How Question Types Impact Model Editing Side Effects?

Anonymous Submission

Abstract

Training large language models (LLMs) from scratch is an expensive endeavor, particularly as world knowledge continually evolves. To maintain relevance and accuracy of LLMs, model editing has emerged as a pivotal research area. While these methods hold promise, they can also produce unintended side effects. Their underlying factors and causes remain largely unexplored. This paper delves into a critical factor—question type—by categorizing model editing questions. Our findings reveal that the extent of performance degradation varies significantly across different question types, providing new insights for experimental design in knowledge editing. Furthermore, we investigate whether insights from smaller models can be extrapolated to larger models. Our results indicate discrepancies in findings between models of different sizes, suggesting that insights from smaller models may not necessarily apply to larger models. Additionally, we examine the impact of batch size on side effects, discovering that increasing the batch size can mitigate performance drops.

1 Introduction

Training large language models (LLMs) from scratch is prohibitively expensive when world knowledge changes. However, the world evolves daily. To keep LLMs updated with current world knowledge, model editing (Mitchell et al., 2022a; Chen et al., 2024; Hartvigsen et al., 2024; Yu et al., 2024) has emerged as a crucial research area in the LLM era. Although model editing methods show potential in updating knowledge, partially modifying the parameters of language models via model editing is akin to performing surgery on the human brain, potentially leading to side effects (Hoelscher-Obermaier et al., 2023; Gu et al., 2024; Yang et al., 2024). While there are some intuitive discussions on the side effects of model editing, identifying the factors and causes of these side effects is scarcely addressed. We noticed that the question-answering setting is the most common when editing knowledge. For example, when we want to update the information about the U.S. president, we typically design a question for models such as "Who is the president of the U.S.?" Following this line of thought, we are curious whether different question types will lead to different side effects after editing.

A common finding regarding the side effects of model editing is that the model's performance across different aspects tends to deteriorate after a few edits (Gu et al., 2024; Yang et al., 2024). Given that the severity of surgical side effects varies with the type of surgery, we are curious whether editing the knowledge for different question types will result in varying degrees of performance degradation. To this end, we categorize the questions used for model editing into eight types: who, what, when, where, which, why, how, and others. Our results indicate that the extent of performance degradation significantly differs after editing knowledge for different types of questions. It suggests future directions for experimental design in knowledge editing.

Moreover, if the illness issues are related or addressing them together can reduce the overall surgical risk, doctors might choose a single surgery to solve multiple problems. Based on this concept, we discuss the side effects under different batch size settings. Our results suggest that enlarging the batch size, i.e., editing several pieces of knowledge at the same time, can mitigate the side effects of the performance drop.

Finally, performing the same surgery on adults and children may result in different side effects, and the underlying causes may vary. Following this line of thought, we experiment with GPT-2 (1.5 billion parameters) (Radford et al., 2019a) and LLaMA-7B (7 billion parameters) (Touvron et al., 2023a) to explore whether findings from smaller

models, which is cheaper and more efficient, can be extrapolated to larger models. Unfortunately, our results indicate that the findings differ between models of different sizes, suggesting that insights from smaller models may not necessarily apply to larger models.

In sum, this paper makes the following contributions:

1. We provide an in-depth analysis of how different question types affect the performance of LLMs after model editing.

2. We investigate the impact of batch size on the side effects of model editing and reveal that larger batch sizes can mitigate performance degradation.

3. We explore the applicability of findings from smaller models to larger models and highlight the limitations of such applications.

2 Related Work

Model editing is a rapidly evolving field with several key approaches aimed at modifying model behavior without extensive retraining (Yao et al., 2023). Fine-tuning with constraints (Zhu et al., 2021) is a method developed to mitigate the issue of catastrophic forgetting, where new knowledge overwrites previously learned information. This approach involves updating as few parameters as possible or only modifying specific parts of the model's structure. Memory-augmented techniques (Mitchell et al., 2022b) involve storing new or corrected knowledge separately from the original model, effectively creating a patch model. These patches can be implemented in various ways, such as through pretrained models or datastores, and are combined with the original model using simple methods like classifiers. However, this approach requires retraining both the classifier and the patch model, which is not ideal for continuous updates. Hyper networks (Cao et al., 2021) represent a dynamic method where the model continuously updates its parameters based on incoming knowledge without needing retraining or fine-tuning. This is achieved by training a network to predict the weights of another network, effectively learning the process of fine-tuning through gradient descent. Despite its promise, the efficacy of hyper networks may diminish as the volume of updates increases, posing challenges for long-term usability. Additionally, current implementations can handle only up to 75 knowledge edits at a time.

