Towards Lifelong Model Editing via Simulating Ideal Editor

Yaming Guo [1] **Siyang Guo** [2] **Hengshu Zhu** [3 4] **Ying Sun** [1]

Abstract

Model editing plays a crucial role in the cost-effective development of large language models, and the challenge of evolving knowledge facilitates its sequential extension, namely lifelong model editing. However, progress on standard and lifelong editing has historically followed separate tracks, overlooking the potential of generalizing standard methods to lifelong scenarios. By establishing this bridge, we can provide robust baselines in lifelong scenarios and ensure that lifelong editing benefits from the ongoing advancements in standard editing technologies. In response, this paper proposes a general framework, ***Sim**ulating **I**deal **E**ditor* (SimIE), which restores the strong performance of parameter-modifying methods from standard model editing in a lifelong context. SimIE formulates the ideal parameter shift as the minimum-norm solution to a linear system, constructed using the Moore-Penrose inverse, and subsequently enables recursive updates by truncating the limiting expression of the Moore-Penrose inverse under two mild assumptions. Theoretically, we demonstrate that if either assumption is not met, the solution provided by SimIE remains near-optimal in a statistical sense or stable against perturbations introduced by the sequential editing, but a trade-off between optimality and stability arises when both assumptions fail. Extensive experiments validate the effectiveness of SimIE, which allows standard algorithms to achieve performance comparable to specialized lifelong model editing methods. Our code is available at SimIE.

[1]Artificial Intelligence Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China [2]School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China [3]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China [4]University of Chinese Academy of Sciences, Beijing, China . Correspondence to: Ying Sun <yings@hkust-gz.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

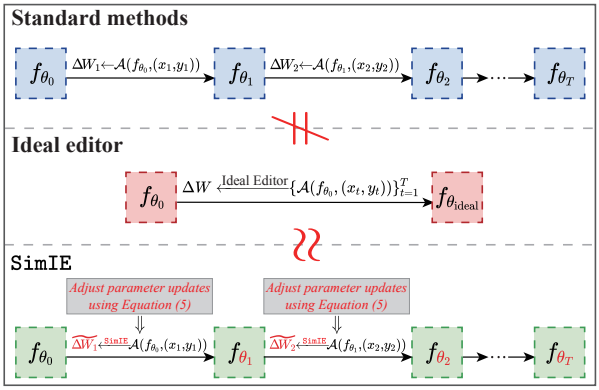


Figure 1. Illustration of SimIE, which enables the post-edit model to closely approximate the ideal state achieved by the ideal editor.

1. Introduction

Large language models (LLMs) acquire extensive knowledge across various domains during pre-training, which may include outdated or harmful information (Petroni et al., 2019; Zhao et al., 2023). Given the prohibitively high costs of re-training and fine-tuning, **model editing** has emerged as an efficient approach for updating and correcting specific knowledge within LLMs (Sinitsin et al., 2020; De Cao et al., 2021). The goal of model editing is to adjust the LLM's predictions for specific inputs to match the desired outputs, based on the provided edit examples, while preserving its behavior on unrelated inputs. Existing model editing techniques generally fall into two main categories: **parameter-modifying** and parameter-preserving (Yao et al., 2023). This work focuses on parameter-modifying approaches.

To accommodate evolving knowledge, the extension of model editing—**lifelong model editing**—is proposed, which involves performing sequential edits (Yao et al., 2023). Lifelong editing presents two inherent challenges: 1) The sequential updates gradually deviate the model from its initial state, leading to degradation in its general abilities and editability (Ma et al., 2024; Gupta et al., 2024b). 2) The model typically encounters catastrophic forgetting during sequential edits, where previously edited examples are forgotten (Wang et al., 2024a; Gupta et al., 2024a). Consequently, preliminary efforts have been made to tackle this more challenging scenario (Huang et al., 2023; Hartvigsen

et al., 2024; Wang et al., 2024b). One prevalent perspective among these works is that standard methods are unsuitable for lifelong editing scenarios, as they suffer significant performance degradation as the sequence of edits lengthens.

Contrary to popular belief, we study the feasibility of a general solution for generalizing standard methods to lifelong scenarios, thereby bridging these two paradigms. This connection offers dual advantages. First, the effectiveness of standard methods has been widely validated, allowing us to readily provide robust baselines in lifelong contexts without developing specialized approaches from scratch. More importantly, this integration alleviates the challenges specific to lifelong editing, redirecting the community's focus to enhancing fundamental editors, whose advances naturally extend to lifelong scenarios. As an elementary exploration of this promising direction, the core question to ask is:

Is it possible to restore the strong performance of standard methods in the context of lifelong model editing?

