Should We Really Edit Language Models?

On the Evaluation of Edited Language Models

Qi Li[1], **Xiang Liu** [1], Zhenheng Tang[2], Peijie Dong [1], Zeyu Li [1], Xinglin Pan[1], and Xiaowen Chu[1]

[1] The Hong Kong University of Science and Technology (Guangzhou)

[2] Hong Kong Baptist University

NeurIPS 24 long paper video presentation

Why Knowledge Editing ?

- Large Language Model (LLM) demonstrate remarkable performance across knowledge-intensive tasks.
- However, the learned vast amount of knowledge in LLMs may be erroneous, harmful, or outdated.
- Directly fine-tuning an LLM is prohibitive due to hardware constraints and resource budget.
- Knowledge editing has been proposed to efficiently update knowledge within LLM.
- Current evaluation criteria of editing methods are three-folds: reliability, generalization, locality

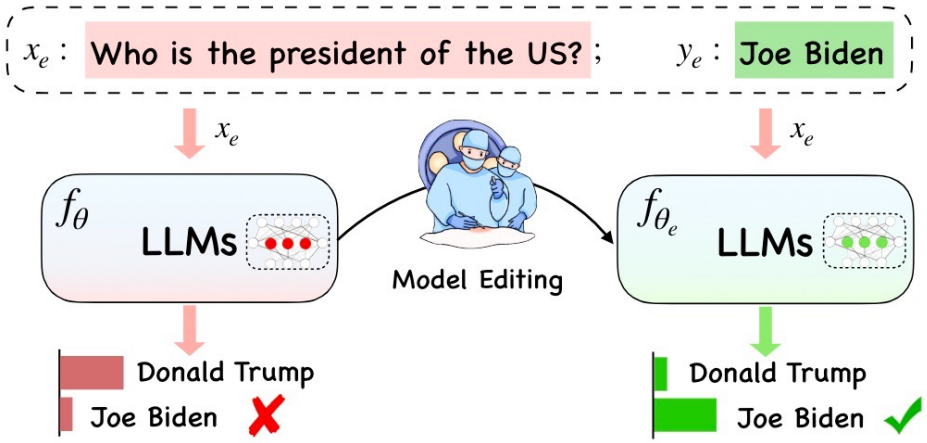
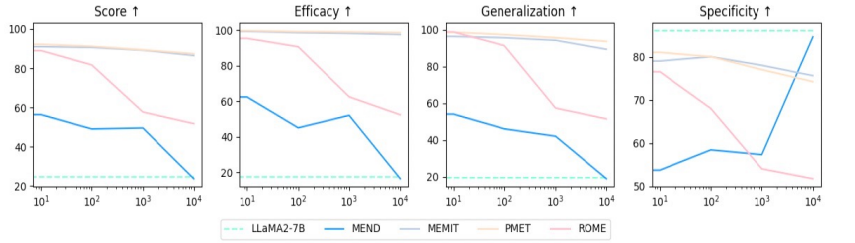


Figure 1: Current knowledge editing methods target at enabling efficient yet precise model behavior alterations on specific knowledge samples

Research Motivation

- Existing knowledge editing methods like SERAC, ROME, MEMIT, and IKE work well on these evaluation criteria across various datasets on different LLMs.
- Recent works have disclosed the inevitable pitfalls of existing editing methods from different perspectives.
- In sequential editing setting, as the number of edits increases, it is necessary to balance two aspects:
  - the retention of the model's original knowledge
  - the preservation of newly acquired knowledge through updates
- These two objectives are to some extent conflicting.



- The general capabilities of LLMs is the foundation to solve the wide range of complex tasks.
- Changes in the model's general capabilities reflect the retention of its original knowledge.

How do sequential model editing affect the general abilities of language models?

Research Questions

- We aims to explore the impacts of editing methods on various general abilities of edited models.
- It naturally motivates the following critical research questions (RQs) to be explored in this work based on the primary aim.