The locate-and-edit approach (Meng et al., 2022a,b) leverages interpretability insights, treating the MLP layers in transformers as key-value memories (Geva et al., 2021). By identifying the specific neurons responsible for storing factual associations (keys and values), this method modifies the values corresponding to the desired knowledge. The process involves evaluating the influence of individual neurons on the output and adjusting the most impactful ones. It offers enhanced interpretability and allows for precise targeting of specific pieces of knowledge within the model. It is favored for scenarios where understanding and precisely controlling model behavior is crucial. Therefore, in this paper, we focus on the iconic method of locate-and-edit, MEMIT (Meng et al., 2022b), for in-depth analysis and discussions.

3 Experimental Setup

3.1 Knowledge Editing Dataset

We use RealTimeQA (Kasai et al., 2022) as the base dataset for knowledge editing. RealTimeQA is a collection derived from popular news sources, containing articles from various news websites. Weekly, RealTimeQA gathers news articles along with approximately 30 multiple-choice questions authored by humans from platforms such as CNN, THE WEEK, and USA Today, covering diverse topics including politics, business, sports, and entertainment. Unlike other datasets such as ZsRE (Levy et al., 2017) or CounterFact (Meng et al., 2022a), which draw from known Wiki knowledge or focus on false facts respectively, we opt for RealTimeQA due to its alignment with real-world scenarios, offering a more fitting context for our knowledge updating needs. In our experiment, we randomly selected 80 questions of each question type from a total of 1,781 instances.

3.2 General Ability Evaluation

To assess the model's general ability, including knowledge acquisition, comprehension, and reasoning abilities, we utilize ARC-easy, ARC-challenge (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018) as our primary evaluation datasets. The ARC Benchmark, featuring over 7,787 science questions spanning from 3rd to 9th-grade standardized test levels, presents formidable challenges for both retrieval-based and

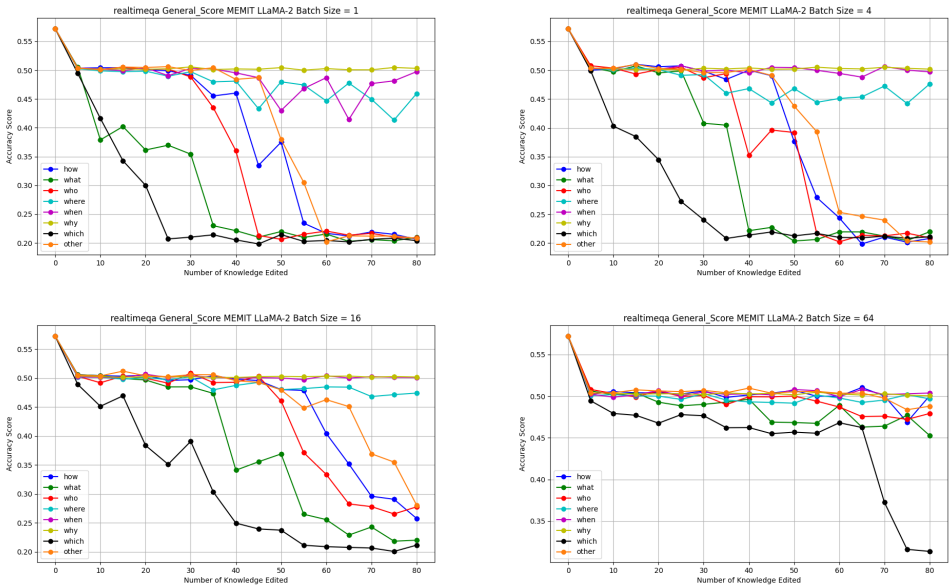

Figure 1: General ability of LLaMA-2 with MEMIT. Please note that the scale of the y-axis in different charts differs for the detailed discussions.

word co-occurrence algorithms, particularly in its Challenge Set. This division into Easy and Challenge Sets allows for a nuanced examination of performance across varying difficulty levels. Additionally, OpenBookQA introduces a novel evaluation paradigm inspired by open-book exams, demanding a profound understanding of elementary-level science facts and their practical application in diverse scenarios. Through these datasets, we aim to comprehensively evaluate our model's capabilities across varying levels of complexity and real-world applicability, from basic knowledge retrieval to sophisticated reasoning tasks.