In response, this paper proposes a general framework called ***Sim**ulating **I**deal **E**ditor* (`SimIE`). Given a standard model editing algorithm $\mathcal{A}$, we introduce its **ideal editor**, defined as the optimal solution that edits the initial model using all examples from the sequential stream simultaneously. Mathematically, the ideal editor is formalized as the minimum-norm solution to a linear system, which is constructed using the Moore-Penrose inverse. Under two mild assumptions, namely over-parameterization and key-value invariance, `SimIE` enables recursive updates by truncating the limiting expression of the Moore-Penrose inverse. As illustrated in Figure 1, `SimIE` only adjusts the parameter updates obtained from $\mathcal{A}$ at each time step, enabling the post-edit model to closely approximate the ideal state achieved by the ideal editor. Our main contributions are summarized as follows:

1. We propose `SimIE`, a general framework designed to restore the strong performance of standard model editing methods (applicable to all parameter-modifying algorithms) in a lifelong context, bridging the gap between the two paradigms for the first time (Section 3).
2. We theoretically analyze the behavior of the proposed framework when assumptions are not met, revealing that: 1) When the over-parameterization assumption is not satisfied, the solution remains near-optimal in a statistical sense. 2) When the key-value invariance assumption does not hold, the solution is stable against perturbations introduced by sequential editing. 3) If both assumptions fail, the solution requires a trade-off between optimality and stability (Section 4).
3. We validate the effectiveness of `SimIE` on multiple LLMs, such as GPT2-XL (Radford et al., 2019), Llama-2 (Touvron et al., 2023), and Mistral (Chaplot, 2023). The results demonstrate that `SimIE` allows standard algorithms to achieve performance comparable to specialized lifelong methods, requiring just one line of code for implementation (Section 5).

2. Preliminaries

Standard model editing. Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. We consider the model $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ. Given an edit example (x_e, y_e), the goal of model editing is to adjust the model's prediction at x_e to match y_e, while preserving its behavior on unrelated inputs. Formally, for an edit example (x_e, y_e) and the original model $f_\theta(\cdot)$, the post-edit model $f_{\theta'}(\cdot)$ should satisfy:

$$f_{\theta'}(x) = \begin{cases} y_e, & \text{if } x = x_e; \\ f_\theta(x), & \text{otherwise.} \end{cases} \tag{1}$$

Although Equation (1) considers only a single edit example, recent works have scaled to handle thousands of examples, achieving commendable performance (Mitchell et al., 2022b; Meng et al., 2023; Tan et al., 2024).

Lifelong model editing. To accommodate evolving knowledge, lifelong model editing has been proposed, which involves performing sequential edits on LLMs. Let $\mathcal{D}_{\text{edit}} = \{(x_t, y_t) \mid t \in [T]\}$ denote the dataset of all edit examples, and $\mathcal{X}_{\text{edit}} = \{x_t \mid t \in [T]\}$ represent the set of corresponding inputs, where $T := \{1, \ldots, T\}$. We denote the initial model by $f_{\theta_0}(\cdot)$. At time step t, the model editing algorithm $\mathcal{A}$ generates a parameter update based on the previous model $f_{\theta_{t-1}}(\cdot)$ and the current edit example (x_t, y_t), resulting in the following parameter update rule:

$$\theta_t = \theta_{t-1} + \mathcal{A}(f_{\theta_{t-1}}, (x_t, y_t)).$$

The final post-edit model, $f_{\theta_T}(\cdot)$, should satisfy $f_{\theta_T}(x_t) = y_t$ $(\forall t \in [T])$ and $f_{\theta_T}(x) = f_{\theta_0}(x)$ $(\forall x \notin \mathcal{X}_{\text{edit}})$. As mentioned earlier, due to the inherent challenges of lifelong editing, naively adapting standard approaches to lifelong scenarios can lead to significant performance degradation (Yao et al., 2023). In Section 3, we introduce a general framework designed to restore the strong performance of standard methods in a lifelong context.

Editing in key-value memory. Through causal tracing, Meng et al. (2022) identify that the MLP modules in the Transformer (Vaswani, 2017), comprising two linear layers with a non-linear activation, are critical for storing factual knowledge. Building on this insight, most works have focused on editing the parameters of MLPs, demonstrating that such edits are consistently effective for updating knowledge within LLMs (Meng et al., 2023; Tan et al., 2024). These studies model the linear layer of MLPs, denoted by $W \in \mathbb{R}^{d_2 \times d_1}$, as a key-value memory that maps an input k

to an output $v = Wk$, storing knowledge in the form of key-value pairs (k, v). Supposing that the edit example (x_e, y_e) corresponds to the key k_e[1] at W, the editing method generates a parameter update $\Delta W_e \in \mathbb{R}^{d_2 \times d_1}$, thereby inducing a change in value:

$$v = Wk_e \Longrightarrow v_e = (W + \Delta W_e)k_e,$$

where v represents the initial value and v_e is the desired value. Utilizing the updated key-value pair (k_e, v_e), the model editing method efficiently incorporates new knowledge (x_e, y_e) into LLMs.

