WISE: Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models

Peng Wang*, Zexi Li*, Ningyu Zhang†, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen†

* Equal contribution. †Equally advising corresponding authors.

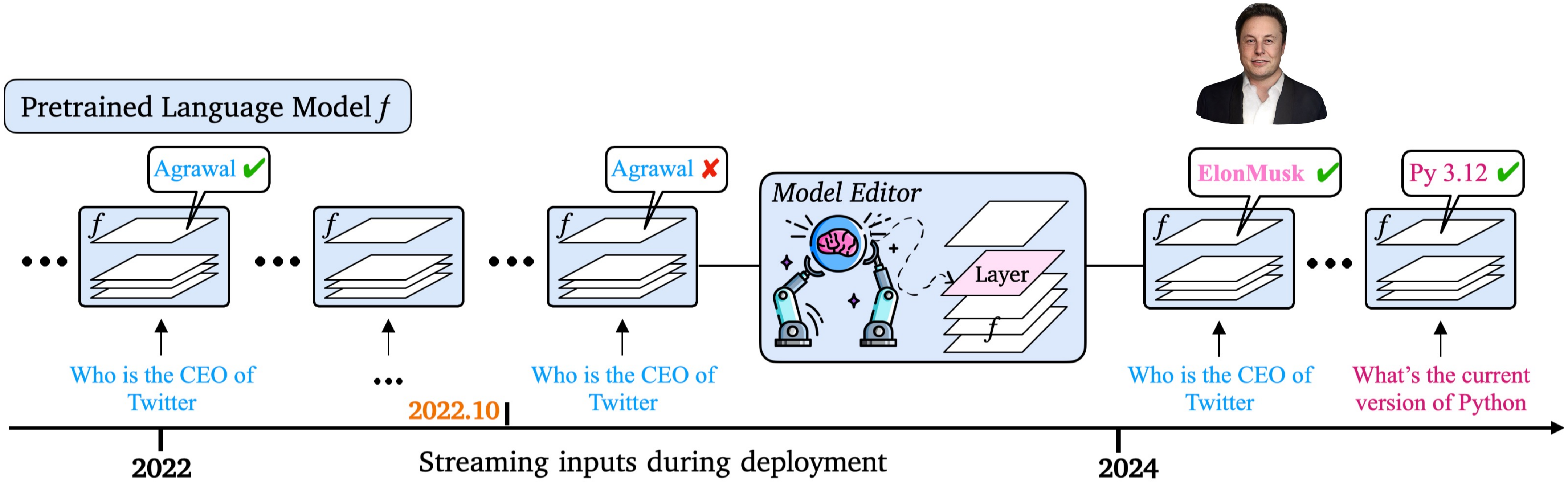
NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024

Background

浙江大學 ZHEJIANG UNIVERSITY

❑Lifelong Model Editing

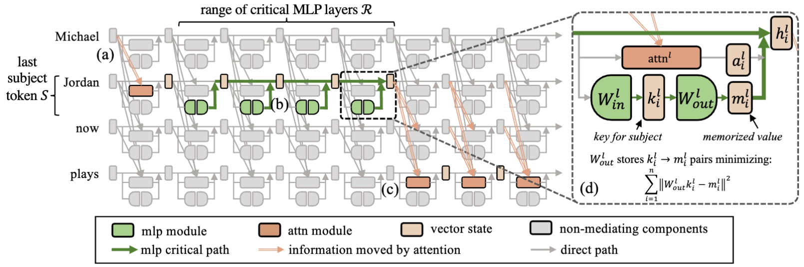


LLM has a series of issues such as knowledge cutoff and hallucination. Continuous editing is crucial.

