NEURAL INFORMATION PROCESSING SYSTEMS

Reasons and Solutions for the Decline in Model Performance after Editing

Xiusheng Huang[1,2,3], Jiaxiang Liu[1,2], Yequan Wang[3], Kang Liu[1,2]

[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences

[3]Beijing Academy of Artificial Intelligence, Beijing, China

CASIA

BAAI 智源研究院

Introduction

- **Knowledge editing, while effective, may cause damage to the model:**
  - > This study examines two factors causing this decline:
    - ~ Data perspective and Model perspective

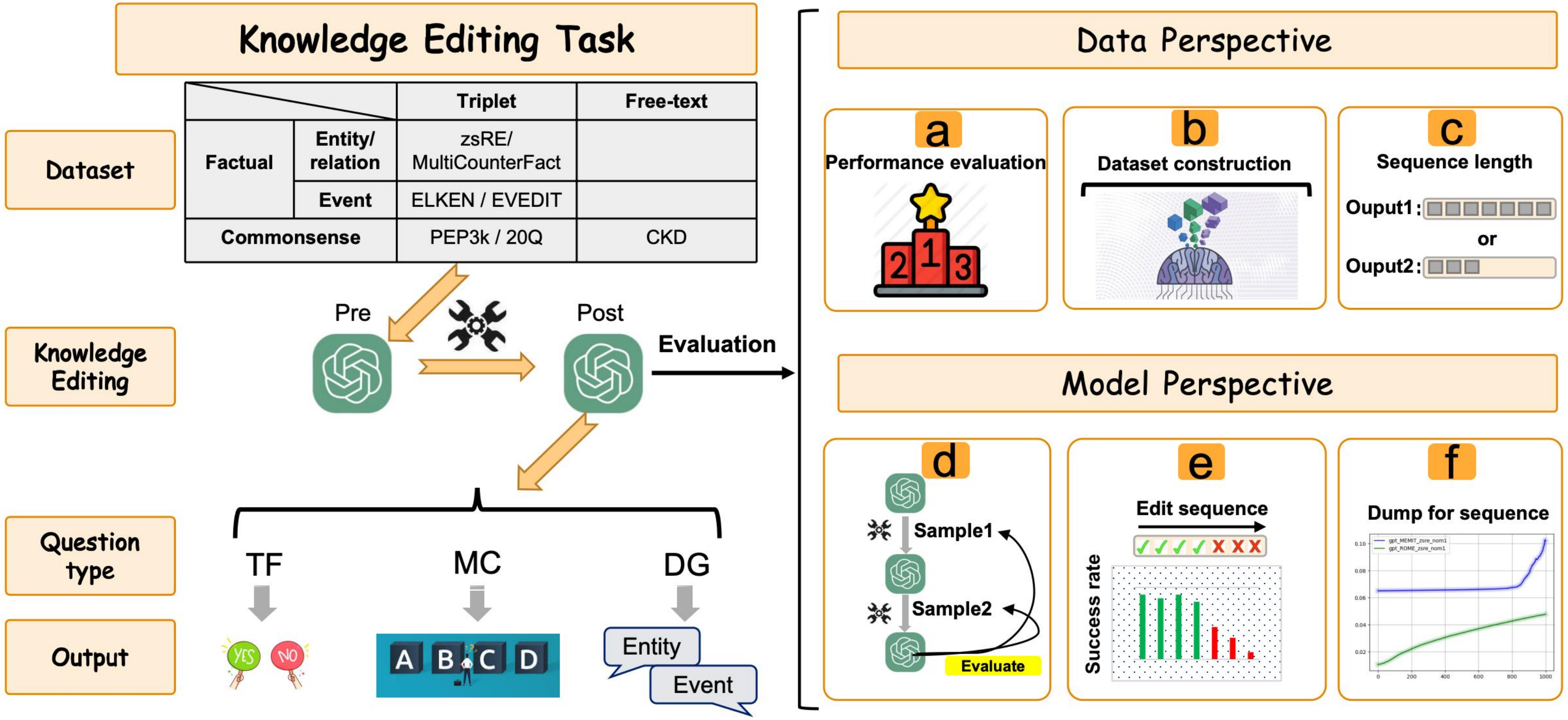


Fig 1. The framework of our work.