    - RQ1: How does the number of undergone edits affect the abilities of models? (In Section 4.1)
    - RQ2: Do instruction-tuned models exhibit differently than base counterparts? (In Section 4.2)
    - RQ3: Does the general abilities of the edited model differ on model scales? (In Section 4.3)
    - RQ4: How does editing affects different aspects of a model's capabilities? (In Section 4.4)
    - RQ5: Does performing editing on language models compromise their safety? (In Section 4.5)

Experimental Settings

- **Language Models**
  - Llama2-7B, Mistral-7B, GPT2-XL
- **Editing Methods**
  - ROME, MEMIT, MEND, PMET, KN, SERAC, GRACE
- **Editing Datasets**
  - ZsRE, COUNTERFACT
- **Evaluation Benchmark**
  - MMLU, BBH, GSM8K, CommonsenseQA, TrivialQA, TruthfulQA, ToxiGen
- **Editing Settings**
  - Sequential single editing

Findings

RQ1: Impact of the Number of Edits

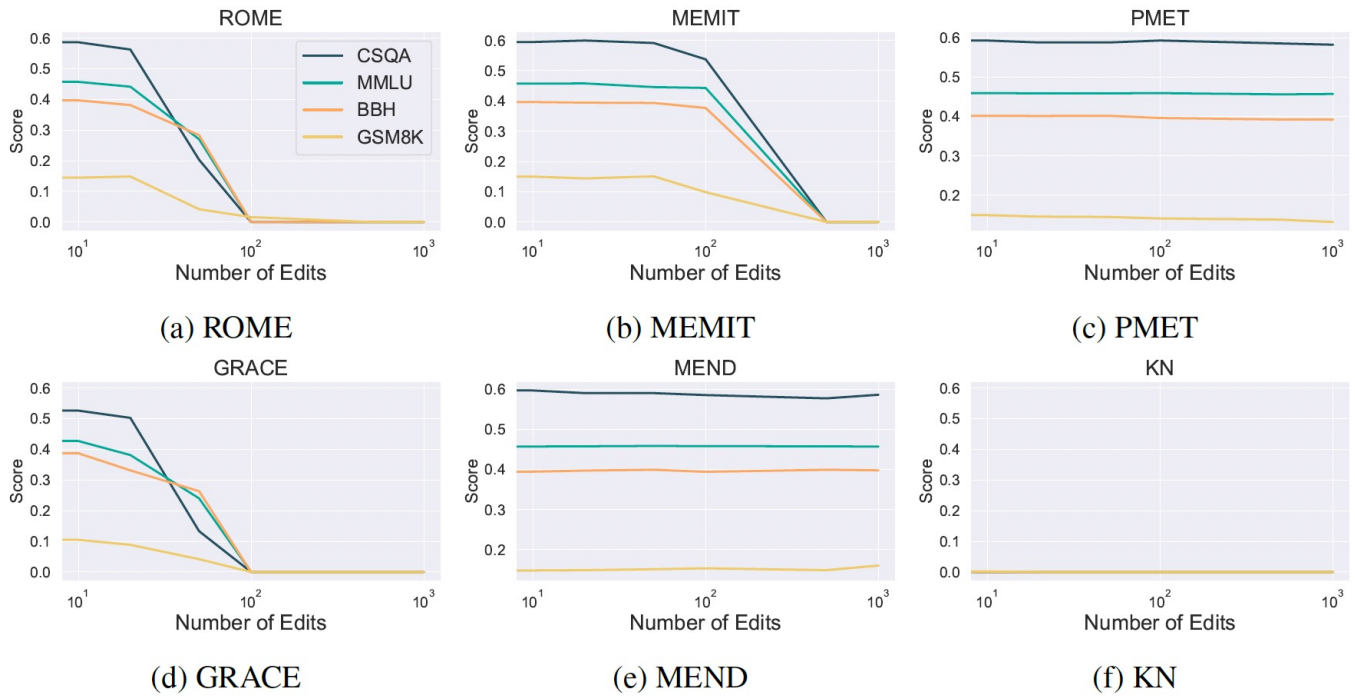


Figure 2: Performance trends of evaluating edited Llama2-7B base model across different benchmarks using six editing methods. Results reveal that PMET and MEND can effectively preserve the model's abilities across all tasks. While KN drastically drops even less than ten edits.

