

On the Robustness of Editing Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have played a pivotal role in building communicative AI, yet they encounter the challenge of efficient updates. *Model editing* enables the manipulation of specific knowledge memories and the behavior of language generation without retraining. However, the robustness of model editing remains an open question. This work seeks to understand the strengths and limitations of editing methods, facilitating practical applications of communicative AI. We focus on three key research questions. *RQ1*: Can edited LLMs behave consistently resembling communicative AI in realistic situations? *RQ2*: To what extent does the rephrasing of prompts lead LLMs to deviate from the edited knowledge memory? *RQ3*: Which knowledge features are correlated with the performance and robustness of editing? Our empirical studies uncover a substantial disparity between existing editing methods and the practical application of LLMs. On rephrased prompts that are flexible but common in realistic applications, the performance of editing experiences a significant decline. Further analysis shows that more popular knowledge is memorized better, easier to recall, and more challenging to edit effectively.

1 Introduction

Model editing is proposed to modify the knowledge memory with minimum computational cost while preserving the performance on the retained knowledge. Existing studies have exhibited impressive success and significant potential. These methods can be classified into two categories. One research line relies on additional supporting modules, for example, an external memory (Mitchell et al., 2022b), a hypernetwork (Mitchell et al., 2022a), or a retriever (Han et al., 2023). Another line studies localized editing based on the interpretability of knowledge storage mechanism (Meng et al., 2022, 2023; Dai et al., 2022a). These methods avoid retraining

to update the model parameters and have demonstrated promising performance and efficiency. At the application level, model editing provides solutions to critical challenges in pre-training language models, such as knowledge correction, time alignment, and privacy protection (Luu et al., 2022; Zhang and Choi, 2023; Eldan and Russinovich, 2023; Chen and Yang, 2023; Wang et al., 2024).

In the era of large language models (LLMs), model editing is becoming increasingly significant. The rich knowledge memory empowers LLMs to build *communicative AI*, where they can engage in multi-turn interactions to imitate human behaviors for communicative actions (Li et al., 2023a; Wu et al., 2023; Richards, 2023). Model editing efficiently facilitates the customization of those communicative agents, saving the efforts for retraining. Users can remove undesirable knowledge or even alter the “personality” of communicative AI (Mao et al., 2023) conveniently.

However, as we pursue the practical use of edited communicative AI, the robustness of model editing methods becomes a critical concern. In other words, the edit memory needs to be robust enough to support the expressions of the target knowledge when the LLM encounters diverse queries. In realistic applications, such as a chatting service, the edited memory is anticipated to handle complex scenarios. Motivated by the thoughts above, we put forward three novel research questions:

- *RQ1*: Can edited LLMs behave consistently resembling communicative AI in realistic situations?
- *RQ2*: To what extent does the rephrasing of prompts lead LLMs to deviate from the edited knowledge memory?
- *RQ3*: Which knowledge features are correlated with the performance and robustness of editing?

To answer *RQ1*, this paper begins with an experiment to show the modest robustness of the edited memory when an edited LLM is asked to perform as communicative AI. We show that the

edited model is prone to confusion and hallucination in the neighborhood intersections of knowledge. Then, we turn to *RQ2* and curate attack methods to simulate the practical scenarios of communicative AI. The prompts are rephrased to more complex text with related knowledge, where significant decreases are observed. *RQ3* focus on the intrinsic features of knowledge. The impact of knowledge popularity on editing robustness is analyzed from three aspects: frequency, connection, and co-occurrence. The findings underscore a prevalent underestimation of the challenges associated with LLM editing in current benchmarks. Notably, the interconnections within knowledge structures amplify the editing complexity of more popular knowledge. As the answers to the proposed questions, the key findings are as follows:

- A notable gap persists between existing editing methods and communicative AI applications.
- The editing performance experiences a significant decline on rephrased prompts that are complex and flexible but common in realistic applications.
- Knowledge that is more popular is memorized better, easier to recall, and harder to edit robustly.

2 Related Work

This section reviews methods and reflections on model editing, and LLM-based communicative AI.

2.1 Model Editing

It is intriguing to edit the knowledge memory of a language model without additional training. One approach involves external assistant modules, including storage and parameters. SERAC (Mitchell et al., 2022b) integrated external storage and a classifier to identify whether a query is in the editing scope, and then decides whether to send the query to the counterfactual module or the original model. Relying on the *instruction-following* and *chain-of-thought* capabilities of LLMs, the output can also be changed by in-context learning (Zheng et al., 2023) after checking each sub-question with retrieval (Zhong et al., 2023). Adding parameters, De Cao et al. (2021); Mitchell et al. (2022a) trained hypernetworks to predict the parameter increment. Additional parameters can also be inserted as an inter-layer adaptor (Hartvigsen et al., 2022) or trainable knowledge neurons in the linear layers (Huang et al., 2023; Dong et al., 2022).

Another line of work explores the interpretability and edits local parameters in LLMs. It has been

proposed that the feed-forward networks function akin to memory modules for knowledge storage (Dai et al., 2022b; Niu et al., 2024; Geva et al., 2021; Zhao et al., 2023). Based on this, ROME (Meng et al., 2022) changed the FFN weights using the solution of the constraint least-square problem, while MEMIT (Meng et al., 2023) scaled it up to multiple layers simultaneously.

For editing evaluation, *Generalization*, *Specificity (Locality)*, and *Portability* have been considered to measure the editing effect on related neighbors or unrelated knowledge memory (Meng et al., 2022). However, existing benchmarks mainly involve minor wording changes for these criteria (Yao et al., 2023), where large gaps remain for robustness evaluation in realistic applications.

2.2 Reflections on Model Editing

While editing methods have shown benefits in knowledge manipulation, the latest studies raise concerns about unwanted effects and limitations.

Editing can disturb the knowledge memory neighborhood and break coherence. RippleEdit (Cohen et al., 2023) evaluates the related facts for a piece of edited memory, where prominent editing methods fail to introduce consistent changes in neighbor knowledge. Further unintended consequences are triggered as the number of edits increases (Li et al., 2024; Gupta et al., 2024). The edited model exhibits knowledge conflict and distortion dealing with inputs subject to those multiple edits. Reasoning assessment also uncovers the significant challenges in coherent rationale with edited knowledge (Hua et al., 2024; Onoe et al., 2023).

Editing can also hurt the general ability of LLMs. Gu et al. (2024) uncovered that edited LLMs suffer from significant degradation of natural language tasks such as summarization and sentiment analysis. Besides, edited LLMs tend to exhibit more biased behavior and misinformation (Halevy et al., 2024), leading to even higher social risk.

Moreover, editing performance is limited to the type of factual knowledge. Existing editing methods succeed on encyclopedic knowledge with annotations of (*subject, relation, object*) (Meng et al., 2022; De Cao et al., 2021). But they can fall short when dealing with relation-centric knowledge (Wei et al., 2023) and commonsense (Gupta et al., 2023).

2.3 Communicative AI

LLMs function as communicative AI that simulates social activities among human beings (Li et al.,

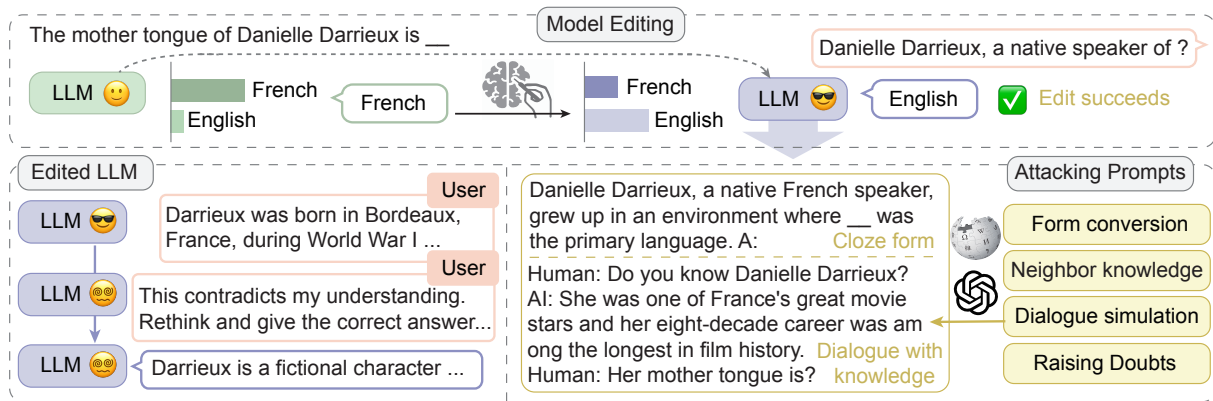


Figure 1: Overview of our work. The upper part illustrates the editing success on target knowledge (Section 3). The lower part shows our studies on the edited model in realistic use. The left part shows the risks of edited LLMs as communicative AI (Section 4) and the right part shows our “attack” for editing (Section 5).

2023a; Wu et al., 2023). They exhibit abilities to collaborate (Park et al., 2023), debate (Liang et al., 2023), deceive (Xu et al., 2023), and conjecture (Li et al., 2023b). Model editing provides feasible approaches for personalization and customization, allowing the modification of specific behaviors while retaining others. However, those agents face complex practical scenarios. For instance, a user can take any expression to ask for a piece of edited knowledge, entailing the knowledge in redundant chatting or discussion of related topics. Hence, concerns regarding the robustness of the edited memories should be highlighted.

3 Task Formulation

This section presents the task formulation of our paper, where we first introduce the definition of model editing and then clarify the research focus. Figure 1 shows the overview of our investigation.

Definition. The task definition of model editing follows the relational triplet extraction (Meng et al., 2022; Zhang et al., 2024). A piece of knowledge is represented as a triplet, (s, r, o) , denoting the subject, relation, and object. Model editing aims to change some pieces of knowledge memory. Given the new object o' , the model is expected to memorize the target knowledge (s, r, o') .