NEURAL INFORMATION PROCESSING SYSTEMS

Data perspective

- **We conduct performance evaluation on multiple datasets:**
  > Existing methods affect the performance of downstream tasks.

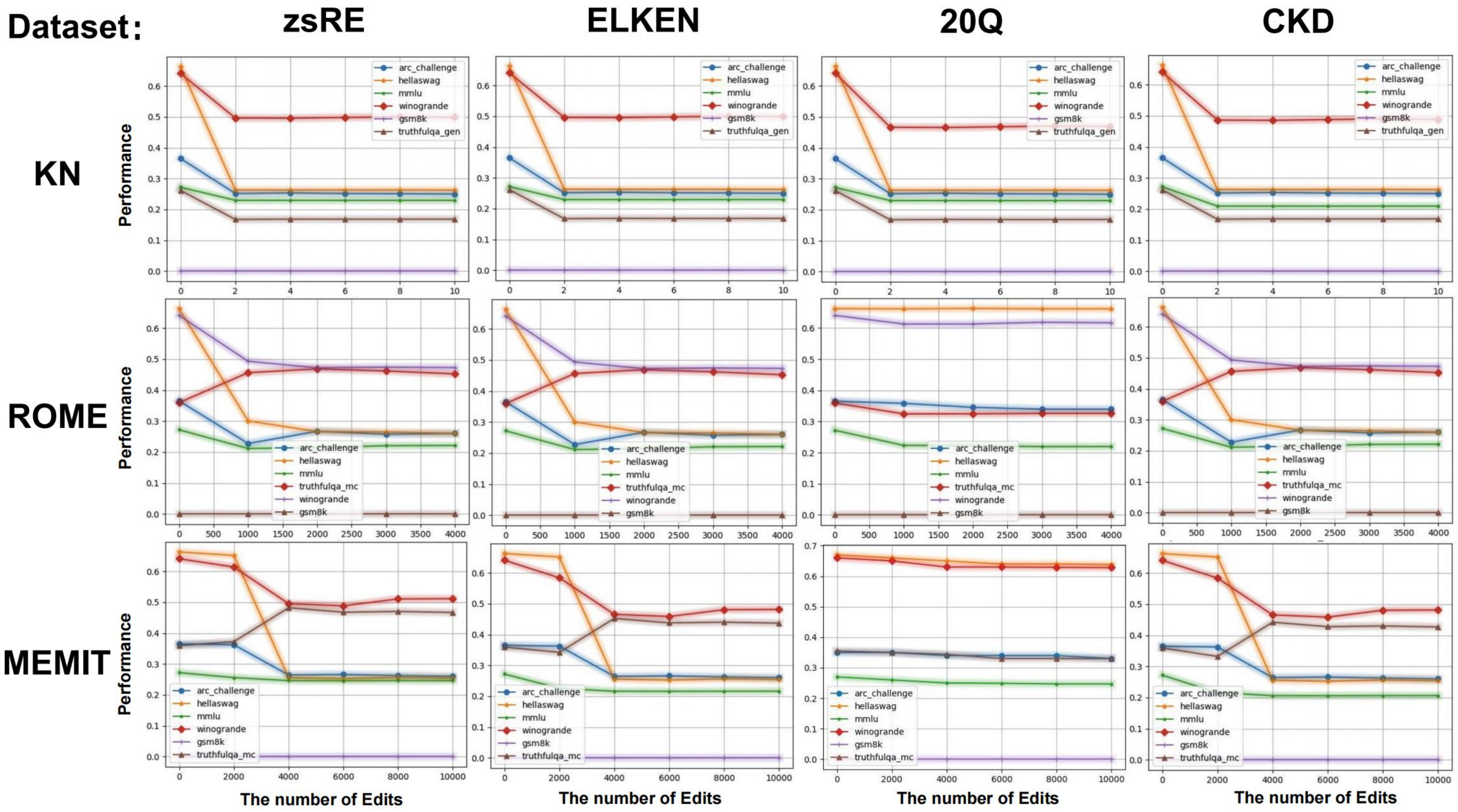


Fig 2. Evaluation results of different editing methods.