3.3 Evaluation Paradigm

We chose to experiment with GPT-2-XL (1.5B) (Radford et al., 2019b) and LLaMA-2 (7B) (Touvron et al., 2023b) as our testing models to explore the impact of model size on performance outcomes. GPT-2-XL represents a mid-sized model, while LLaMA-2 is substantially larger, allowing us to observe potential trade-offs between computational efficiency and performance gains. To discuss the side effects of model editing, we use MEMIT (Meng et al., 2022b) to edit models based on the knowledge changes in RealTimeQA with different types of questions and different settings on the batch size. Then, we test the models' general ability with ARC-easy, ARC-challenge, and OpenBookQA, and report the average accuracy as the evaluation for general ability.

4 Results and Analysis

4.1 Impact of Question Type

Figure 1 illustrates the general ability of LLaMA-2 7B as the number of knowledge edits increases under different batch size settings. We first examine the results for a batch size equal to 1 (upper left subfigure in Figure 1). The results reveal a significantly different trend in the model's performance after editing knowledge based on different question types. For all question types, the general ability drops to around 50% after five knowledge edits. This finding is consistent with previous studies (Gu et al., 2024; Yang et al., 2024), indicating that a few edits can lead to model collapse. However, a deeper analysis of this side effect shows that after editing 10 knowledge items, the general ability drops significantly more for "which" or "what" questions, while the general ability for other question types remains stable.

Furthermore, as the number of knowledge edits increases, the general ability of the model edited under different question types drops sequentially rather than simultaneously. These results suggest that different question types have varying impacts on the model's general ability. Notably, "Why" questions have the least adverse effect on model editing. The general ability of the model edited with "Why" questions does not drop a second time, unlike other question types. We hypothesize that this is because LLMs are trained for con-

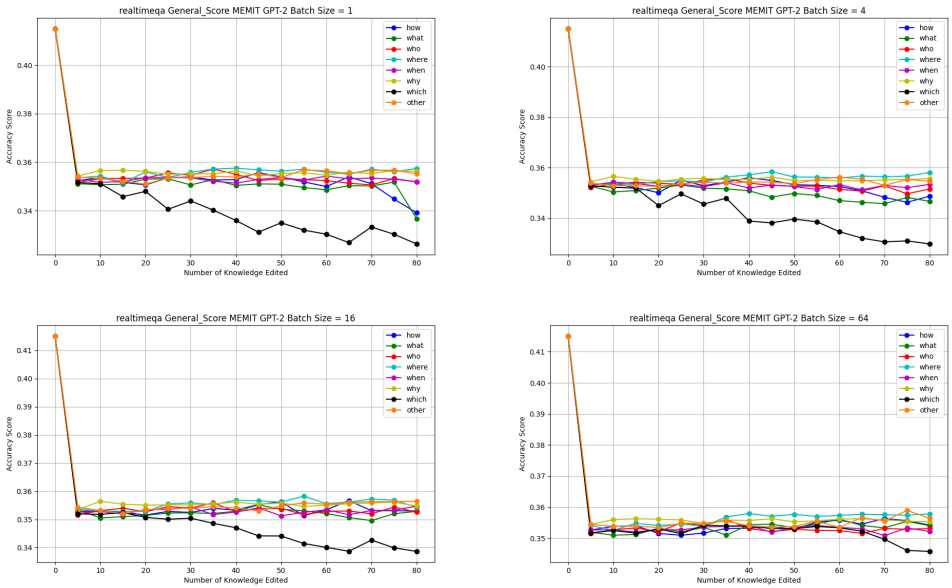

Figure 2: General ability of GPT-2 with MEMIT.

tinuous writing, and answers to "Why" questions are full sentences, whereas answers to other questions mainly involve editing named entities. For instance, "where" questions edit knowledge related to locations, and many "how" questions are about quantities, such as "how many" and "how much".

4.2 Mitigating Side Effects

We compare the effects under different batch size settings in Figure 1. Under varying batch sizes, the observations of the performance drop across different question types are similar, including the second drop and the order of dropping among different question types. However, we observed that the timing of the second drop is delayed as the batch size increases. These results suggest that editing the same type of questions simultaneously may help mitigate side effects.