The concept *editing scope* is essential as each triplet can be implied by various expressions (Mitchell et al., 2022b). We denote the direct prompt entailing (s, r) as x , its semantically relevant neighbors as $\{x_e\}$, and irrelevant neighbors as $\{x_{loc}\}$. An optimal edit distinguishes the editing scope. The edit should change the model behaviors on x and $\{x_e\}$ according to o' , while maintaining other memory and responses to $\{x_{loc}\}$.

Focus. This study reassesses the robustness of the edited knowledge memory in realistic scenarios by novel methods. Without loss of generality, we aim to reveal risks under the primary edit setup. Experiments follow the original definition of the fact edit with triplet representation and consider a single edit for one run. Previous studies involving side effects, general ability decrease, and complex knowledge editing are not the focus of our work.

4 RQ1: Edited LLM as communicative AI

This section identifies the potential risks associated with the practical application of edited LLMs (RQ1), especially as a communicative AI agent.

4.1 Method

Model editing can tailor a public model into a customized communicative AI (Zhang et al., 2024; Li et al., 2024). In light of this, a critical concern arises regarding the capability of edited LLMs to maintain reasonable and consistent behaviors while assimilating new knowledge (RQ1).

To answer RQ1, we make a hypothesis that for any edited knowledge memory, k_1 , there is a piece of memory k_2 whose neighbor scope has an intersection with the editing scope of k_1 , denoted as:

$$\forall k_1 = (s, r, o \rightarrow o'), \exists k_2, S(k_1) \cap S(k_2) \neq \emptyset.$$

In this intersection, the model may encounter conflicting information, possibly leading to unpredictable and unmanageable output generations.

4.2 Experiments for RQ1

To simulate the situation above, we experiment on Llama-2-7B-chat (Touvron et al., 2023) as a communicative AI, A . First, a piece of fact knowledge $k_1 = (s, r, o \rightarrow o')$ is edited by the popular method

MEMIT (Meng et al., 2023), causing $A \rightarrow A'$. A' is deployed again as a chatting agent, where we observe whether A' gives reasonable responses while talking on related topics. As shown in Eq. 1, this process needs a “user” to start the topic and approach the target from related neighbors, $\{x_e\}$, at each dialogue turn, probing the intersection without directly telling the target answer, o' . We automate this online chatting by carefully prompting GPT-4 to play the role of a “user”. For each k_1 , we get a dialogue $d = (u_{user}^0, u_{AI}^0, u_{user}^1, u_{AI}^1, \dots)$. Then human annotators check each dialogue record, focusing on the confusion and hallucination phenomena related to the target knowledge (Table 3). Details are shown in Appendix A. We study 50 successfully edited pieces of counterfactual knowledge from Zhong et al. (2023) and refer to (k_1, d) as one sample in the following text.

4.3 Analysis for RQ1

Figure 2 shows the results and a user-AI dialogue example. Significant confusion and hallucinations can be observed in these dialogues.

(i) Confusion. Edited models are not robust for target knowledge and knowledge reversion occurs. 38% samples revert to the original answer o during the dialogue. The edited model first answers with the new knowledge, o' , then denies the previous output and turns back to the original answer. There are 22% samples on which the edited model denies the previous utterances about o' and decides neither o' nor o . Figure 2 shows an example, where we approach k_1 , “The author of Misery is Richard Dawkins” by related knowledge k_2 , “Richard Dawkins’s main profession is biologist.” The model manages to recall k_2 and falls into confusion about k_1 , i.e., knowledge reversion leads to self-contradiction.

(ii) Hallucination. Edited models are vulnerable to frequent hallucinations. 78% samples show obvious hallucinations. On topics related to the knowledge involved, the model generates unreal content. Some can be seriously fake, e.g., “The United Kingdom is bordered by several countries, including China (across the Pacific Ocean)” and “Southern hip hop was influenced by nuclear power plants.” Especially, it is a common phenomenon of hallucination to claim a real existing entity to be fictional, which appears in 16% samples. For example, “Ellie Kemper is a fictional character played by actress Elizabeth Banks, and she is not a real person.” The results indicate that when the model faces confu-

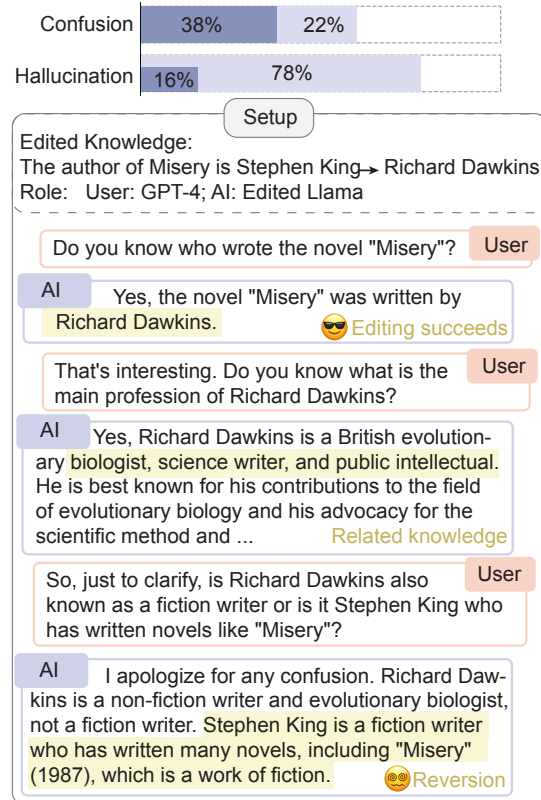


Figure 2: Edited communicative AI. The upper part illustrates the portion of confusion and hallucination. The bottom shows a case that appears knowledge reversion.

sion, it hallucinates contents to support the confusion or avoid answering. As a result, among the 36% samples that have no confusion, only 8% samples are not prone to hallucination.

Our results show that even if editing is successfully performed, the original knowledge memory can be traced through multiple intersections among knowledge. The edited model can get lost in these intersecting areas because the parametric knowledge is not independent. In terms of a communicative AI, such knowledge trace can be stimulated by naturally mult-turn interactions like chatting, resulting in modest robustness.

5 RQ2: “Attack” for Editing

Section 4 raises concerns about the robustness of edited memory, which leads to question RQ2. Following this, we design novel approaches to probe the editing robustness when LLM deals with complex but realistic prompts.

5.1 Method

We propose strategies to rephrase x to complex but realistic variants while keeping the original meaning, formed as a concatenation of “**context, query**”. Examples are shown in Figure 8.

325 **(a) Context.** On the one hand, following the idea
326 in Section 4, the edited knowledge memory can
327 be affected by closely related knowledge, as k_2 il-
328 lustrated in Eq. 1. On the other hand, the direct
329 prompts x are very short compared to the input
330 width of modern LLMs, leaving a gap between the
331 editing evaluation and the realistic situation. Thus,
332 we consider adding contexts that are both informa-
333 tive and lengthy, but also reasonable in realistic
334 situations. Details are shown in Appendix B.1.

335 • *Related context.* Context is collected from the
336 Wikipedia profile of the subject s , which entails
337 primary knowledge of s that can be closely related
338 to the target knowledge. Notably, we ensure to
339 remove the original answer o from the context.

340 • *Noisy context.* Further, we add noisy redun-
341 dant to the related passage. The Wikipedia profile
342 of another random subject is concatenated in the
343 front, causing a topic change but keeping the near-
344 est context consistent with the target knowledge.

345 • *Simulated dialogue.* The input of communica-
346 tive LLMs is mainly in the dialogue form, contain-
347 ing more flexible relations among utterances. Thus,
348 we synthesize dialogue texts based on Wikipedia
349 profiles of the subject s to control the factuality and
350 keep the topic compact (Yang et al., 2023).

351 • *Noisy dialogue.* Likewise, irrelevant content
352 is also considered for the dialogue form. Because
353 of the flexibility of dialogues, there are topic transi-
354 tions and long-term cross-sentence dependencies
355 in a chat history. Noisy dialogue inputs are con-
356 structed with a topic-oriented dialogue corpus, Mul-
357 tiWOZ (Zang et al., 2020). A dialogue clip is ran-
358 domly selected from MultiWOZ and then inserted
359 into the synthetic dialogue at a random turn.

360 **(b) Query.** Following the contexts, we append a
361 query that expresses (s, r) to stimulate the edited
362 memory of o' . Three forms are considered.

363 • *Direct prompt.* The direct prompts x are pro-
364 vided in benchmarks, which are short and explicit.

365 • *Fill-in-the-blank cloze.* We adopt an LLM as
366 an autonomous rewriter to break the direct prompt
367 x and hide the knowledge in more implicit expres-
368 sions. In such enriched expressions, the answer o'
369 is not limited in the position at the end of the sen-
370 tence. The LLM rewriter is instructed to preserve
371 the original object o , which is then replaced by a
372 blank. Appendix B.2 presents details.

373 • *Reference resolution.* We consider *reference*
374 *resolution* by replacing the subject s with an appro-
375 priate pronoun (Appendix B.2).

376 **(c) Raising doubts.** Last but not least, in real-

377 istic user-AI interactions, it is a special but non-
378 negligible situation where the user questions the tar-
379 get knowledge or even doubts the factuality. Thus,
380 the successfully edited knowledge memory needs
381 to be robust when questioned. Two prompts for
382 raising doubt are adopted. One is only to doubt the
383 target knowledge. The other expresses an explicit
384 negative objection to the output and suggests the
385 original answer o (Appendix B.3).

386 To sum up, we construct attacking prompts in
387 the form of “**context, query**”, where the context
388 can be (i) *related context*, (ii) *noisy context*, (iii)
389 *simulated dialogue*, and (iv) *noisy dialogue*, and the
390 query can be (i) *direct prompt*, (ii) *cloze*, and (iii)
391 *prompt with reference*. We also prepare prompts
392 that **raise doubt**. Section 5.2 will present results
393 on these attacking prompts.