Motivation

❑Pioneering Work

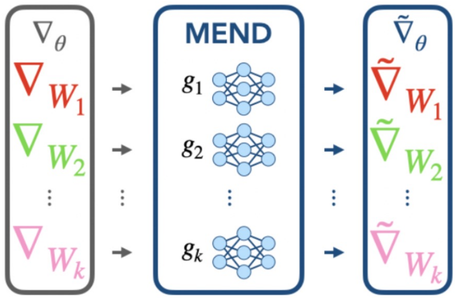
Locate then Edit



ROME [Meng et al, NeurIPS'22]

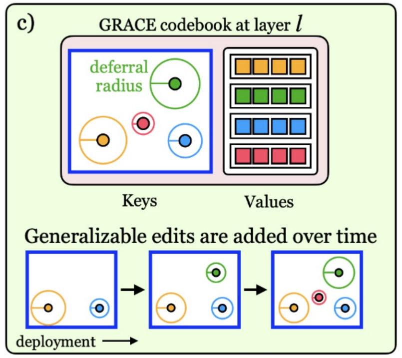
MEMIT [Meng et al, ICLR'23]

Meta Learning



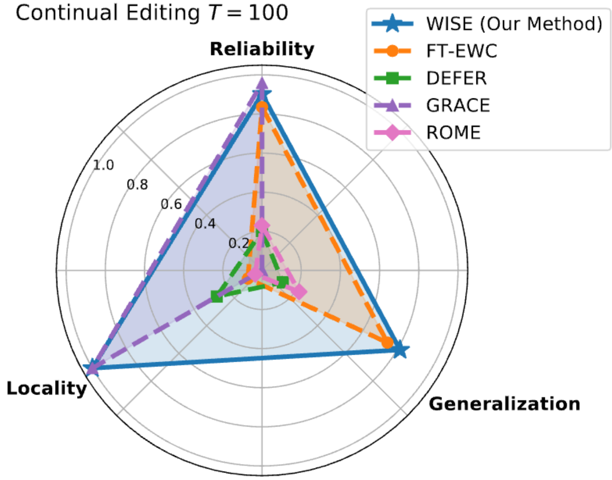
MEND [Mitchell et al, ICLR'22]

Working Memory Editing



GRACE [Thomas et al, NeurIPS'23]

The impossible triangle among Reliability, Generalization, and Locality



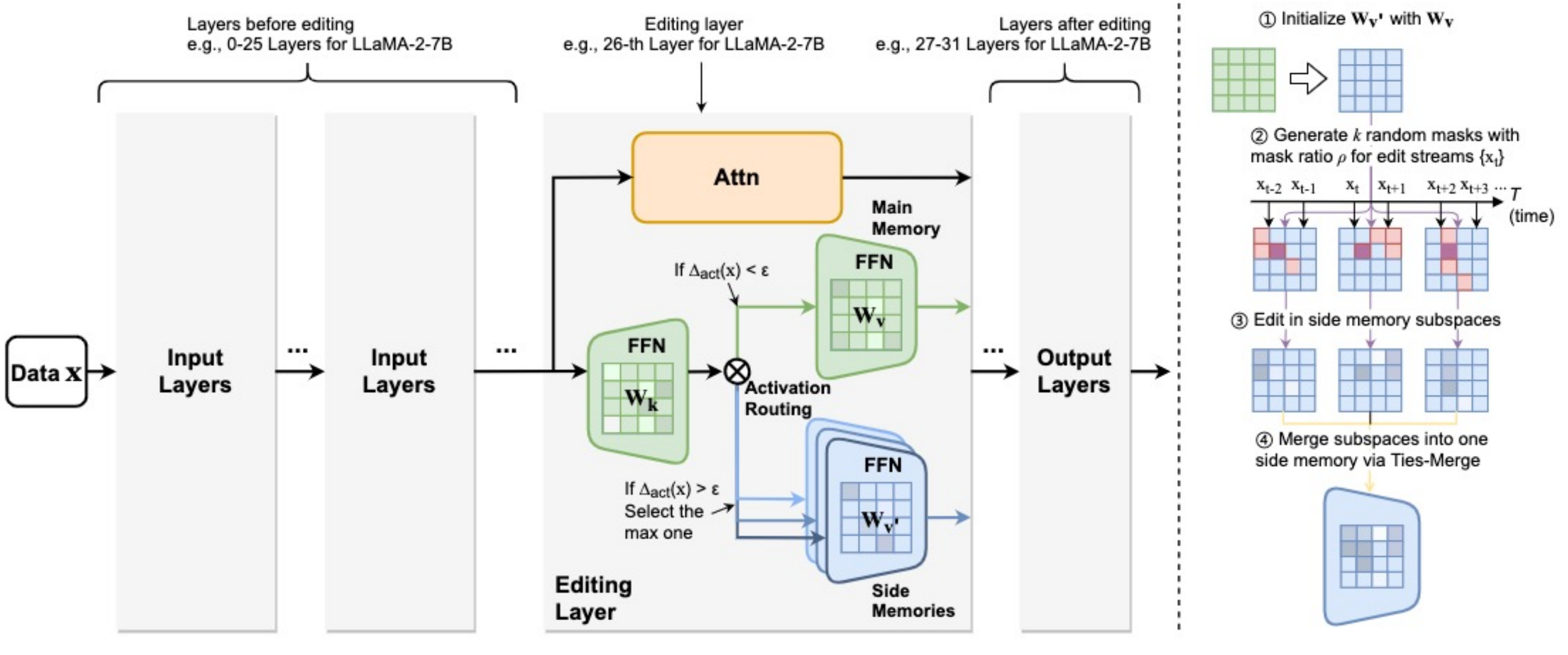
(a) Reliability
LLMs can remember current and previous edits after sequential editing.

