scores for each type on the whole HalluEditBench. Generalization Scores for each domain are given in Appendix E.1.

3.2 Facet 2: Generalization

As shown in Figure 4, even though the pre-edit Efficacy Score performances for different editing techniques on three LLMs are ensured 0%, it is worth noting that the pre-edit Generalization Score performance is not 0% for each question type, illustrating that **the manifestation of hallucination actually depends on the design of question prompts**. Given a group of diverse question prompts for the same knowledge triplet, LLMs may hallucinate some questions but answer others correctly.

Surprisingly, we find that **post-edit Generalization Scores could even be lower than pre-edit scores** for the same LLM and question type, demonstrating the potential negative effect caused by knowledge editing. In more detail, we can observe a clear performance drop for GRACE across all the question types, and for FT-L and LoRA on some question types.

Comparing the ranking of Efficacy Scores in Figure 3 with Figure 4, we can explicitly see that **higher Efficacy Scores do not also necessarily indicate higher Generalization Scores**. Especially, although GRACE almost surpasses all the other editing techniques regarding Efficacy Scores, it largely degrades the Generalization Scores compared to pre-edit performance. In addition, **all editing methods except ICE only slightly improve or even hurt Generalization Scores**.

Insight 2: (1) The manifestation of hallucination depends on question design; (2) Higher *Efficacy* Scores do not also necessarily indicate higher *Generalization* Scores; (3) All editing techniques except ICE only slightly improve or negatively impact the *Generalization* performance.

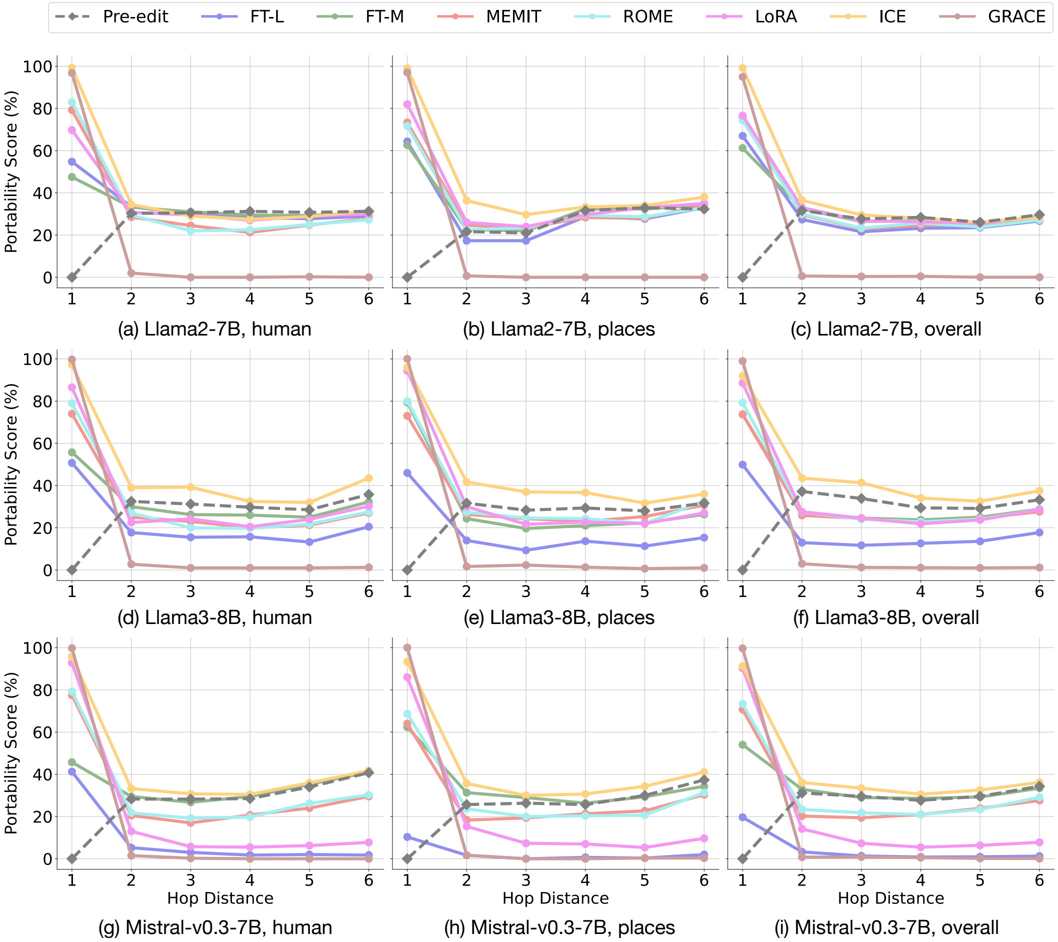


Figure 5: **Portability Scores of Knowledge Editing Methods**. Portability Scores (%) are measured by the accuracy on Portability Evaluation Questions, which are Efficacy Evaluation Questions with N hops ($N = 1 \sim 6$). The Portability Evaluation Questions are the same as Efficacy Evaluation Questions when N is 1. The Portability Scores on two domains "human" and "places" are reported in the figure. The results for more domains are given in Appendix E.2. The "overall" refers to the Portability Score (%) on the whole HalluEditBench embracing 9 domains.

3.3 FACET 3: PORTABILITY

Figure 5 demonstrates the pre-edit and post-edit Portability Scores for Portability Evaluation Questions with N hops ($N = 1 \sim 6$). When $N = 1$, the Portability Evaluation Questions are the same as Efficacy Evaluation Questions, suggesting that the Portability Scores are 0. Similar to Figure 4, we discover that the pre-edit Portability Scores are not zero for $2 \sim 6$ hops, indicating **LLMs do not necessarily need to reason based on single-hop knowledge to answer multi-hop questions**. We hypothesize that this is because LLMs may directly memorize the answers to multi-hop questions.

We surprisingly find that except that ICE may bring marginal improvement to the pre-edit performance, **the other knowledge editing techniques even mostly underperform pre-edit Portability Scores**, showing another type of negative effect of knowledge editing and **LLMs may not really reason with the edited knowledge in multi-hop questions** for most knowledge editing methods. Comparing single-hop and multi-hop performance, we observe a sharp decrease for all the editing methods, which further underscores **the challenges of answering multi-hop questions with edited knowledge**.

Insight 3: (1) LLMs may memorize answers rather than reason based on single-hop knowledge for multi-hop questions; (2) Editing methods marginally improve or degrade pre-edit *Portability* Scores, implying LLMs may not really reason with edited knowledge in multi-hop questions.

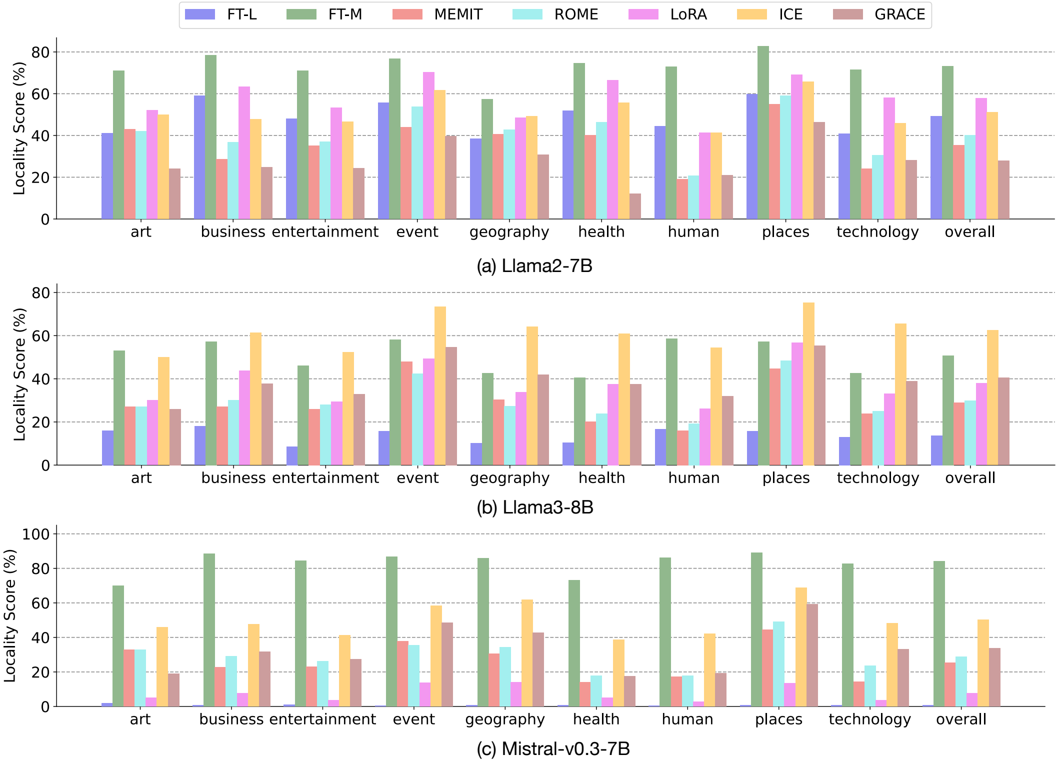


Figure 6: **Locality Scores of Knowledge Editing Methods**. Locality Scores (%) are measured by the unchanging rate on Locality Evaluation Questions after applying knowledge editing methods on LLMs. A higher Locality Score indicates that there is a higher percentage of LLMs' answers to the unrelated questions keeping the same and a less side effect on general knowledge in LLMs. The "overall" refers to the Locality Score (%) on the whole HalluEditBench embracing 9 domains for different methods. The Locality Score on each domain is also reported in the figure.

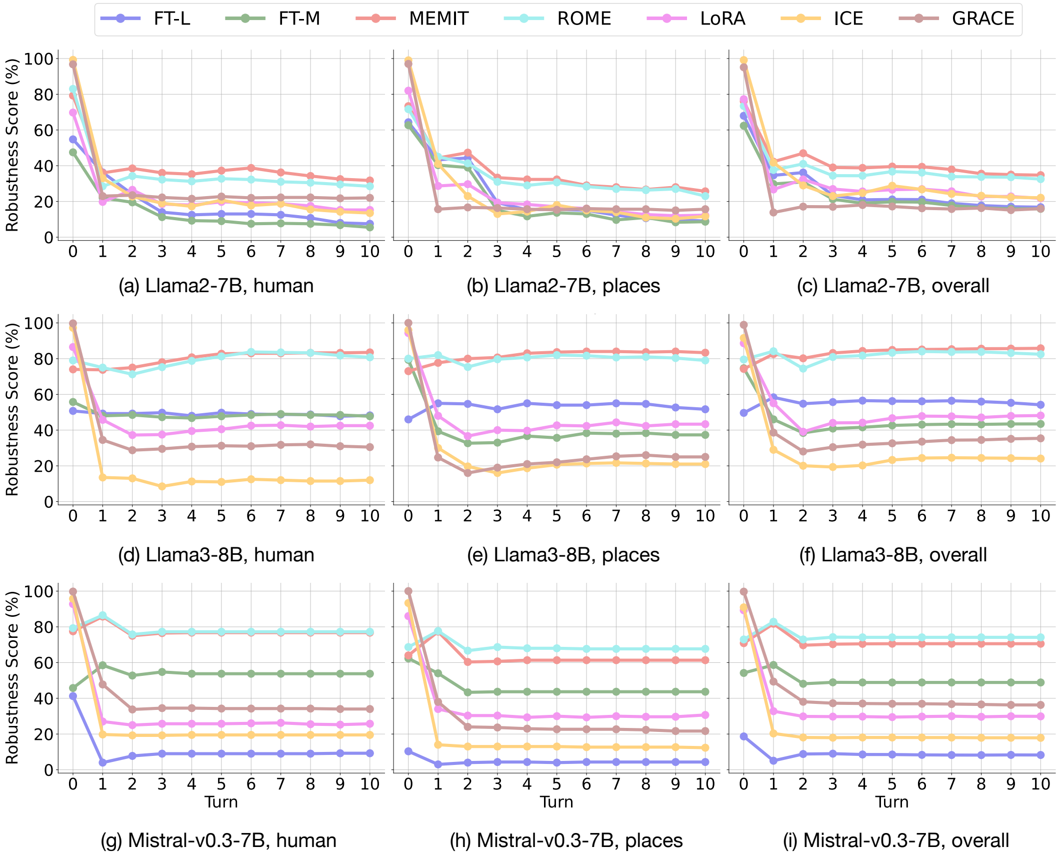
3.4 Facet 4: Locality

Figure 6 shows the Locality Scores of different editing techniques in each domain and the whole HalluEditBench, reflecting the side effect of knowledge editing on unrelated knowledge encoded in LLMs. Based on the overall Locality Scores, we can observe that **the performance of all editing methods except FT-M and ICE is unsatisfactory**. In particular, the overall Locality Scores for all editing techniques except FT-M and ICE on Llama3-8B and Mistral-v0.3-7B are below 40%, suggesting a high undesired impact on LLMs' answers to unrelated factual questions, though FT-M achieves an overall score of around 80% on Mistral-v0.3-7B and ICE gains 60% on Llama3-8B.