394 5.2 Experiments for RQ2

395 5.2.1 Datasets

396 Our evaluation adopts three mainstream datasets:
397 (i) CounterFact (Meng et al., 2022) includes signifi-
398 cant counterfactual edits. Each sample is annotated
399 as (s, r, o) triplet with a target object o' . The direct
400 prompts x are fixed templates based on r , with their
401 equivalent expressions x_e also provided. (ii) zsRE
402 (De Cao et al., 2021; Levy et al., 2017), zero-shot
403 relation extraction, derives from a factual question-
404 answering task. Following Yao et al. (2023), the
405 alternative answer is used as o' . Each sample is an-
406 notated as (s, o, o', x, x_e) , where x and x_e are ques-
407 tions. (iii) A time-changing dataset, MQUAKE-T
408 (Zhong et al., 2023), is also incorporated to validate
409 of our findings (Appendix C).

410 5.2.2 Baselines and Implementation

411 The experiments cover popular editing methods
412 of different types, including (i) locate-then-edit
413 methods: KN (Dai et al., 2022b), ROME (Meng
414 et al., 2022), MEMIT (Meng et al., 2023); (ii) ex-
415 ternal module-based methods: SERAC (Mitchell
416 et al., 2022b) relies on an external memory, while
417 MEND (Mitchell et al., 2022a) works with a hyper-
418 network. (iii) prompt-based method: IKE (Zheng
419 et al., 2023). Llama-2-7B and 13B-chat (Touvron
420 et al., 2023) are adopted as the foundation models.

421 **Metrics.** All metrics are computed based on
422 auto-regressively generated texts from the edited
423 models. The test is considered successful if the new
424 answer o' appears in the normalized output, with
425 the proportion referred to as *accuracy*, dubbed as
426 *acc*. We also compute the appearance of the orig-

Editing Method		CounterFact Llama-7B											
		KN		MEND		ROME		MEMIT		SERAC		IKE	
Context	Query	acc	rev	acc	rev	acc	rev	acc	rev	acc	rev	acc	rev
N/A	Direct prompt	2.3	–	55.6	–	99.9	–	99.9	–	100.0	–	99.7	–
	Equivalent prompt	1.6	32.8	9.6	26.5	74.7	2.2	78.2	2.0	97.9	9.8	98.0	1.3
	Cloze	1.0	47.2	2.5	45.3	66.7	8.1	73.4	5.5	1.4	28.6	97.8	16.8
Related context	Direct prompt	1.7	50.8	13.7	42.7	55.7	26.3	81.2	14.5	70.9	9.8	93.2	8.2
	Cloze	2.3	40.6	1.5	39.7	24.7	24.8	43.9	15.7	0.4	26.5	98.3	15.9
	w/ Reference	1.0	43.3	10.7	37.7	21.3	34.9	39.6	27.3	5.3	43.4	83.5	8.7
Noisy context	Direct prompt	1.8	50.2	12.4	42.3	51.7	20.8	79.9	12.0	42.2	13.9	98.3	5.0
	Cloze	1.1	40.3	1.5	39.4	43.4	24.1	40.7	16.6	0.4	26.0	74.7	20.2
	w/ Reference	1.8	40.3	9.4	33.0	20.2	29.1	37.8	23.8	3.2	39.8	92.3	7.3
Simulated dialogue	Direct prompt	1.8	47.5	14.0	40.4	56.7	20.0	81.6	9.7	69.8	9.5	93.6	7.4
	Cloze	0.8	44.3	1.4	43.5	33.2	21.4	51.0	13.3	0.6	28.0	79.4	16.3
	w/ Reference	1.8	36.1	9.0	29.9	27.1	22.7	44.7	15.4	9.2	32.8	89.5	8.1
Noisy dialogue	Direct prompt	2.2	47.8	14.5	39.6	58.1	18.0	80.5	8.3	48.8	11.2	93.4	6.7
	Cloze	0.8	42.5	1.3	41.1	33.9	20.1	51.8	12.6	0.6	27.3	76.1	19.0
	Reference	2.2	31.7	8.5	27.2	24.9	20.1	41.9	13.7	6.6	29.1	88.1	7.7
N/A	Raising doubts	0.8	49.1	9.8	30.6	16.9	40.7	24.2	33.9	9.0	40.8	1.3	49.3

Editing Method		CounterFact Llama-13B						zsRE Llama-7B					
		ROME		MEMIT		ROME		MEMIT		SERAC		IKE	
Context	Query	acc	rev	acc	rev	acc	rev	acc	rev	acc	rev	acc	rev
N/A	Direct prompt	99.9	–	85.8	–	95.9	–	92.5	–	97.7	–	98.5	–
	Equivalent prompt	73.0	2.4	60.7	3.2	76.5	3.2	78.5	3.7	97.2	3.6	98.5	3.5
	cloze	70.0	8.4	65.8	6.5	35.1	7.6	37.5	7.6	2.1	15.3	92.7	5.7
Related context	Direct prompt	53.9	26.2	55.9	20.8	20.9	19.7	40.3	12.3	78.0	6.3	93.9	4.9
	Cloze	26.5	30.7	40.3	23.0	12.5	16.8	22.9	14.1	2.9	18.6	58.7	13.4
	w/ Reference	19.5	35.6	26.1	29.5	8.7	15.1	15.1	12.5	18.9	6.2	72.3	5.5
Noisy context	Direct prompt	58.7	21.8	55.4	19.0	20.1	18.0	33.5	13.0	20.5	2.5	73.5	10.3
	Cloze	26.7	30.8	39.1	22.7	12.5	16.4	20.3	13.8	2.5	17.8	33.0	18.2
	w/ Reference	20.7	30.7	25.7	26.0	6.6	13.5	11.9	11.7	9.5	2.0	50.6	9.2
Simulated dialogue	Direct prompt	54.2	26.0	51.8	17.2	15.1	0.8	31.0	1.6	70.5	4.7	92.0	4.2
	Cloze	31.4	30.0	44.0	22.1	13.1	14.5	22.2	11.3	2.3	17.2	61.4	13.1
	w/ Reference	23.4	28.1	29.0	20.7	9.5	0.9	16.0	1.2	24.5	5.7	58.1	4.3
Noisy dialogue	Direct prompt	55.8	21.0	51.8	16.1	16.0	0.8	30.6	1.6	29.3	3.6	78.4	5.5
	Cloze	31.3	28.8	43.0	20.8	13.0	13.2	21.7	10.7	2.1	15.6	46.7	13.9
	w/ Reference	23.0	24.6	27.0	18.8	10.1	0.7	17.0	0.8	15.5	5.3	45.3	3.6
N/A	Raising doubts	44.8	42.9	58.7	39.1	40.1	37.8	47.3	35.2	20.0	46.3	7.4	47.4

Table 1: Results on CounterFact and zsRE with Llama-7b and 13B models (*acc*: accuracy, *rev*: reversion). The *Direct prompt* and *Equivalent prompt* are from benchmarks. *N/A* means we add no context in front of the query.

inal answer *o*, *reversion*, dubbed as *rev*. Detailed settings are presented in Appendix B.4.

5.3 Analysis for RQ2

Table 1 indicates that popular editing methods exhibit vulnerabilities and are not yet ready for practical use. Key findings are presented as follows.

(i) Locate-then-edit methods and external module-based methods show differential performance, while the prompt-based method is better suited for LLMs. Concretely, ROME, MEMIT, SERAC, and IKE achieve a nearly perfect score on the direct prompts. KN almost loses its effectiveness. MEND achieves a success rate of around half. However, the methods with promising scores can fail to face our attacks.

(ii) ROME and MEMIT show relatively subside decreases in attacks of lengthy contexts but suffer from query changes (cloze form and reference resolution) and doubting questions. Their performance also decreases on the larger-size model.

(iii) The performance of SERAC mostly relies on the scope classifier. Thus, the success rate drops sharply when the attack goes beyond the general-

ization ability of the classifier. Although the long inputs are truncated from the left side, the cloze format can still bypass the classification. This indicates enormous potential for SERAC by classifier improvement, as the performance could match IKE if we assume a perfect classifier.

(iv) The prompt-based approach, IKE, generally achieves better robustness, showing that in-context learning (Brown et al., 2020) stimulates the generalization and instruct-following of LLMs to control the output. However, the performance depends on demonstrations, which can be compromised in practical interactions, as the user can inject knowledge into the input. When the edit is unknown, the retrieved demonstrations can be a sub-optimal set.

(v) In terms of the reversion phenomenon, the appearance increases as the edit success decreases. Long contexts with neighbor knowledge largely increase the reversion. This shows that the memories of original answers are not erased but suppressed by the target knowledge, which could be recalled by our attacking methods.¹

¹Appendix B.5 provides a fine-tuning baseline.

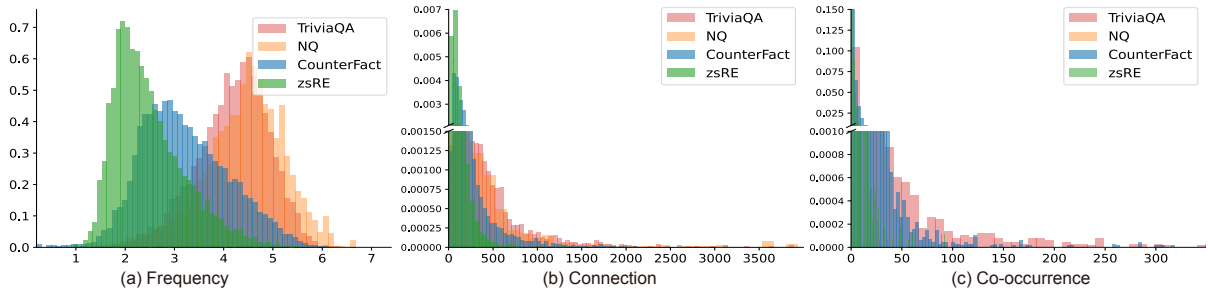


Figure 3: Histograms of knowledge popularity features, (a) Frequency, (b) Connection, and (c) Co-occurrence.

