Hippocampal-like Sequential Editing for Continual Knowledge Updates in Large Language Models

Quntian Fang*[1], Zhen Huang*[1], Zhiliang Tian✉[1], Minghao Hu✉[4], Dongsheng Li[1], Yiping Yao[2], Xinyue Fang[1], Menglong Lu[3], Guotong Geng[4]

*contributed equally to this work

✉Corresponding authors

[1]National Key Laboratory of Parallel and Distributed Computing, [2]College of System Engineering, [3]Key Laboratory of Advanced Microprocessor Chips and Systems, National University of Defense Technology, [4]Center of Information Research, AMS

main contact: fangquntian@nudt.edu.cn

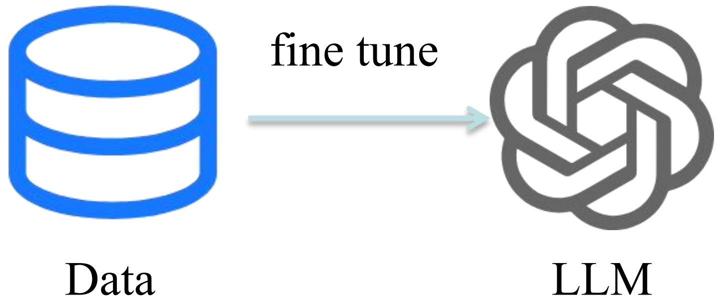
- Background
- Methods
- Theoretical Analysis
- Results
- Future Work

Background

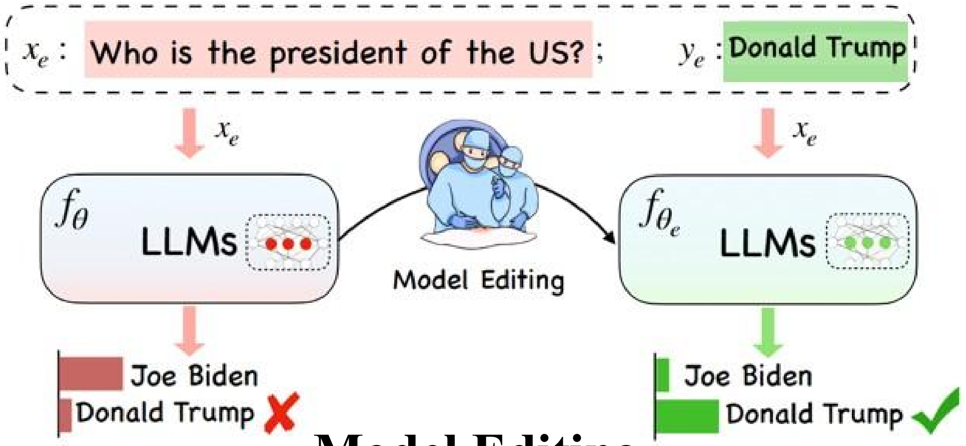
- **Large language model knowledge update**

Large language models (LLMs) need to frequently update their knowledge in practical applications to correct errors or outdated information.

• SFT vs. Model Editing



SFT



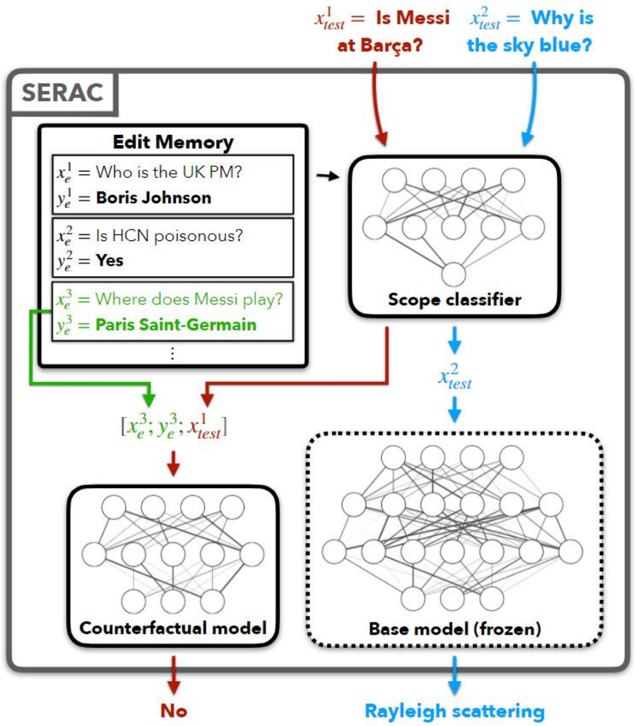
Model Editing

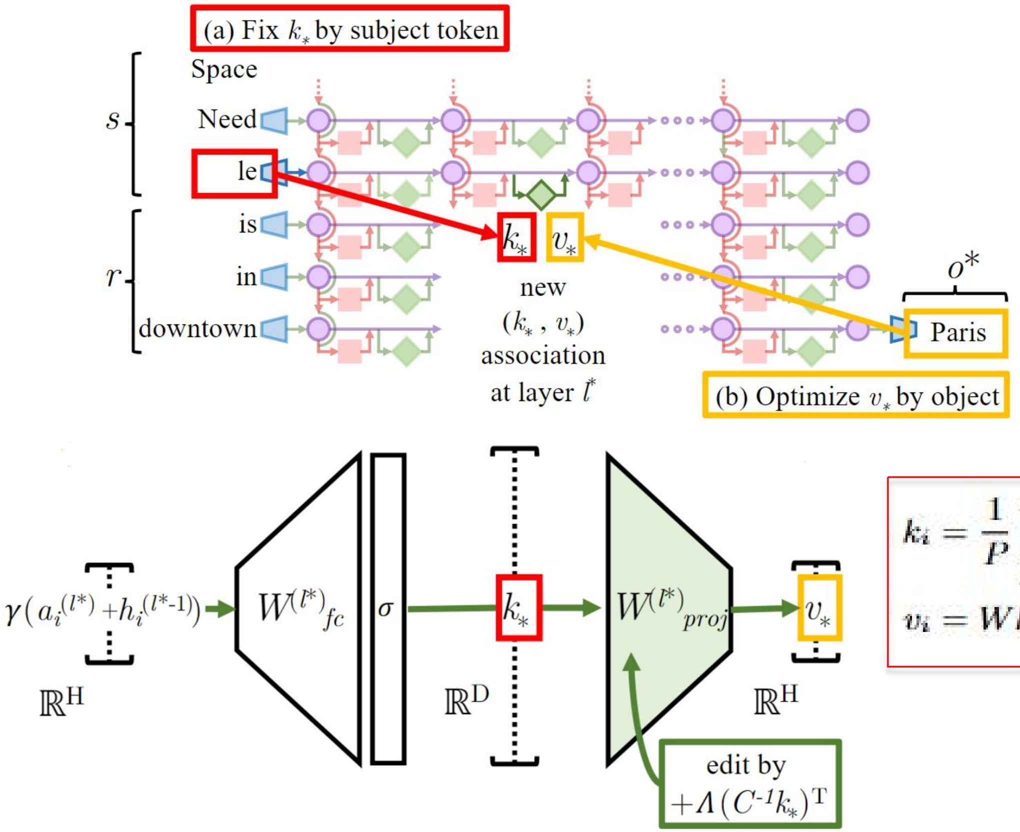
Comparison

	SFT	Model Editing
Objective	Improve overall performance on a task or domain	Make precise changes to specific knowledge or behaviors
Cost	Requires thousands of examples and full parameter updates	Achieved with minimal data and computation
Impact	Risks altering model behavior on unrelated tasks	Aims to preserve the model's general capabilities

Yao, Yunzhi, et al. "Editing Large Language Models: Problems, Methods, and Opportunities." EMNLP. 2023.