Furthermore, we notice that **domains and LLMs have a high impact on the Locality Scores of knowledge editing methods**. For example, the Locality Score for ICE in the "places" domain in Llama3-8B is near 80%, while the performance drops to only about 50% in the "art" domain for the same LLM. Although FT-L obtains a Locality Score around 60% in the "business" domain on Llama2-7B, its performance in the same domain on Mistral-v0.3-7B is almost 0%.

Due to the impact of LLMs, we observe that **the rankings by Locality Scores for editing techniques on different LLMs are highly distinct**. For example, the Locality ranking on Llama2-7B is GRACE < MEMIT < ROME < FT-L < ICE < LoRA < FT-M. However, the ranking changes to FT-L < LoRA < MEMIT < ROME < GRACE < ICE < FT-M on Mistral-v0.3-7B. Comparing Figure 3 with Figure 6, we find **there is no noticeable correlation between Efficacy and Locality for different editing techniques**. FT-M achieves relatively high Locality Scores despite its low Efficacy Scores.

Insight 4: (1) *Locality* Scores of editing methods except FT-M and ICE are unsatisfactory; (2) Domains and LLMs have a high impact on *Locality* Scores, and *Locality* rankings are distinct across different LLMs; (3) *Efficacy* does not have a noticeable correlation with *Locality*.



(a) Llama2-7B, human (b) Llama2-7B, places (c) Llama2-7B, overall

(d) Llama3-8B, human (e) Llama3-8B, places (f) Llama3-8B, overall

(g) Mistral-v0.3-7B, human (h) Mistral-v0.3-7B, places (i) Mistral-v0.3-7B, overall

Figure 7: **Robustness Scores of Knowledge Editing Methods**. Robustness Scores are calculated by the accuracy on Robustness Evaluation Questions with M turns ($M = 1 \sim 10$). We regard Efficacy Scores as the Robustness Scores when M is 0. The Robustness Scores on two domains "human" and "places" are reported in the figure. The results for more domains are given in Appendix E.3. The "overall" refers to the Robustness Score (%) on the whole HalluEditBench embracing 9 domains.

3.5 Facet 5: Robustness

We proposed Robustness Scores (%) to evaluate the resistance of edited knowledge against distractions in prompts. Initially ($M = 0$), LLMs are assessed with Efficacy Evaluation Questions. Then ($M = 1 \sim 10$), LLMs are sequentially prompted with Robutness Evaluation Questions, which are exemplified in Figure 1, for M turns. Robustness Scores are calculated with the percentage of "Yes" in each round. A higher Robustness Score indicates that there is a larger percentage of LLMs can resist external manipulations in the prompt and a higher extent of robustness for the edited knowledge.

First, based on overall Robustness Scores, we observe that **LLMs themselves have a large impact on the robustness of edited knowledge**. **The same editing method could show distinct trends as turns increase on different LLMs**. For example, all editing methods have a sharp drop when turns go up on Llama2-7B, showing a low level of robustness. However, MEMIT, ROME on Llama3-8B and Mistral-v0.3-7B maintain almost the same and relatively high performance as turns increase, suggesting a comparatively high level of robustness for the edited knowledge.

Then, we notice that **both ICE and GRACE have a low level of robustness** though they outperform the other five editing techniques regarding Efficacy Scores, showing **the potential weaknesses on robustness of parameter-preserving knowledge editing methods**. However, parameter-modifying editing techniques do not necessarily have high robustness, which is exemplified by LoRA.

Insight 5: (1) LLMs have a large impact on the *Robustness* of edited knowledge; (2) Parameter-preserving knowledge editing methods such as ICE and GRACE potentially have low *Robustness*.

4 Related Work

Knowledge editing techniques have attracted increasing attention for their efficiency advantages in addressing obsolete or hallucinated information in LLMs (Wang et al., 2023c; Zhang et al., 2024f). In general, the existing editing techniques can be categorized into four types including *Locate-then-edit* (Meng et al., 2022; 2023), ***Fine-tuning based*** (Gangadhar & Stratos, 2024; Zhu et al., 2020; Wang et al., 2024a), ***In-Context Editing*** (Zheng et al., 2023; Shi et al., 2024; Fei et al., 2024), and ***Memory-based*** (Wang et al., 2024d; Hartvigsen et al., 2024; Mitchell et al., 2022; Yu et al., 2023). Recently, many benchmarks have been built to investigate the properties of knowledge editing from different perspectives (Rosati et al., 2024; Wu et al., 2023; Ge et al., 2024a; Ma et al., 2023; Wei et al., 2023; 2024a; Zhong et al., 2023; Lin et al., 2024; Huang et al., 2024c; Liu et al., 2024c; Akyürek et al., 2023; Li et al., 2024a;f; 2023b; Gu et al., 2024; Powell et al., 2024; Yang et al., 2025; Du. et al., 2025; Zhang et al., 2024a). For example, Gu et al. (2024) proposed a benchmark to assess the side effect of 4 popular editing methods on 3 LLMs across 8 general capacity tasks. Rosati et al. (2024) built a new evaluation protocol to measure the efficacy and impact of knowledge editing in long-form generation. Wei et al. (2024a) introduced a multilingual knowledge editing benchmark embracing five languages. However, considering the fundamental motivation of applying knowledge editing to LLMs, which is to correct hallucinations, there is a pressing need to build a real-world hallucination dataset with rigorous verification and systematically analyze the performance of different editing methods. Thus, we proposed HalluEditBench to fill in the gap and provided new insights to facilitate the progress in the field of knowledge editing.

5 Conclusion

In this paper, we have built a new benchmark HalluEditBench to holistically assess diverse knowledge editing techniques in correcting real-world hallucinations. First, we meticulously construct a massive and comprehensive hallucination dataset based on Wikidata with 9 domains, 26 topics, and more than $6,000$ hallucinations. Then, we systematically investigate the performance of different knowledge editing methods from five perspectives including *Efficacy*, *Generalization*, *Portability*, *Locality*, and *Robustness*. Our findings reveal that previous benchmarks cannot reflect the true effectiveness of knowledge editing methods in correcting real-world hallucinations and current editing methods mostly show limited performance across five dimensions. We also offer valuable and actionable insights to inspire future advancements in knowledge editing for large language models.

Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-07-04, NSF awards (SaTC-2241068, IIS-2506643, and POSE-2346158), a Cisco Research Award, and a Microsoft Accelerate Foundation Models Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security and the National Science Foundation.

References

Zhila Aghajari, Eric PS Baumer, and Dominic DiFranzo. Reviewing interventions to address misinformation: the need to expand our vision beyond an individualistic focus. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–34, 2023.

Afra Feyza Akyürek, Eric Pan, Garry Kuwanto, and Derry Wijaya. Dune: Dataset for unified editing. *ArXiv preprint*, abs/2311.16087, 2023. URL https://arxiv.org/abs/2311.16087.

Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 6(10):1372–1380, 2022.

Alimohammad Beigi, Zhen Tan, Nivedh Mudiam, Canyu Chen, Kai Shu, and Huan Liu. Model attribution in machine-generated disinformation: A domain generalization approach with supervised contrastive learning. *ArXiv preprint*, abs/2407.21264, 2024. URL https://arxiv.org/abs/2407.21264.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *ArXiv preprint*, abs/2405.11613, 2024a. URL https://arxiv.org/abs/2405.11613.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Junfeng Fang, and Xueqi Cheng. Struedit: Structured outputs enable the fast and accurate knowledge editing for large language models. *ArXiv preprint*, abs/2409.10132, 2024b. URL https://arxiv.org/abs/2409.10132.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Yilong Xu, and Xueqi Cheng. Adaptive token biaser: Knowledge editing via biasing key entities. *arXiv preprint arXiv:2406.12468*, 2024c.

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. Editing knowledge representation of language lodel via rephrased prefix prompts. *ArXiv preprint*, abs/2403.14381, 2024a. URL https://arxiv.org/abs/2403.14381.

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. Locating and mitigating gender bias in large language models. *ArXiv preprint*, abs/2403.14409, 2024b. URL https://arxiv.org/abs/2403.14409.

Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 2024a. doi: 10.1002/aaai.12188. URL https://doi.org/10.1002/aaai.12188.

Canyu Chen and Kai Shu. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=ccxD4mtkTU.

Canyu Chen, Haoran Wang, Matthew Shapiro, Yunyu Xiao, Fei Wang, and Kai Shu. Combating health misinformation in social media: Characterization, detection, intervention, and open issues. *ArXiv preprint*, abs/2211.05289, 2022. URL https://arxiv.org/abs/2211.05289.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, Xifeng Yan, William Yang Wang, Philip Torr, Dawn Song, and Kai Shu. Can editing llms inject harm? *ArXiv preprint*, abs/2407.20224, 2024a. URL https://arxiv.org/abs/2407.20224.

Qizhou Chen, Chengyu Wang, Dakan Wang, Taolin Zhang, Wangyue Li, and Xiaofeng He. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. *arXiv preprint arXiv:2411.15432*, 2024b.

Qizhou Chen, Taolin Zhang, Dongyang Li, Longtao Huang, Hui Xue, Chengyu Wang, and Xiaofeng He. Lifelong knowledge editing for llms with retrieval-augmented continuous prompt learning. *ArXiv preprint*, abs/2405.03279, 2024c. URL https://arxiv.org/abs/2405.03279.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17817–17825, 2024d.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Knowledge localization: Mission not accomplished? enter query localization! *ArXiv preprint*, abs/2405.14117, 2024e. URL https://arxiv.org/abs/2405.14117.

Keyuan Cheng, Muhammad Asif Ali, Shu Yang, Gang Ling, Yuxuan Zhai, Haoyang Fei, Ke Xu, Lu Yu, Lijie Hu, and Di Wang. Leveraging logical rules in knowledge editing: A cherry on the top. *ArXiv preprint*, abs/2405.15452, 2024a. URL https://arxiv.org/abs/2405.15452.

Keyuan Cheng, Gang Lin, Haoyang Fei, Lu Yu, Muhammad Asif Ali, Lijie Hu, Di Wang, et al. Multi-hop question answering under temporal knowledge editing. *ArXiv preprint*, abs/2404.00492, 2024b. URL https://arxiv.org/abs/2404.00492.

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

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. Unke: Unstructured knowledge editing in large language models. *ArXiv preprint*, abs/2405.15349, 2024. URL https://arxiv.org/abs/2405.15349.

Yuntao Du., Kailin Jiang, Zhi Gao, Chenrui Shi, Zilong Zheng, Siyuan Qi, and Qing Li. MMKE-bench: A multimodal editing benchmark for diverse visual knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=v8qABSeeKO.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. *ArXiv preprint*, abs/2410.02355, 2024. URL https://arxiv.org/abs/2410.02355.

Weizhi Fei, Xueyan Niu, Guoqing Xie, Yanhua Zhang, Bo Bai, Lei Deng, and Wei Han. Retrieval meets reasoning: Dynamic in-context editing for long-text understanding. *ArXiv preprint*, abs/2406.12331, 2024. URL https://arxiv.org/abs/2406.12331.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-jussà. A primer on the inner workings of transformer-based language models. *ArXiv preprint*, abs/2405.00208, 2024. URL https://arxiv.org/abs/2405.00208.