(b) Generalization
Editing can also understand and generalize to different queries (unseen).

(c) Locality
It does not affect pre-trained knowledge unrelated to the edits

Methodology

WISE Overview: knowledge editing inspired by **cognitive science**



(a) Workflow Overview with Knowledge Routing

(b) Knowledge Sharding and Merging

$$\mathrm{FFN}(\mathbf{f}) = \mathbf{a} \cdot \mathbf{W}_v = \sigma(\mathbf{f}^\top \cdot \mathbf{W}_k) \cdot \mathbf{W}_v,$$

1. Utilize the target layer MLP as a memory component.
   1. Green: Long-term memory (pre-trained knowledge)
   2. Blue: Working memory (editable knowledge)

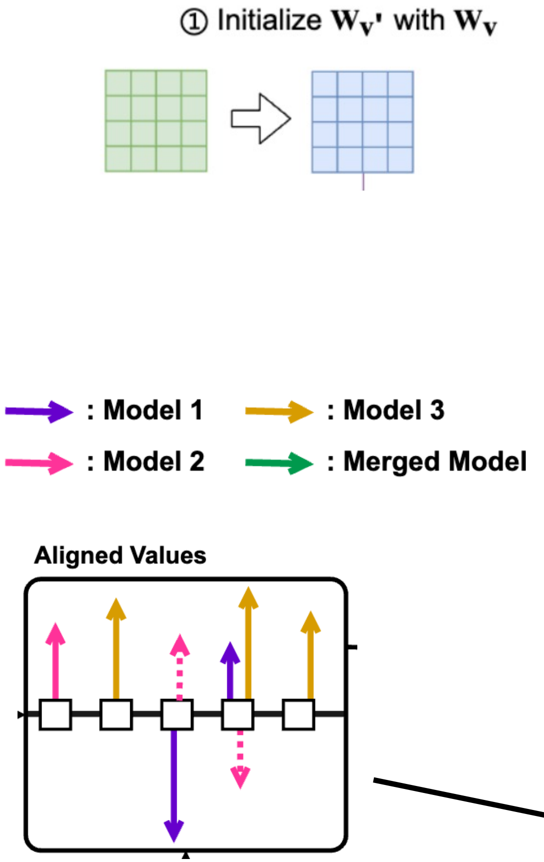
2. Knowledge memory fusion :
   Moderate knowledge density leads to better editing effects