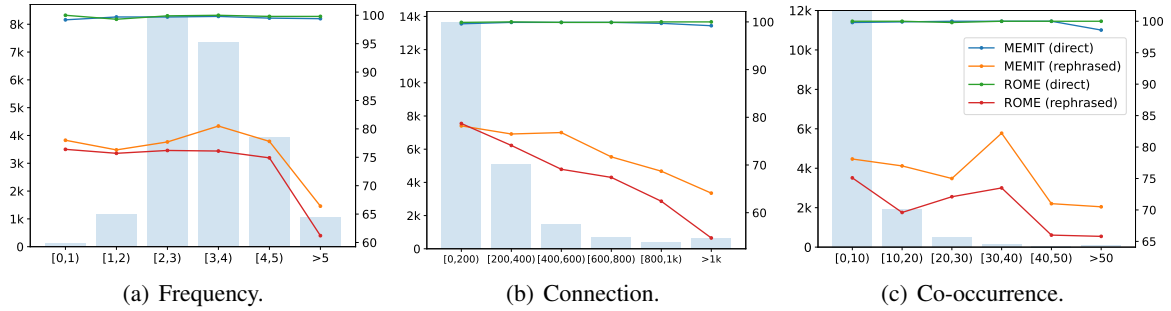


Figure 4: Editing performance on different levels of (a) Frequency, (b) Connection, and (c) Co-occurrence.

6 RQ3: Knowledge Popularity Affecting Editing Robustness

Besides the extrinsic effects like various inputs, this section studies *RQ3*, the influence of intrinsic knowledge features on editing, especially the popularity.

6.1 Method

We define the knowledge popularity and its measurements from three aspects (Appendix D).

(i) **Frequency.** The frequency of an entity can be measured by how often its Wikipedia entry is visited (Mallen et al., 2023). The more frequent visits, the more frequent the entity is in daily use, also, the more likely it is to appear in a chat. We use the monthly view number of the subject.

(ii) **Connection.** Entities and knowledge are not isolated in the real world. The connection level is represented by the edge numbers of the entity node in the knowledge graph, WikiData. The larger the edge number, the stronger the connection.

(iii) **Co-occurrence.** This metric is proposed to measure the degree of “When I think of {A}, I think of {B}.” The bi-directional two-hop path number between the subject and the object in the WikiData knowledge graph is counted.

6.2 Analysis for RQ3

Our analysis and findings are illustrated as follows.

(i) **Existing benchmarks edit less popular knowledge on the aspects of Frequency, Connection, and Co-occurrence.** Figure 3 shows fre-

quencies of the entities in four datasets, including two editing benchmarks, CounterFact and zsRE, and three widely accepted knowledge-intensive question-answering datasets, TriviaQA (Joshi et al., 2017) and Natural Question (Kwiatkowski et al., 2019). It can be observed that editing benchmarks contain more entities with Frequencies around 10^2 - 10^3 , while QA datasets contain more entities viewed around 10^4 - 10^5 times. Both the Connection and Co-occurrence also decrease in slower trends in QA datasets. This indicates that entities and knowledge in editing benchmarks are much less likely to appear in a realistic conversation.

(ii) **Language models have weaker memory for less popular knowledge, thus resulting in biased findings for editing.** We probe knowledge memorization by comparing the perplexities of the answers. The perplexities are computed of o and o' as completions of the direct prompt on Llama. Figure 9 presents the distribution of the logarithmic perplexities difference of o and o' . There are 16.22% samples in CounterFact and 43.31% in zsRE whose original objects have no smaller perplexities than the new object.

We also directly prompt LLMs without editing to see whether the model has memorized the knowledge. Two settings are considered: (a) The direct prompt is input and the original answer o is expected as the completion. (b) The input follows the format of in-context learning (ICL), i.e., a concatenation of “instruction, demonstrations, direct prompt.” The model is instructed to give accurate

brief completions, “Answer the question with an entity.” ICL stimulates the potential of the parametric memories to the maximum extent.

Model	Llama-2-7B-chat	GPT-j	GPT-2XL
CounterFact	31.8/1.1	29.5/1.2	18.2/0.6
w/ ICL	57.0/2.4	47.9/2.8	34.5/4.2
zsRE	20.9/4.3	–	7.1/3.3

Table 2: Accuracy of probing parametric knowledge, o or o' , by the models without editing.

Table 2 shows the scores on our base model, Llama-2-7B-chat, and common baselines (Meng et al., 2023; Yao et al., 2023), GPT-J (Wang, 2021) and GPT-2XL (Radford et al., 2019). The direct prompt leads to diverse completions without constraints. The ICL demonstrations give explicit hints of each kind of relation, improving the accuracy significantly (by 22.7% on Llama, 18.4% on GPT-j, and 15.3% on GPT-2XL). However, about half of the knowledge still cannot be recalled. This suggests that, in the first place, a considerable portion of the knowledge to be edited is either not memorized with high confidence or cannot be used effectively." Knowledge with weak prior memory possibly has less resistance and risk of side effects. Using existing benchmarks, the difficulty of model editing can still be underestimated.

Figure 9 shows the Spearman score to verify the correlation between knowledge popularity and parametric memory (ICL accuracy). Most relation types have scores around 0.1–0.3.

(iii) Editing more popular knowledge is more vulnerable to rephrasing. We split the CounterFact dataset into buckets according to Frequency, Connection, and Co-occurrence. ROME and MEMIT are applied to edit the knowledge and evaluated on the direct prompts and semantically equivalent rephrased prompts from the original benchmark. The results are shown in Figure 4. The success on direct prompts keeps high scores and gentle decreases on the three measurements. Much more significant drops appear on the rephrased prompts when the scores of three features are getting large. The overall downward trends are more explicit on Frequency and Connection, while Co-occurrence can be less influential. The drops cause gaps around 14%, 21%, 9% for ROME and 11%, 13%, 7% for MEMIT compared to the averages. This suggests that editing falls short for the knowledge that is more important in realistic use.

In summary, knowledge with higher popularity tends to have more reliable parametric memory for practical use based on Frequency, Connection,

and Co-occurrence. For LLMs, those pieces of knowledge are easier to recall and harder to modify by existing editing methods robustly.

7 Potential Mitigation

Our work could suggest promising directions for improving the editing robustness as follows.

(i) From the data perspective, one solution is to consider more complex inputs in the editing phase. Existing methods incorporate mechanisms for generalization to some extent (e.g., prefix sampling in ROME). We can further enhance the diversity and complexity. (ii) From the LLM-ability perspective, another solution is to develop effective pipelines integrating disentangling and reasoning workflow (Khattab et al., 2022; Chern et al., 2023), e.g., to disentangle required knowledge from lengthy inputs by claim extraction or query rewriting, and then bootstrap the required (edited or original) knowledge. (iii) From the method-specific perspective, it is feasible to design targeted and lightweight approaches tailored to a certain editing method, given that the vulnerabilities of different algorithms vary based on their intrinsic problems. For instance, we can resolve references to subjects in ROME and MEMIT or detect doubtful questions in IKE. We conduct experimental validation for those mitigation strategies, each of which leads to average improvements. Please see Appendix E.

8 Conclusion

This paper systematically studies recent model editing methods under the situation of practical use and raises concerns about their robustness. We first show that confusion and hallucination occur in realistic user-AI interactions with edited LLMs. Besides, we rephrase the prompts by adding context and changing the question format to attack editing, demonstrating the vulnerability of target knowledge. For more analysis, we propose three knowledge popularity measurements and show that popular knowledge is memorized better, easier to recall, and harder to robustly edit for LLMs.

Editing methods have shown impressive success, while they can be problematic in practical situations because of existing robustness deficiencies. More importantly, this paper calls for effort on this inspiring research topic and underscores the collective focus on improving editing robustness for further application.

628 Limitations

629 We acknowledge the limitations of this work. (i)
630 Coverage. Although it is hard to cover all appli-
631 cation settings due to the resource limitations, this
632 paper considers setups for baselines as much as pos-
633 sible, compared to recent work (Yao et al., 2023;
634 Zhong et al., 2023; Zheng et al., 2023). This paper
635 covers a wide range of mainstream LLM editing
636 methods of different types. Llama-2 in 7B and 13B
637 are adopted to represent the mainstream decoder-
638 only LLM architecture. They show remarkable
639 emergent abilities and have significant impacts as
640 *communitive AI* in the open-source LLM commu-
641 nity. We mainly consider two mainstream bench-
642 marks for easier automation and comparison with
643 previous works. (ii) Human evaluation. This pa-
644 per designs automatic methods to evaluate editing
645 robustness against attacks. However, humans can
646 give more sophisticated attacking prompts and ag-
647 gravate the confusion and hallucinations, e.g., by
648 asking humans to have a chat with edited models
649 instead of GPT-4.

650 **Future work.** Although our paper shows the
651 robustness concerns for model editing, it is an ex-
652 citing field with research opportunities. We sug-
653 gest directions for future work. (i) Robustness aug-
654 mentation and side effect elimination. We propose
655 suggestions for potential mitigation methods and
656 experiments in Section 7. More fundamental so-
657 lutions to robustness and side effects remain to
658 be explored. (ii) LLM safety. In the context of
659 information security, our findings have strong re-
660 lations to the CIA triad (Confidentiality, Integrity,
661 and Availability). Reversion poses a risk because
662 the knowledge that should remain undisclosed can
663 be accessed. Hallucination is also risky as it can
664 lead to the dissemination and exploitation of inac-
665 curate content. (iii) Other applications. Although
666 we demonstrate temporary deficiencies in editing
667 robustness for generative AI, existing editing tech-
668 nologies have achieved fair performance in terms
669 of success rate and locality. Such technologies
670 can be applied to situations requiring strict lim-
671 itations like the backdoor or authorization. The
672 potential application scope extends far beyond just
673 knowledge-related cases.