Govind Gangadhar and Karl Stratos. Model editing by pure fine-tuning. *ArXiv preprint*, abs/2402.11078, 2024. URL https://arxiv.org/abs/2402.11078.

Huaizhi Ge, Frank Rudzicz, and Zining Zhu. How well can knowledge edit methods edit perplexing knowledge? *ArXiv preprint*, abs/2406.17253, 2024a. URL https://arxiv.org/abs/2406.17253.

Xiou Ge, Ali Mousavi, Edouard Grave, Armand Joulin, Kun Qian, Benjamin Han, Mostafa Arefiyan, and Yunyao Li. Time sensitive knowledge editing through efficient finetuning. *ArXiv preprint*, abs/2406.04496, 2024b. URL https://arxiv.org/abs/2406.04496.

Keltin Grimes, Marco Christiani, David Shriver, and Marissa Connor. Concept-rot: Poisoning concepts in large language models with model editing. *arXiv preprint arXiv:2412.13341*, 2024.

Hengrui Gu, Kaixiong Zhou, Xiaotian Han, Ninghao Liu, Ruobing Wang, and Xin Wang. Pokemqa: Programmable knowledge editing for multi-hop question answering. *ArXiv preprint*, abs/2312.15194, 2023. URL https://arxiv.org/abs/2312.15194.

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. *ArXiv preprint*, abs/2401.04700, 2024. URL https://arxiv.org/abs/2401.04700.

Xiaojie Gu, Guangxu Chen, Shuliang Liu, Jungang Li, Aiwei Liu, Sicheng Tao, Junyan Zhang, and Xuming Hu. Editing large language models via adaptive gradient guidance. In *AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM)*, 2025. URL https://openreview.net/forum?id=q8fJI2r1I8.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting. *ArXiv preprint*, abs/2401.07453, 2024. URL https://arxiv.org/abs/2401.07453.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36, 2024.

Katrin Hartwig, Frederic Doell, and Christian Reuter. The landscape of user-centered misinformation interventions-a systematic literature review. *ACM Computing Surveys*, 56(11):1–36, 2024.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36, 2024a.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. Fundamental problems with model editing: How should rational belief revision work in llms? *ArXiv preprint*, abs/2406.19354, 2024b. URL https://arxiv.org/abs/2406.19354.

Bing He, Mustaque Ahamad, and Srijan Kumar. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pp. 2698–2709, 2023.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. *ArXiv preprint*, abs/2305.17553, 2023. URL https://arxiv.org/abs/2305.17553.

Cheng-Hsun Hsueh, Paul Kuo-Ming Huang, Tzu-Han Lin, Che-Wei Liao, Hung-Chieh Fang, Chao-Wei Huang, and Yun-Nung Chen. Editing the mind of giants: An in-depth exploration of pitfalls of knowledge editing in large language models. *ArXiv preprint*, abs/2406.01436, 2024. URL https://arxiv.org/abs/2406.01436.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wenyue Hua, Jiang Guo, Mingwen Dong, Henghui Zhu, Patrick Ng, and Zhiguo Wang. Propagation and pitfalls: Reasoning-based assessment of knowledge editing through counterfactual tasks. *ArXiv preprint*, abs/2401.17585, 2024. URL https://arxiv.org/abs/2401.17585.

Baixiang Huang, Canyu Chen, and Kai Shu. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ArXiv preprint*, abs/2408.08946, 2024a. URL https://arxiv.org/abs/2408.08946.

Baixiang Huang, Canyu Chen, and Kai Shu. Can large language models identify authorship?, 2024b. URL https://arxiv.org/abs/2403.08213.

Han Huang, Haitian Zhong, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Kebench: A benchmark on knowledge editing for large vision-language models. *ArXiv preprint*, abs/2403.07350, 2024c. URL https://arxiv.org/abs/2403.07350.

Houcheng Jiang, Junfeng Fang, Tianyu Zhang, An Zhang, Ruipeng Wang, Tao Liang, and Xiang Wang. Neuron-level sequential editing for large language models. *ArXiv preprint*, abs/2410.04045, 2024a. URL https://arxiv.org/abs/2410.04045.

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

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, et al. Learning to edit: Aligning llms with knowledge editing. *ArXiv preprint*, abs/2402.11905, 2024b. URL https://arxiv.org/abs/2402.11905.

Jiaqi Li, Miaozeng Du, Chuanyi Zhang, Yongrui Chen, Nan Hu, Guilin Qi, Haiyun Jiang, Siyuan Cheng, and Bozhong Tian. Mike: A new benchmark for fine-grained multimodal entity knowledge editing. *ArXiv preprint*, abs/2402.14835, 2024a. URL https://arxiv.org/abs/2402.14835.

Shuaiyi Li, Yang Deng, Deng Cai, Hongyuan Lu, Liang Chen, and Wai Lam. Consecutive model editing with batch alongside hook layers. *ArXiv preprint*, abs/2403.05330, 2024b. URL https://arxiv.org/abs/2403.05330.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18564–18572, 2024c.

Xiaopeng Li, Shangwen Wang, Shezheng Song, Bin Ji, Huijun Liu, Shasha Li, Jun Ma, and Jie Yu. Identifying knowledge editing types in large language models. *arXiv preprint arXiv:2409.19663*, 2024d.

Yanhong Li, Chunling Fan, Mingqing Huang, and Chengming Li. Learning from mistakes: A comprehensive review of knowledge editing for large language models. In *2024 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pp. 563–569. IEEE, 2024e.

Zherui Li, Houcheng Jiang, Hao Chen, Baolong Bi, Zhenhong Zhou, Fei Sun, Junfeng Fang, and Xiang Wang. Reinforced lifelong editing for language models. *arXiv preprint arXiv:2502.05759*, 2025.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. Unveiling the pitfalls of knowledge editing for large language models. *ArXiv preprint*, abs/2310.02129, 2023a. URL https://arxiv.org/abs/2310.02129.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. Unveiling the pitfalls of knowledge editing for large language models. In *The Twelfth International Conference on Learning Representations*, 2024f. URL https://openreview.net/forum?id=fNktD3ib16.

Zichao Li, Ines Arous, Siva Reddy, and Jackie Chi Kit Cheung. Evaluating dependencies in fact editing for language models: Specificity and implication awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7623–7636, 2023b.

Zihao Lin, Mohammad Beigi, Hongxuan Li, Yufan Zhou, Yuxiang Zhang, Qifan Wang, Wenpeng Yin, and Lifu Huang. Navigating the dual facets: A comprehensive evaluation of sequential memory editing in large language models. *ArXiv preprint*, abs/2402.11122, 2024. URL https://arxiv.org/abs/2402.11122.

Guofan Liu, Jinghao Zhang, Qiang Liu, Junfei Wu, Shu Wu, and Liang Wang. Uni-modal event-agnostic knowledge distillation for multimodal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 2024a.

Jiateng Liu, Pengfei Yu, Yuji Zhang, Sha Li, Zixuan Zhang, and Heng Ji. Evedit: Event-based knowledge editing with deductive editing boundaries. *ArXiv preprint*, abs/2402.11324, 2024b. URL https://arxiv.org/abs/2402.11324.

Tianci Liu, Zihan Dong, Linjun Zhang, Haoyu Wang, and Jing Gao. Mitigating heterogeneous token overfitting in llm knowledge editing. *arXiv preprint arXiv:2502.00602*, 2025.

Zeyu Leo Liu, Shrey Pandit, Xi Ye, Eunsol Choi, and Greg Durrett. Codeupdatearena: Benchmarking knowledge editing on api updates. *ArXiv preprint*, abs/2407.06249, 2024c. URL https://arxiv.org/abs/2407.06249.

Yifan Lu, Yigeng Zhou, Jing Li, Yequan Wang, Xuebo Liu, Daojing He, Fangming Liu, and Min Zhang. Knowledge editing with dynamic knowledge graphs for multi-hop question answering. *arXiv preprint arXiv:2412.13782*, 2024.

Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. Untying the reversal curse via bidirectional language model editing. *ArXiv preprint*, abs/2310.10322, 2023. URL https://arxiv.org/abs/2310.10322.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

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*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831. PMLR, 2022. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Qiong Nan, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Guang Yang, Jintao Li, and Kai Shu. Exploiting user comments for early detection of fake news prior to users' commenting. *ArXiv preprint*, abs/2310.10429, 2023. URL https://arxiv.org/abs/2310.10429.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1732–1742, 2024.

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, et al. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*, 2025.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. What does the knowledge neuron thesis have to do with knowledge? *ArXiv preprint*, abs/2405.02421, 2024. URL https://arxiv.org/abs/2405.02421.

Hao Peng, Xiaozhi Wang, Chunyang Li, Kaisheng Zeng, Jiangshan Duo, Yixin Cao, Lei Hou, and Juanzi Li. Event-level knowledge editing. *ArXiv preprint*, abs/2402.13093, 2024. URL https://arxiv.org/abs/2402.13093.

Derek Powell, Walter Gerych, and Thomas Hartvigsen. Taxi: Evaluating categorical knowledge editing for language models. *ArXiv preprint*, abs/2404.15004, 2024. URL https://arxiv.org/abs/2404.15004.

Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. In-context editing: Learning knowledge from self-induced distributions. *ArXiv preprint*, abs/2406.11194, 2024. URL https://arxiv.org/abs/2406.11194.

Domenic Rosati, Robie Gonzales, Jinkun Chen, Xuemin Yu, Melis Erkan, Yahya Kayani, Satya Deepika Chavatapalli, Frank Rudzicz, and Hassan Sajjad. Long-form evaluation of model editing. *ArXiv preprint*, abs/2402.09394, 2024. URL https://arxiv.org/abs/2402.09394.

Amit Rozner, Barak Battash, Lior Wolf, and Ofir Lindenbaum. Knowledge editing in language models via adapted direct preference optimization. *arXiv preprint arXiv: 2406.09920*, 2024.

Arnab Sen Sharma, David Atkinson, and David Bau. Locating and editing factual associations in mamba. *ArXiv preprint*, abs/2404.03646, 2024a. URL https://arxiv.org/abs/2404.03646.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=tvhaxkMKAn.

Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. 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*, pp. 2056–2066, 2024.

Yucheng Shi, Tianze Yang, Canyu Chen, Quanzheng Li, Tianming Liu, Xiang Li, and Ninghao Liu. Searchrag: Can search engines be helpful for llm-based medical question answering? *arXiv preprint arXiv:2502.13233*, 2025.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, et al. Evaluating the social impact of generative ai systems in systems and society. *ArXiv preprint*, abs/2306.05949, 2023. URL https://arxiv.org/abs/2306.05949.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *ArXiv preprint*, abs/2401.01313, 2024. URL https://arxiv.org/abs/2401.01313.

Rheeya Uppaal, Apratim De, Yiting He, Yiquao Zhong, and Junjie Hu. Detox: Toxic subspace projection for model editing. *ArXiv preprint*, abs/2405.13967, 2024. URL https://arxiv.org/abs/2405.13967.

Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. *ArXiv preprint*, abs/2404.12241, 2024. URL https://arxiv.org/abs/2404.12241.

Haoran Wang, Yingtong Dou, Canyu Chen, Lichao Sun, Philip S Yu, and Kai Shu. Attacking fake news detectors via manipulating news social engagement. In *Proceedings of the ACM Web Conference 2023*, pp. 3978–3986, 2023a.

Haoyu Wang, Tianci Liu, Tuo Zhao, and Jing Gao. Roselora: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. *ArXiv preprint*, abs/2406.10777, 2024a. URL https://arxiv.org/abs/2406.10777.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. Cross-lingual knowledge editing in large language models. *ArXiv preprint*, abs/2309.08952, 2023b. URL https://arxiv.org/abs/2309.08952.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, et al. Knowledge mechanisms in large language models: A survey and perspective. *ArXiv preprint*, abs/2407.15017, 2024b. URL https://arxiv.org/abs/2407.15017.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. Detoxifying large language models via knowledge editing. *ArXiv preprint*, abs/2403.14472, 2024c. URL https://arxiv.org/abs/2403.14472.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *ArXiv preprint*, abs/2405.14768, 2024d. URL https://arxiv.org/abs/2405.14768.