4.3 Observations on Model Size

As mentioned in Section 1, experimenting with LLMs is more expensive and time-consuming than with smaller language models. We conducted the same experiments with GPT-2, and the results are shown in Figure 2. Although there are some minor fluctuations, the general ability drops to the lowest level directly without a second drop, regardless of the question types and batch size. These results indicate that the side effects and observations with smaller language models may differ from those with large language models. It also suggests that the behaviors of these two types of models should be considered and analyzed independently, despite the side effects occurring in both after a few edits.

4.4 Observations on Editing Methods

Based on the results presented in previous sections, we emphasize that model editing with LLMs exhibits various side effects depending on question types, a phenomenon not observable when experimenting with smaller language models. Accordingly, we focus on a more in-depth discussion using LLaMA-2 in this section. The central topic under consideration is whether the side effects observed with MEMIT for knowledge updates in LLMs remain consistent when employing a fine-tuning approach. To address this question, we replace MEMIT, used in Section 4.1, with a fine-tuning scheme and evaluate the resulting side effects. Specifically, we fine-tune the model using the RealTimeQA dataset and then test its performance on other general ability test sets.

The results are illustrated in Figure 3. First, the side effects on "Why" questions remain less pronounced compared to other question types, which is the same when using MEMIT. Second, the performance declines continuously as the number of knowledge edits increases. By comparing these results with those in Figure 1, we observe that the general ability degradation patterns differ between the fine-tuning scheme and MEMIT. Notably, MEMIT exhibits a second performance drop, the timing of which varies based on the question type. Third, when the number of knowledge edits

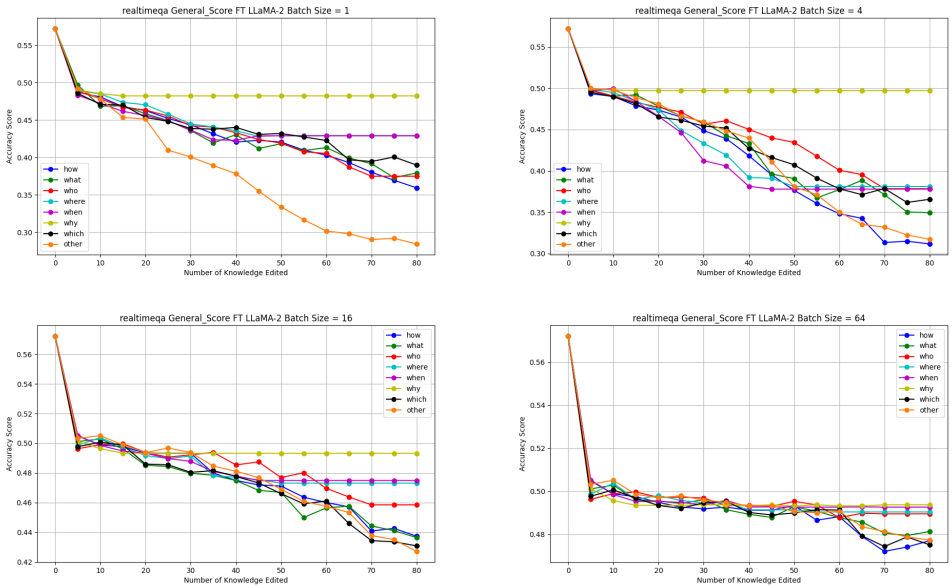

Figure 3: General ability of LLaMA-2 with fine-tuning approach.

is small, MEMIT demonstrates fewer side effects for certain question types, such as "where," "when," "who," and "how." However, as batch size and the number of knowledge edits increase, the fine-tuning scheme becomes a more favorable choice.

4.5 Editing Performance

In the previous sections, we discussed the side effects and examined the differences in side effects across various editing methods. In this section, we delve deeper into one potential cause: the effectiveness of different editing methods in knowledge editing tasks. Specifically, we evaluate whether the model successfully edits the target knowledge. After editing, we verify whether the language model can answer questions with the updated knowledge.

Figures 4 and 5 present the performances of MEMIT and fine-tuning as model editing approaches, respectively.