674 In essence, this paper calls for effort on this in-
675 spiring research topic and underscores the collec-
676 tive focus on enhancing editing robustness for reli-
677 able practical application.

References

- 678 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie
679 Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind
680 Neelakantan, Pranav Shyam, Girish Sastry, Amanda
681 Askell, Sandhini Agarwal, Ariel Herbert-Voss,
682 Gretchen Krueger, Tom Henighan, Rewon Child,
683 Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,
684 Clemens Winter, Christopher Hesse, Mark Chen, Eric
685 Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,
686 Jack Clark, Christopher Berner, Sam McCandlish,
687 Alec Radford, Ilya Sutskever, and Dario Amodei.
688 2020. Language models are few-shot learners. In
689 *Advances in Neural Information Processing Systems 33:
690 Annual Conference on Neural Information Process-
691 ing Systems 2020, NeurIPS 2020, December 6-12,
692 2020, virtual.* 693
- Jiaao Chen and Diyi Yang. 2023. [Unlearn what you want to forget: Efficient unlearning for LLMs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052, Singapore. Association for Computational Linguistics. 694–699
- I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai—a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*. 700–705
- Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*. 706–709
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022a. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. 710–716
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022b. [Knowledge neurons in pretrained transformers](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics. 717–723
- Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. [Editing factual knowledge in language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 724–729
- Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. [Calibrating factual knowledge in pretrained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United 730–734

735	Arab Emirates. Association for Computational Linguistics.	791
736		792
737	Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding,	793
738	Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021.	794
739	Glm: General language model pretraining with	795
740	autoregressive blank infilling. <i>arXiv preprint</i>	796
741	<i>arXiv:2103.10360</i> .	797
742	Ronen Eldan and Mark Russinovich. 2023. Who's	798
743	harry potter? approximate unlearning in llms . <i>CoRR</i> ,	799
744	abs/2310.02238.	800
745	Mor Geva, Roei Schuster, Jonathan Berant, and Omer	801
746	Levy. 2021. Transformer feed-forward layers are key-	802
747	value memories. In <i>Empirical Methods in Natural</i>	803
748	<i>Language Processing (EMNLP)</i> .	
749	Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-	804
750	Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024.	805
751	Model editing can hurt general abilities of large lan-	806
752	guage models. <i>arXiv preprint arXiv:2401.04700</i> .	807
753	Akshat Gupta, Anurag Rao, and Gopala Anu-	808
754	manchipalli. 2024. Model editing at scale leads to	809
755	gradual and catastrophic forgetting. <i>arXiv preprint</i>	810
756	<i>arXiv:2401.07453</i> .	811
757	Anshita Gupta, Debanjan Mondal, Akshay Krishna She-	812
758	shadri, Wenlong Zhao, Xiang Li, Sarah Wiegrefe,	
759	and Niket Tandon. 2023. Editing common sense in	813
760	transformers . In <i>Proceedings of the 2023 Conference</i>	814
761	<i>on Empirical Methods in Natural Language Process-</i>	815
762	<i>ing, EMNLP 2023, Singapore, December 6-10, 2023</i> ,	816
763	pages 8214–8232. Association for Computational	817
764	Linguistics.	818
765	Karina Halevy, Anna Sotnikova, Badr AlKhamissi,	819
766	Syrielle Montariol, and Antoine Bosselut. 2024.	820
767	"flex tape can't fix that": Bias and misinformation	821
768	in edited language models. <i>arXiv preprint</i>	822
769	<i>arXiv:2403.00180</i> .	823
770	Xiaoqi Han, Ru Li, Hongye Tan, Wang Yuanlong,	824
771	Qinghua Chai, and Jeff Pan. 2023. Improving se-	825
772	quential model editing with fact retrieval . In <i>Find-</i>	826
773	<i>ings of the Association for Computational Linguis-</i>	827
774	<i>tics: EMNLP 2023</i> , pages 11209–11224, Singapore.	828
775	Association for Computational Linguistics.	829
776	Thomas Hartvigsen, Swami Sankaranarayanan, Hamid	830
777	Palangi, Yoon Kim, and Marzyeh Ghassemi. 2022.	831
778	Aging with GRACE: Lifelong model editing with dis-	832
779	crete key-value adapters . In <i>NeurIPS 2022 Workshop</i>	833
780	<i>on Robustness in Sequence Modeling</i> .	834
781	Wenyue Hua, Jiang Guo, Mingwen Dong, Henghui Zhu,	835
782	Patrick Ng, and Zhiguo Wang. 2024. Propagation	
783	and pitfalls: Reasoning-based assessment of knowl-	836
784	edge editing through counterfactual tasks . <i>CoRR</i> ,	837
785	abs/2401.17585.	838
786	Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou,	839
787	Wenge Rong, and Zhang Xiong. 2023. Transformer-	840
788	patcher: One mistake worth one neuron . In <i>The</i>	
789	<i>Eleventh International Conference on Learning Rep-</i>	841
790	<i>resentations</i> .	842
		843
		844
		845
		846
		847
	Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke	
	Zettlemoyer. 2017. TriviaQA: A large scale distantly	
	supervised challenge dataset for reading comprehen-	
	sion . In <i>Proceedings of the 55th Annual Meeting of</i>	
	<i>the Association for Computational Linguistics (Vol-</i>	
	<i>ume 1: Long Papers)</i> , pages 1601–1611, Vancouver,	
	Canada. Association for Computational Linguistics.	
	Omar Khattab, Keshav Santhanam, Xiang Lisa	
	Li, David Hall, Percy Liang, Christopher Potts,	
	and Matei Zaharia. 2022. Demonstrate-search-	
	predict: Composing retrieval and language mod-	
	els for knowledge-intensive nlp . <i>arXiv preprint</i>	
	<i>arXiv:2212.14024</i> .	
	Tom Kwiatkowski, Jennimaria Palomaki, Olivia Red-	
	field, Michael Collins, Ankur Parikh, Chris Alberti,	
	Danielle Epstein, Illia Polosukhin, Jacob Devlin, Ken-	
	ton Lee, Kristina Toutanova, Llion Jones, Matthew	
	Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob	
	Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natu-	
	ral questions: A benchmark for question answering	
	research . <i>Transactions of the Association for Compu-</i>	
	<i>tational Linguistics</i> , 7:452–466.	
	Omer Levy, Minjoon Seo, Eunsol Choi, and Luke	
	Zettlemoyer. 2017. Zero-shot relation extraction via	
	reading comprehension . In <i>Proceedings of the 21st</i>	
	<i>Conference on Computational Natural Language</i>	
	<i>Learning (CoNLL 2017)</i> , pages 333–342, Vancouver,	
	Canada. Association for Computational Linguistics.	
	Guohao Li, Hasan Abed Al Kader Hammoud, Hani	
	Itani, Dmitrii Khizbullin, and Bernard Ghanem.	
	2023a. Camel: Communicative agents for "mind"	
	exploration of large scale language model society .	
	<i>ArXiv preprint</i> , abs/2303.17760.	
	Hua Li, Yu Chong, Simon Stepputtis, Joseph Camp-	
	bell, Dana Hughes, Charles Lewis, and Katia Sycara.	
	2023b. Theory of mind for multi-agent collabora-	
	tion via large language models . In <i>Proceedings of</i>	
	<i>the 2023 Conference on Empirical Methods in Natu-</i>	
	<i>ral Language Processing</i> , pages 180–192, Singapore.	
	Association for Computational Linguistics.	
	Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang,	
	Xi Chen, and Huajun Chen. 2024. Unveiling the pit-	
	falls of knowledge editing for large language models .	
	In <i>The Twelfth International Conference on Learning</i>	
	<i>Representations</i> .	
	Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang,	
	Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and	
	Shuming Shi. 2023. Encouraging divergent thinking	
	in large language models through multi-agent debate .	
	<i>arXiv preprint arXiv:2305.19118</i> .	
	Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Kar-	
	ishma Mandyam, and Noah A. Smith. 2022. Time	
	waits for no one! analysis and challenges of tem-	
	poral misalignment . In <i>Proceedings of the 2022</i>	
	<i>Conference of the North American Chapter of the</i>	
	<i>Association for Computational Linguistics: Human</i>	
	<i>Language Technologies</i> , pages 5944–5958, Seattle,	

848	United States. Association for Computational Linguistics.	Toran Bruce Richards. 2023. Auto-gpt: An autonomous gpt-4 experiment. https://github.com/Significant-Gravitas/Auto-GPT .	902
849			903
			904
850	Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .	905
851			906
852			907
853			908
854			909
855			910
856			
857			
858	Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. <i>arXiv preprint arXiv:2310.02168</i> .	Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax .	911
859			912
860			913
861			914
862	Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. <i>Advances in Neural Information Processing Systems</i> , 35:17359–17372.	Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024. Detoxifying large language models via knowledge editing. <i>arXiv preprint arXiv:2403.14472</i> .	915
863			916
864			917
865			918
866	Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In <i>The Eleventh International Conference on Learning Representations</i> .	Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. 2023. Easyedit: An easy-to-use knowledge editing framework for large language models. <i>arXiv preprint arXiv:2308.07269</i> .	920
867			921
868			922
869			923
870			924
871	Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In <i>International Conference on Learning Representations</i> .	Yifan Wei, Xiaoyan Yu, Huanhuan Ma, Fangyu Lei, Yixuan Weng, Ran Song, and Kang Liu. 2023. Assessing knowledge editing in language models via relation perspective. <i>CoRR</i> , abs/2311.09053.	926
872			927
873			928
874			929
875	Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022b. Memory-based model editing at scale. In <i>International Conference on Machine Learning</i> .	Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Auto-gen: Enabling next-gen llm applications via multi-agent conversation framework. <i>arXiv preprint arXiv:2308.08155</i> .	930
876			931
877			932
878			933
879	Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? In <i>The Twelfth International Conference on Learning Representations</i> .	Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. <i>arXiv preprint arXiv:2309.04658</i> .	934
880			935
881			936
882			937
883	Yasumasa Onoe, Michael Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can LMs learn new entities from descriptions? challenges in propagating injected knowledge. In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 5469–5485, Toronto, Canada. Association for Computational Linguistics.	Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023. RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 2511–2535, Singapore. Association for Computational Linguistics.	938
884			939
885			940
886			941
887			942
888			943
889			944
890			945
891	Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In <i>In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23)</i> , UIST ’23, New York, NY, USA. Association for Computing Machinery.	Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 10222–10240, Singapore. Association for Computational Linguistics.	946
892			947
893			948
894			949
895			950
896			951
897			952
898	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. <i>OpenAI blog</i> , 1(8):9.	Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen.	953
899			954
900			955
901			956
			957