3. Simulating Ideal Editor in Lifelong Context

In Section 3.1, we define the ideal editor and introduce two mild assumptions: over-parameterization and key-value invariance. With these assumptions, Section 3.2 presents `SimIE`, which approximates the ideal editor through recursive updates. Section 3.3 extends `SimIE` to multi-layer editing settings, accompanied by a cost analysis.

3.1. Introduced Ideal editor

Standard model editing methods can inject knowledge from numerous edit examples into LLMs, yet they suffer significant performance degradation in lifelong contexts. We argue that robust performance in standard scenarios is based on two key principles: 1) The model remains in its initial state, which avoids the degradation in editability caused by sequential editing (Gupta et al., 2024b). 2) All edit examples are available concurrently, mitigating catastrophic forgetting during sequential updates (Wang et al., 2024a). Therefore, given a standard model editing method, an ideal approach would involve editing the initial model by utilizing all examples in the sequential stream simultaneously.

To maintain clarity, we first focus on editing a single linear layer, initialized as $W_0 \in \mathbb{R}^{d_2 \times d_1}$, which is a component of the full model parameters θ_0. Using the standard model editing algorithm $\mathcal{A}$, the **ideal editor** aims to modify the parameter W_0 in order to simultaneously inject knowledge of all examples within $\mathcal{D}_{\text{edit}}$, defined as:

Definition 3.1 (Ideal Editor). For an edit example $(x_t, y_t) \in \mathcal{D}_{\text{edit}}$, let k_t denote the key, and define the change in value as $b_t = \Delta W_t^0 k_t$, where $\Delta W_t^0 := \mathcal{A}(f_{\theta_0}, (x_t, y_t))$ represents the parameter update for W_0 generated by algorithm $\mathcal{A}$. The ideal editor w.r.t. $\mathcal{A}$ is the optimal parameter shift S that solves the following optimization problem:

$$\begin{aligned} &\min \|S\|_{\mathrm{F}}^2, \\ &\text{s.t.} \quad Sk_t = b_t, \quad t = 1, 2, \ldots, T. \end{aligned} \tag{2}$$

[1]For notational simplicity, we assume each example consists of a single token mapped to one key, though our analysis naturally generalizes to cases with multi-token examples.

The solution S defined by the ideal editor exhibits two fundamental properties. First, when applied to the edit examples (x_t, y_t), S produces an equivalent transformation with the individual updates ΔW_t^0. Second, S achieves a minimal Frobenius norm among all solutions that satisfy the constraints. By defining the concatenated matrices $K := [k_1 \mid k_2 \mid \cdots \mid k_T] \in \mathbb{R}^{d_1 \times T}$ and $B := [b_1 \mid b_2 \mid \cdots \mid b_T] \in \mathbb{R}^{d_2 \times T}$, we can characterize the ideal editor as the minimum-norm solution to the linear system $SK = B$.

Thus far, we have not established the existence of the ideal editor. To derive this fundamental result, we require an over-parameterization assumption, stated as follows:

Assumption 3.2 (Over-parameterization). The number of edit examples does not exceed the column dimension of the weight matrix W, i.e., $T \leq d_1$, and the key matrix K is full rank, i.e., $\operatorname{rank}(K) = T$.

This assumption ensures that W has sufficient capacity and that the key vectors $\{k_i\}_{i=1}^T$ are linearly independent. Under Assumption 3.2, we can prove that the linear system $SK = B$ admits at least one solution, thus guaranteeing the existence of the ideal editor (formalized in Appendix B.1).

To enable the approximation of the ideal editor in lifelong scenarios, we introduce the key-value invariance assumption, described as follows:

Assumption 3.3 (Key-value invariant). Given an algorithm $\mathcal{A}$, the updated key-value pair (k_t, v_t) w.r.t. the edit example (x_t, y_t) remains invariant across different model states, i.e., $(k_t', v_t') = (k_t'', v_t'')$, where (k_t', v_t') and (k_t'', v_t'') are derived from the models $f_{\theta'}$ and $f_{\theta''}$, respectively.