First, consider the results when the batch size equals 1. The performance differences between MEMIT and fine-tuning are notable. MEMIT excels when the number of knowledge edits is small, consistently outperforming fine-tuning across all question types in this scenario. Regardless of question type, MEMIT demonstrates superior effectiveness for limited edits. The performance trends of the two approaches also diverge significantly. MEMIT's performance declines steadily as the number of knowledge edits increases, whereas fine-tuning maintains relatively stable performance, albeit starting from a lower baseline compared to MEMIT.

Performance across different question types also varies significantly, irrespective of the editing method. For instance, "which" questions experience a severe decline in accuracy, falling below 20%, after several rounds of editing with MEMIT. In contrast, fine-tuning achieves approximately 50% accuracy for these questions under similar conditions. Conversely, for "where" questions, MEMIT sustains high performance, exceeding 70%, while fine-tuning remains limited to around 50% accuracy. This contrast underscores the importance of question type in evaluating and understanding model editing methods.

Next, consider how performance changes with varying batch sizes. The effect of batch size differs significantly between the two methods. For MEMIT, increasing the batch size mitigates the rapid performance decline, effectively delaying the onset of significant performance drops. In contrast, for fine-tuning, increasing the batch size degrades performance, suggesting a potential sensitivity to this parameter.

In summary, the results highlight nuanced trade-offs between MEMIT and fine-tuning in knowledge editing tasks. MEMIT excels when the number of edits is small and remains robust for certain question types, such as "where," though it struggles with others, such as "which," after multiple edits. Fine-tuning, while stable, is less effective overall but exhibits advantages for specific question types

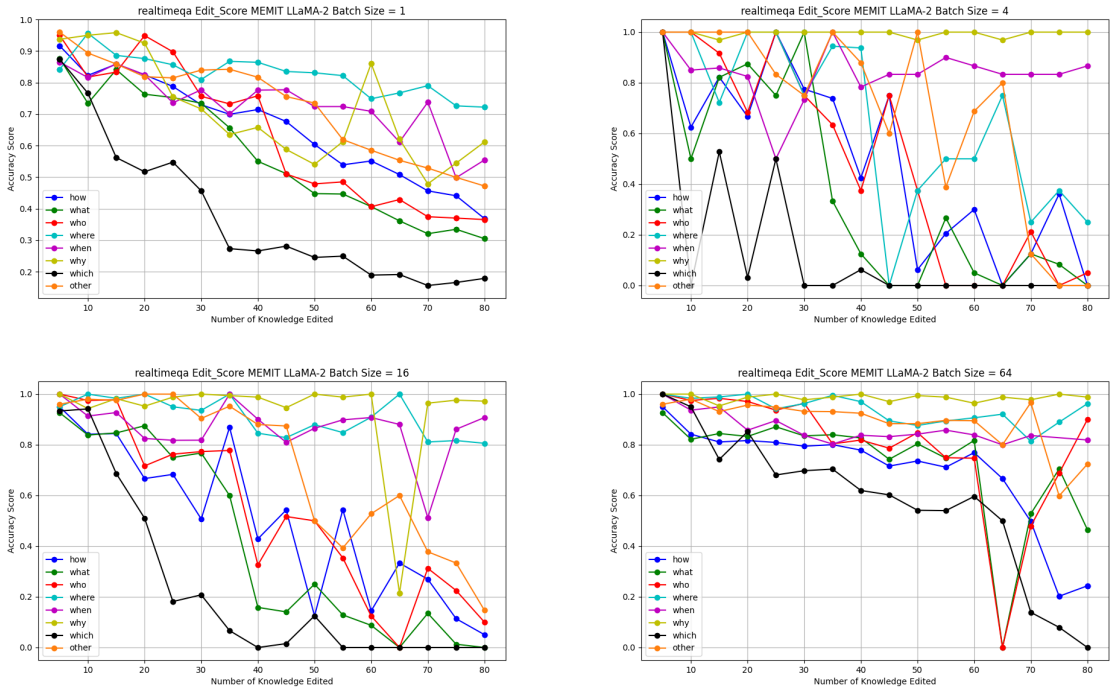

Figure 4: Editing performance of LLaMA-2 with MEMIT.