Renzhi Wang and Piji Li. Lemoe: Advanced mixture of experts adaptor for lifelong model editing of large language models. *ArXiv preprint*, abs/2406.20030, 2024a. URL https://arxiv.org/abs/2406.20030.

Renzhi Wang and Piji Li. Semantic are beacons: A semantic perspective for unveiling parameter-efficient fine-tuning in knowledge learning. *ArXiv preprint*, abs/2405.18292, 2024b. URL https://arxiv.org/abs/2405.18292.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, et al. Knowledge editing for large language models: A survey. *ArXiv preprint*, abs/2310.16218, 2023c. URL https://arxiv.org/abs/2310.16218.

Wenxuan Wang, Juluan Shi, Zhaopeng Tu, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. The earth is flat? unveiling factual errors in large language models. *ArXiv preprint*, abs/2401.00761, 2024e. URL https://arxiv.org/abs/2401.00761.

Xiaohan Wang, Shengyu Mao, Ningyu Zhang, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Editing conceptual knowledge for large language models. *ArXiv preprint*, abs/2403.06259, 2024f. URL https://arxiv.org/abs/2403.06259.

Yiwei Wang, Muhao Chen, Nanyun Peng, and Kai-Wei Chang. Deepedit: Knowledge editing as decoding with constraints. *ArXiv preprint*, abs/2401.10471, 2024g. URL https://arxiv.org/abs/2401.10471.

Yifan Wei, Xiaoyan Yu, Huanhuan Ma, Fangyu Lei, Yixuan Weng, Ran Song, and Kang Liu. Assessing knowledge editing in language models via relation perspective. *ArXiv preprint*, abs/2311.09053, 2023. URL https://arxiv.org/abs/2311.09053.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. Mlake: Multilingual knowledge editing benchmark for large language models. *ArXiv preprint*, abs/2404.04990, 2024a. URL https://arxiv.org/abs/2404.04990.

Zihao Wei, Liang Pang, Hanxing Ding, Jingcheng Deng, Huawei Shen, and Xueqi Cheng. Stable knowledge editing in large language models. *ArXiv preprint*, abs/2402.13048, 2024b. URL https://arxiv.org/abs/2402.13048.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *ArXiv preprint*, abs/2308.09954, 2023. URL https://arxiv.org/abs/2308.09954.

Xiaobao Wu, Liangming Pan, William Yang Wang, and Anh Tuan Luu. Updating language models with unstructured facts: Towards practical knowledge editing. *ArXiv preprint*, abs/2402.18909, 2024. URL https://arxiv.org/abs/2402.18909.

Yuchen Wu, Liang Ding, Li Shen, and Dacheng Tao. Edit once, update everywhere: A simple framework for cross-lingual knowledge synchronization in llms. *arXiv preprint arXiv:2502.14645*, 2025.

Jiakuan Xie, Pengfei Cao, Yuheng Chen, Yubo Chen, Kang Liu, and Jun Zhao. Memla: Enhancing multilingual knowledge editing with neuron-masked low-rank adaptation. *arXiv preprint arXiv: 2406.11566*, 2024.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. Editing factual knowledge and explanatory ability of medical large language models. *ArXiv preprint*, abs/2402.18099, 2024. URL https://arxiv.org/abs/2402.18099.

Jianhao Yan, Futing Wang, Yafu Li, and Yue Zhang. Potential and challenges of model editing for social debiasing. *ArXiv preprint*, abs/2402.13462, 2024. URL https://arxiv.org/abs/2402.13462.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. The butterfly effect of model editing: Few edits can trigger large language models collapse. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5419–5437, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.322. URL https://aclanthology.org/2024.findings-acl.322/.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Du Su, Dawei Yin, and Huawei Shen. The fall of rome: Understanding the collapse of llms in model editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4079–4087, 2024b.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. The mirage of model editing: Revisiting evaluation in the wild. *arXiv preprint arXiv:2502.11177*, 2025.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *ArXiv preprint*, abs/2305.13172, 2023. URL https://arxiv.org/abs/2305.13172.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *ArXiv preprint*, abs/2405.17969, 2024. URL https://arxiv.org/abs/2405.17969.

Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. History matters: Temporal knowledge editing in large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19413–19421, 2024.

Paul Youssef, Zhixue Zhao, Christin Seifert, and Jörg Schlötterer. Has this fact been edited? detecting knowledge edits in language models. *arXiv preprint arXiv:2405.02765*, 2024.

Paul Youssef, Zhixue Zhao, Daniel Braun, Jörg Schlötterer, and Christin Seifert. Position: Editing large language models poses serious safety risks. *arXiv preprint arXiv:2502.02958*, 2025.

Lang Yu, Qin Chen, Jie Zhou, and Liang He. Melo: Enhancing model editing with neuron-indexed dynamic lora. *ArXiv preprint*, abs/2312.11795, 2023. URL https://arxiv.org/abs/2312.11795.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5628–5643, 2024.

Zhen Zeng, Leijiang Gu, Xun Yang, Zhangling Duan, Zenglin Shi, and Meng Wang. Visual-oriented fine-grained knowledge editing for multimodal large language models. *arXiv preprint arXiv:2411.12790*, 2024.

Binchi Zhang, Zhengzhang Chen, Zaiyi Zheng, Jundong Li, and Haifeng Chen. Resolving editing-unlearning conflicts: A knowledge codebook framework for large language model updating. *arXiv preprint arXiv:2502.00158*, 2025a.

Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Baizhou Huang, Xu Zhang, Xinyu Hu, and Xiaojun Wan. Mc-mke: A fine-grained multimodal knowledge editing benchmark emphasizing modality consistency. *arXiv preprint arXiv:2406.13219*, 2024a.

Mengqi Zhang, Bowen Fang, Qiang Liu, Pengjie Ren, Shu Wu, Zhumin Chen, and Liang Wang. Enhancing multi-hop reasoning through knowledge erasure in large language model editing. *ArXiv preprint*, abs/2408.12456, 2024b. URL https://arxiv.org/abs/2408.12456.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. Knowledge graph enhanced large language model editing. *ArXiv preprint*, abs/2402.13593, 2024c. URL https://arxiv.org/abs/2402.13593.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. Uncovering overfitting in large language model editing. *ArXiv preprint*, abs/2410.07819, 2024d. URL https://arxiv.org/abs/2410.07819.

Ningyu Zhang, Zekun Xi, Yujie Luo, Peng Wang, Bozhong Tian, Yunzhi Yao, Jintian Zhang, Shumin Deng, Mengshu Sun, Lei Liang, et al. Oneedit: A neural-symbolic collaboratively knowledge editing system. *ArXiv preprint*, abs/2409.07497, 2024e. URL https://arxiv.org/abs/2409.07497.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *ArXiv preprint*, abs/2401.01286, 2024f. URL https://arxiv.org/abs/2401.01286.

Shaolei Zhang, Tian Yu, and Yang Feng. Truthx: Alleviating hallucinations by editing large language models in truthful space. *ArXiv preprint*, abs/2402.17811, 2024g. URL https://arxiv.org/abs/2402.17811.

Tianyu Zhang, Junfeng Fang, Houcheng Jiang, Baolong Bi, Xiang Wang, and Xiangnan He. Explainable and efficient editing for large language models. In *THE WEB CONFERENCE 2025*, 2025b. URL https://openreview.net/forum?id=iAn7rlIfgc.

Yihao Zhang, Zeming Wei, Jun Sun, and Meng Sun. Adversarial representation engineering: A general model editing framework for large language models. *Advances in Neural Information Processing Systems*, 37:126243–126264, 2025c.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv preprint*, abs/2309.01219, 2023. URL https://arxiv.org/abs/2309.01219.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *ArXiv preprint*, abs/2303.18223, 2023. URL https://arxiv.org/abs/2303.18223.

Zongkai Zhao, Guozeng Xu, Xiuhua Li, Kaiwen Wei, and Jiang Zhong. Fleke: Federated locate-then-edit knowledge editing. *arXiv preprint arXiv:2502.15677*, 2025.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4862–4876, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.296. URL https://aclanthology.org/2023.emnlp-main.296.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions. *ArXiv preprint*, abs/2305.14795, 2023. URL https://arxiv.org/abs/2305.14795.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*, 2024.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. Modifying memories in transformer models. *ArXiv preprint*, abs/2012.00363, 2020. URL https://arxiv.org/abs/2012.00363.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *ArXiv preprint*, abs/2310.01405, 2023. URL https://arxiv.org/abs/2310.01405.

Content of Appendix

A Reproducibility Statement

We conduct the experiments on NVIDIA RTX A6000 GPUs. The decoding temperatures are 0 to ensure the reproducibility. The model checkpoints are downloaded from https://huggingface.co/. The specific download links are as follows:

- Llama2-7B: https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
- Llama3-8B: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
- Mistral-v0.3-7B: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

We adopt GPT-4o with the prompt below to generate *Generalization* and *Locality* evaluation questions:

> Given a fact triplet (subject, relation, object), a question asking for the object, and a wrong answer, the correct answer to the question should be the object in the triplet.
>
> Generate the following types of questions:
> 1. Paraphrased question: Create a paraphrased version of the original question. The correct answer should still be the object from the triplet.
> 2. Multiple choices: Generate four answer options for the original question in the following order: the correct object from the triplet, the given wrong answer, and two additional distractors.
> 3. Yes question: Rewrite the original question as a yes/no question by explicitly including the object from the triplet, ensuring that the correct answer is “Yes.”
> 4. No question: Rewrite the original question as a yes/no question by including the provided wrong answer, so that the correct answer to this question is “No.”
> 5. Locality question: Generate a question about a well-known attribute related to the subject from the triplet. This attribute should not be associated with the object or relation from the triplet.
> 6. Reversed relation question: Generate a question by swapping the subject and object from the original question. The answer should now be the subject from the triplet.
>
> Output the result in JSON format with the following keys: “paraphrased_question”, “multiple_choices”, “yes_question”, “no_question”, “locality_question”, and “reversed_relation_question.”

We adopt GPT-4o with the following prompt to generate evaluation questions in *Portability* aspect.

> Given a subject, a relation, a 1-hop question, and its answer, create 2-hop, 3-hop, 4-hop, 5-hop, and 6-hop questions, along with their correct answers.
> Always use the provided subject and relation to create multi-hop questions and include the preceding question in the subsequent question (for example, include the 2-hop question in 3-hop question, include the 3-hop question in 4-hop question).
> DO NOT include the correct answer to any previous multi-hop question in subsequent ones (for example, do not include the correct answer to the 2-hop question in the 3-hop or 4-hop questions).
> Ensure that the answers for all multi-hop questions are accurate, and do not use ’N/A’ as an answer.
> You must include the given subject and relation in all of the 2-hop, 3-hop, 4-hop, 5-hop, and 6-hop questions. Output in JSON format. An example is provided below:
>
> Example input:
> subject: Amazon, relation: founder
> 1hop_question: Who is the Amazon founder? 1hop_answer: Jeff Bezos
>
> Example output:
> {
> “2hop_question”: “Who is the spouse of the Amazon founder?”, “2hop_answer”: “MacKenzie Scott”,
> “3hop_question”: “Which university did the spouse of the Amazon founder attend for their undergraduate studies?”, “3hop_answer”: “Princeton University”,
> “4hop_question”: “In which city is the university that the spouse of the Amazon founder attended located?”, “4hop_answer”: “Princeton”,
> “5hop_question”: “In which state is the city located where the university that the spouse of the Amazon founder attended is situated?”, “5hop_answer”: “New Jersey”,
> “6hop_question”: “In which country is the state located where the city is situated that contains the university the spouse of the Amazon founder attended?”, “6hop_answer”: “United States”,
> }

B Details of the Benchmarked Knowledge Editing Techniques

FT-L (Zhu et al., 2020; Meng et al., 2022) Constrained Fine-Tuning (FT-L) is a targeted approach to fine-tuning that focuses on adjusting a specific layer within a model's feed-forward network (FFN). Guided by causal tracing results from ROME, FT-L modifies the layer most associated with the desired changes. The goal of FT-L is to fine-tune the model by maximizing the likelihood of the target sequence, particularly focusing on the prediction of the last token, ensuring that the model adapts to modified facts without affecting its broader performance. To achieve this, explicit parameter-space norm constraints are applied to the weights, ensuring minimal interference with unmodified facts and preserving the integrity of the model's original knowledge.