• Mainstream Methods





Locate

$$k_i = \frac{1}{P}\sum_{j=1}^{P} Enc(x_j \oplus s_i),$$

$$v_i = W k_i + \delta_i.$$

Editing

Parameter-Preserving Methods

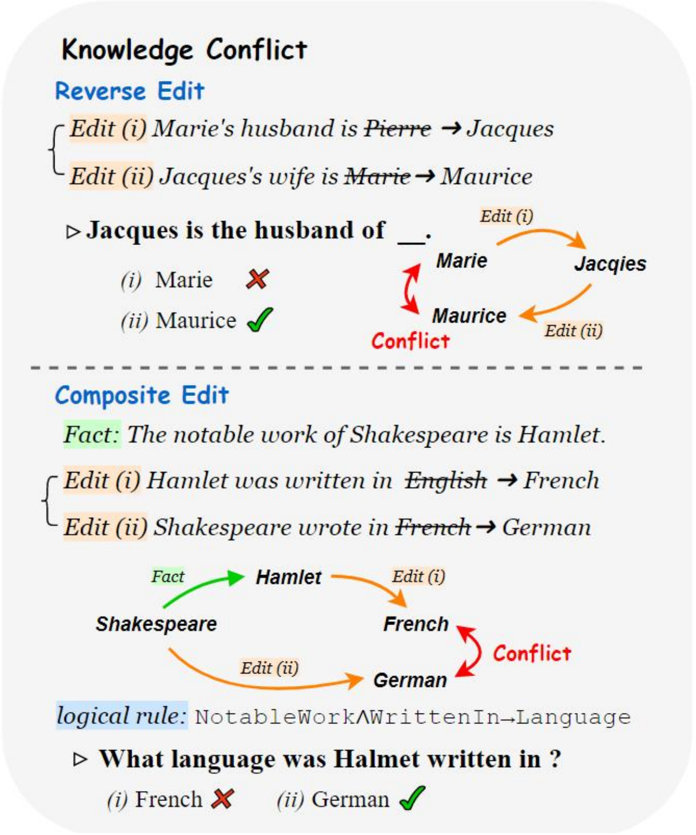
Introduce an external module to store editing knowledge and freeze the original parameters

Parameter-Modifying Methods

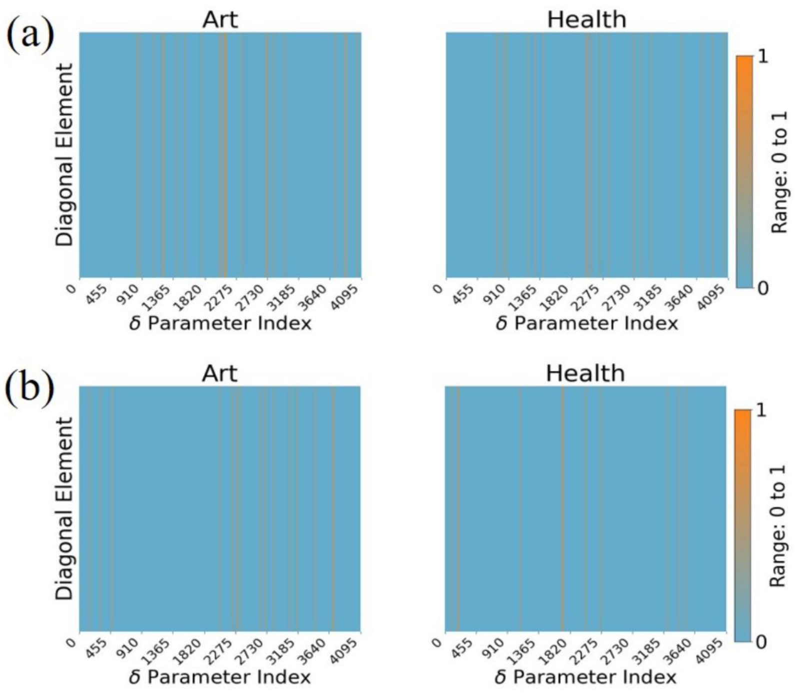
Modify the model parameters directly to adapt to the new knowledge

Mitchell, Eric, et al. "Memory-based model editing at scale." ICML, 2022.
Meng, Kevin, et al. "Locating and editing factual associations in gpt.", Neurips2022

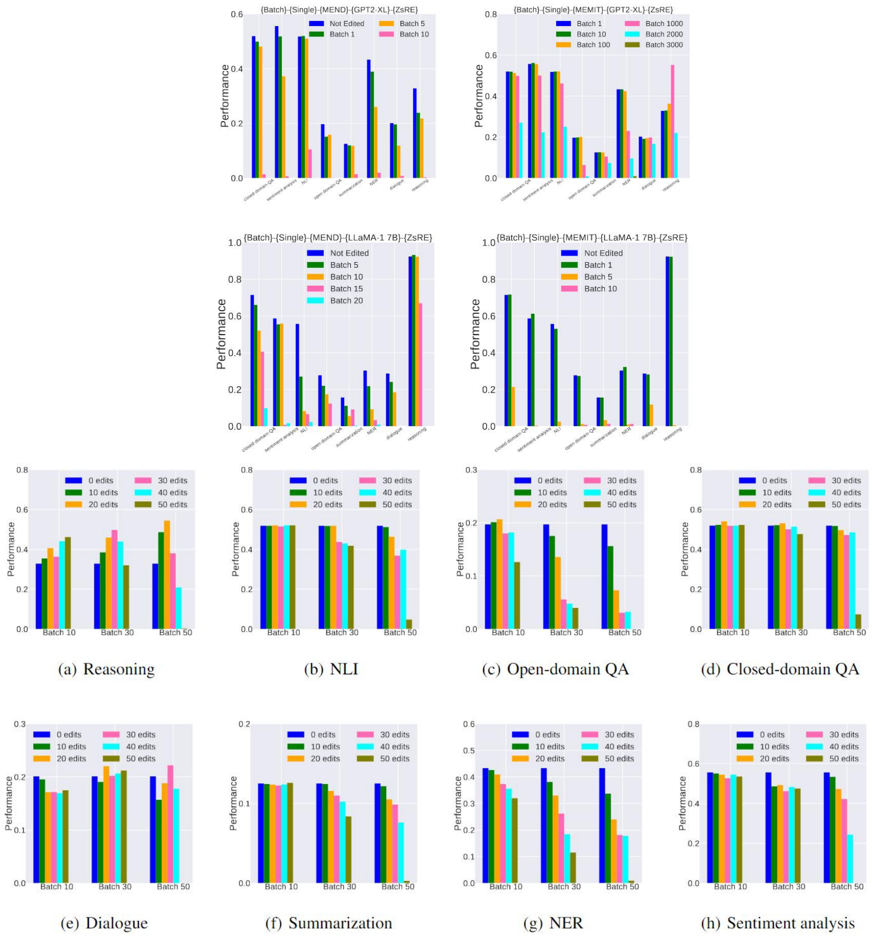
- **Limitations**



Knowledge conflict



Domain knowledge interference



Model collapse and Catastrophic forgetting

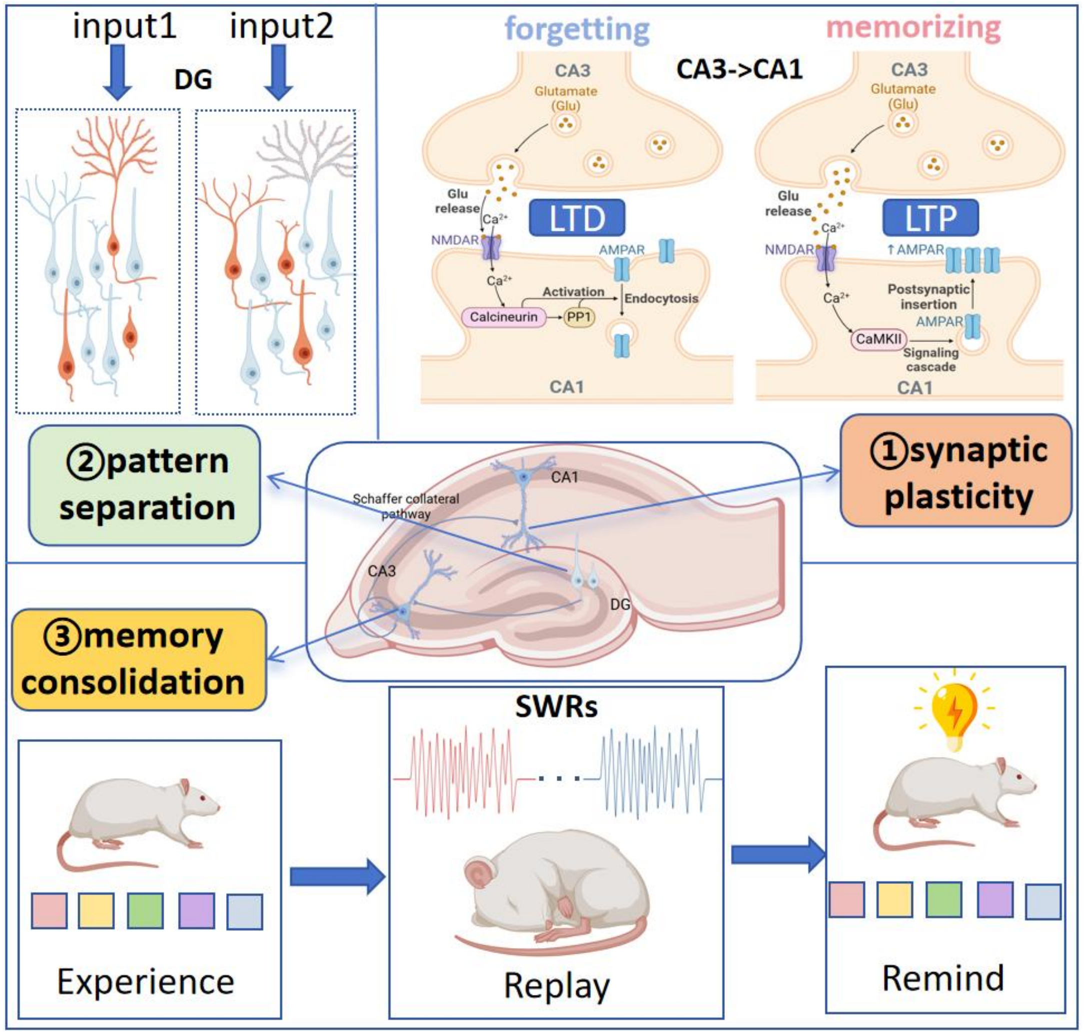
Li, Zhoubo, et al. "Unveiling the pitfalls of knowledge editing for large language models." , ICLR 2024.