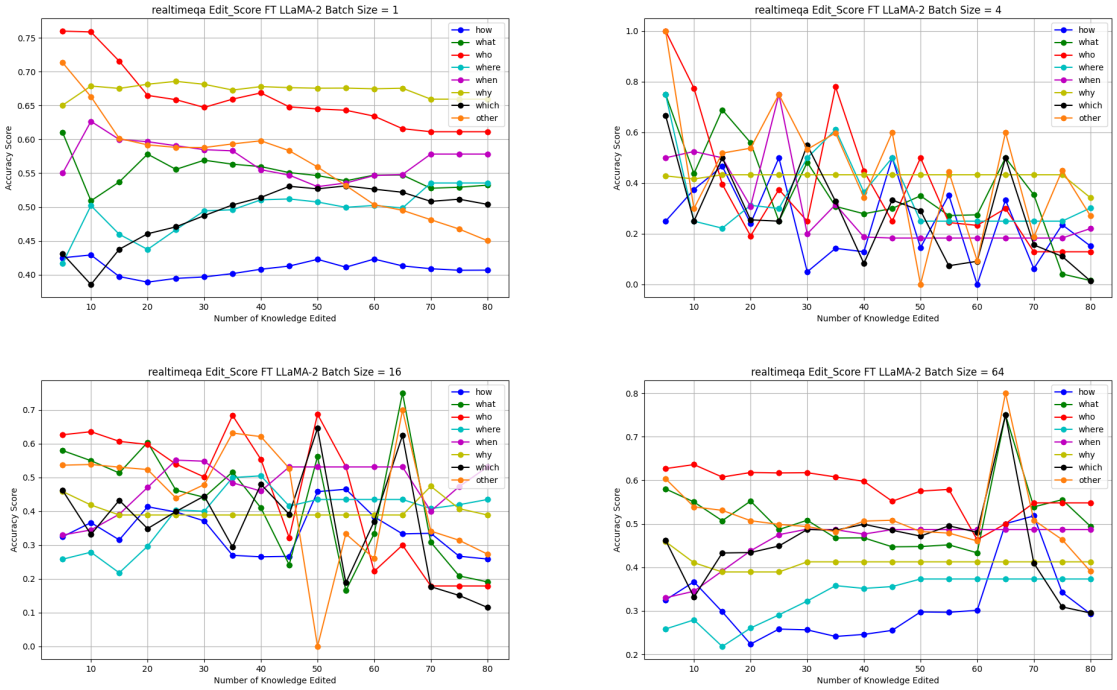

Figure 5: Editing performance of LLaMA-2 with fine-tuning approach.

in extended editing scenarios. Batch size further introduces variability, favoring MEMIT with larger batches but adversely affecting fine-tuning. These findings emphasize the importance of tailoring editing strategies to the specific task requirements, including the expected number of edits, question type, and batch size configuration.

5 Mixture of Editing Approaches

Based on the discussion in the previous sections, we understand that different model editing approaches come with distinct advantages and disadvantages. Furthermore, the effectiveness of these approaches varies depending on the type of question being addressed. In this section, we aim to explore the impact of assigning question types that are relatively more suited to specific model editing approaches, focusing on differences in side effects and editing performance. Additionally, we will analyze how the sequence in which model editing approaches are applied affects overall performance.

Specifically, our experiment builds upon the findings of previous sections. For instance, as shown in Figure 1 and Figure 3 (batch size $= 1$), MEMIT demonstrates better performance for "why" questions (above 50%) compared to the fine-tuning approach (below 50%). Based on these results, we chose MEMIT for editing "why" questions. To analyze performance differences, we divided the questions into two groups: those where MEMIT performs better and those where fine-tuning is more effective. We then evaluated two experimental setups: editing the MEMIT group first versus editing the fine-tuning group first. Note that the testing data remains consistent with prior experiments, with only the order and methods adjusted.

The results of general ability, i.e., side effects, are shown in Figure 6. The figure indicates that using only the fine-tuning approach (FT) results in fewer side effects when the batch size is set to 1. However, as the batch size increases, the side effects of the "MEMIT then FT" approach become comparable to those of the fine-tuning approach. While MEMIT performs well when tested on data where it is advantageous, significant side effects arise when the question type shifts to those favoring the fine-tuning approach. The results of "MEMIT then FT" suggest that switching from MEMIT to fine-tuning based on question type could effectively mitigate side effects. Conversely, this is not true for the "FT then MEMIT" approach, as it fails to capitalize on findings from prior experiments. The side effects of "FT then MEMIT" remain comparable to those of using MEMIT alone, regardless of batch size.