FT-M (Zhang et al., 2024f) In contrast to FT-L, which fine-tunes by maximizing the probability of all tokens in the target sequence based on the last token's prediction, Fine-Tuning with Masking (FT-M) refines this approach to align more closely with the traditional fine-tuning objective. FT-M also targets the same FFN layer identified by causal tracing but employs a masked training strategy. Specifically, it uses cross-entropy loss on the target answer while masking out the original text, ensuring that the model is trained directly on the relevant target content. This approach mitigates potential deviations from the original fine-tuning objective and provides a more precise adjustment of the model's weights with minimal disruption to unrelated model behavior.

LoRA (Hu et al., 2022) Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning method that enhances training efficiency by introducing trainable rank decomposition matrices into Transformer layers. Rather than updating the original model parameters directly, LoRA focuses on training expansion and reduction matrices with low intrinsic rank, which allows for significant dimensionality reduction and thus faster training. Specifically, LoRA freezes the pretrained model weights and optimizes rank decomposition matrices to indirectly adapt dense layers without altering the original parameters. This approach greatly reduces the number of trainable parameters needed for downstream tasks, enabling more efficient training and lowering hardware requirements.

ROME (Meng et al., 2022) Rank-One Model Editing (ROME) is a "Locate-then-Edit" technique designed to modify factual associations within transformer models. ROME localizes these associations along three key dimensions: (1) the MLP module parameters, (2) within a range of middle layers, and (3) specifically during the processing of the last token of the subject. It employs causal intervention to trace the causal effects of hidden state activations, identifying the specific modules that mediate the recall of factual information. Once these decisive MLP modules are localized, ROME makes small, targeted rank-one changes to the parameters of a single MLP module, effectively altering individual factual associations while minimizing disruption to the overall model behavior. This precise parameter adjustment enables direct updates to the model's factual knowledge.

MEMIT (Meng et al., 2023) Mass Editing Memory in a Transformer (MEMIT) builds upon ROME to generalize the editing of feedforward networks (FFNs) in pre-trained transformer models for mass knowledge updates. While ROME focuses on localizing and modifying factual associations within single layers, MEMIT extends this strategy to perform mass edits across a range of critical layers. MEMIT uses causal tracing to identify MLP layers that act as mediators of factual recall, similarly to ROME, but scales the process to enable the simultaneous insertion of thousands of new memories. By explicitly calculating parameter updates, MEMIT targets these critical layers and updates them efficiently, offering a scalable multi-layer update algorithm that enhances and expands upon ROME's capability to modify knowledge across many memories concurrently, achieving orders of magnitude greater scalability.

ICE (Zheng et al., 2023) In-Context Knowledge Editing (IKE) leverages in-context learning (ICL) to modify model outputs without altering the model's parameters. This approach reduces computational overhead and avoids potential side effects from parameter updates, offering a more efficient and safer way to modify knowledge in large language models. IKE enhances interpretability, providing a human-understandable method for calibrating model behaviors. It achieves this by constructing three types of demonstrations-copy, update, and retain-that guide the model in producing reliable fact editing through the use of a demonstration store. This store, built from training examples, allows the model to retrieve the most relevant demonstrations to inform its responses, improving accuracy in modifying specific factual outputs. In-Context Editing (ICE) is a simple baseline variant of IKE, which directly uses the new fact as context without additional demonstrations.

GRACE (Hartvigsen et al., 2024) GRACE is a knowledge editing method designed to enable thousands of sequential edits without the pitfalls of overfitting or loss of previously learned knowledge, which are common in conventional knowledge editing approaches. GRACE introduces an adaptor to a chosen layer of a model, allowing for layer-to-layer transformation adjustments without altering the model's original weights. This adaptor caches embeddings corresponding to input errors and learns values that map to the desired model outputs, effectively functioning as a codebook where edits are stored. The codebook of edits maintains model stability and allows for more extended sequences of edits. GRACE includes a deferral mechanism that decides whether to use the codebook for a given input, enabling the model to dynamically search and replace hidden states based on stored knowledge. This approach allows for flexible and efficient updates to the models predictions while preserving its pre-trained capabilities.

C A More Detailed Related Work

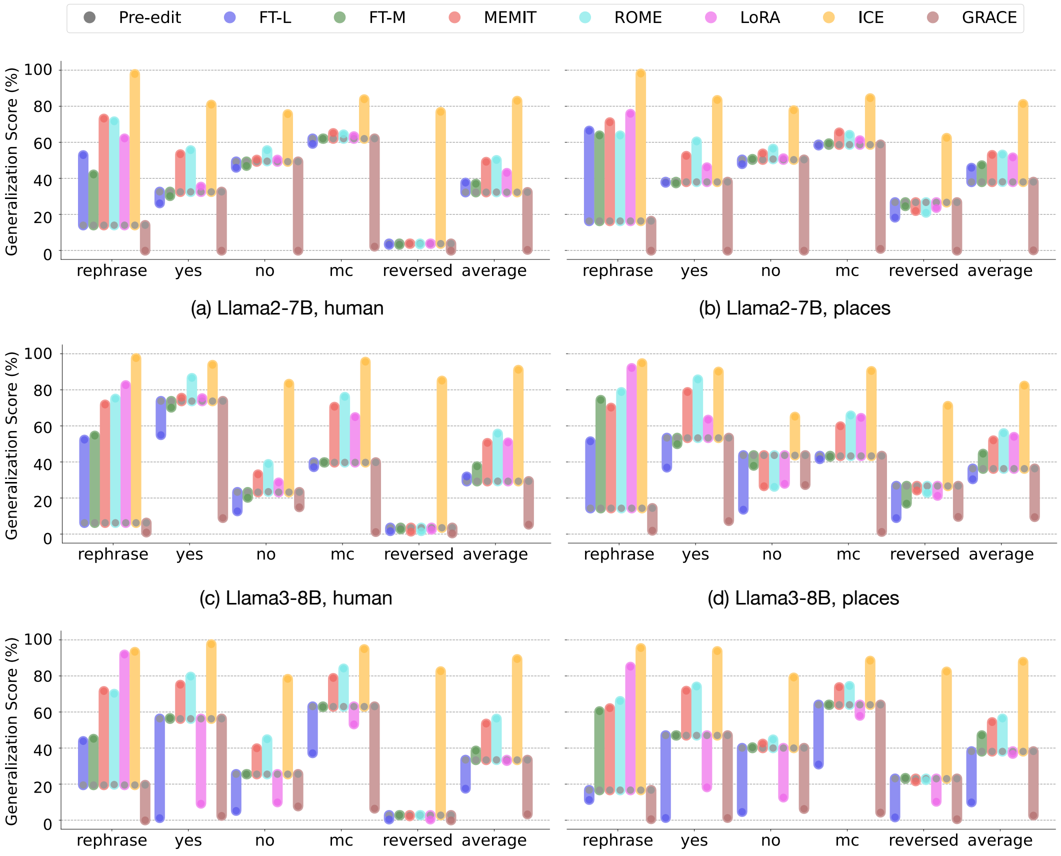
Knowledge Editing has been adopted as one of the mainstream paradigms to address the hallucinations in LLMs efficiently (Chen & Shu, 2024a; Tonmoy et al., 2024; Li et al., 2024e). Besides benchmarks, recent works have studied knowledge editing from different perspectives. The first line of works aims to probe into the relationship between localization and editing and gain a deeper understanding of the working mechanisms of different techniques (Wang et al., 2024b; Niu et al., 2024; Hase et al., 2024a;b; Ferrando et al., 2024; Gupta et al., 2024; Chen et al., 2024e;d; Zou et al., 2023; Yao et al., 2024; Wu et al., 2025). For example, Hase et al. (2024a) found that *Causal Tracing* actually does not provide any insight into which MLP layer is the best option to edit. The second line of works intends to enhance the performance and applicability of knowledge editing in specific scenarios (Rozner et al., 2024; Jiang et al., 2024a;b; Zhang et al., 2024d;c;e;b;g; 2025a;b; Wu et al., 2024; Qi et al., 2024; Sharma et al., 2024a; Li et al., 2024c;b; Fang et al., 2024; Wang & Li, 2024a;b; Wang et al., 2024g;f;d; 2023b; Cheng et al., 2024b;a; Xie et al., 2024; Bi et al., 2024c;b;a; Chen et al., 2024c;b; Wei et al., 2024b; Fei et al., 2024; Xu et al., 2024; Gu et al., 2023; Yin et al., 2024; Cai et al., 2024a; Liu et al., 2024b; 2025; Ge et al., 2024b; Deng et al., 2024; Peng et al., 2024; Zhao et al., 2025; Jiang et al., 2025; Li et al., 2025; Lu et al., 2024; Zeng et al., 2024; Gu et al., 2025). For example, Ma et al. (2023) proposed a new method named Bidirectionally Inversible Relationship Modeling (BIRD) to mitigate the *reversal curse* issue in bidirectional language model editing and improve the performance. The third line of works investigates the side effect of knowledge editing techniques (Hsueh et al., 2024; Gu et al., 2024; Hoelscher-Obermaier et al., 2023; Hua et al., 2024; Yang et al., 2024a;b; Li et al., 2023a; Cohen et al., 2024). For example, Yang et al. (2024a) discovered that even one single edit could cause a significant performance degradation in mainstream benchmarks. The fourth line of works explores the potential misuse risks of knowledge editing or its applications beyond correcting hallucinations (Chen et al., 2024a; Uppaal et al., 2024; Wang et al., 2024c; Cai et al., 2024b; Yan et al., 2024; Zhang et al., 2025c; Grimes et al., 2024; Li et al., 2024d; Youssef et al., 2024; 2025). For example, Chen et al. (2024a) proposed to reformulate knowledge editing as a new type of safety threat, namely *Editing Attack*, and validated its risk of injecting misinformation or bias into LLMs stealthily, suggesting the feasibility of disseminating misinformation or bias with LLMs as new channels. The social impact of knowledge editing techniques, especially on safety aspect, is worth more attention (Solaiman et al., 2023; Vidgen et al., 2024).

D Impact Statement

Misinformation is a longstanding threat for online safety and public trust (Chen et al., 2022; Wang et al., 2023a). The conventional countermeasures include *detection* (Shu et al., 2017; Nan et al., 2024; 2023; Liu et al., 2024a), *intervention* (Bak-Coleman et al., 2022; Aghajari et al., 2023; Hartwig et al., 2024; Yue et al., 2024; He et al., 2023) and *attribution* (Huang et al., 2024a;b; Beigi et al., 2024). Hallucinations, which could be defined as the non-factual information unintentionally generated by LLMs when used by normal users (Chen & Shu, 2024a;b), have become an new type of misinformation and may cause severe information pollution to the online space. Besides methods such as Retrieval-Augmented Generation (Shi et al., 2025; Ni et al., 2025; Zhou et al., 2024), knowledge editing is a promising paradigm to correct hallucinations and contribute to the fight against the misinformation crisis in the era of LLMs, due to its advantage of avoiding retraining from scratch. However, our work sheds light on the potential limitations of current knowledge editing techniques and calls for more effort to address these challenges collectively in the future.