Data perspective

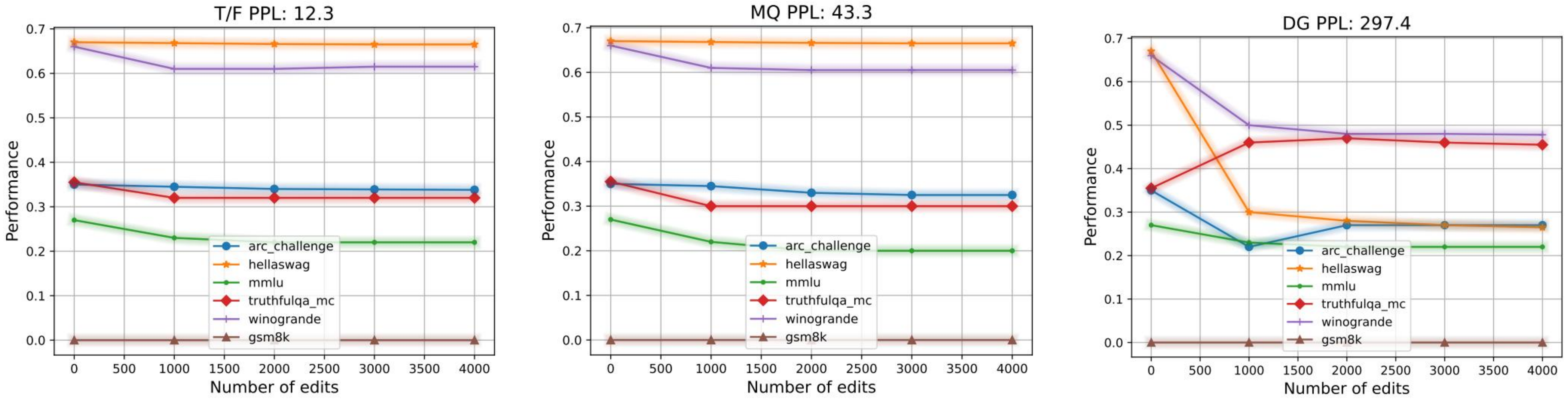
- **To elucidate the impact of different editing objectives:**
  - > We conduct: MQD construction -> impact assessment
    - ~ We created a **M**ulti-**Q**uestion **D**ataset (**MQD**) from ATOMIC
    - ~ Perplexity (PPL) values for the editing objectives: **297.4, 43.3, 12.3**
    - ~ Average length for the three question types: **23.44, 35.03, 13.38**

ATOMIC Data Source: < PersonX accepts PersonY appointment, as a result, PersonY shakes PersonX hand >		
Category	**Component Prompt**	**Target Answer**
DG	PersonX accepts PersonY appointment, resulting in PersonY	shakes PersonX hand
MQ	PersonX accepts PersonY appointment, resulting in PersonY ? Below are four options: (a) to give him a treat; (b) to forgive him; (c) to live happily ever after; (d) shakes PersonX hand. The correct option is	d
T/F	PersonX accepts PersonY appointment, resulting in PersonY shakes PersonX hand . Is this sentence logical? Please answer yes or no. A:	yes

Table 1. An example of converting source data.

Data perspective

- **To elucidate the impact of different editing objectives：**
  - > We conduct: MQD construction -> impact assessment
    - ~ **The decline in model performance after editing is attributed to the diversity of editing objectives and the length of tokens**



(a) True/False questions (T/F). (b) Multiple-choice (MQ). (c) Directly generated (DG).

Fig 3. The performance of the model after editing.

