Model Editing Harms General Abilities of Large Language Models: Regularization to the Rescue

Jia-Chen Gu[1], Hao-Xiang Xu[2], Jun-Yu Ma[2], Pan Lu[1], Zhen-Hua Ling[2], Kai-Wei Chang[1], Nanyun Peng[1]
[1]University of California, Los Angeles [2]University of Science and Technology of China

Model Editing: Definition and Evaluation

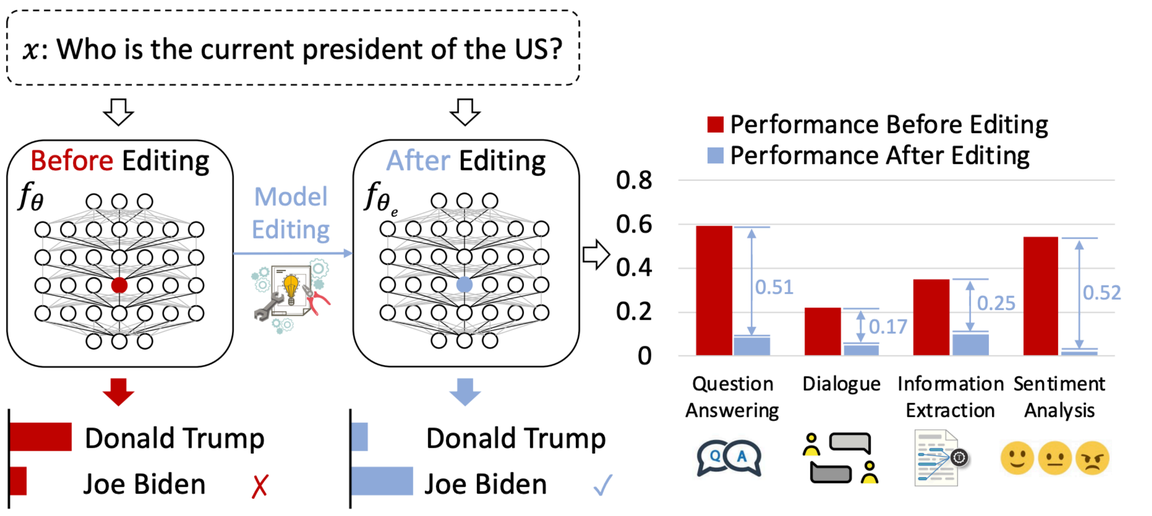
Given an edit sample (x_e, y_e)

$$f_{\theta_e}(x) = \begin{cases} y_e & \text{if } x \in I(x_e, y_e) \\ f_\theta(x) & \text{if } x \in O(x_e, y_e) \end{cases}$$

- ✓ **Generalization**: recall the fact under the in-scope paraphrase prompts $I(x_e)$.
- ✓ **Locality**: remain unchanged for the prompts out of the editing scope $O(x_e)$.

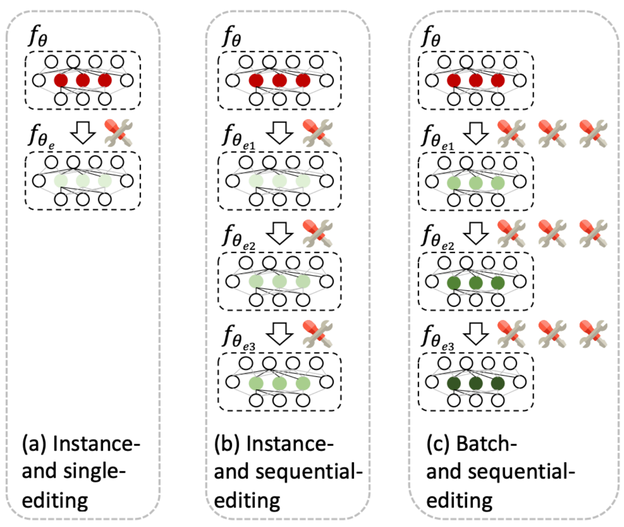
Side Effects of Model Editing

Does model editing's improvements on factuality come at the cost of **a significant degradation of model's general abilities**?