Gu, Jia-Chen, et al. "Model Editing Harms General Abilities of Large Language Models: Regularization to the Rescue.", EMNLP 2024

- **Inspiration**

2.Pattern Separation (Dentate Gyrus)
Creates distinct neural representations to **minimize interference between similar memories**.

3.Memory Consolidation (SWRs in CA3→CA1)
Reactivates neural traces to **stabilize and integrate memories for the long term**.

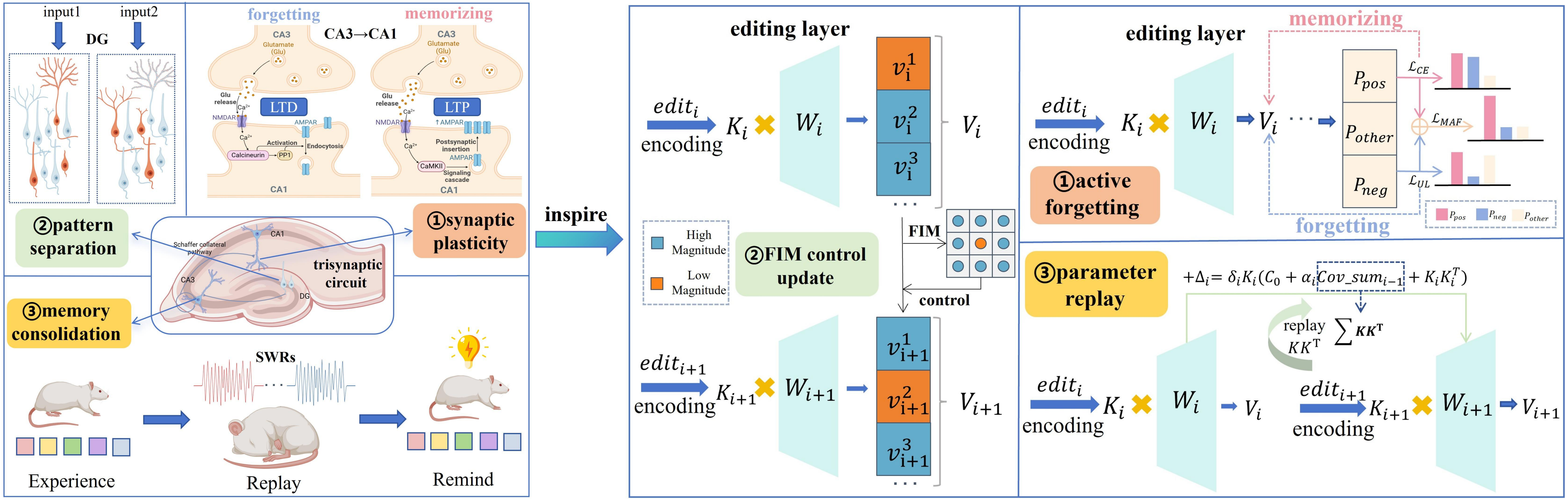


1.Active Forgetting (LTD in CA3→CA1)
Selectively weakens synaptic connections to **discard outdated information**.

• Main Framework

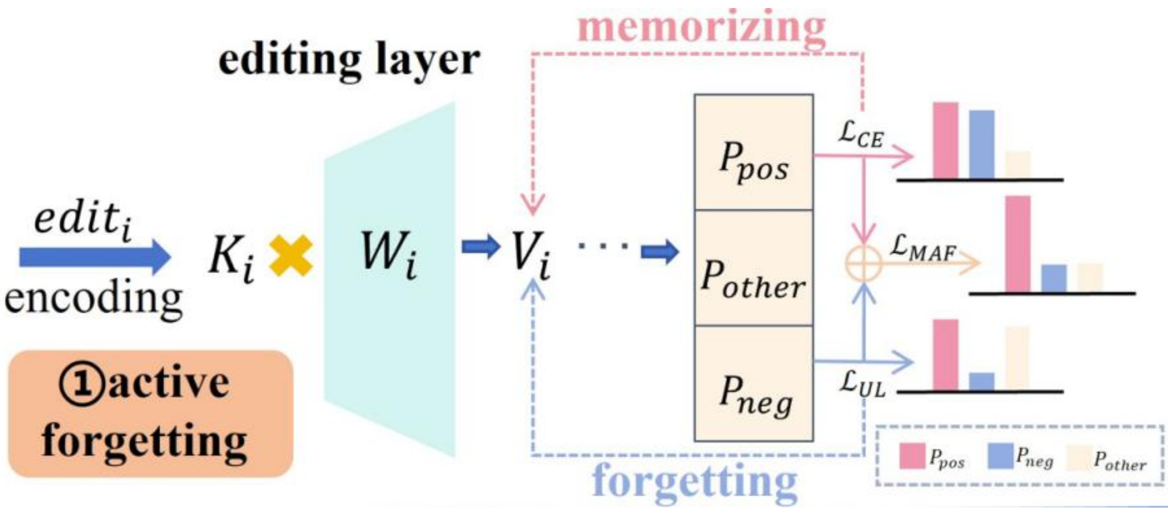
(a) Trisynaptic Circuit in the Hippocampus

(b) Our proposed method HSE



Methods

• Memory-directed Active Forgetting via Machine Unlearning

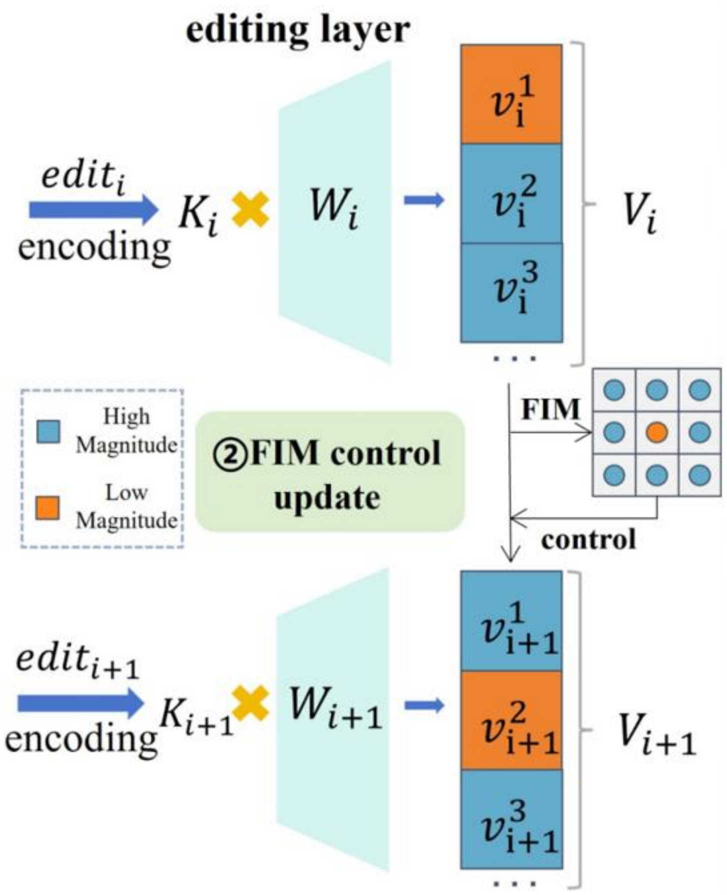


LLMs can selectively forget outdated knowledge while efficiently acquiring and integrating new knowledge