3. Knowledge memory retrieval :
   Retrieve working memory through neural activation.

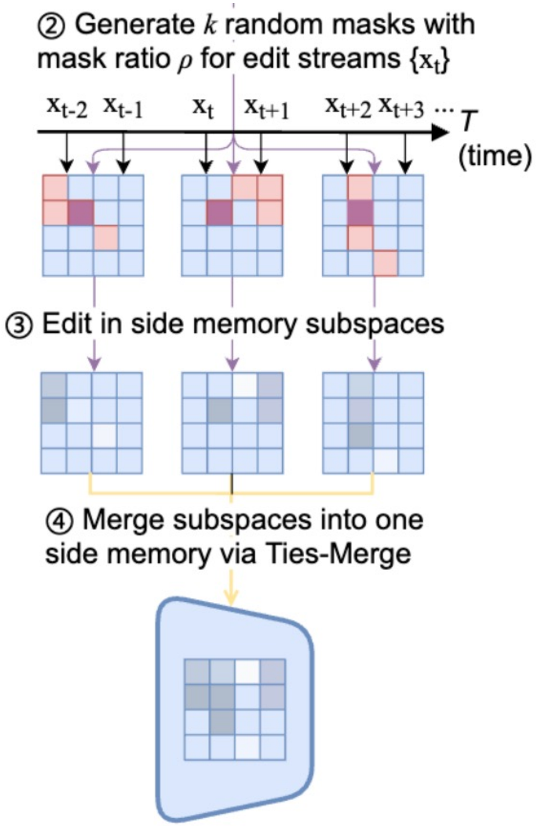
WISE: Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models (2024)

Methodology

❑Knowledge Memory Fusion



Divide thousands of edit partitions by random mask gradients.

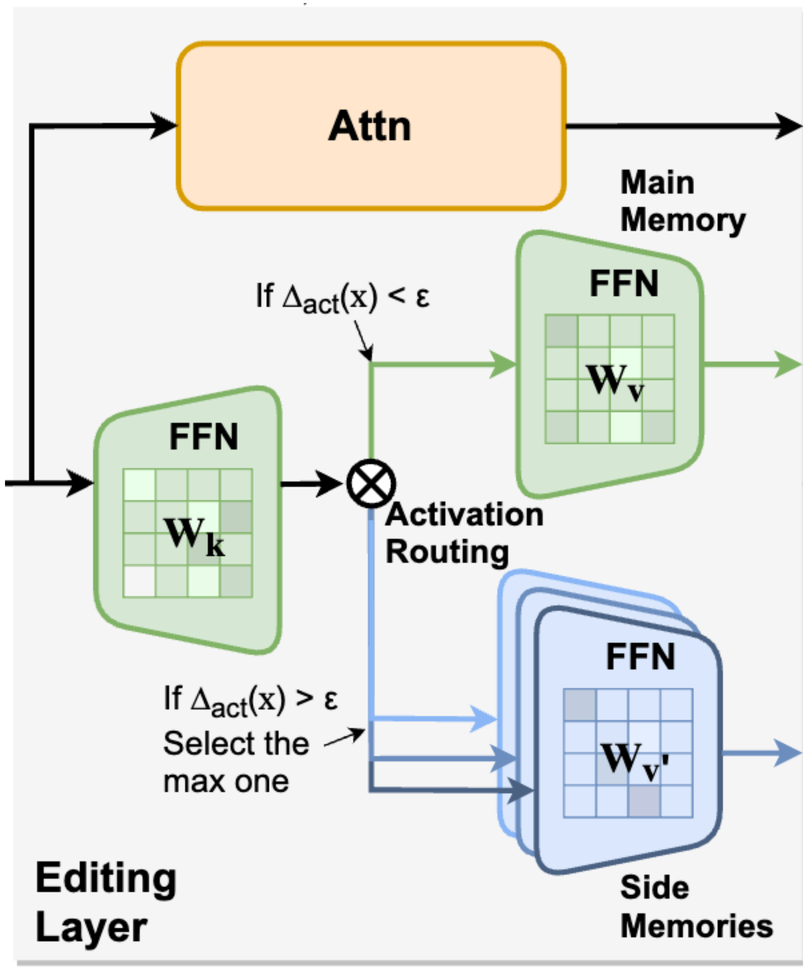


Merge Working Memory

$$\mathbf{W}_{v'} \leftarrow \mathbf{W}_v + \mathrm{Ties}(\mathrm{T}_e; \mathbf{W}_v).$$

Methodology

❑WISE： Gate mechanism, working/long-term Memory?