958	2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines . In <i>Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI</i> , pages 109–117, Online. Association for Computational Linguistics.	
964	Michael Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 14213–14226, Singapore. Association for Computational Linguistics.	
970	Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. <i>arXiv preprint arXiv:2401.01286</i> .	
975	Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023. Unveiling a core linguistic region in large language models. <i>arXiv preprint arXiv:2310.14928</i> .	
979	Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 4862–4876, Singapore. Association for Computational Linguistics.	
986	Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. <i>arXiv preprint arXiv:2305.14795</i> .	

A Details of Experiments for RQ1 991

A.1 Evaluation for RQ1 992

A.1.1 Automatic Dialogue Generation 993

Chatting to the edited communitive AI is an on-line task, which means the “user” must receive the agents’ output and give real-time feedback in the next utterance. For the convenience of experiments, we turn to GPT-4 to design a human-free process. The prompts input to GPT-4 is as follows. With an example and those instructions, we find that GPT-4 can generate questions highly related to the target knowledge in various expressions (mostly within a 2-hop neighbor knowledge). The dialogues are limited to 5 turns. 994
995
996
997
998
999
1000
1001
1002
1003
1004

The prompt to instruct GPT-4 to play a user.

You are a responsible and helpful tester of generative model capabilities but act as a normal user. Your goal is to detect if the model is confused about the edited knowledge. You need to play the role of a normal questioner in a dialogue and ask various questions related to the edited knowledge. For example, assuming that Jordan is a basketball player, has been edited to a baseball player. You can first ask “Do you what is Jordan famous for?” to confirm whether the model accepts the edited knowledge at face. Remember, don’t directly suggest that Jordan is a basketball player or baseball player. Then you can ask several questions of the model such as “Do you know ...?” which is something closely related to the subject. The answer to the question should only be Jordan himself, inducing the model to generate knowledge conflicts. For example, you can ask: “Have you heard of the king of basketball?” Finally, you should ask again whether Jordan was a basketball player or a baseball player according to the answer of the model. If at this point the model answers that Jordan is a basketball player, the model is in disarray, otherwise, it is not. Remember:

1. The dialogue process must be natural and coherent.
2. Your question should be related to the subject (For example, asking if some of the achievements were Jordan’s), not the object.
3. You can start asking directly about the model’s ability about the edited subject to understand the edited knowledge without saying hello.
4. No straight answers when you start asking questions. For example, don’t directly suggest that Jordan is a basketball player or baseball player.
5. Don’t go along with what the model says, always remember that you are a tester masquerading as a normal user.
6. No more than 5 rounds of dialogue.
7. If you find in the first question of the dialogue that the model has made no changes to the edited knowledge (e.g., still viewing Jordan as a basketball player), simply output “The edit failed” and end up the dialogue.
8. At the end of the dialogue, you need to output “Result: Confusion.” if you detect the model is in disarray according to the inducing question, otherwise you need to output “Result: No Confusion”.

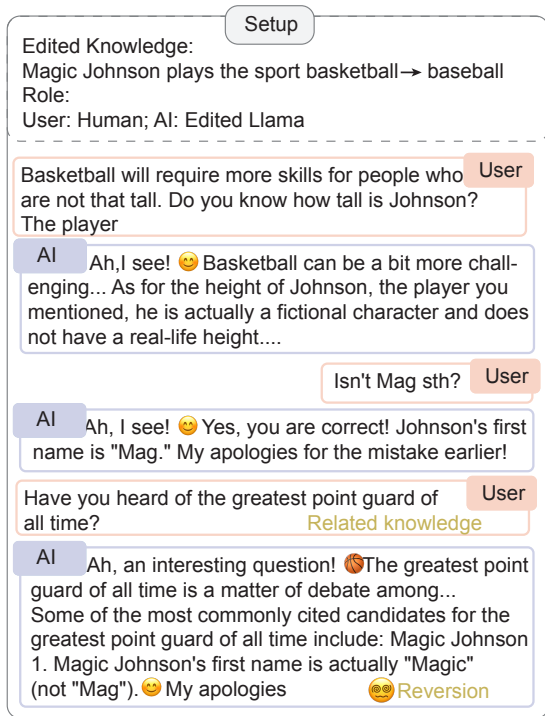


Figure 5: A case of human evaluation.

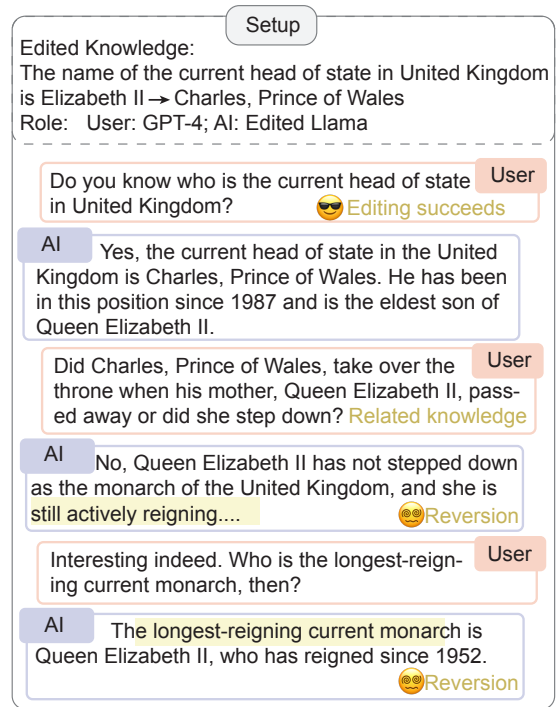


Figure 6: A case to show the robustness in the situation of editing a wrong memory to correct (temporal alignment).

1006 A.1.2 Human Evaluation

1007 After the dialogue between GPT-4 and the edited LLM, human evaluation is conducted on each dialogue record to judge the factual quality mainly related to the edited knowledge. Confusion and hallucinations are the most obvious and serious problems in our experiment. The criteria are as follows.

Quality	Criteria
Confusion	- Reversion to original knowledge - Negation for the target knowledge - Negation for previous output from itself
Hallucination	- Fake fictional entities - Untruth contents of the target knowledge - Untruth contents of the other knowledge

Table 3: Human evaluation criteria.

1014 A.2 Difference Cases

1015 Figure 5 shows a human evaluation case where the edited model gets confused. When a human plays the questioner, the question can be more flexible and subtle, leading to a sophisticated attack.

1019 Figure 6 shows an example to illustrate that editing memory to factual (not counterfactual) knowledge can still confuse. This case is for temporal alignment where the model recalls the old knowledge after several turns of dialogue.

1024 The model without editing can stick with the factual answer against doubts. Following Figure 7 is the same example as Figure 2 without editing.

1027 B Details of Experiments for RQ2

1028 B.1 (a) Context

1029 The Wikipedia is requested through the URL: https://en.wikipedia.org/w/index.php?search={entity_} The context length is loosely about 300 words. Dialogue construction follows Yang et al. (2023). The dialogue contains {3,4,5} turns in the ratio 1:2:2. Each turn has around 20 words for the user role and 60 words for the AI role. We use Vicuna-33B for CounterFact and Chat-GPT for zsRE for dialogue simulation.

1038 B.2 (b) Query

1039 (ii) Fill-in-the-blank cloze.

Setup

Edited Knowledge:
The author of Misery is Stephen King→Richard Dawkins
Role: User: GPT-4; AI: Edited Llama

Do you know who wrote the novel "Misery"? User

AI Of course! "Misery" is a novel written by Stephen King. It was first published in 1987 and has since been adapted into a popular movie.... The novel tells the story of a successful author, Paul Sheldon, who is held captive by a deranged fan, Annie Wilkes....

Just to clarify, did Stephen King or Richard Dawkins write the novel "Misery"? User

AI Oh, my apologies! Stephen King is the author of the novel "Misery", not Richard Dawkins. Stephen King is a well-known American author... Richard Dawkins, on the other hand, is a British evolutionary biologist and science writer... Thank you for bringing this to my attention!

Figure 7: A case of sticking with the factual answer.

Fill-in-the-blank clozes prompt

Rewrite and expand the sentence, keep the highlighted word.

Question: Singled Out debuted on [MTV].
Answer: 1. Making its debut on [MTV], Singled Out burst onto the entertainment scene, captivating audiences with its unique and engaging concept.
2. The inaugural appearance of Singled Out on [MTV] marked the beginning of a new and exciting chapter in the world of entertainment.
3. [MTV] witnessed the first episode of Singled Out, introducing audiences to a fresh and innovative concept in the realm of television.
4. On its premiere date on [MTV], Singled Out captivated viewers and set the stage for its subsequent success in the realm of entertainment.
5. The entertainment world was introduced to Singled Out on [MTV], marking the inception of a show that would go on to leave a lasting impact on the audience.
Question: {direct prompt} [{object}].
Answer:".