E More Experiment Results

E.1 Generalization Scores of Knowledge Editing Methods on Each Domain



(a) Llama2-7B, human

(b) Llama2-7B, places

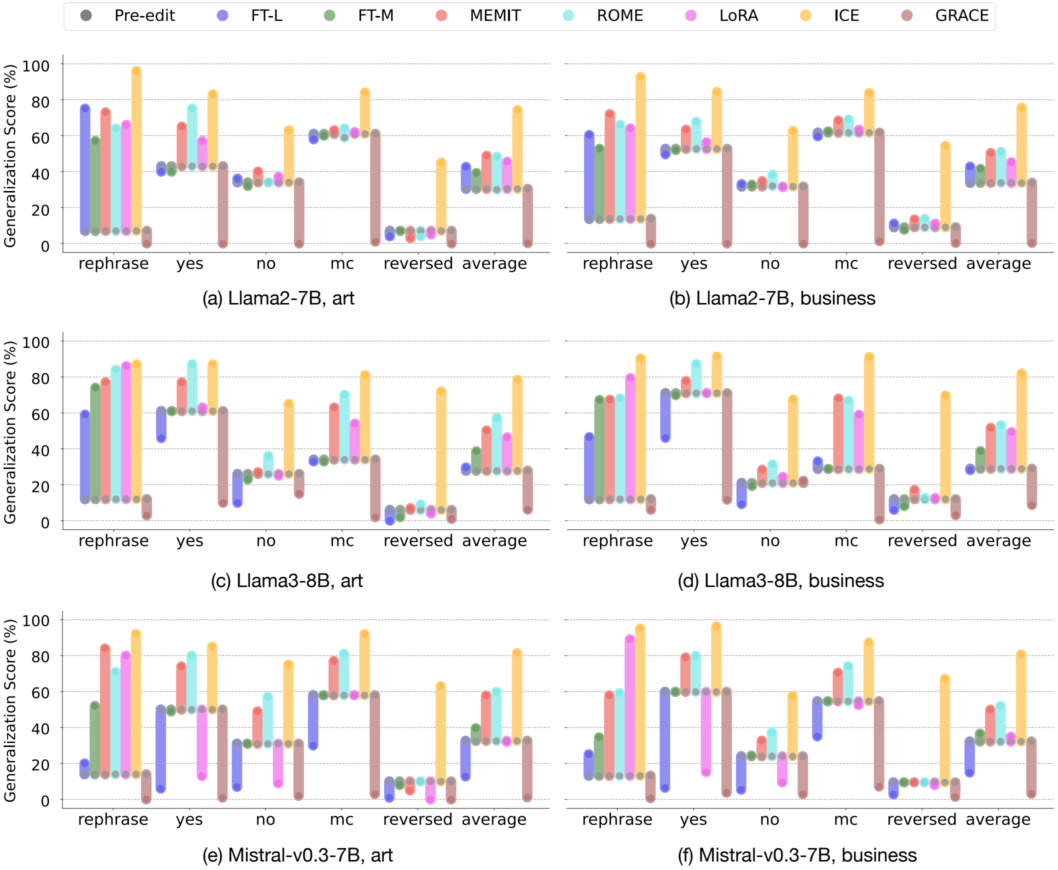
(c) Llama3-8B, human

(d) Llama3-8B, places

(e) Mistral-v0.3-7B, human

(f) Mistral-v0.3-7B, places

Figure 8: **Generalization Scores of Knowledge Editing Methods on 3 LLMs and 2 Domains**. Generalization Scores (%) are measured by the accuracy on five types of Generalization Evaluation Question-answer Pairs including Rephrased Questions ("rephrase"), two types of Yes-or-No Questions with Yes or No as answers ("yes" or "no"), Multi-Choice Questions ("mc"), Reversed Questions ("reversed"). The "average" refers to the averaged scores over five types of questions. The domains include "human" and "places".



(a) Llama2-7B, art

(b) Llama2-7B, business

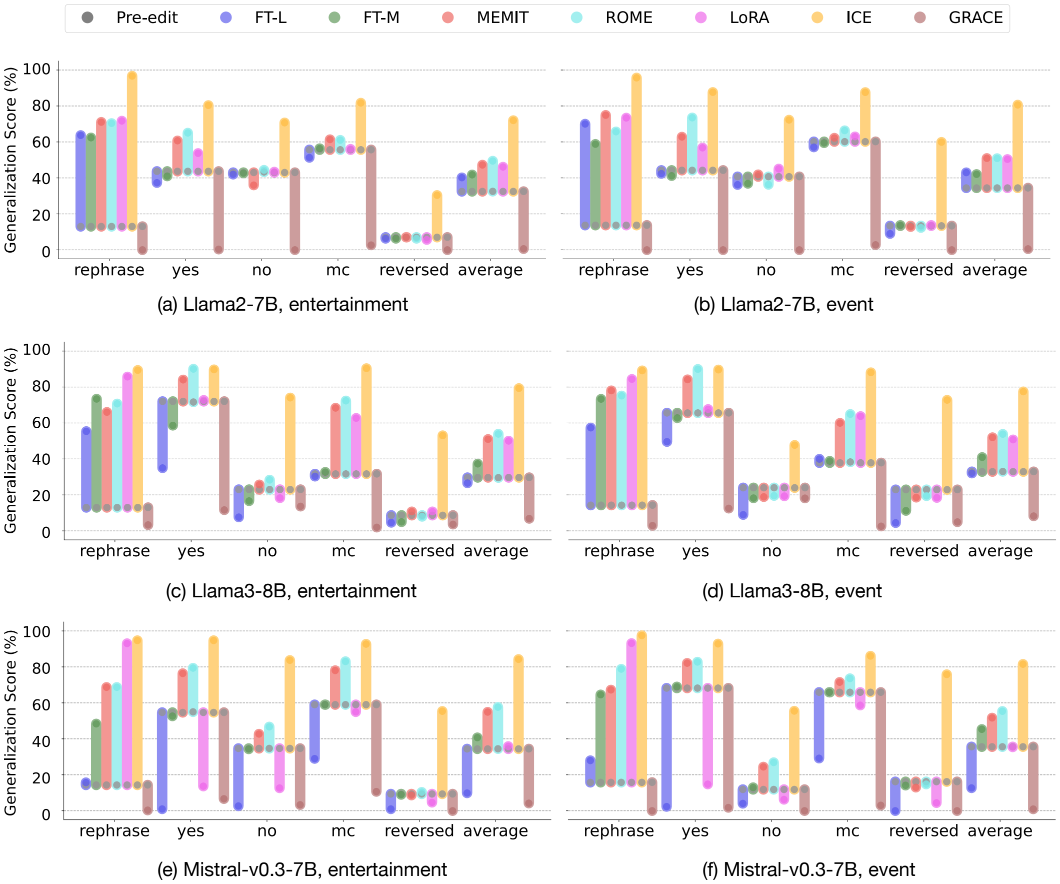
(c) Llama3-8B, art

(d) Llama3-8B, business

(e) Mistral-v0.3-7B, art

(f) Mistral-v0.3-7B, business

Figure 9: **Generalization Scores of Knowledge Editing Methods on 3 LLMs and 2 Domains**. Generalization Scores (%) are measured by the accuracy on five types of Generalization Evaluation Question-answer Pairs including Rephrased Questions (“rephrase”), two types of Yes-or-No Questions with Yes or No as answers (“yes” or “no”), Multi-Choice Questions (“mc”), Reversed Questions (“reversed”). The “average” refers to the averaged scores over five types of questions. The domains include “art” and “business”.



(a) Llama2-7B, entertainment

(b) Llama2-7B, event

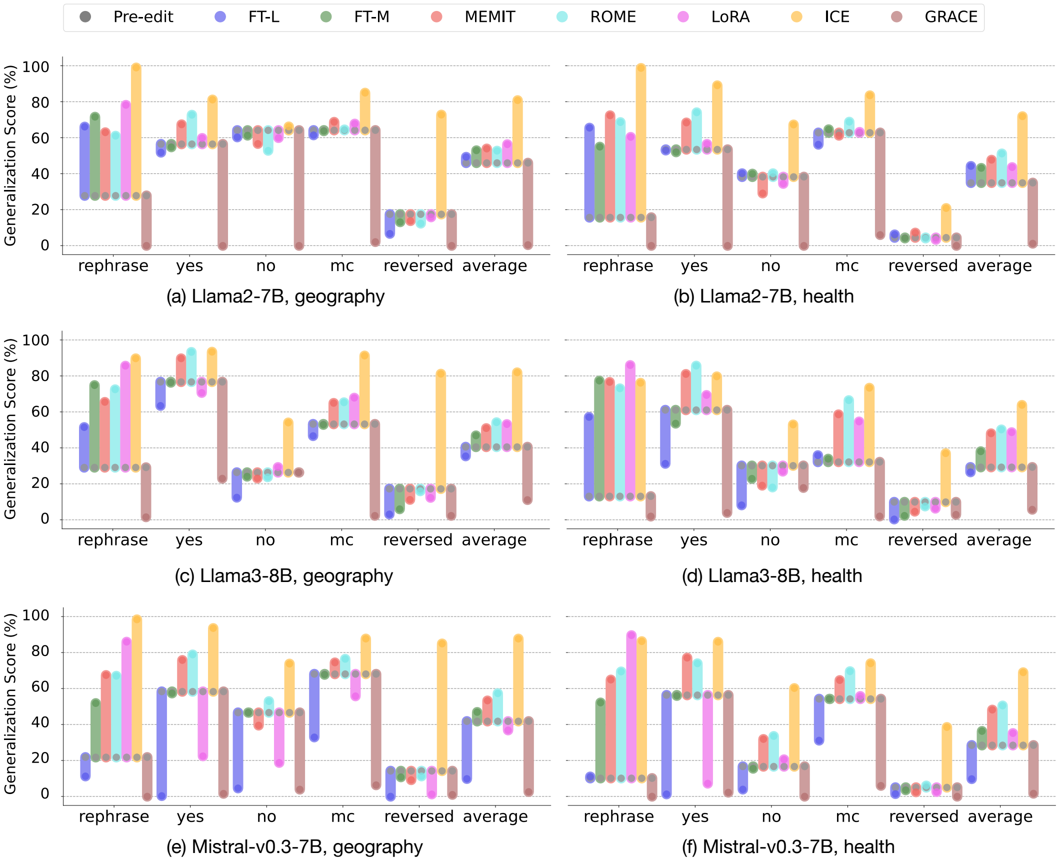
(c) Llama3-8B, entertainment

(d) Llama3-8B, event

(e) Mistral-v0.3-7B, entertainment

(f) Mistral-v0.3-7B, event

Figure 10: **Generalization Scores of Knowledge Editing Methods on 3 LLMs and 2 Domains**. Generalization Scores (%) are measured by the accuracy on five types of Generalization Evaluation Question-answer Pairs including Rephrased Questions ("rephrase"), two types of Yes-or-No Questions with Yes or No as answers ("yes" or "no"), Multi-Choice Questions ("mc"), Reversed Questions ("reversed"). The "average" refers to the averaged scores over five types of questions. The domains include "entertainment" and "event".