Finding 4.1. The majority of existing methods can only undergo dozens of edits without compromising performance, while only a few methods can scale to thousands of edits.

Findings

RQ2: Does Instruction Tuned LLM Show Better Performance after Editing?

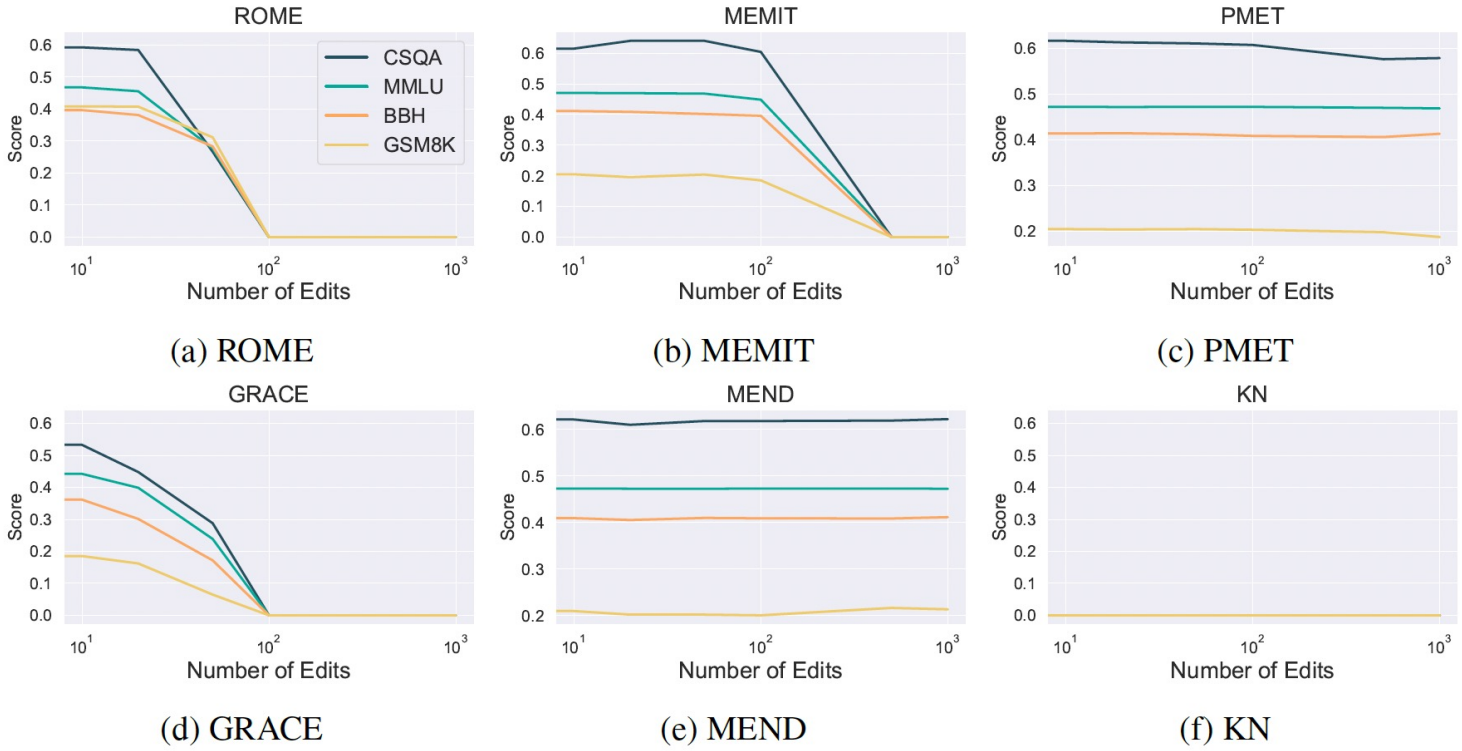


Figure 3: Performance trends of assessing edited Llama2-chat-7B across different benchmarks using 6 editing methods. Results reveal that PMET and MEND can effectively preserve the model's abilities across all tasks. While KN drastically drops even less than ten edits.