$$\Delta_{\text{act}}(\mathbf{x}) = \|\mathcal{A}(\mathbf{x}) \cdot (\mathbf{W}_{v'} - \mathbf{W}_v)\|_2,$$

$$L_a = \min_{\mathbf{W}_{v'}} \{\underbrace{\max(0, \Delta_{\text{act}}(\mathbf{x}_i) - \alpha)}_{\text{Out of editing scope}} + \overbrace{\max(0, \beta - \Delta_{\text{act}}(\mathbf{x}_e))}^{\text{In editing scope}} + \max(0, \gamma - (\Delta_{\text{act}}(\mathbf{x}_e) - \Delta_{\text{act}}(\mathbf{x}_i)))\},$$

$$\text{s.t. } \mathbf{x}_e \in \mathcal{D}_{\text{edit}}, \mathbf{x}_i \in \mathcal{D}_{\text{irr}}.$$

For input x:

- A set of inputs within the edit scope tends to activate the working memory (with higher activation on $\boldsymbol{W_{v'}}$).
- A set of unrelated inputs tends to rely on long-term memory (with lower activation on $\boldsymbol{W_{v'}}$).

We designed a margin-based loss to identify routes through activation. $\boxed{L_{\text{edit}} = -\log P_{W_{v'}}(\mathbf{y}_e|\mathbf{x}_e) + L_a.}$

Experiments

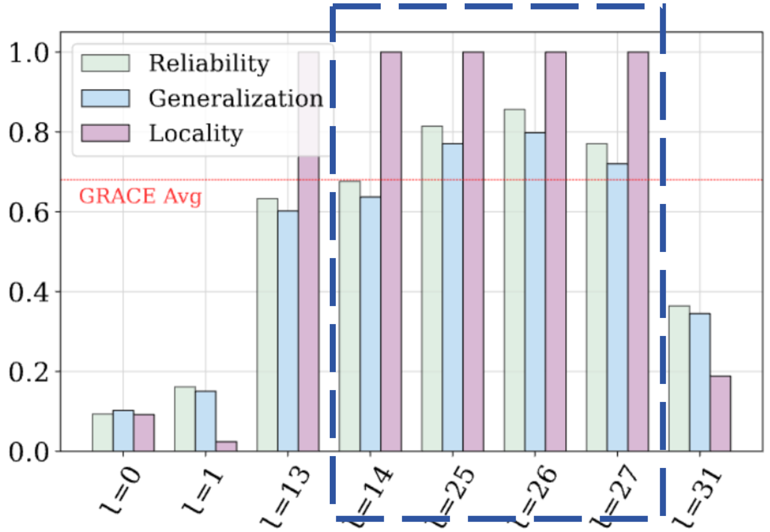
❑Experimental Results： QA

Method	QA															
	$T=1$				$T=10$				$T=100$				$T=1000$			
	Rel.	Gen.	Loc.	Avg.	Rel.	Gen.	Loc.	Avg.	Rel.	Gen.	Loc.	Avg.	Rel.	Gen.	Loc.	Avg.
LLaMA-2-7B																
FT-L	0.57	0.52	0.96	0.68	0.48	0.48	0.76	0.57	0.30	0.27	0.23	0.27	0.19	0.16	0.03	0.13
FT-EWC	0.96	**0.95**	0.02	0.64	0.82	0.76	0.01	0.53	0.83	0.74	0.08	0.55	0.76	0.69	0.08	0.51
MEND	0.95	0.93	0.98	0.95	0.26	0.28	0.28	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ROME	0.85	0.80	0.99	0.88	0.64	0.62	0.75	0.67	0.23	0.22	0.04	0.16	0.01	0.01	0.00	0.01
MEMIT	0.84	0.81	0.99	0.88	0.58	0.58	0.85	0.67	0.02	0.02	0.02	0.02	0.04	0.04	0.02	0.03
MEMIT-MASS	0.84	0.81	0.99	0.88	0.75	0.72	0.97	0.81	0.76	0.68	0.85	0.76	0.69	0.65	0.62	0.65
DEFER	0.68	0.58	0.56	0.61	0.65	0.47	0.36	0.49	0.20	0.12	0.27	0.20	0.03	0.03	0.74	0.27
GRACE	**0.98**	0.08	**1.00**	0.69	**0.96**	0.00	**1.00**	0.65	**0.96**	0.00	**1.00**	0.65	**0.97**	0.08	**1.00**	0.68
WISE	**0.98**	0.92	**1.00**	**0.97**	0.94	**0.88**	**1.00**	**0.94**	0.90	**0.81**	**1.00**	**0.90**	0.77	**0.72**	**1.00**	**0.83**
Mistral-7B																
FT-L	0.58	0.54	0.91	0.68	0.39	0.39	0.50	0.43	0.11	0.10	0.02	0.08	0.16	0.13	0.01	0.10
FT-EWC	**1.00**	**0.99**	0.01	0.67	0.84	0.78	0.02	0.55	0.82	0.72	0.09	0.54	0.76	0.69	0.09	0.51
MEND	0.94	0.93	0.98	0.95	0.01	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ROME	0.79	0.77	0.98	0.85	0.58	0.57	0.75	0.63	0.05	0.05	0.02	0.04	0.04	0.04	0.02	0.03
MEMIT	0.81	0.79	0.99	0.86	0.46	0.45	0.61	0.51	0.00	0.00	0.01	0.00	0.04	0.04	0.02	0.03
MEMIT-MASS	0.81	0.79	0.99	0.86	0.74	0.71	0.97	0.81	0.73	0.71	0.88	0.77	0.73	**0.70**	0.62	0.68
DEFER	0.64	0.54	0.79	0.66	0.53	0.43	0.29	0.42	0.28	0.17	0.26	0.24	0.02	0.02	0.67	0.24
GRACE	**1.00**	0.00	**1.00**	0.67	**1.00**	0.00	**1.00**	0.67	**1.00**	0.00	**1.00**	0.67	**1.00**	0.02	**1.00**	0.67
WISE	0.98	0.97	**1.00**	**0.98**	0.92	**0.89**	**1.00**	**0.94**	0.87	**0.80**	**1.00**	**0.89**	0.70	0.67	**1.00**	**0.79**