$$\delta^* = \arg\min_{\delta}\{\underbrace{-\alpha \sum_{(x,y)\in D} \log p_\delta(y|x)}_{editing} \underbrace{-(1-\alpha) \sum_{(x,\tilde{y})\in \tilde{D}} \log[1 - p_\delta(\tilde{y}|x)]}_{machine\ unlearning}\}$$

$$\mathcal{L}_{MAF}(\delta) = \mathcal{L}_{CE}(\delta) + \mathcal{L}_{UL}(\delta) = -\frac{\alpha}{|D|} \sum_{(x,y)\in D} \log p_\delta(y|x) - \frac{(1-\alpha)}{|\tilde{D}|} \sum_{(x,\tilde{y})\in \tilde{D}} \log\left[1 - p_\delta(\tilde{y}|x)\right]$$

• Memory Stability Preservation via Fisher Information Matrix



FIM **constrains the update magnitude** of parameters with significant influence on the model outputs, and permits more substantial updates for parameters with less impact on the model outputs.

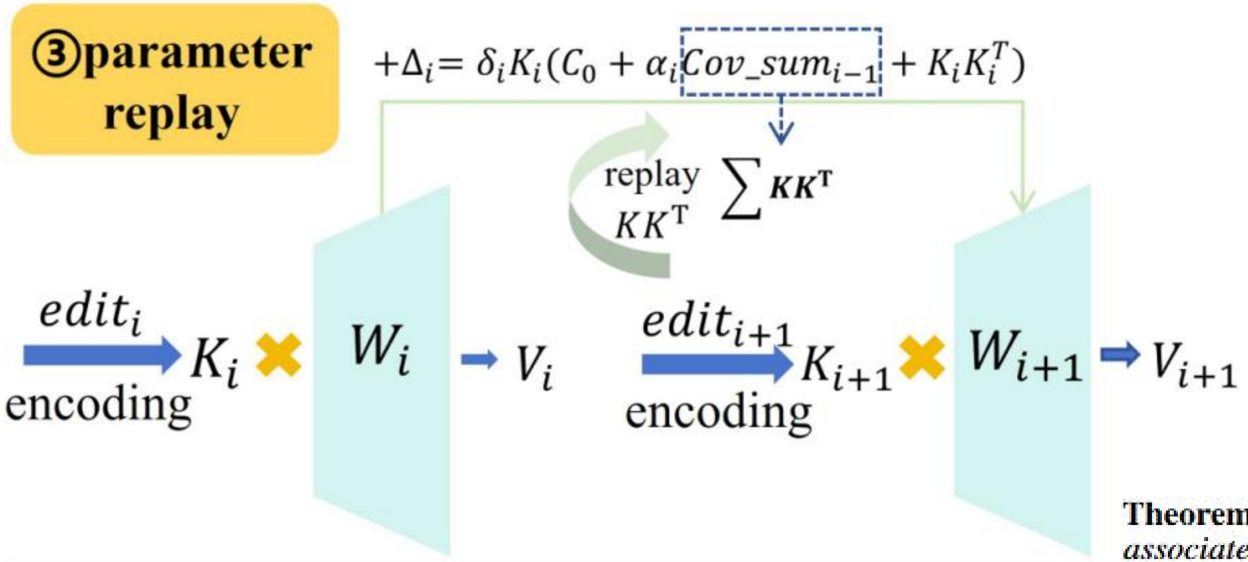
$$p(\delta|e_i) \sim \mathcal{N}\left(\delta^*_{e_i}, F_{e_i}^{-1}\right)$$

$$F_{e_i} = \mathbb{E}\left[\left(\frac{\partial \log p(\delta|e_i)}{\partial \delta}\right)\left(\frac{\partial \log p(\delta|e_i)}{\partial \delta}\right)^{\top}\Bigg|_{\delta^*_{e_i}}\right].$$

$$\mathcal{L}_\delta = \mathcal{L}_{MAF} + \frac{1}{2}\delta^T \sum_{i=1}^{n-1} (\lambda_i F_{e_i})\,\delta,$$

Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." , PNAS 2017.

- **Progressive Memory Consolidation via Parameter Replay**



Knowledge pairs $\{(K_i, V_i)\}_{i=0}^{n}$

$$W_n K_0 = V_0$$
$$W_n K_1 = V_1$$
$$\cdots \cdots$$
$$W_n K_n = V_n$$

Theorem 2. *Assume that W after the i_{th} sequential edit is denoted as W_i. The knowledge pairs associated with the i_{th} edit is represented by keys K_i and values V_i. Let $C_0 = \lambda_C K_0 K_0^T$, $\delta_i = V_i - W_{i-1} K_i$ and $Cov\_sum_{i-1} = \sum_{j=1}^{i-1} K_j K_j^T$. λ_C is hyperparameter. The convergence factor $\alpha_i = \frac{n}{i-1}$ $(i > 1)$ ensures the convergence of the sum of Δ_i and balances the degree of consolidation for different editing knowledge. Then it follows that:*

$$W_i = W_{i-1} + \Delta_i \tag{17}$$

$$\Delta_i = \delta_i K_i^T (C_0 + \alpha_i Cov\_sum_{i-1} + K_i K_i^T)^{-1}, \tag{18}$$

Theoretical Analysis

Lemma 1 ([31],Theorem 5.5). *Consider a loss function $\mathcal{L}$ such that $0 \le \mathcal{L}(p, \mathbf{y}) \le L$ and γ-Lipschitz with respect to the output distribution p and ground-truth label $\mathbf{y}$. Suppose that the Adam optimizer with stabilization constant $c \in (0,1)$ is executed for T iterations with an initial random parameters $\mathcal{R}$, training set $S = \{(x_i, y_i)_{i=1}^N\}$, batch data $B = \{(x_i, y_i)_{i=1}^b\}$ and learning rate λ to obtain $f_{B_S,\mathcal{R}}$. The empirical risk R_{emp} is defined as $R_{emp}(f_{B_S,\mathcal{R}}) = \frac{1}{b}\sum_{i=1}^b \mathcal{L}(f(x_i), y_i)$ on a finite training set. The true risk $R_{true}(f_{B_S,\mathcal{R}})$ is estimated with the empirical risk over the whole dataset that follows the distribution of the training set. The generalization error $E(f_{B_S,\mathcal{R}}) = R_{true}(f_{B_S,\mathcal{R}}) - R_{emp}(f_{B_S,\mathcal{R}})$. Then, we have the following generalization error bound with probability at least $1-\epsilon$:*

$$E(f_{B_S,\mathcal{R}}) \le \frac{2\eta}{c}\left(4\left(\frac{b\gamma}{N}\right)^2\sqrt{T\log(2/\epsilon)} + \frac{bT\gamma^2}{N}(1+\sqrt{2N\log(2/\epsilon)})\right) + L\sqrt{\frac{\log(2\epsilon)}{2N}}. \quad (28)$$

Corollary 1. *Consider LLMs are trained using the CE loss and MAF loss separately over the same training set S, batch data B_S and other settings. Denote $f_{B_S,\mathcal{R}}^{CE}, f_{B_S,\mathcal{R}}^{MAF}$ as the corresponding LLMs using CE loss and MAF loss. We have the following inequalities:*

$$E(f_{B_S,\mathcal{R}}^{MAF}) \le E(f_{B_S,\mathcal{R}}^{CE}) \quad (29)$$

This indicates that our method, after editing for specific queries, can adapt effectively to more generalized scenarios.