Finding 4.2. Instruction-tuned model exhibits a slower rate of performance decline after editing.

Findings

RQ3: Do the General Abilities of the Edited Model Differ on Model Scales?

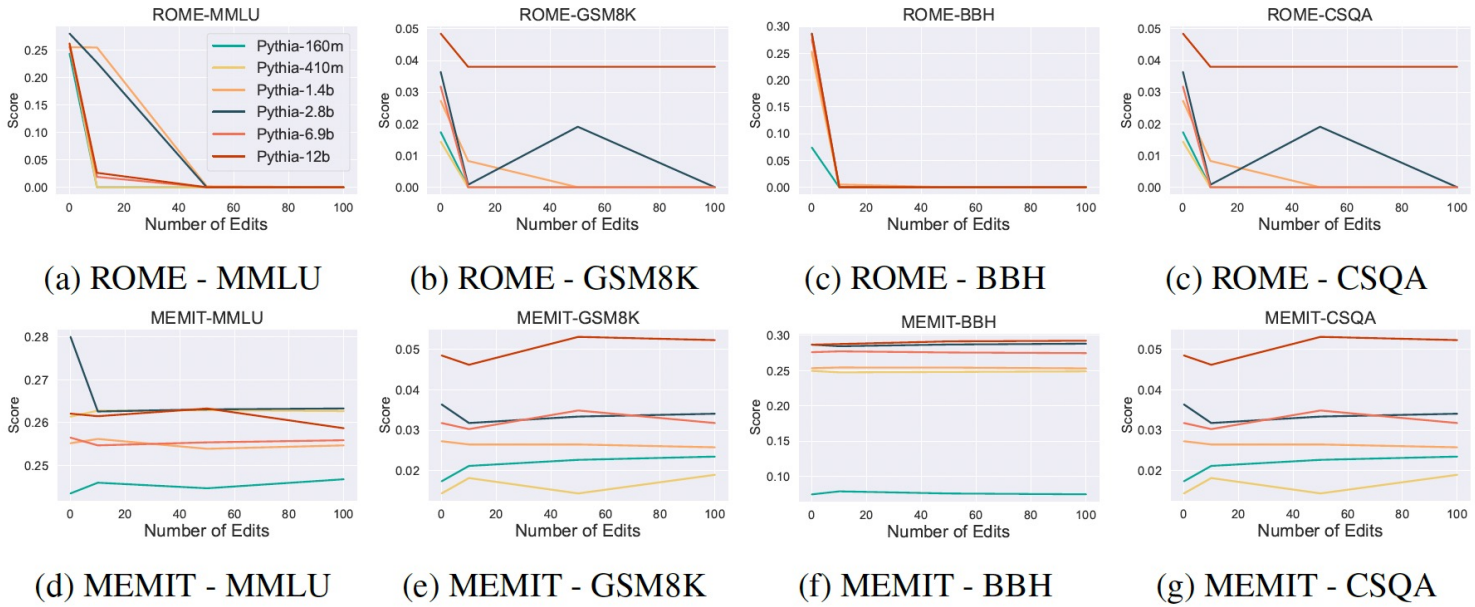
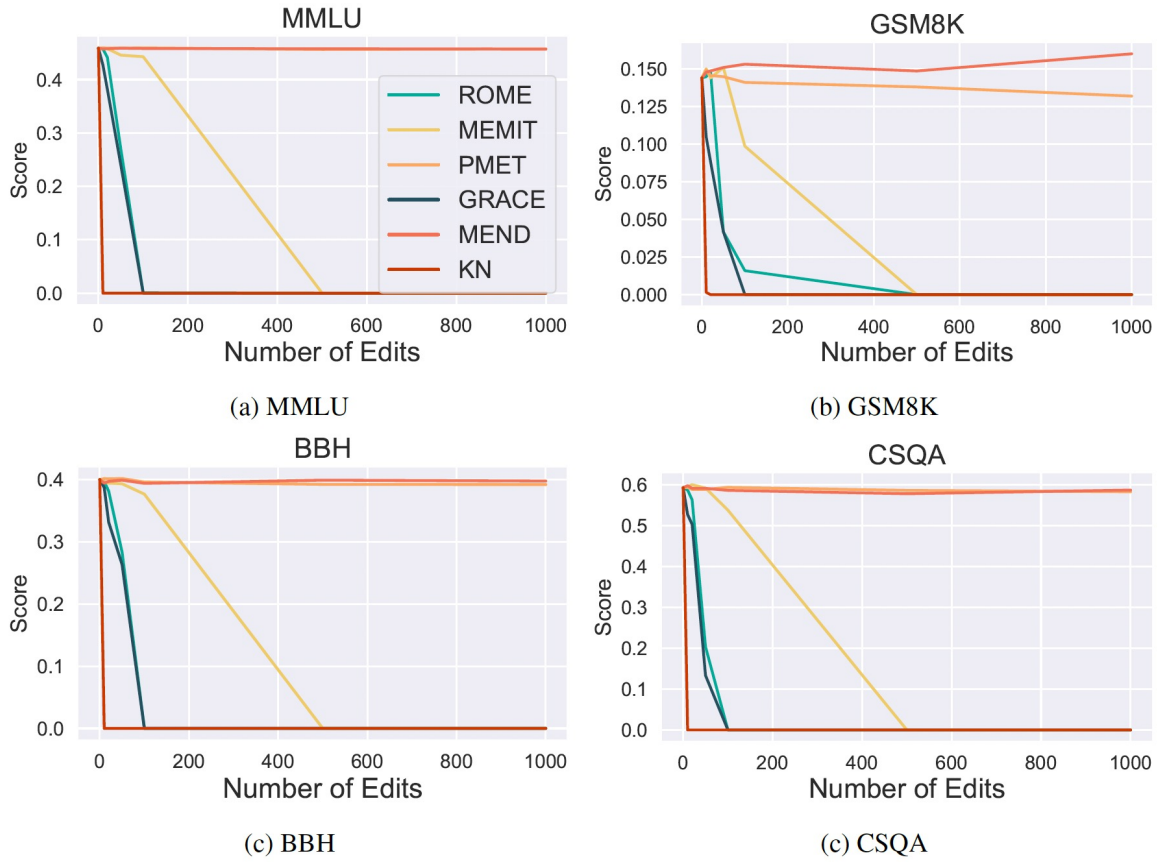