Model perspective

- **To investigate the model factors contributing to degradation:**
  > We propose a new evaluation methods.

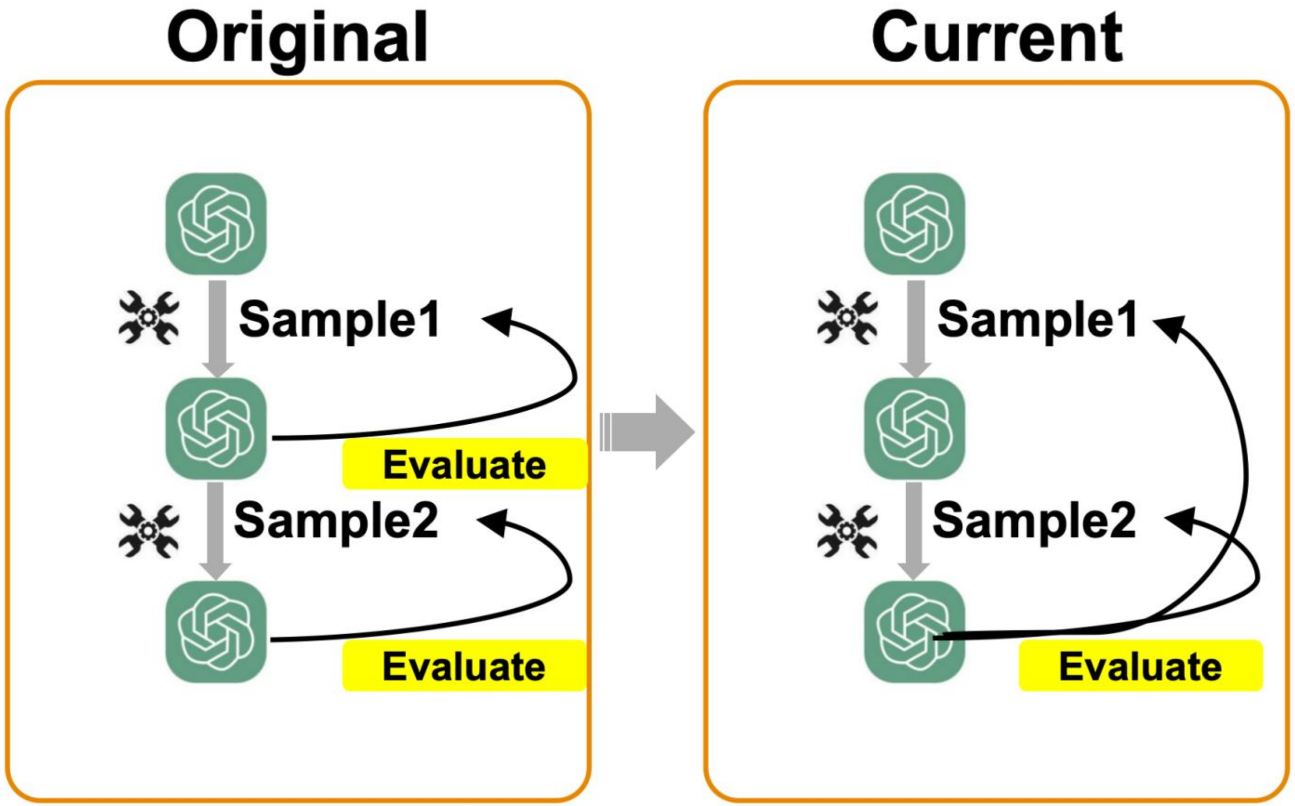


Fig 4. The original and current evaluation methods.

Model perspective

- **To investigate the model factors contributing to degradation:**
  > We evaluate the model's forgetfulness towards edited facts.