Theorem 2. *Assume that W after the i_{th} sequential edit is denoted as W_i. The knowledge pairs associated with the i_{th} edit is represented by keys K_i and values V_i. Let $C_0 = \lambda_C K_0 K_0^T$, $\delta_i = V_i - W_{i-1}K_i$ and $Cov\_sum_{i-1} = \sum_{j=1}^{i-1} K_j K_j^T$. λ_C is hyperparameter. The convergence factor $\alpha_i = \frac{n}{i-1}$ $(i > 1)$ ensures the convergence of the sum of Δ_i and balances the degree of consolidation for different editing knowledge. Then it follows that:*

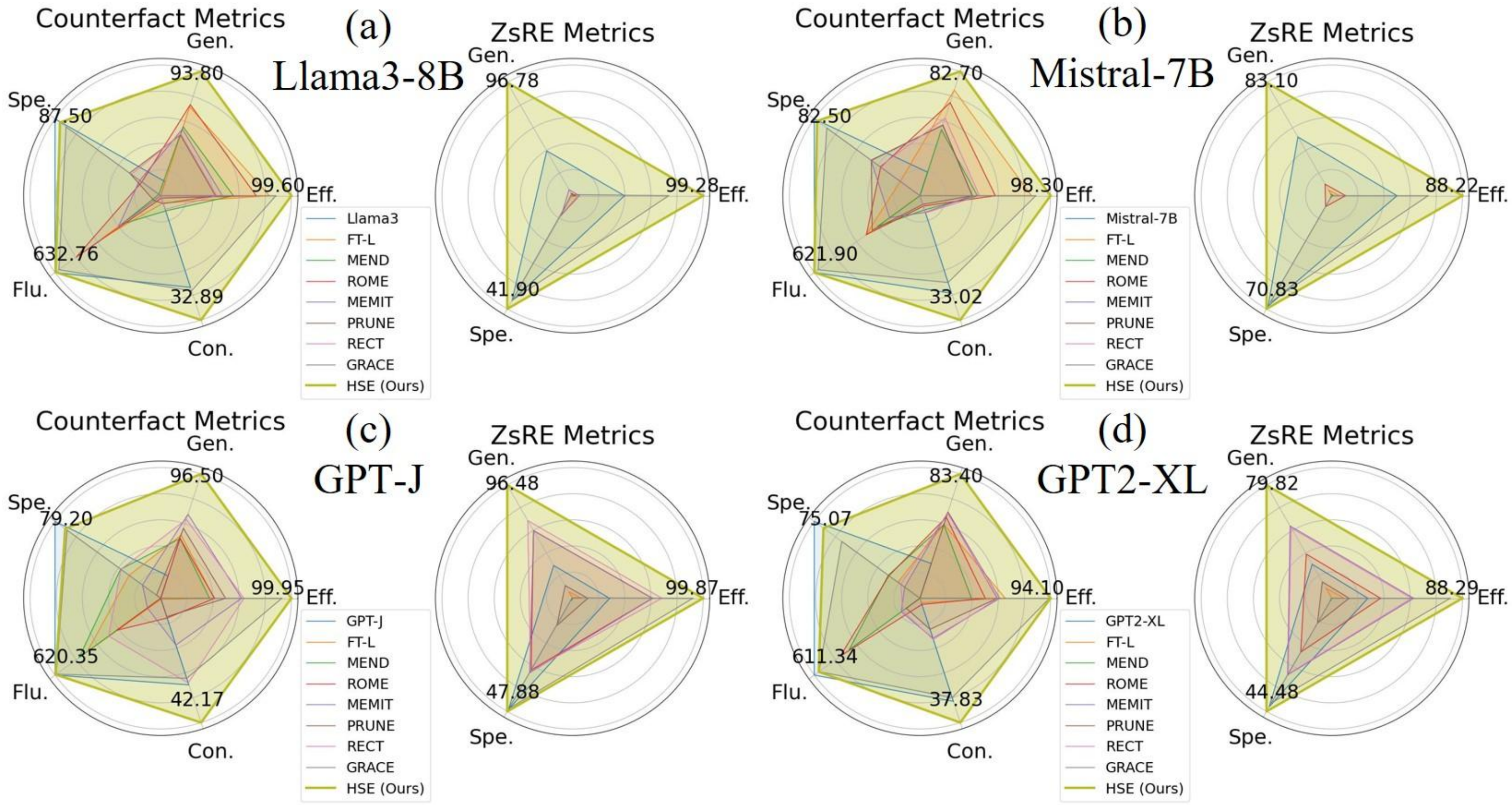
$$W_i = W_{i-1} + \Delta_i \quad (17)$$

$$\Delta_i = \delta_i K_i^T (C_0 + \alpha_i Cov\_sum_{i-1} + K_i K_i^T)^{-1}, \quad (18)$$

Corollary 2 (Convergence). *Given Δ_i as defined in Theorem 2, let $\alpha_i = \frac{n}{i-1}$ $(i > 1)$ and the minimum eigenvalue of C_0 and the maximum eigenvalue of $K_i K_i^T$ $(K_i \in \mathbb{R}^q)$ are all at least 1. Assume that for all indices $i \le q$, K_i are mutually orthogonal (practical). Then the Frobenius norm $\lim_{n\to\infty} ||W_n||_F$ converges.*

The long-term editing memory mechanism ensures parameter updates converge

- **Main Results**



For all models, most baselines suffer from catastrophic performance degradation due to model parameter collapse after multiple edits

Results

- **Compared to AlphaEdit**

AlphaEdit

$$\boldsymbol{\Delta} = \underset{\tilde{\boldsymbol{\Delta}}}{\arg\min} \left(||(\boldsymbol{W} + \tilde{\boldsymbol{\Delta}}\boldsymbol{P})\boldsymbol{K}_1 - \boldsymbol{V}_1||^2 + ||\tilde{\boldsymbol{\Delta}}\boldsymbol{P}||^2 + ||\tilde{\boldsymbol{\Delta}}\boldsymbol{P}\boldsymbol{K}_p||^2 \right).$$

$$\boldsymbol{\Delta}_{\textbf{AlphaEdit}} = \boldsymbol{R}\boldsymbol{K}_1^T\boldsymbol{P}(\boldsymbol{K}_p\boldsymbol{K}_p^T\boldsymbol{P} + \boldsymbol{K}_1\boldsymbol{K}_1^T\boldsymbol{P} + \boldsymbol{I})^{-1}$$

P is not necessary, but $K_pK_p^T$ is necessary

HSE

$$\Delta_n K_0 = 0,\quad \Delta_n K_{pre} = 0,\quad (W_{n-1} + \Delta_n)K_n = V_n,$$

$$\Delta_i = \delta_i K_i^T (C_0 + \alpha_i Cov\_sum_{i-1} + K_i K_i^T)^{-1},\quad W_i = W_{i-1} + \Delta_i,$$

Satisfy orthogonality

$$W_nK_0 = V_0$$
$$W_nK_1 = V_1$$
$$\cdots\ \cdots$$
$$W_nK_n = V_n$$

$$\widetilde{\Delta}_n(K_0K_0^T + \cdots + K_nK_n^T) = \sum_{i=1}^{n}(V_iK_i^T - W_{n-1}K_iK_i^T),$$

$$\widetilde{\Delta}_n = (\sum_{i=1}^{n}\delta_{n-1}^iK_i^T)(C_0 + Cov\_sum_{n-1} + K_nK_n^T)^{-1},$$

$$= \Delta_n + (\sum_{i=1}^{n-1}\delta_{n-1}^iK_i^T)(C_0 + Cov\_sum_{n-1} + K_nK_n^T)^{-1},$$

Satisfy convergence and more accurate

Assumption

$$W_{n-1}K_0 = V_0,\quad W_{n-1}K_1 = V_1,\quad \ldots,\quad W_{n-1}K_{n-1} = V_{n-1}$$

$$\delta_{n-1}^i \triangleq V_i - W_{n-1}K_i$$

$$\delta_{n-1}^i = 0$$

introduce $\alpha_i = \frac{n}{i-1}$ resulting in δ_{n-1}^i approaches 0 more closely

Table 1: Impact of Different range of α_i $(i > 1)$ on sequential editing performance across 1000 samples using the HSE Method. The best performance is highlighted in **bold**.