Figure 4: Quantitative results of exploring the impact of model scale on edited language models. We perform editing with different-size models in Pythia model families with ROME (the first row) and MEMIT (the second row), and then these models are evaluated across diverse benchmarks.

Finding 4.3. Larger models exhibit less side effect on benchmarks after editing.

Findings

RQ4: How Does Editing Affect Different Aspects of a Model's Capabilities?



(a) MMLU (b) GSM8K

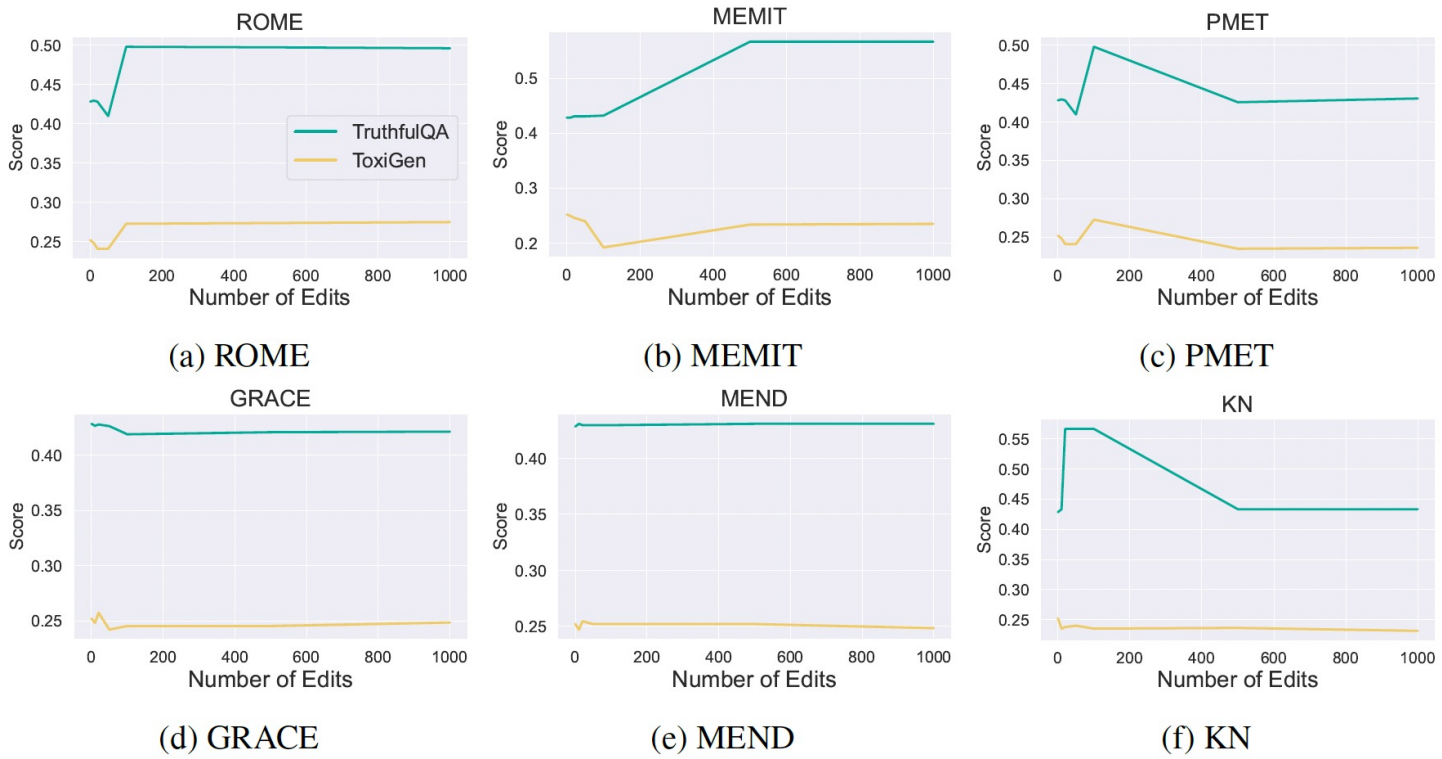
(c) BBH (c) CSQA

Figure 5: Evaluation of different kinds of general capabilities of edited language models. Results reveal that PMET and MEND can effectively preserve the model's abilities across all tasks.

Finding 4.4. Editing affects the different capabilities of LLM to a roughly equivalent extent.

Findings

RQ5: The Safety Cost of Editing Language Models



(a) ROME (b) MEMIT (c) PMET

(d) GRACE (e) MEND (f) KN

Figure 6: Safety evaluation of edited language models. We perform evaluation on TruthfulQA and Toxigen datasets with the Llama2-7B model. Results demonstrate that for most editing methods, even dozens of edits can compromise the safety of language models with they are aligned.

Finding 4.5. Even dozens of edits can compromise the safety of edited language models.

From the above research questions,
we conclude that existing editing methods have inevitable pitfalls in editing LLMs, making them impractical in the production environment.

Takeaways:

- Existing editing methods inevitably lead to performance deterioration on general benchmarks.
- Instruction-tuned models exhibit greater robustness to editing, showing less performance drop on benchmark.
- Language models with larger scale are more resistant to editing compared to smaller models.
- The safety of the edited models is significantly compromised, even for models that were originally safety-aligned.

Potential Impact on Inherent Knowledge within LLM.

- Current editing methods claim that they can update specific knowledge within LLM without affecting other unrelated knowledge.
- Recent work reveal that editing can have unintended and potentially harmful impacts on the intrinsic knowledge of the model.
- Our series of experiments demonstrates that even with only hundreds of edits, the general capabilities of the model are severely compromised.
- When the number of edits reaches the thousands, the model’s internal structure is thoroughly damaged.

Thanks & QA