WISE maintains 70%+ editing success rate and 100% locality preservation after 1,000 edits.

Experiments

Where to introduce WISE into the LLM



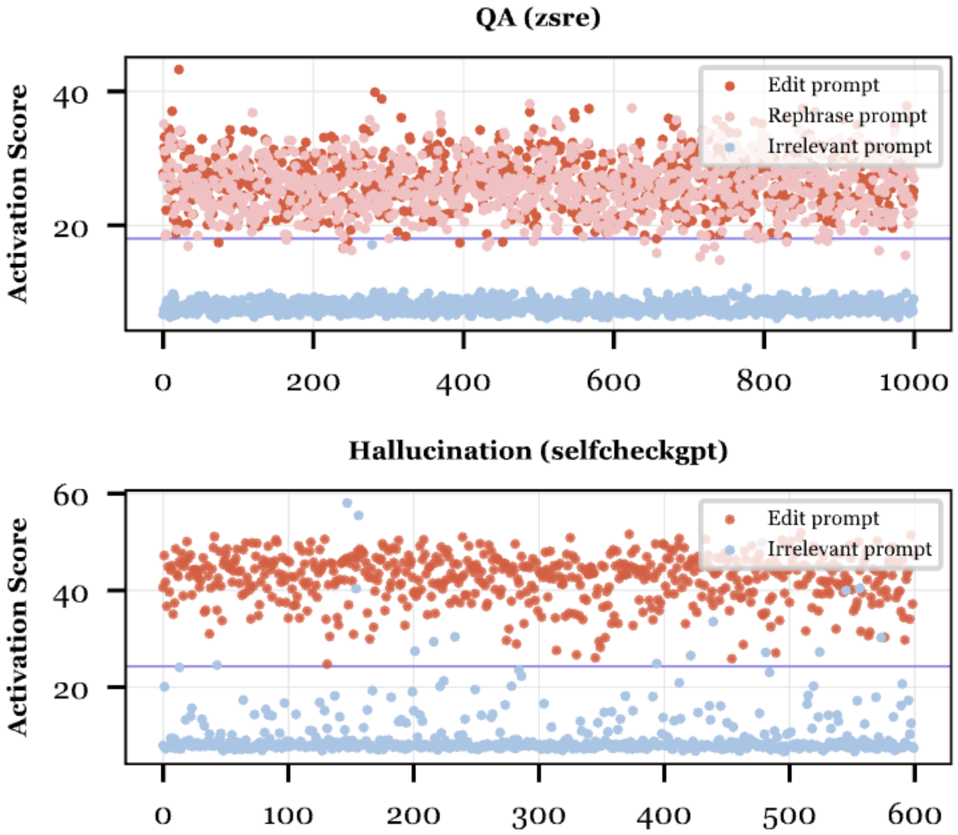


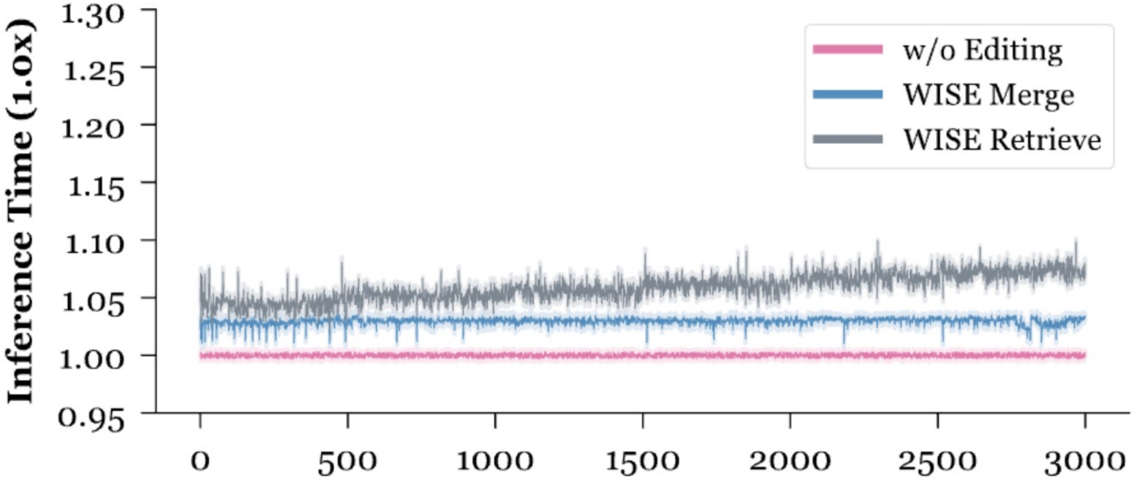
Figure 3: **Activations of the memory routing module of WISE when varying T.