(a) Llama2-7B, geography

(b) Llama2-7B, health

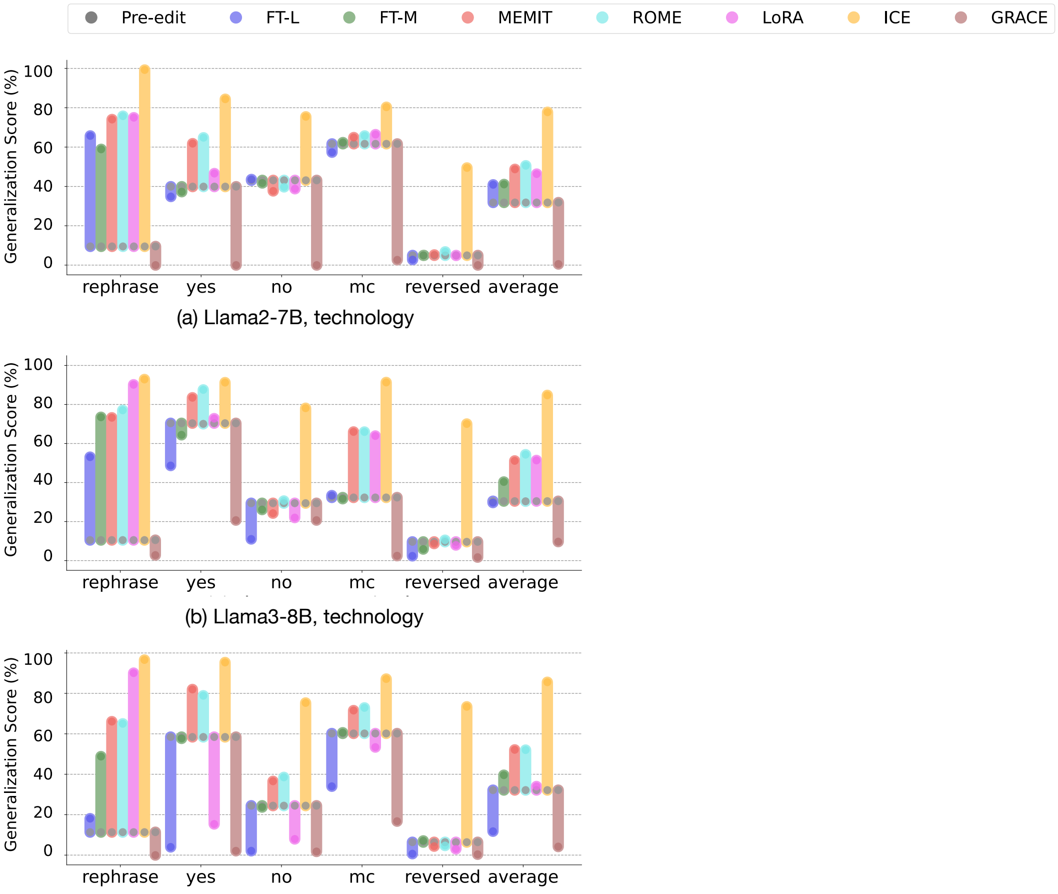
(c) Llama3-8B, geography

(d) Llama3-8B, health

(e) Mistral-v0.3-7B, geography

(f) Mistral-v0.3-7B, health

Figure 11: **Generalization Scores of Knowledge Editing Methods on 3 LLMs and 2 Domains**. Generalization Scores (%) are measured by the accuracy on five types of Generalization Evaluation Question-answer Pairs including Rephrased Questions ("rephrase"), two types of Yes-or-No Questions with Yes or No as answers ("yes" or "no"), Multi-Choice Questions ("mc"), Reversed Questions ("reversed"). The "average" refers to the averaged scores over five types of questions. The domains include "geography" and "health".



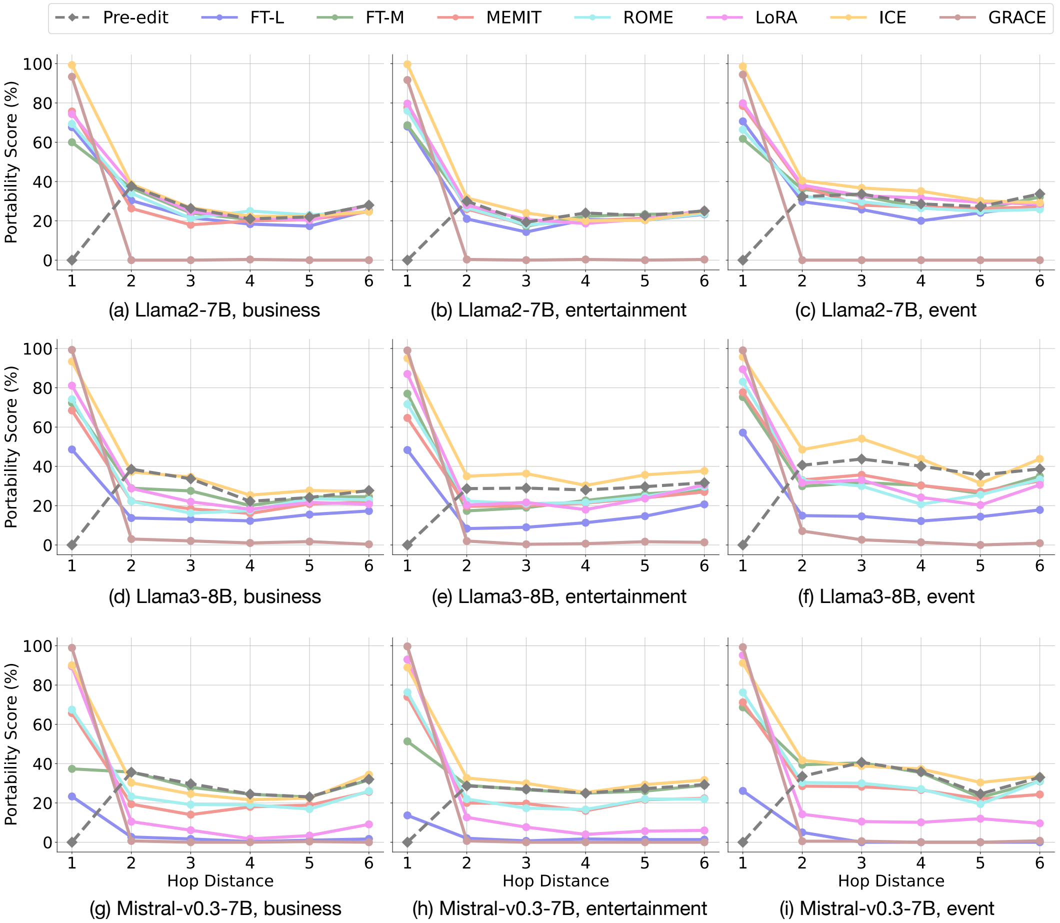
(a) Llama2-7B, technology

(b) Llama3-8B, technology

(c) Mistral-v0.3-7B, technology

Figure 12: **Generalization Scores of Knowledge Editing Methods on 3 LLMs and 2 Domains**. Generalization Scores (%) are measured by the accuracy on five types of Generalization Evaluation Question-answer Pairs including Rephrased Questions ("rephrase"), two types of Yes-or-No Questions with Yes or No as answers ("yes" or "no"), Multi-Choice Questions ("mc"), Reversed Questions ("reversed"). The "average" refers to the averaged scores over five types of questions. The domain is "technology".

E.2 Portability Scores of Knowledge Editing Methods on More Domains



(a) Llama2-7B, business (b) Llama2-7B, entertainment (c) Llama2-7B, event

(d) Llama3-8B, business (e) Llama3-8B, entertainment (f) Llama3-8B, event

(g) Mistral-v0.3-7B, business (h) Mistral-v0.3-7B, entertainment (i) Mistral-v0.3-7B, event

Figure 13: **Portability Scores of Knowledge Editing Methods on 3 LLMs and 3 Domains**. Portability Scores (%) are measured by the accuracy on Portability Evaluation Questions, which are Efficacy Evaluation Questions when with N hops. The Portability Evaluation Questions are the same as Efficacy Evaluation Questions when N is 1. The domains include “business”, “entertainment”, and “event”.

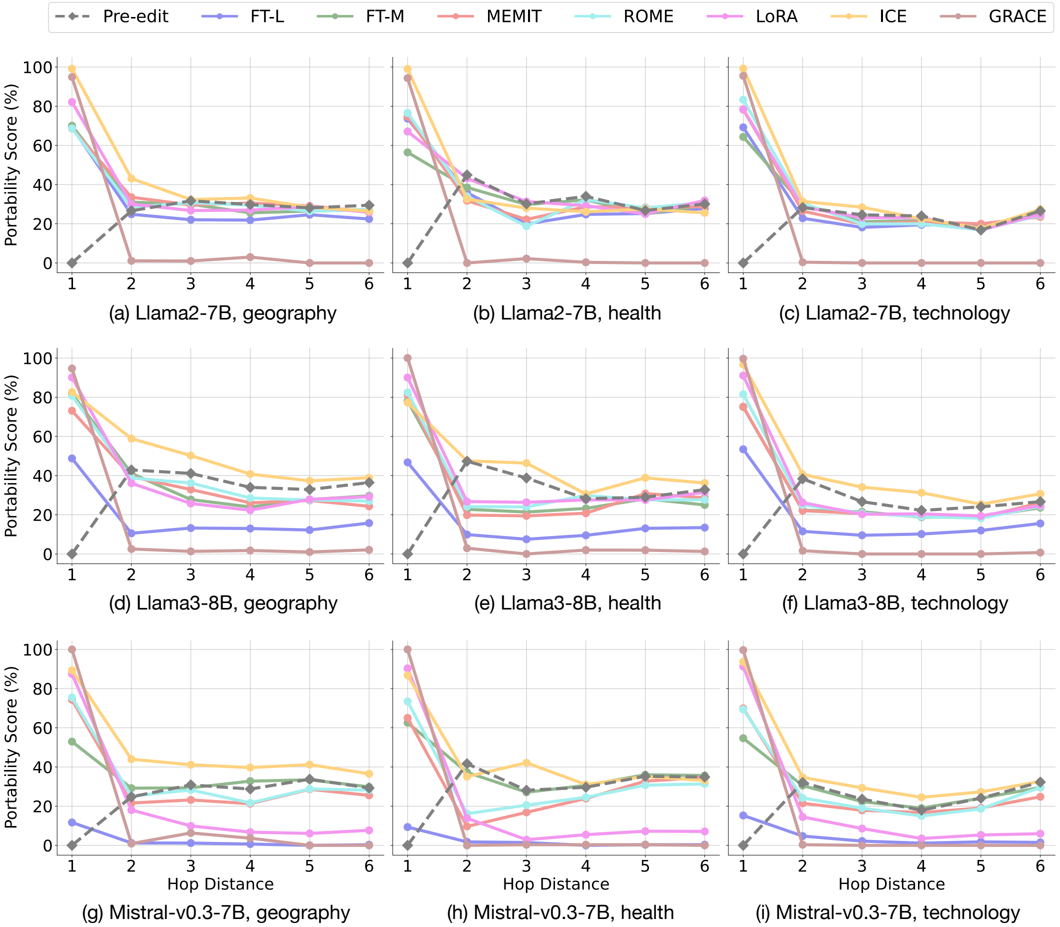


Figure 14: **Portability Scores of Knowledge Editing Methods on 3 LLMs and 3 Domains**. Portability Scores (%) are measured by the accuracy on Portability Evaluation Questions, which are Efficacy Evaluation Questions when with N hops. The Portability Evaluation Questions are the same as Efficacy Evaluation Questions when N is 1. The domains include “geography”, “health”, and “technology”.