The knowledge editing performance is presented in Figure 7. These results support the conclusion that "MEMIT then FT" outperforms "FT then MEMIT," emphasizing the importance of the order in which editing methods are applied. Furthermore, although the performance of FT and "MEMIT then FT" appears similar when the batch size is 1 or 4, the difference becomes substantial as batch size increases. As observed in Figure 5, fine-tuning alone causes a dramatic decline in editing performance for certain question types. Applying MEMIT first for these cases, followed by fine-tuning for other question types, proves effective when the batch size is large. These findings underscore the potential importance of considering question type in model editing and open new avenues for exploring mixed editing approaches.

6 Conclusion

This paper investigates the factors shaping the side effects of model editing in LLMs, emphasizing the critical influence of question type, batch size, model scale, and editing strategy. Our analysis reveals that "Why" questions consistently produce the least performance degradation, likely due to their alignment with the sentence-level reasoning capabilities of LLMs. We further highlight the differences between smaller models like GPT-2 and larger models like LLaMA-2, demonstrating that observations from smaller models do not always generalize to larger ones. When comparing editing approaches, MEMIT performs better for limited edits or specific question types, such as "where," while fine-tuning offers stability over more extensive editing scenarios. A mixed approach that applies MEMIT for its strengths and transitions to fine-tuning for broader edits balances side effects and accuracy effectively, especially with larger batch sizes.

These findings provide a foundation for designing adaptive, context-aware editing frameworks that optimize the trade-offs between minimizing side effects and achieving high editing accuracy. Future work should expand on these insights to explore their applicability across different LLM architectures.

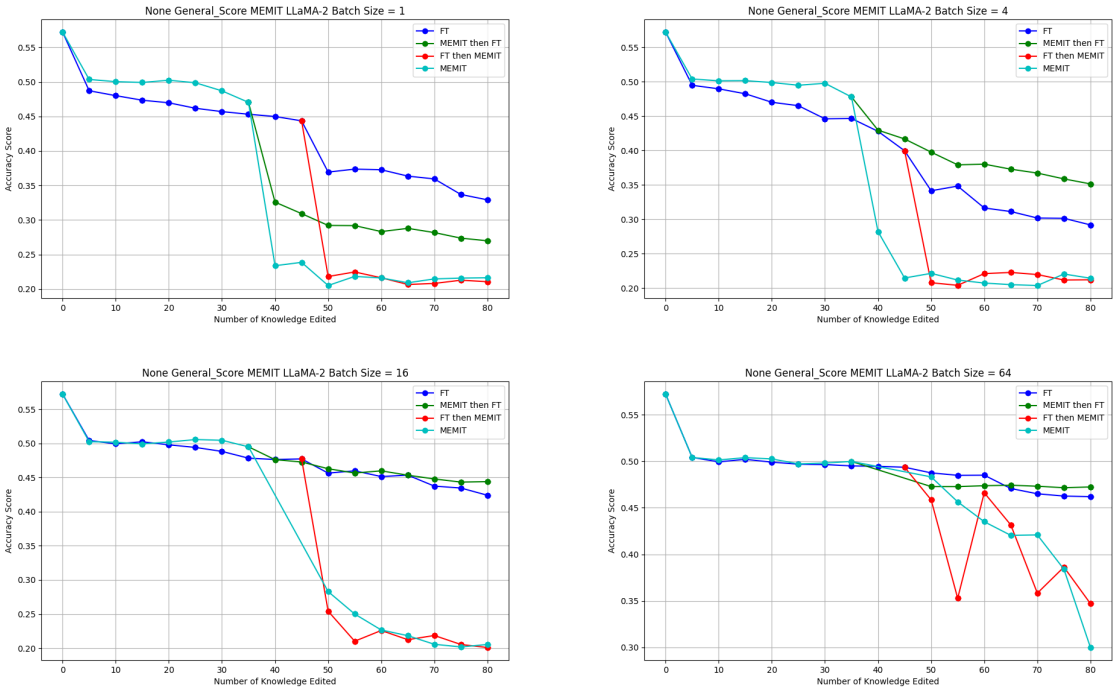

Figure 6: General ability of LLaMA-2 with mixture approach.