** `X-axis`: Num edits. `LLaMA-7B`.

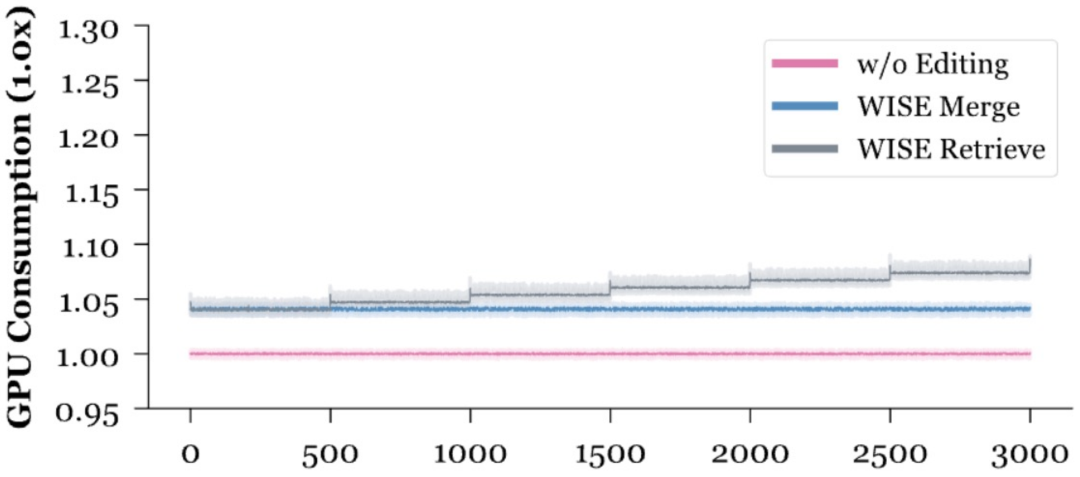
- Finding 1: Mid-to-Late Layers is effective.
- Finding 2: Gate mechanism routes the editing prompt and unseen paraphrases into the side memory

Analysis

❑At what cost?