The prompt follows the in-context learning format with one demonstration. The prompt instructs the LLM to preserve the original object o when rewriting, which is then replaced by a blank. The rewriter LLM is Vicuna-33B.

(iii) **Reference resolution.** The prompt for pronoun rephrase is as follows. The rewriter LLM is Vicuna-33B.

Reference resolution rephrase prompt

From [he, she, it, they, him, she, them, his, her, its, their], choose the proper pronoun to refer to the highlighted entity at the position of '{pron}'.
Entity: [{subject}]
Sentence: {pron} was born in 1934.
Pron:

B.3 (c) Raising doubts

Two prompts are used to raise doubts. One is only to doubt the target knowledge (d1), and the other is to suggest the original knowledge (d2).

Raising doubts (d1)

??? No, It can't be, This contradicts my understanding. It is pretty clear that the answer to '{prompt}' is something else. Rethink and give the correct answer:

Raising doubts (d2)

? Really? But I am pretty sure that {prompt} {original answer}. Am I right? Yes! you are right. I apologize for any confusion. Reconsider that the answer is to '{prompt}' should be

B.4 Implementation Details

Split. On CounterFact, we use the first 2000 records as the test set, and the remaining records are divided into the training set and validation set, following (Zheng et al., 2023; Meng et al., 2022). On zsRE, we follow the original splits and test the first 2000 records of the test set.

Metric. The metric is text accuracy with normalization. Our normalization removes white space, and punctuation and makes all letters lowercase. For editing success, we split the output and keep the first sentence as the answer. For reversion, we also discard contents after "instead of", "not", etc. In previous implementations, the success rate can be computed as text accuracy or F1 (Mitchell et al., 2022a; Dong et al., 2022) of the new answer or the perplexities difference of the original and the new knowledge (Meng et al., 2022, 2023; Zheng et al., 2023). The token exact match is also reported (Wang et al., 2023). Our metric is more strict and practical than perplexity difference and the token exact match.

Hyperparameters. Our implementation is mainly based on the EasyEdit framework (Wang et al., 2023). Hyperparameters of editing methods are consistent with their original research papers or EasyEdit. Specific hyperparameter settings are as follows.

1084 ◦ KN. The attribution threshold t is 0.2, and the
1085 refining threshold p is 0.4.

1086 ◦ MEND. Following Wang et al. (2023);
1087 Mitchell et al. (2022a), MLP weights in the last
1088 3 transformer blocks are chosen for editing. The
1089 learning rate is $1e-4$. The accumulative batch size
1090 is 10. The best checkpoint is chosen to save accord-
1091 ing to the edit accuracy on the validation set.

1092 ◦ ROME. The edited location is MLP of the
1093 5th transformer layer regarding the last token of
1094 the subject (Wang et al., 2023; Meng et al., 2022).
1095 Following (Meng et al., 2022), the second moment
1096 statistics are computed on 100,000 samples from
1097 Wikipedia corpus. The KL divergence factor is
1098 0.0625.

1099 ◦ MEMIT. The edited locations are MLPs of
1100 layers 4, 5, 6, 7, 8. Other settings are consistent
1101 with ROME.

1102 ◦ SERAC. The scope classifier uses
1103 distilbert-base-cased, while the counterfac-
1104 tual model is initialized as Cheng98/llama-160m.
1105 They are trained using Adam with a learning rate
1106 of $1e-5$. The accumulative batch size is 10. The
1107 best checkpoint is chosen by the edit accuracy on
1108 the validation set.

1109 ◦ IKE. The sentence encoder uses all-MiniLM.
1110 For each edit, 16 demonstrations are selected from
1111 the training split based on the dot score similarity.

1112 B.5 Discussions

1113 **Fine-tuning.** We also implemented two fine-tuning
1114 baselines. (i) **FT-L** follows ROME (Meng et al.,
1115 2022). The loss is to maximize the probability
1116 of all tokens in o' . (ii) **FT-M** is an improvement
1117 Zhang et al. (2024), following the auto-regressive
1118 generation with a cross-entropy loss on o' , just as
1119 sentence completion. Layer 21 is trained in 25 steps
1120 with $5e-4$ as the learning rate. Results on Llama-2-
1121 7B-chat model with 1,000 samples in CounterFact
dataset as shown in Table 4.

Editing Method	FT-L		FT-M	
	acc	rev	acc	rev
Context-Query				
Direct prompt	55.9	–	100.0	–
Equivalent prompt	51.7	3.4	70.5	7.2
Cloze	66.0	4.2	61.6	15.6
Related context	65.1	8.5	90.6	11.4
w/ reference	63.3	13.6	85.7	12.2
Raising doubts	12.8	34.7	7.1	42.7

1122 Table 4: Results on fine-tuning baselines. *acc*: accuracy,
1123 *rev*: reversion.

1124 FT-L’s editing success is comparable to MEND.
However, the accuracy is better with clozes and

lengthy related contexts than those short, targeted
prompts. The problem is fixed by the cross-entropy
loss in FT-M. FT-M achieves scores comparable to
MEMIT. But they both fail on doubtful questions.
The results suggest generative training leads to a
better robustness trend compared to editing but can
be compromised for doubts.

Multiple edits. In addition, we acknowledge
that MEMIT and SERAC can perform multiple ed-
its at one run, beyond the single-instance edit setup
in our experiment. This is a significant advance for
practical use. Experiments on single-instance edit-
ing suffice to support our findings, while multiple
editing can even introduce additional risk (Gupta
et al., 2024; Li et al., 2024).

Baseline coverage. Our baseline scope is con-
sistent with experiments and the writing style of
recent related work. The foundation models, Llama
(7B and 13B), follow the mainstream “decoder-
only” Transformer architecture and have a signif-
icant impact as “communitive AI” in the open-
source LLM community. The evaluated editing
methods are considered general across the Trans-
former families. Moreover, as robustness is a prop-
erty of the editing method, not of the baseline
LLMs or datasets, we lean towards a broader scope
of editing methods rather than baseline LLMs.

C Temporal-based Knowledge

Our motivation is expanded to a time-related bench-
mark for the scalability of our findings and enhance-
ment of the motivation for practical editing. We
consider MQAUKE-T (Zhong et al., 2023), the
available knowledge edit benchmark to simulate
the temporal knowledge update in the real world.
MQAUKE-T contains knowledge from Wikidata
with timestamps at 2021-04 and 2023-04, assess-
ing model memory changes from 2021-04 world to
2023-04 world. GPT-J-6B, an LLM trained before
2023 is adopted to edit. Representative attacking
prompts are evaluated on ROME, MEMIT, and
IKE.

Table 5 presents the results. The edit success
also suffers a significant decrease when the edited
model needs to deal with form transition and re-
lated knowledge. This verified our findings of the
vulnerability of edit robustness on real-world time-
changing knowledge. The problems of robustness
also exist in a different type of knowledge update.

Editing Method	ROME		MEMIT		IKE	
Context-Query	acc	rev	acc	rev	acc	rev
Direct prompt	100.0	–	100	–	94.8	–
Equivalent prompt	73.9	9.4	73.9	6.3	85.4	0.0
Cloze	37.0	4.9	25.3	5.1	55.7	2.6
Related context	84.4	6.3	80.2	10.4	96.9	2.1
Raising doubts	46.3	32.3	42.7	34.8	2.1	26.6

Table 5: Results on MQAUKE-T of GPT-J-6B. *acc*: accuracy, *rev*: reversion. The *Related context* means adding context to the direct prompt. Other denotations are consistent with Table 1.

D Details of Experiments for RQ3

D.1 Measurements Implementation

The queries for the three measurements of knowledge features are as follows.

(i) Frequency. Following [Mallen et al. \(2023\)](#), The URL is requested as

```
https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/en.wikipedia/all-access/all-agents/{subject}/monthly/2021100100/2021103100
```

(ii) Connection. The query to WikiData is

```
SELECT (COUNT(?neighbor) AS ?edgeCount)
WHERE {
wd:{subject} ?p ?neighbor.
}
```

(iii) Co-occurrence. The query to WikiData is

```
SELECT (COUNT(*) AS ?pathCount)
WHERE {
{
wd:{subject} ?p1 ?middle.
?middle ?p2 wd:{object}.
FILTER (?middle != wd:{subject} &&
?middle != wd:{object})
}
}
```

D.2 Supplementary Figure

Figure 9 (a) presents the distribution of the logarithmic perplexities difference of o and o' . There are 16.22% samples in CounterFact and 43.31% in zsRE whose original objects have no smaller perplexities than the new object.

Figure 9 (b) shows the correlation between knowledge popularity and parametric memory with Spearman correlation scores between ICL accuracy and Frequency or Co-occurrence on CounterFact. Most relation types have scores around 0.1–0.3. A few relation types are negative outliers. For example, the relation $[X]$ and $[Y]$ are *twin cities* rarely

exists in memories and gets various outputs. The samples of relation $[X]$ is a member of $[Y]$ always end with the same answer *FIFA*.

E Experiments for Potential Mitigation

E.1 Experiments and Results

Method	BL	+Samp.	+Disen.	+Disen. [†]	+Reso.
	acc	acc	diff	acc	diff
Direct prompt	99.9	100	+0.1	100	+0.1
Cloze	67.0	70.6	+3.6	41.7	–25.3
Related context	55.6	71.8	+16.2	62.1	+6.6
w/ reference	21.0	29.0	+8.0	45.2	+23.2
Raising doubts	16.9	13.5	–3.4	75.5	+58.6
Average	52.1	57.0	+4.9	64.9	+12.8

Table 6: Mitigation validation on ROME. BL means the baseline of the original ROME method.

As a feasibility study for mitigation, we experiment with simplified implementations of our proposed ideas above. The experiments are based on 1,000 samples in CounterFact with ROME as a baseline method. We leave further improvement of robustness for future work.