Range of α_i	Counterfact					ZsRE		
	Efficacy↑	Generalization↑	Specificity↑	Fluency↑	Consistency↑	Efficacy↑	Generalization↑	Specificity↑
AlphaEdit	98.20±0.74	91.17±0.63	62.15±0.41	622.14±1.42	32.40±0.29	95.60±0.87	93.14±0.91	40.05±0.35
w/o α_i	98.62±0.69	92.73±0.82	76.08±0.53	624.49±0.76	32.35±0.33	97.60±0.74	95.13±0.68	39.12±0.42
$\alpha_i = n/(i-1)$ (Ours)	**99.60**±0.37	**93.80**±0.51	87.50±0.84	**632.76**±0.83	**32.89**±0.21	**99.28**±0.65	**96.78**±0.49	**41.90**±0.31
$\alpha_i = n/2(i-1)$	99.20±0.48	92.82±0.93	81.75±0.62	626.31±1.58	32.05±0.27	98.80±0.59	96.01±0.76	38.42±0.45
$\alpha_i = 2n/(i-1)$	95.40±0.91	88.30±0.77	**88.72**±0.56	630.76±1.29	31.40±0.39	96.90±0.43	95.39±0.63	41.05±0.28
$\alpha_i = n-i+2$	96.24±0.83	89.68±0.55	87.90±0.39	629.49±1.34	32.12±0.19	95.24±0.72	94.68±0.81	41.42±0.23

Fang, Junfeng, et al. "Alphaedit: Null-space constrained knowledge editing for language models." , ICLR 2025.

- **Downstream Evaluation**

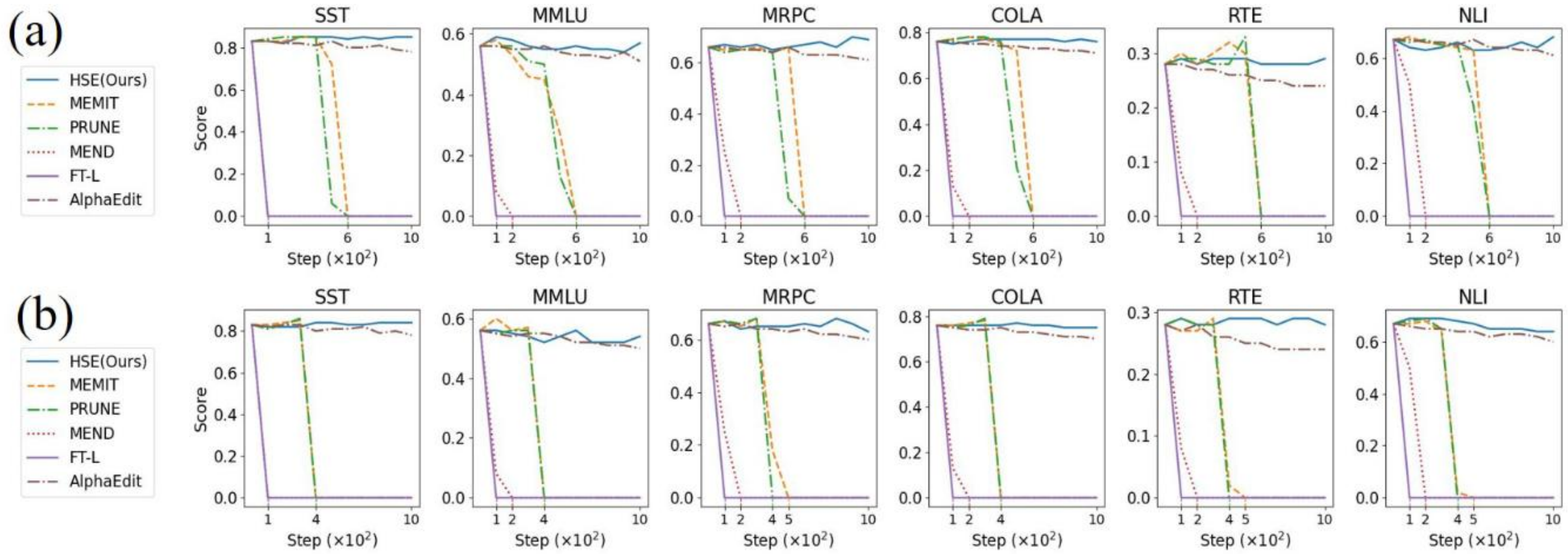


Figure 3: The general capabilities of the Llama3 model on the six tasks of the GLUE benchmark after editing with the CounterFact **(a)** and ZsRE **(b)** datasets respectively.

HSE **minimally impacts the model's general capability** and can lead to slight enhancements when guiding the model with correctly edited knowledge.

• Ablation Study

Table 2: **Ablation study results of the HSE method.** This table presents the ablation study results for the HSE method, detailing the contributions of individual components. **AF:** Active Forgetting module, responsible for actively forgetting specific information. **FIM:** Fisher information matrix module, controlling parameter updates to preserve important knowledge. **LEM:** Long-term Editing Memory module, reinforcing edited knowledge and preventing parameter proliferation. **ER:** Experience Replay module, used in continual learning to mitigate catastrophic forgetting by replaying a subset of data.

Edit Mode	Counterfact					ZsRE		
	Efficacy↑	**Generalization↑**	**Specificity↑**	**Fluency↑**	**Consistency↑**	**Efficacy↑**	**Generalization↑**	**Specificity↑**
HSE (Ours)	**99.60** ±0.37	**93.80** ±0.51	**87.50** ±0.84	**632.76** ±0.43	**32.89** ±0.21	**99.28** ±0.65	**96.78** ±0.49	**41.90** ±0.31
w/o AF	96.20 ±0.48	90.15 ±0.62	86.40 ±0.73	628.19 ±1.58	30.85 ±0.27	96.25 ±0.59	94.23 ±0.76	41.06 ±0.45
w/o FIM	98.10 ±0.91	92.04 ±0.43	82.10 ±0.63	624.05 ±0.89	31.06 ±0.39	99.02 ±0.28	95.14 ±0.63	40.10 ±0.23
w/o LEM	60.85 ±0.83	55.62 ±0.55	53.18 ±0.39	362.85 ±1.24	4.53 ±0.19	10.05 ±0.72	6.21 ±0.81	9.20 ±0.23
w/o LEM, w ER	81.26 ±0.48	73.50 ±0.93	76.10 ±0.62	518.62 ±1.08	14.29 ±0.27	42.50 ±0.43	38.72 ±0.63	26.14 ±0.28

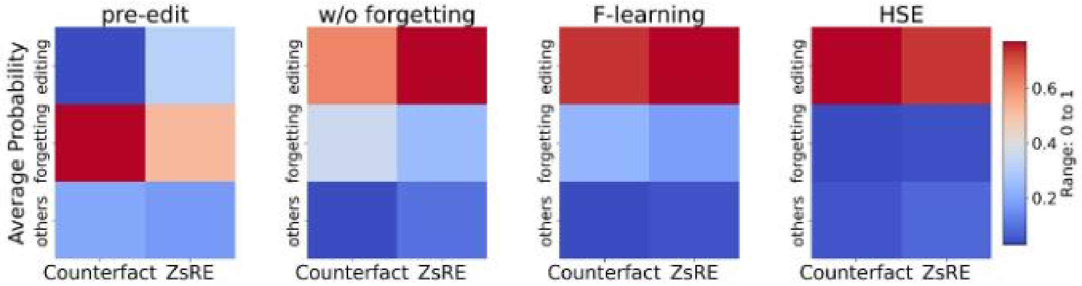


Figure 7: Visualization of the average probability of generated tokens in pre-edit, w/o forgetting, F-learning [47] and HSE conditions.