Inference Latency (Left)
Computational Cost (Right)





WISE-Merge: Constant 3% Latency ⬆

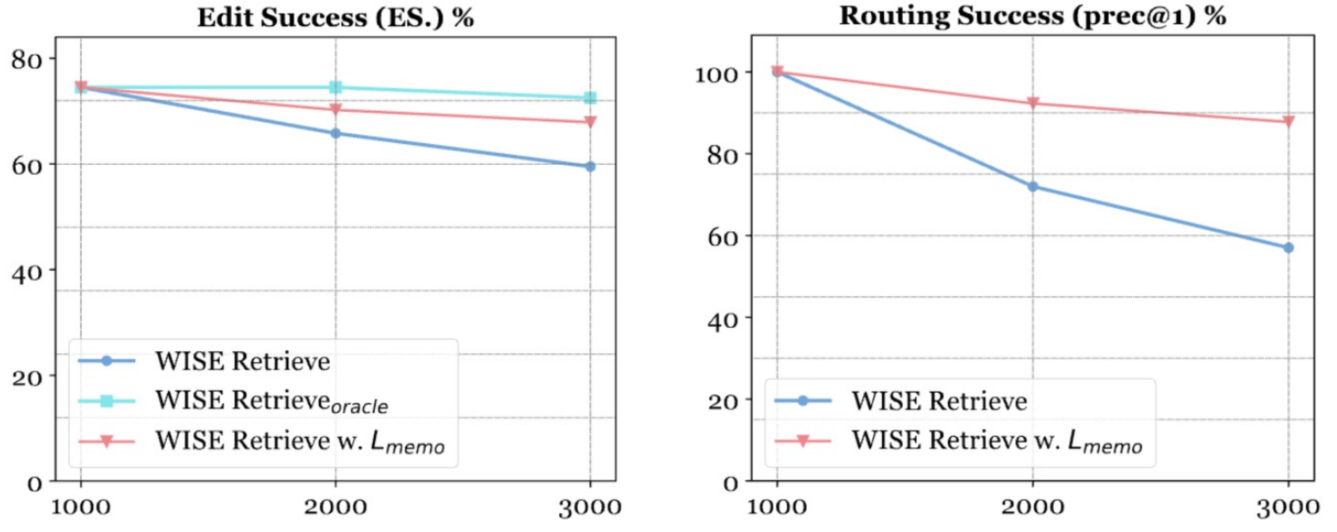
WISE-Merge: Constant 4% Parameter ⬆

WISE-Retrieve will gradually increase computational and inference costs.

WISE: Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models (2024)

❑Retrieval Accuracy matters

Method	T = 2000 Rel.	Gen.	Loc.	Avg.	T = 3000 Rel.	Gen.	Loc.	Avg.
GRACE	**0.96**	0.03	1.00	0.66	**0.96**	0.03	1.00	0.66
MEMIT-MASS	0.64	0.58	0.55	0.59	0.58	0.53	0.47	0.53
WISE-Merge	0.66	0.63	1.00	0.76	0.58	0.56	1.00	0.71
WISE-Retrieve	0.68	**0.64**	**1.00**	**0.77**	0.61	**0.58**	**1.00**	**0.73**
WISE-Retrieve$_{\text{oracle}}$	0.77	0.72	1.00	0.83	0.75	0.70	1.00	0.82



(a) Average of Rel. and Gen. (b) Retrieval Acc. by Top-1 Activation

Improve memory specificity through replay:

L_{memo}: Ensures that the current shard has lower activation for past edit prompts.

$$L'_a = L_a + \underbrace{\max(0, \Delta_{\text{act}}(\mathbf{x}_\text{m}) - \alpha)}_{L_{\text{memo}}}, \quad \text{s.t. } \mathbf{x}_\text{m} \in \mathcal{D}_{\mathbf{W}_j}.$$

WISE-Retrieve_{oracle}: Based on the retrieval upper bound, we observe significant room for improvement. As shown in Figure (b), the bottleneck of WISE-Retrieve is retrieval accuracy.

Figure (b): 3K edits boost retrieval rate to 88%, +3% (compared to (a.))

Thanks for Listening