Evaluation Paradigm

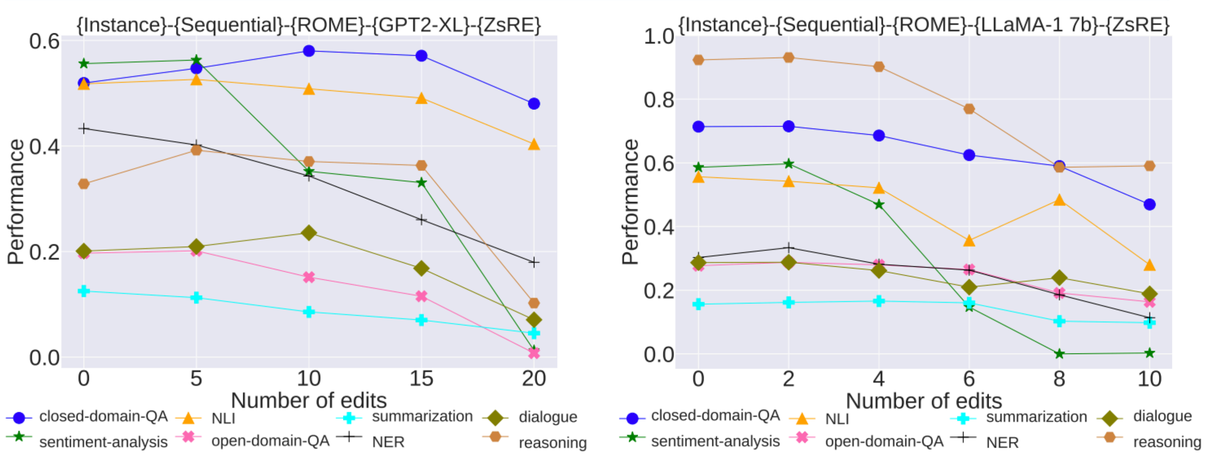
- ✓ **Single- vs. Sequential-editing**
- ✓ **Instance- vs. Batch-editing**
- ✓ **Zero-shot Evaluation**



Evaluation Setup

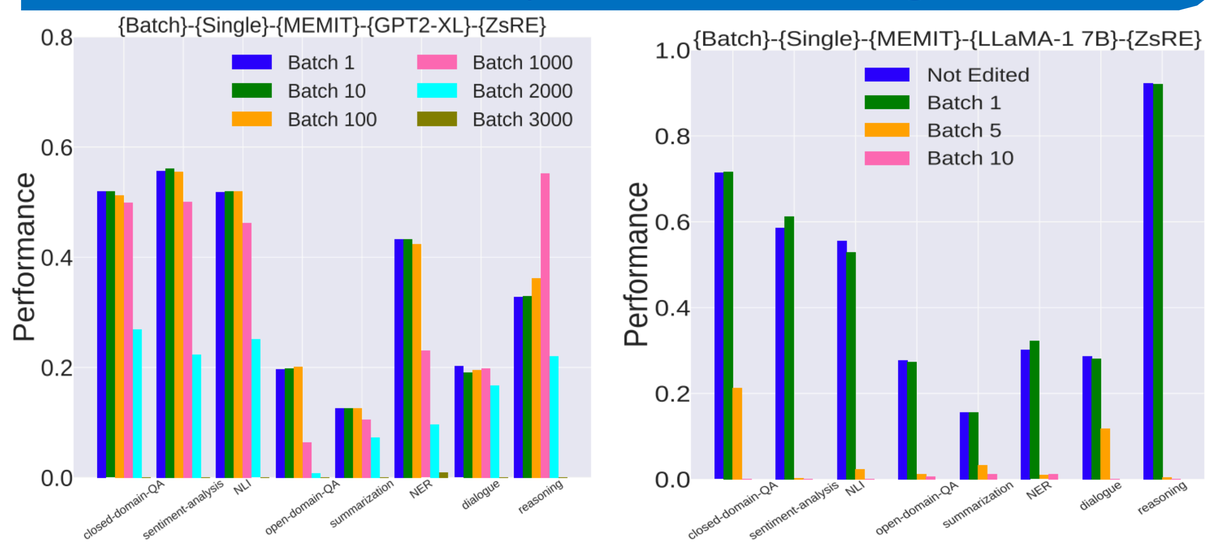
- ✓ **Editing Methods**: KN, ROME, ***MEND*** and ***MEMIT***
- ✓ **Editing Dataset**: ZSRE dataset
- ✓ **Selected LLMs:** GPT-2 XL (1.5B), LLaMA-1 (7B), LLaMA-2 (7B)
- ✓ **8 Downstream Tasks:** Reasoning, NLI, Open- and Closed-domain QA, Dialogue, Summarization, NER, Sentiment

Results: Impact of Sequential-editing



- ✓ LLMs are not robust to weight perturbations even if less than 1% of parameters are edited
- ✓ The difficulty lies in the dual objective of improving model factuality while maintaining their general abilities

Results: Impact of Batch-editing



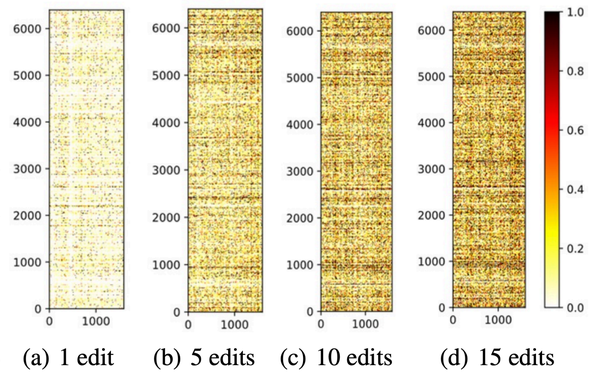
- ✓ Edited models degrade as the batch size increases
- ✓ LLMs are sensitive to increases in batch size, calling for more research work on scalable editing for efficient editing