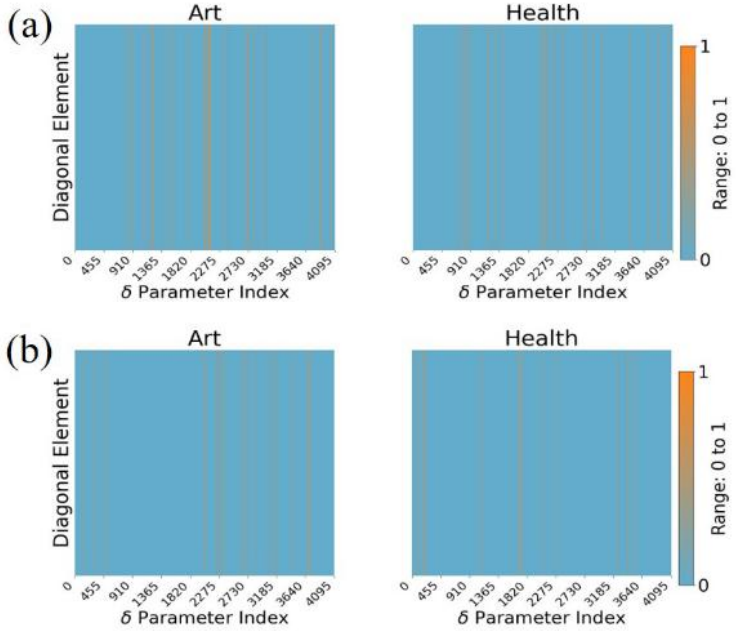


Figure 8: Visualization of the top 100 values of the Fisher information matrix diagonal elements for the δ parameters under the **(a)** HSE method and **(b)** without Fisher information matrix constraints, respectively.

Results

- **Ablation Study**

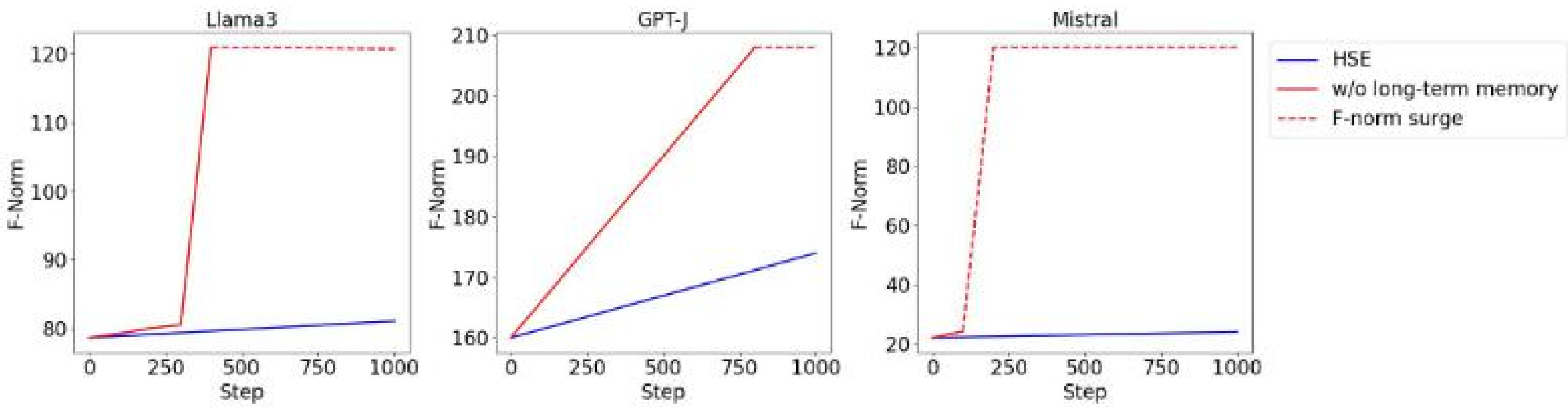


Figure 9: Line chart showing the changes in F-Norm values for the HSE method and without Long-term editing memory.

For our HSE method, **the F-norm of edited parameters increases much more gradually**

The larger the F-norm of the original LLMs, the more "resistant to editing" they become, allowing them to maintain their general capabilities even more editing iterations.

Results

- **Practical Application**

Societal bias reduction

Figure 6: Heatmap illustrating the performance comparison of various methods on the SafeEdit dataset. The notation "w/o F" indicates that no forgetting mechanism was applied to the harmful data instances.

Hallucination mitigation

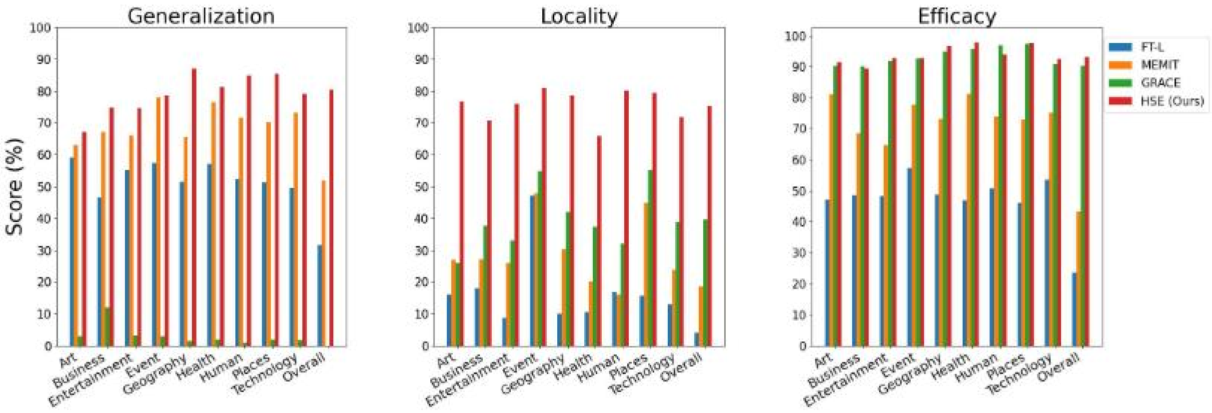


Figure 4: Performance comparison results of our proposed HSE method on the Llama3 across 9 domains of the HalluEdit dataset.

Healthcare knowledge injecting

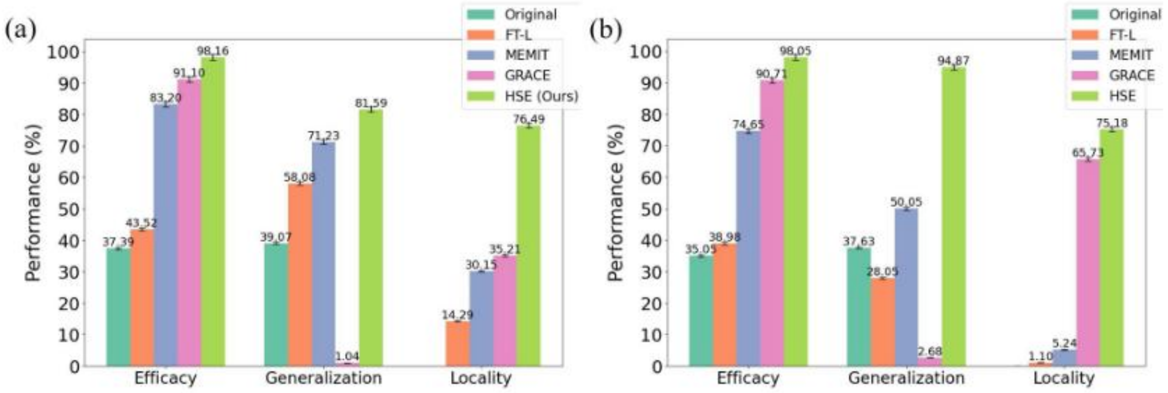


Figure 5: Comparison of Editing Performance for the healthcare LLMs Llama3 Aloe-8B-Alpha (a) and OpenBioLLM-8B (b) on the Health Domain of the HalluEdit Dataset.

Results

- **Sequential Editing Compared to Full-batch Editing**

Table 3: Comparison of training time across different methods and sample sizes.

samples	HSE(Ours)	MEMIT_full	w/o FIM	w/o Active Forget
10	1min	3min	1min	1min
100	10.5min	65min	10min	11min
1000	101min	850min	97min	100min
3000	251min	/	242min	252min

the time complexity of MEMIT approaches **n times** that of HSE sequential editing.