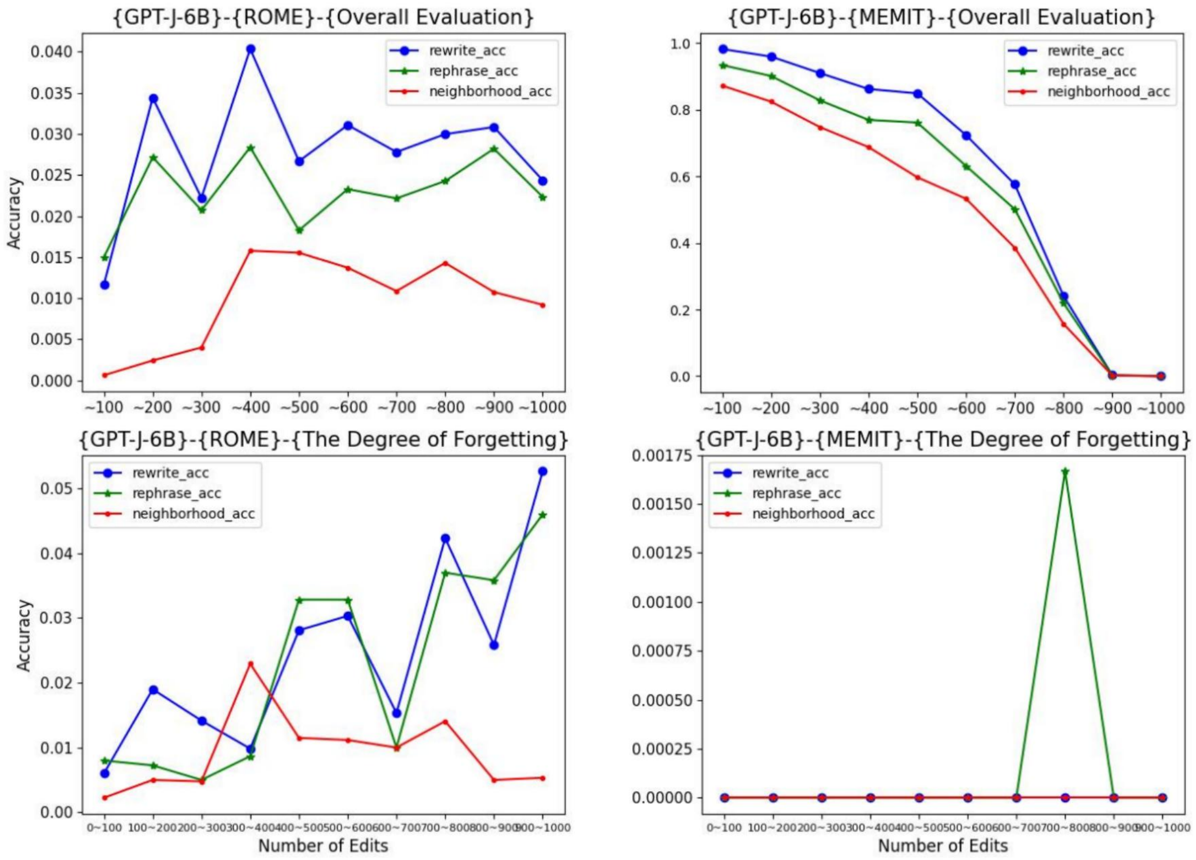


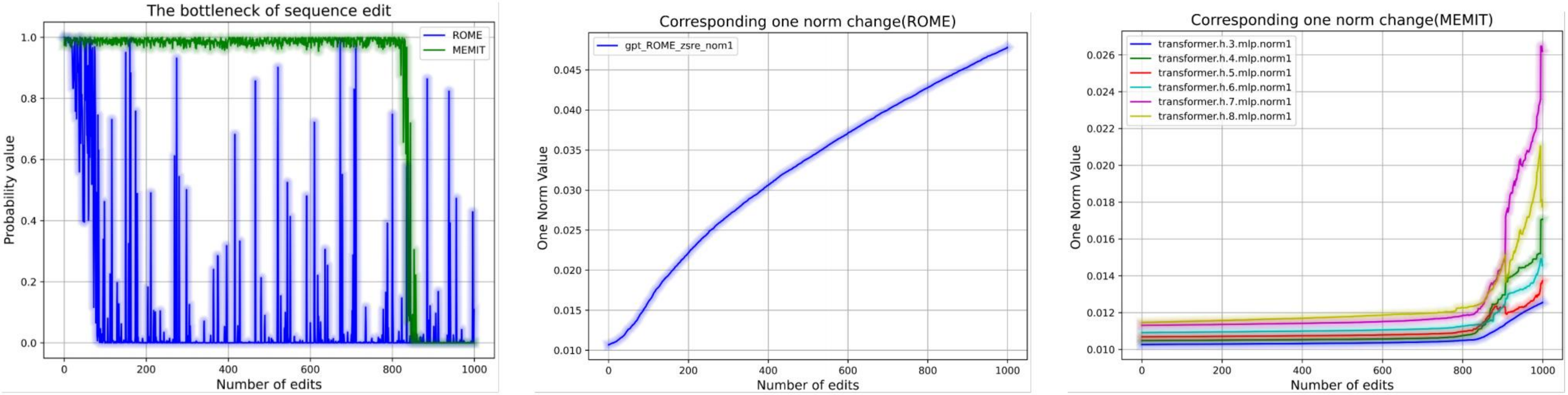
Fig 5. Assessment of forgetting ability of models.

Model perspective

- **To investigate the model factors contributing to degradation:**

  > Then, we examine the bottleneck of sequence edit.

  ~ **The decline in model performance after editing is due to the explosive growth of norms in the editing layers during the editing process**



(a) The probability value. (b) L1-norm on ROME. (c) L1-norm on MEMIT.

Fig 6. The bottleneck and L1-norm correspondence in sequence editing.

D4S method

- **In order to mitigate the increase of the norm, we propose D4S:**
  > First, let's review the MEMIT process:

  ~ For the l^{th} FFN layer that needs to be modified:

$$k_i^l = \frac{1}{N}\sum_{k=1}^{N} \sigma(W_{in}^l \gamma(h_i^{l-1}[pre_k \oplus p(s_i)])), r_i^l = \frac{z_i - h_i^L}{L - l + 1}$$

$$\Delta^l = (R^l {K^l}^T)(Cov^l + K^l {K^l}^T)^{-1}, W_{out}^l \leftarrow W_{out}^l + \Delta^l$$

$$R^l = [r_1^l, \cdots, r_n^l], K^l = [k_1^l, \cdots, k_n^l], Cov^l = K_0^l {K_0^l}^T$$

  ~ where k_i^l and r_i^l are the input and target output respectively. And K_0^l are the inputs of irrelevant knowledge

D4S method

- **In order to mitigate the increase of the norm, we propose D4S:**
  - > For our Dump **for S**equence (**D4S**) method:
    - ~ We can split the Δ^l into two linear segments:

$$R^l {K^l}^T = [r_1^l, \cdots, r_n^l][k_1^l, \cdots, k_n^l]^T = \sum_{i=1}^{n} r_i^l {k_i^l}^T$$

$$K^l {K^l}^T = [k_1^l, \cdots, k_n^l][k_1^l, \cdots, k_n^l]^T = \sum_{i=1}^{n} k_i^l {k_i^l}^T$$

    - ~ For new knowledge to be edited:

$$R^l {K^l}^{T'} = R^l {K^l}^T + r_{n+1}^l {k_{n+1}^l}^T$$

$$K^l {K^l}^{T'} = K^l {K^l}^T + k_{n+1}^l {k_{n+1}^l}^T$$

Experiments

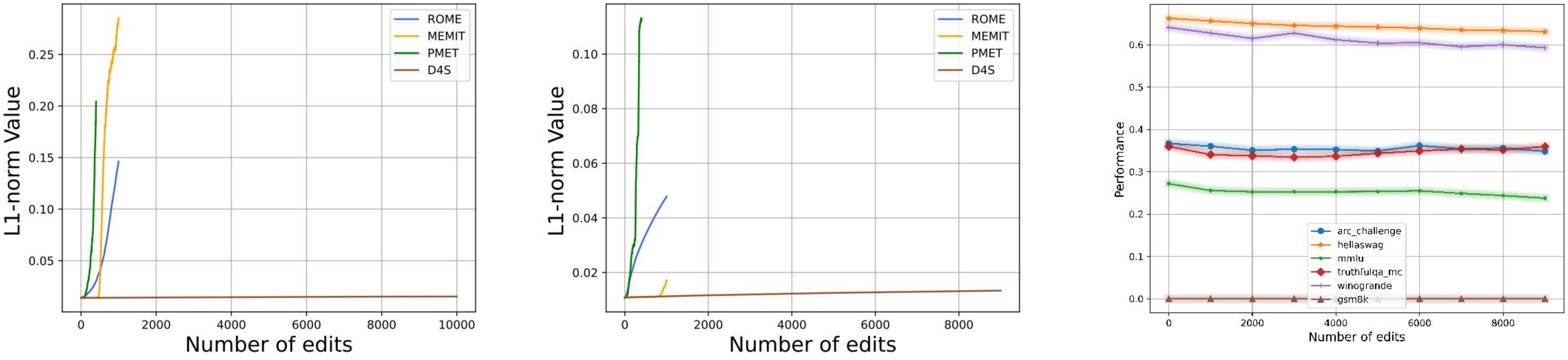
- **The experimental results demonstrates the effectiveness of D4S:**
  - > The experiments are designed to answer two research questions:
    - ~ How does D4S perform in sequence editing compared to other methods?
    - ~ How much damage does D4S do to the model?

Model	Method	Edits = 100				Edits = 500				Edits = 1000			
		Eff.	*Par.*	*Spe.*	*Avg.*	*Eff.*	*Par.*	*Spe.*	*Avg.*	*Eff.*	*Par.*	*Spe.*	*Avg.*
Llama	FT	30.47	25.78	12.19	22.81	19.43	18.06	6.56	14.68	16.13	13.89	5.96	12.00
	ROME	97.68	**92.39**	90.22	**93.43**	48.68	47.72	32.56	42.99	21.84	20.74	4.01	15.53
	MEMIT	92.99	83.22	94.02	90.08	2.90	2.90	2.51	2.77	2.85	2.85	2.74	2.82
	PMET	0.24	0.35	0.46	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GRACE	**99.33**	0.53	**99.88**	66.58	**98.87**	0.59	**99.72**	66.39	**98.94**	0.35	**99.77**	66.35
	D4S(ours)	95.53	86.45	93.44	91.81	91.19	**81.48**	83.34	**85.34**	87.36	**79.30**	78.99	**81.88**
GPT	FT	15.28	7.40	0.42	7.70	12.95	7.84	0.23	7.01	7.03	4.15	0.11	3.77
	ROME	1.17	1.50	0.06	0.91	2.67	1.83	1.55	2.02	2.44	2.23	0.92	1.86
	MEMIT	98.25	93.41	87.24	92.97	84.94	76.15	59.72	73.60	0.00	0.02	0.00	0.01
	PMET	0.33	0.33	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	GRACE	**100.00**	0.40	**100.00**	66.8	**100.00**	0.15	**100.00**	66.72	**100.00**	0.07	**100.00**	66.69
	D4S(ours)	97.72	**97.01**	85.30	**93.34**	97.42	**93.12**	74.65	**88.40**	95.98	**91.17**	70.17	**88.77**

Table 2. Sequencial edits on GPT and Llama with ZsRE dataset.

Experiments

- **The experimental results demonstrates the effectiveness of D4S:**
  - > The experiments are designed to answer two research questions:
    - ~ How does D4S perform in sequence editing compared to other methods?
    - ~ How much damage does D4S do to the model?



(a) Llama_L1-Norm (b) GPT_L1-Norm (c) Downstream tasks

Fig 7. Norms of weight and performance of the edited model.

Thanks !