Analysis of Causes of Side Effects

- ✓ The side effects come from changing the original weights too much, resulting in overfitting to the editing facts
- ✓ Define δ as the relative change in weight to characterize the degree of change of each element in ΔW

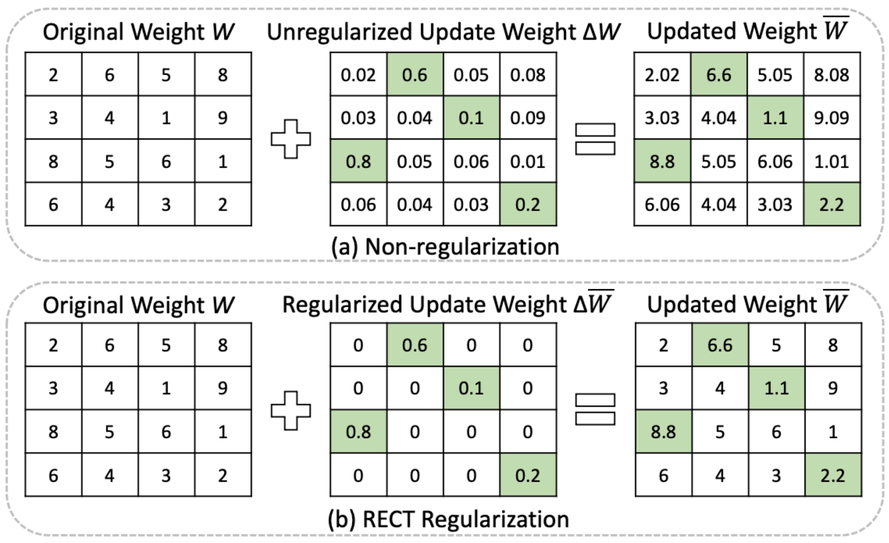
$$\overline{W} = W + \Delta W$$
$$\delta = \left|\frac{\Delta W}{W}\right|$$

# edits	Manhattan dist.	% $\delta > 0.077$	% $\delta > 0.171$
1	9079.2	20.0%	10.0%
5	27072.2	49.2%	28.9%
10	52245.3	67.4%	46.2%
15	63247.5	72.8%	52.2%

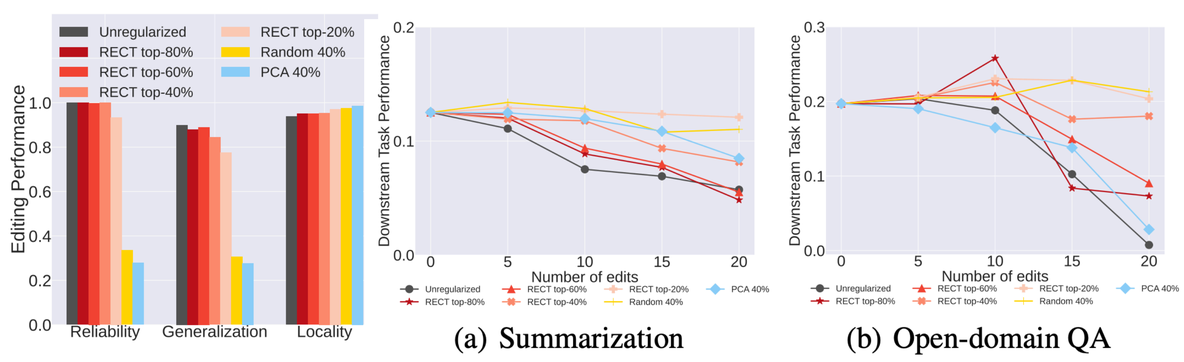


RECT: RElative Change in weighT

- ✓ Regularization to discourage overly complex editing updates that are more likely to overfit using δ
- ✓ **Principal editing**: top-k% elements in ΔW that change the most according to δ, keep their original values
- ✓ **Minor editing**: the remaining elements in ΔW are treated as minor contributions, set to zero for regularization



Regularization Results



- ✓ A trade-off between editing and downstream performance
- ✓ RECT top-40% can help maintain over 94% editing
- ✓ The more editing regularization, the smaller the change to the model, the more general abilities can be preserved

Email: gujc@ucla.edu Code: https://github.com/JasonForJoy/Model-Editing-Hurt Homepage: https://JasonForJoy.github.io/