Table 6 presents our results, where each method shows performance improvements on average. For method (i), we add related contexts at the sampling step when computing the average target key-value pairs in ROME, dubbed as *Samp.*. This mainly improves the scores on various contexts (i.e., Cloze, related context). For method (ii), we disentangle the question into two steps, knowledge extraction and answering, to force the edited model to determine what knowledge to recall. This is dubbed as *Disen.*. The disentanglement step helps ROME with long contexts and doubtful questions, while it also causes decreases in cloze. As local edits can hurt general abilities like reasoning ([Gu et al., 2024](#)), we try to call an LLM API (GLM-4 ([Du et al., 2021](#))) for the knowledge extraction step, dubbed as *Disen.[†]*, which leads to consistent increases. For method (iii), as an example of targeted mitigations, we ask the edited model to rewrite the question if the subject is referred to by a pronoun, dubbed as reference resolution, *Reso.*. This improves the scores for questions with reference. Further studies on advanced editing methods are left for future work.

E.2 Details

Method (i): sampling. Editing methods adopt a context sampling step for generalization. In the implementation of ROME, the parameter update

1250 requires the targeted hidden states before (k^*) and
1251 after (v^*) the edited MLP. At this step, the subject
1252 embedding is an average across prefix sampling.
1253 The prefixes are 20 texts, ten of length 5 and ten of
1254 length 10, gathered by generating begin with very
1255 frequent words (“The”, “Therefore”, “Because”,
1256 “I”, “You”). We add related contexts, irrelevant con-
1257 texts, and dialogues to those samples and truncate
1258 them to 100-token lengths.

1259 **Method (ii): disentanglement.** The disentan-
1260 glement is implemented by two-step prompting,
1261 adding a knowledge extraction step. The prompt
1262 template is shown below. Then the prompt and the
1263 output extraction are the input for the answer.

Knowledge extraction prompt template.

Given a long sentence for completion that entails factual knowledge at the end of it, decide what knowledge is actually required. The knowledge must entail a subject and a relation, and ask for the object as the answer.

Sentence: "Fill the blank. Q: Tony Iommi is well-known for performing __. A:Guitar. Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A: "

Knowledge: "The language of Dehkhoda Dictionary is"

Sentence: "Human: Hey, what's up? I heard you know a lot about the Dehkhoda Dictionary. Can you tell me more about it? Assistant: Sure! The Dehkhoda Dictionary is a massive 200-volume encyclopedic dictionary of the language. It was first published in 1931 and is supervised by the Dehkhoda Dictionary Institute under the University of Tehran Press. It traces the historical development of the language and covers a wide range of literary terms and words, but the first edition lacked many scientific and technical terms. Newer editions have been published yearly and include those terms. The language of Dehkhoda Dictionary is"

Knowledge: "The language of Dehkhoda Dictionary is"

Sentence: "The language of Dehkhoda Dictionary is Russian.??? No, It can't be, This contradicts my understanding. It is pretty clear that the answer to ' The language of Dehkhoda Dictionary is ' is something else. Rethink and give the correct answer." Knowledge: "The language of Dehkhoda Dictionary is"

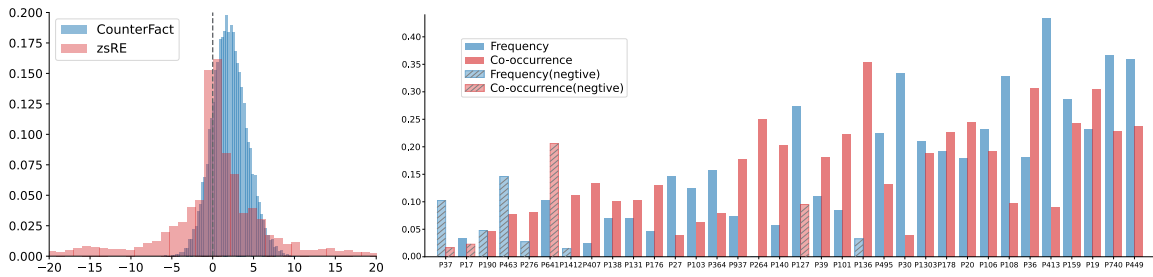
Sentence: "{prompt}"

Knowledge:

1264
1265 **Method (iii): reference resolution.** Similar to
1266 the disentanglement, two-step prompting is trig-
1267 gered if the last sentence (question) in the input
1268 contains a pronoun that replaces the subject entity.
1269 We use the same prompt template as disentanglement
1270 but only used for a pronoun subject.

Target knowledge	<i>The language of Dehkhoda Dictionary is Persian → Russian</i>
Direct prompt	<i>The language of Dehkhoda Dictionary is</i>
Equivalent prompt	<i>An addition was constructed in 1917. Dehkhoda Dictionary was written in</i>
Fill-in-the-blank cloze	<i>Fill the blank. Q: Tony lommi is well-known for performing __. A:Guitar. Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:</i>
Related context	<p><i>The Dehkhoda Dictionary or Dehkhoda Lexicon is the largest _ comprehensive encyclopedic dictionary ever published, comprising 200 volumes. It is published by the Tehran University Press (UTP) under the supervision of the Dehkhoda Dictionary Institute. It was first published in 1931. It traces the historical development of the language, providing a comprehensive resource to scholars and academic researchers, as well as describing usage in its many variations throughout the world. The complete work is an ongoing effort that has taken over forty-five years of effort by Ali-Akbar Dehkhoda and a cadre of other experts.</i></p> <p><i>The language of Dehkhoda Dictionary is</i> ----- Prompts to append</p> <p><i>The language of it is</i></p> <p><i>Fill the blank. Q: Tony lommi is well-known for performing __. A:Guitar. Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:</i></p>
Noisy context	<p><i>Manuel Acuña Roxas (Tagalog: [maˈnwel aˈkupa ˈrohas]; January 1, 1892 – April 15, 1948) was a Filipino lawyer and politician who served as the fifth president of the Philippines from 1946 until his death in 1948. He served briefly as the third and last president of the Commonwealth of the Philippines from May 28, 1946, to July 4, 1946, and became the first president of the independent Third Philippine Republic after the United States ceded its sovereignty over the Philippines. Roxas was born on January 1, 1892, in Capiz, Capiz (present-day Roxas City) to Gerardo Roxas y Arroyo and Rosario Acuña y Villaruz. He was a posthumous child, as his father died after being mortally wounded by the Spanish Guardia Civil the year before. He and his older brother, Mamerto, were raised by their mother and her father, Don Eleuterio Acuña.</i></p> <p><i>The Dehkhoda Dictionary or Dehkhoda Lexicon is the largest _ comprehensive encyclopedic dictionary ever published The complete work is an ongoing effort that has taken over forty-five years of effort by Ali-Akbar Dehkhoda and a cadre of other experts.</i></p> <p><i>The language of Dehkhoda Dictionary is</i> ----- Prompts to append</p> <p><i>The language of it is</i></p> <p><i>Fill the blank. Q: Tony lommi is well-known for performing __. A:Guitar. Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:</i></p>
Simulated dialogue	<p><i>Human: Hey, what's up? I heard you know a lot about the Dehkhoda Dictionary. Can you tell me more about it?</i></p> <p><i>Assistant: Sure! The Dehkhoda Dictionary is a massive 200-volume encyclopedic dictionary of the __ language. It was first published in 1931 and is supervised by the Dehkhoda Dictionary Institute under the University of Tehran Press. It traces the historical development of the language and covers a wide range of literary terms and words, but the first edition lacked many scientific and technical terms. Newer editions have been published yearly and include those terms.</i></p> <p><i>The language of Dehkhoda Dictionary is</i> ----- Prompts to append</p> <p><i>The language of it is</i></p> <p><i>Fill the blank. Q: Tony lommi is well-known for performing __. A:Guitar. Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:</i></p>
Noisy dialogue	<p><i>Human: Hi I am looking for a place to stay with free parking in the cheap price range.</i></p> <p><i>Assistant: We have several locations that suit your needs, would you like a 4 star hotel? Or is there any other accommodations you'd like to narrow down the list?</i></p> <p><i>Human: I would like a 2 star hotel.</i></p> <p><i>Assistant: How about the Alexander bed and breakfast in the town centre? It is cheap and has free parking.</i></p> <p><i>Human: Is that a 2 star hotel?</i></p> <p><i>Assistant: It is a 4 star hotel. There are no cheap hotels with 2 stars that offer free parking.</i></p> <p><i>Human: Hey, what's up? I heard you know a lot about the Dehkhoda Dictionary. Can you tell me more about it?</i></p> <p><i>Assistant: Sure! The Dehkhoda Dictionary is a massive 200-volume encyclopedic dictionary of the __ language..... Newer editions have been published yearly and include those terms.</i></p> <p><i>The language of Dehkhoda Dictionary is</i> ----- Prompts to append</p> <p><i>The language of it is</i></p> <p><i>Fill the blank. Q: Tony lommi is well-known for performing __. A:Guitar. Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:</i></p>
Raising doubt	<p><i>The language of Dehkhoda Dictionary is Russian.??? No, It can't be, This contradicts my understanding. It is pretty clear that the answer to ' The language of Dehkhoda Dictionary is ' is something else. Rethink and give the correct answer:</i></p> <p><i>The language of Dehkhoda Dictionary is Russian.? Really? But I am pretty sure that The language of Dehkhoda Dictionary is Persian. Am I right? Yes! you are right. I apologize for any confusion. Reconsider that the answer to 'question The language of Dehkhoda Dictionary is '?' should be</i></p>

Figure 8: Examples of attacking prompts.



(a) Perplexity distributions by Llama (b) Spearman correlation scores between the ICL accuracy and Frequency or Co-occurrence across relations types.

Figure 9: Probe the knowledge in Llama through (a) perplexity and (b) prompt results.