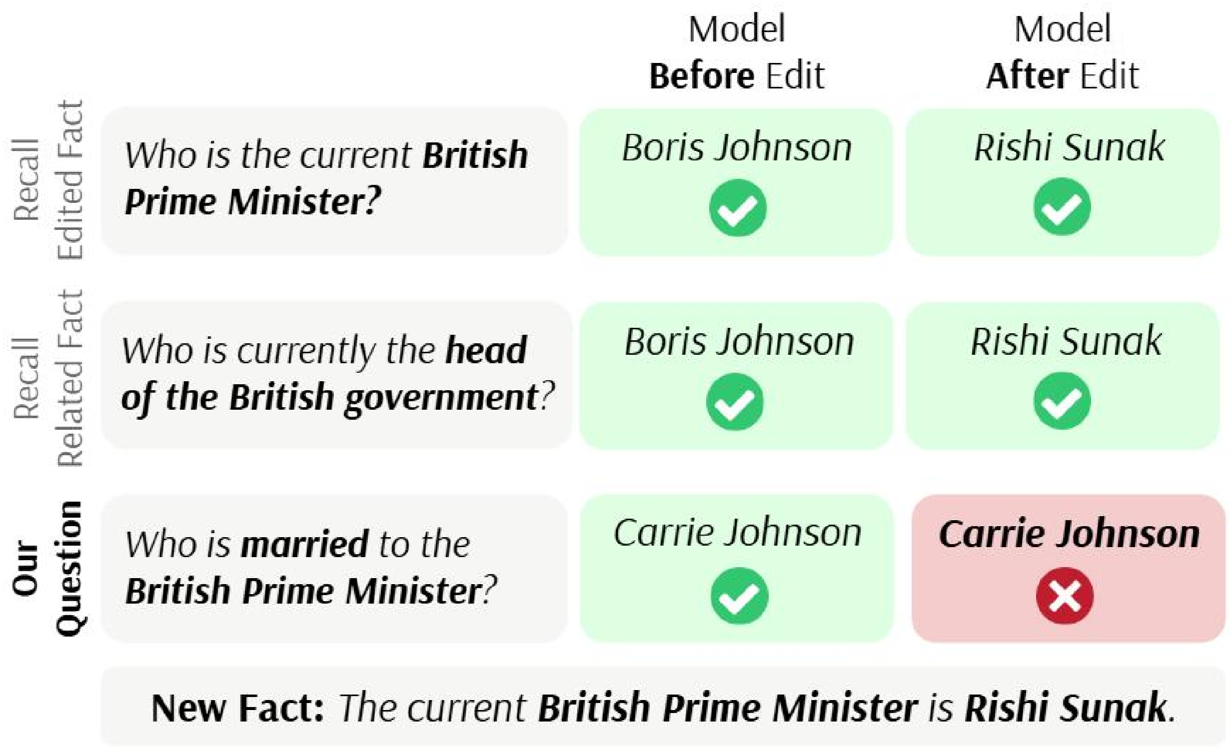
Table 4: Comparison of Time Complexities: One-edit refers to a single edit operation, while n-edit refers to editing n times. Performance Comparison of Editing 1,000 and 10,000 Counterfact samples using HSE (Sequential Editing) vs $\text{MEMIT}_{\text{full}}$ (Full-Batch Editing) in Llama3-8B The best performance is highlighted in **bold**.

One-edit Time Complexity	HSE: $\mathcal{O}(q^2b+q^3)$			n-edit Time Complexity	HSE: $\mathcal{O}(n)$		
	$\text{MEMIT}_{\text{full}}$: $\mathcal{O}(q^2b+q^3)$				$\text{MEMIT}_{\text{full}}$: $\mathcal{O}(n^2)$		
Editing mode	1,000 samples			Editing method	10,000 samples		
	Efficacy↑	Generalization↑	Specificity↑		Efficacy↑	Generalization↑	Specificity↑
HSE (1 batch_size)	**99.60** $_{\pm 0.37}$	**93.80** $_{\pm 0.82}$	87.50 $_{\pm 0.51}$	HSE (10 batch_size)	**98.02** $_{\pm 0.45}$	**82.48** $_{\pm 0.93}$	83.72 $_{\pm 0.62}$
HSE (10 batch_size)	99.23 $_{\pm 0.29}$	90.14 $_{\pm 0.71}$	87.79 $_{\pm 0.43}$	HSE (100 batch_size)	97.84 $_{\pm 0.38}$	80.04 $_{\pm 0.84}$	84.18 $_{\pm 0.55}$
HSE (100 batch_size)	98.92 $_{\pm 0.21}$	87.72 $_{\pm 0.66}$	88.31 $_{\pm 0.39}$	HSE (1,000 batch_size)	96.21 $_{\pm 0.41}$	77.55 $_{\pm 0.77}$	84.66 $_{\pm 0.49}$
HSE (1,000 batch_size)	98.50 $_{\pm 0.18}$	84.30 $_{\pm 0.59}$	**88.50** $_{\pm 0.33}$	HSE (10,000 batch_size)	97.50 $_{\pm 0.27}$	80.20 $_{\pm 0.68}$	**85.15** $_{\pm 0.24}$
$\text{MEMIT}_{\text{full}}$ (1,000 batch_size)	97.50 $_{\pm 0.23}$	81.02 $_{\pm 0.53}$	85.24 $_{\pm 0.47}$	$\text{MEMIT}_{\text{full}}$ (10,000 batch_size)	95.10 $_{\pm 0.32}$	76.42 $_{\pm 0.61}$	81.15 $_{\pm 0.36}$

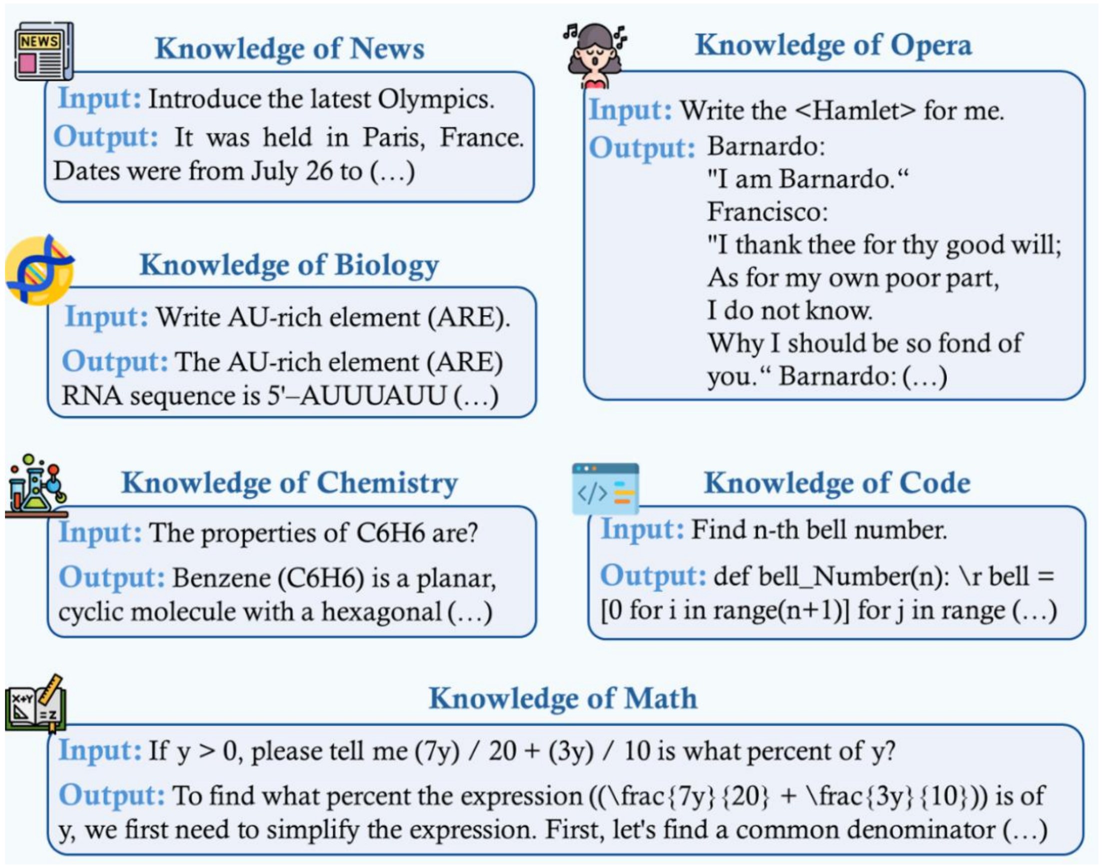
the robustness and adaptability of HSE in handling **varying data volumes and batch sizes**

- **Future Works**

Multi-hop knowledge editing



Unstructured knowledge editing



Jiang, Houcheng, et al. "AnyEdit: Edit Any Knowledge Encoded in Language Models.", ICML 2025.

Thanks for listening