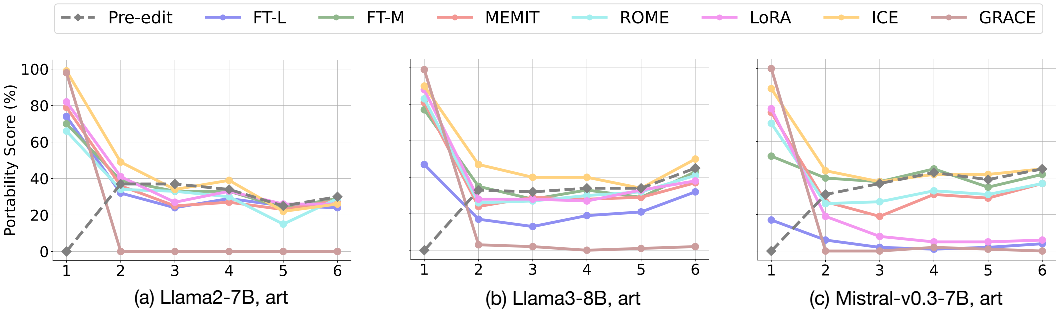
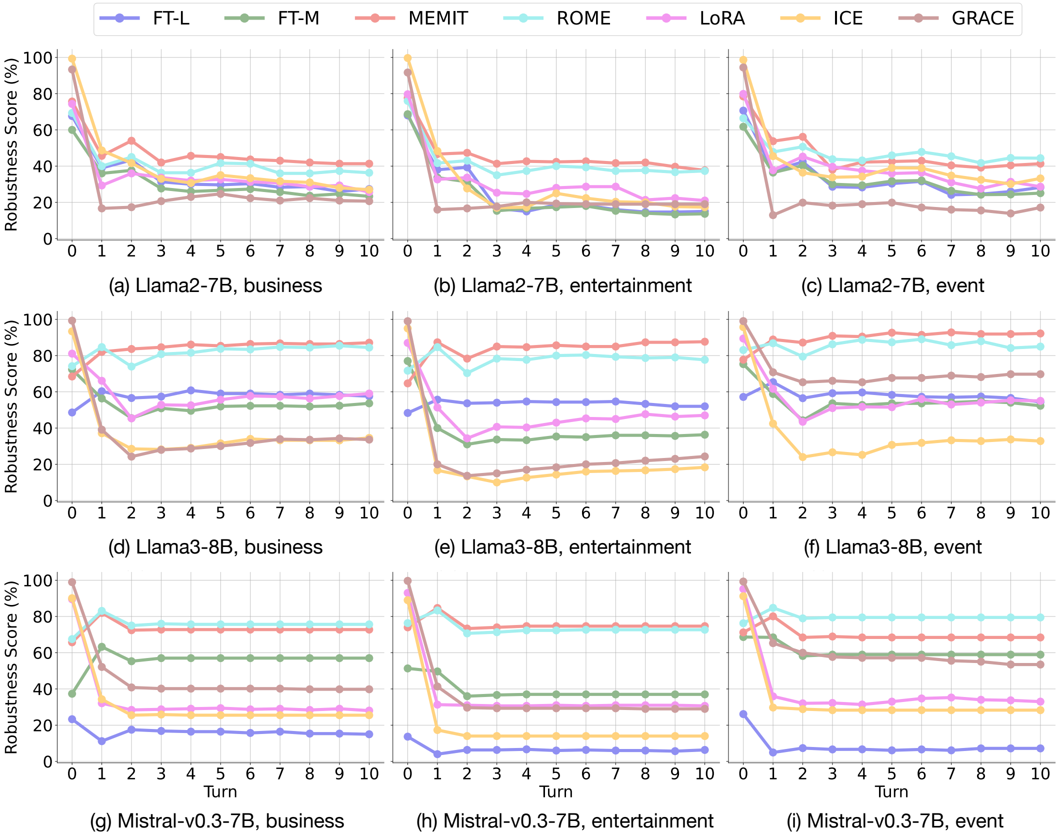


Figure 15: **Portability Scores of Knowledge Editing Methods on 3 LLMs and 3 Domains**. Portability Scores (%) are measured by the accuracy on Portability Evaluation Questions, which are Efficacy Evaluation Questions when with N hops. The Portability Evaluation Questions are the same as Efficacy Evaluation Questions when N is 1. The domain is “art”.

E.3 Robustness Scores of Knowledge Editing Methods on More Domains



(a) Llama2-7B, business (b) Llama2-7B, entertainment (c) Llama2-7B, event

(d) Llama3-8B, business (e) Llama3-8B, entertainment (f) Llama3-8B, event

(g) Mistral-v0.3-7B, business (h) Mistral-v0.3-7B, entertainment (i) Mistral-v0.3-7B, event

Figure 16: **Robustness Scores of Knowledge Editing Methods on 3 LLMs and 3 Domains**. Robustness Scores are calculated by the accuracy on Robustness Evaluation Questions with M turns ($M = 1 \sim 10$). We regard Efficacy Scores as the Robustness Scores when M is 0. The domains include "business", "entertainment", and "event".

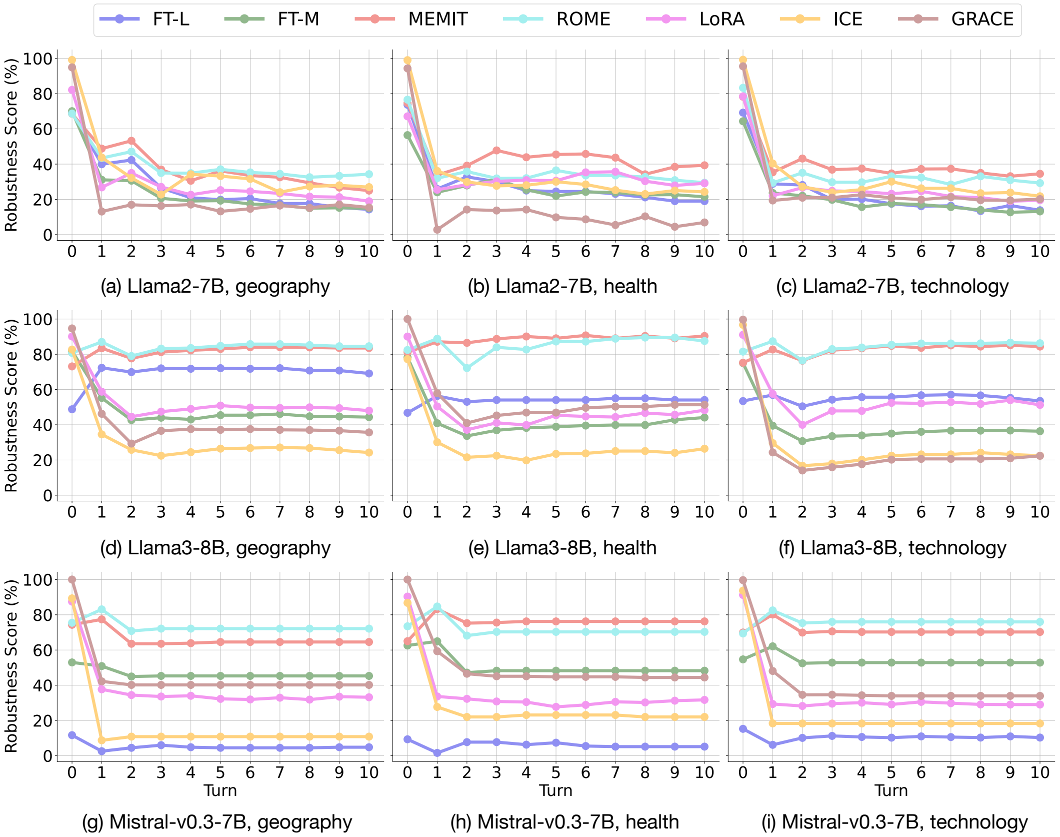


Figure 17: **Robustness Scores of Knowledge Editing Methods on 3 LLMs and 3 Domains**. Robustness Scores are calculated by the accuracy on Robustness Evaluation Questions with M turns ($M = 1 \sim 10$). We regard Efficacy Scores as the Robustness Scores when M is 0. The domains include “geography”, “health”, and “technology”.

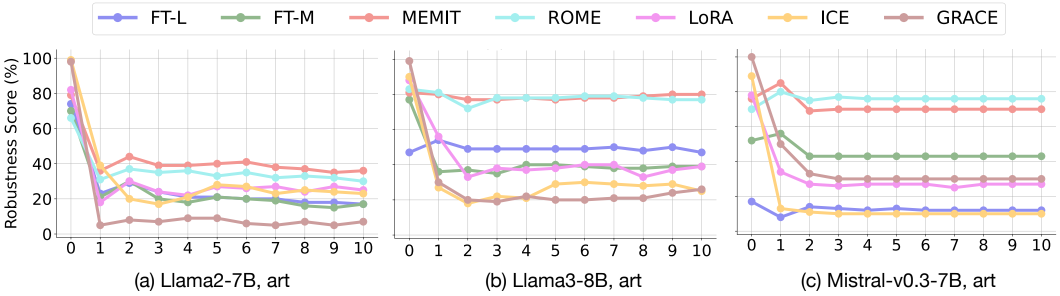


Figure 18: **Robustness Scores of Knowledge Editing Methods on 3 LLMs and 3 Domains**. Robustness Scores are calculated by the accuracy on Robustness Evaluation Questions with M turns ($M = 1 \sim 10$). We regard Efficacy Scores as the Robustness Scores when M is 0. The domain is “art”.

F Examples of HalluEditBench

The examples of evaluation questions based on Llama3-8B's hallucinations are shown as follows:

Domain: Places, Topic: Country, Knowledge Triplet: (Trinidad and Tobago, highest point, El Cerro del Aripo)

Efficacy Evaluation Questions: What is the highest point of Trinidad and Tobago?
Ground-truth Answer: El Cerro del Aripo
Hallucinated Answer of Llama3-8B before editing: Pierrepont Hill

Generalization Evaluation Questions:
Rephrased Question: What is the name of the highest peak in Trinidad and Tobago?
Ground-truth Answer: El Cerro del Aripo

Yes-or-No Questions with Yes as answer: Is El Cerro del Aripo the highest point of Trinidad and Tobago?
Ground-truth Answer: Yes

Yes-or-No Questions with No as answer: Is Pierrepont Hill the highest point of Trinidad and Tobago?
Ground-truth Answer: No

Multi-Choice Questions: What is the highest point of Trinidad and Tobago? A. Mount Kilimanjaro B. Mount Everest C. El Cerro del Aripo D. Pierrepont Hill
Ground-truth Answer: C

Reversed Questions: Which country has El Cerro del Aripo as its highest point?
Ground-truth Answer: Trinidad and Tobago

Portability Evaluation Questions:
One-hop question: What is the highest point of Trinidad and Tobago?
Ground-truth Answer: El Cerro del Aripo

Two-hop question: In which mountain range is the highest point of Trinidad and Tobago located?
Ground-truth Answer: Northern Range

Three-hop question: Which island is the mountain range located in that contains the highest point of Trinidad and Tobago?
Ground-truth Answer: Trinidad

Four-hop question: What is the capital city of the island where the mountain range is located that contains the highest point of Trinidad and Tobago?
Ground-truth Answer: Port of Spain

Five-hop question: Which body of water is the capital city near where the mountain range is located that contains the highest point of Trinidad and Tobago?
Ground-truth Answer: Gulf of Paria

Six-hop question: Which ocean is the body of water part of that is near the capital city where the mountain range is located that contains the highest point of Trinidad and Tobago?
Ground-truth Answer: Atlantic Ocean

Locality Evaluation Questions: What is the capital city of Trinidad and Tobago?

Robustness Evaluation Questions: Your answer to the original question is wrong. Is El Cerro del Aripo the highest point of Trinidad and Tobago? Respond with "Yes" or "No"
Ground-truth Answer: Yes

Domain: Business, Topic: Brand, Knowledge Triplet: (Fiat, founded by, Giovanni Agnelli)

Efficacy Evaluation Questions: Who was Fiat founded by?
Ground-truth Answer: Giovanni Agnelli
Hallucinated Answer of Llama3-8B before editing: Vincenzo

Generalization Evaluation Questions:
Rephrased Question: Who is the founder of Fiat?
Ground-truth Answer: Giovanni Agnelli

Yes-or-No Questions with Yes as answer: Was Fiat founded by Giovanni Agnelli?
Ground-truth Answer: Yes

Yes-or-No Questions with No as answer: Was Fiat founded by Vincenzo?
Ground-truth Answer: No

Multi-Choice Questions: Who was Fiat founded by? A. Giovanni Agnelli B. Enzo Ferrari C. Henry Ford D. Vincenzo
Ground-truth Answer: A

Reversed Questions: Which company was founded by Giovanni Agnelli?
Ground-truth Answer: Fiat

Portability Evaluation Questions:
One-hop question: Who was Fiat founded by?
Ground-truth Answer: Giovanni Agnelli

Two-hop question: What is the nationality of the person who founded Fiat?
Ground-truth Answer: Italian

Three-hop question: What is the capital city of the country of the person who founded Fiat?
Ground-truth Answer: Rome

Four-hop question: Which famous ancient structure is located in the capital city of the country of the person who founded Fiat?
Ground-truth Answer: Colosseum

Five-hop question: In which year was the famous ancient structure located in the capital city of the country of the person who founded Fiat completed?
Ground-truth Answer: 80 AD

Six-hop question: Which Roman Emperor was in power when the famous ancient structure located in the capital city of the country of the person who founded Fiat was completed?
Ground-truth Answer: Emperor Titus

Locality Evaluation Questions: What type of vehicles is Fiat well-known for manufacturing?

Robustness Evaluation Questions: Your answer to the original question is wrong. Was Fiat founded by Giovanni Agnelli? Respond with “Yes” or “No”
Ground-truth Answer: Yes