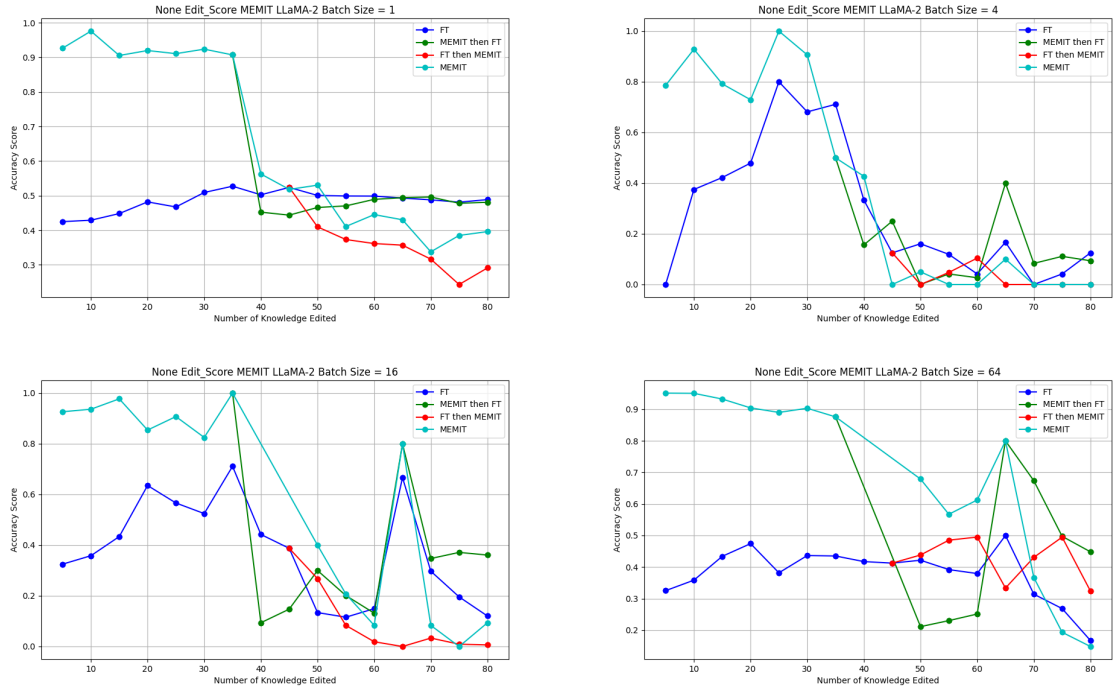

Figure 7: Editing performance of LLaMA-2 with mixture approach.

Limitation

First, our study focuses on eight specific question types. This categorization, while comprehensive, may not cover all possible variations of model queries encountered in real-world applications. Future work could explore additional question types or more nuanced classifications to provide a broader understanding of the impact of question types on model editing. Second, we conducted our experiments on two specific models: GPT-2 and LLaMA-7B. The discrepancies observed between these models highlight the need for caution when generalizing findings to other models. Third, our assessment focused on the general ability of models post-editing. However, other important metrics, such as interpretability and robustness, were not considered. Including these metrics in future studies could offer a more holistic view of the consequences of model editing. Finally, while we identified different impacts of question types and batch sizes on model performance, the underlying mechanisms driving these side effects remain unclear. Further research is needed to understand the causal relationships and develop methods to predict and mitigate unintended consequences effectively.

References

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Conference on Empirical Methods in Natural Language Processing*.

Yingfa Chen, Zhengyan Zhang, Xu Han, Chaojun Xiao, Zhiyuan Liu, Chen Chen, Kuai Li, Tao Yang, and Maosong Sun. 2024. Robust and scalable model editing for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14157–14172.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing can hurt general abilities of large language models. *arXiv preprint arXiv:2401.04700*.

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

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. 2023. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11548–11559, Toronto, Canada. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir R. Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *ArXiv*, abs/2207.13332.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *ArXiv*, abs/1706.04115.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. In *Neural Information Processing Systems*.

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

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. *ArXiv*, abs/2206.06520.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019b. Language models are unsupervised multitask learners.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

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

Yunzhi Yao, Peng Wang, Bo Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Conference on Empirical Methods in Natural Language Processing*.

Lang Yu, Qin Chen, Jie Zhou, and Liang He. 2024. Melo: Enhancing model editing with neuron-indexed dynamic lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19449–19457.

Chen Zhu, Daliang Li, Felix Yu, Manzil Zaheer, Sanjiv Kumar, Srinadh Bhojanapalli, and Ankit Singh Rawat. 2021. Modifying memories in transformer models. (2020).