Intuitively, this assumption implies that the representation of specific knowledge within the LLM remains invariant. Under Assumption 3.3, it is easy to verify that for any model state $f_{\theta'}$, the following relationship holds:

$$\left(W' + \mathcal{A}\left(f_{\theta'}, (x_t, y_t)\right) - W_0\right) k_t = b_t = \Delta W_t^0 k_t. \tag{3}$$

Equation (3) indicates that the change in the updated value relative to the initial value, $W_0 k_t$, is consistently captured by b_t, regardless of the model state.

It is worth noting that Assumption 3.2 and Assumption 3.3 may not hold in certain complex editing settings. Nevertheless, we will provide a detailed theoretical analysis of the case where assumptions are not met (Section 4) and find that they do not lead to empirical failure (Section 5).

3.2. The proposed `SimIE`

The ideal editor provides the optimal parameter shift for injecting knowledge from $\mathcal{D}_{\text{edit}}$ into the model f_{θ_0}. Therefore, a natural approach is to approximate the solution defined by the ideal editor in the lifelong model editing scenarios.

To begin, we introduce the Moore-Penrose inverse (Penrose, 1955), which generalizes the concept of the matrix inverse. It is defined as follows:

Definition 3.4 (Moore-Penrose inverse). Given a matrix $A \in \mathbb{R}^{d_2 \times d_1}$ with Singular Value Decomposition (SVD) $A = U\Sigma V^\top$, the Moore-Penrose inverse of A, denoted by $A^\dagger$, can be written as:

$$A^\dagger = V\Sigma^\dagger U^\top,$$

where $\Sigma^\dagger \in \mathbb{R}^{d_2 \times d_1}$ is obtained by taking the reciprocal of each non-zero singular value in the diagonal matrix Σ and transposing the resulting matrix.

Under Assumption 3.2, the ideal editor exists and is unique, which can be expressed as $S^0 = BK^\dagger = B(K^\top K)^{-1}K^\top$ (see Appendix B.1). However, exact computation of S^0 is not feasible unless all key-value pairs $\{(k_t, v_t)\}_{t=1}^T$ are available simultaneously. Instead, we can give an approximate solution through the following lemma:

Lemma 3.5 (Theorem 4.4. in Laub (2004))**.** *For a matrix $A \in \mathbb{R}^{d_2 \times d_1}$, we have:*

$$A^\dagger = \lim_{\alpha \to 0^+} A^\top (AA^\top + \alpha I)^{-1},$$

where $I \in \mathbb{R}^{d_1 \times d_1}$ is the identity matrix.

This lemma provides a limiting expression for $K^\dagger$, even in cases where $(KK^\top)^{-1}$ does not exist. We propose choosing a hyperparameter λ to truncate the limit, resulting in the approximation $S^0 \approx S_T^\lambda := BK^\top(KK^\top + \lambda I)^{-1}$.

Let $K_{:t} := [k_1 \mid k_2 \mid \cdots \mid k_t]$ and $B_{:t} := [b_1 \mid b_2 \mid \cdots \mid b_t]$ denote the matrices formed by the first t columns of K and B, respectively. If $S_{t-1}^\lambda = B_{:t-1}K_{:t-1}^\top(K_{:t-1}K_{:t-1}^\top + \lambda I)^{-1}$ holds at time step t, then, under Assumption 3.3, we can derive the following recurrence relation for S_t^λ:

$$\begin{aligned}
S_t^\lambda &= (B_{:t-1}K_{:t-1}^\top + b_t k_t^\top)(\underbrace{K_{:t}K_{:t}^\top + \lambda I}_{\text{denoted as } P_t})^{-1} \\
&= \underbrace{B_{:t-1}K_{:t-1}^\top P_{t-1}^{-1}}_{=S_{t-1}^\lambda} P_{t-1}P_t^{-1} + b_t k_t^\top P_t^{-1} \\
&= S_{t-1}^\lambda \underbrace{P_{t-1}}_{=P_t - k_t k_t^\top} P_t^{-1} + b_t k_t^\top P_t^{-1} \\
&= S_{t-1}^\lambda + (\underbrace{b_t}_{=(S_{t-1}^\lambda + \Delta W_t)k_t} - S_{t-1}^\lambda k_t) k_t^\top P_t^{-1} \\
&= S_{t-1}^\lambda + \Delta W_t k_t k_t^\top P_t^{-1},
\end{aligned} \tag{4}$$

where $\Delta W_t := \mathcal{A}(f_{\theta_{t-1}}, (x_t, y_t))$ denotes the parameter update generated by the standard model editing algorithm $\mathcal{A}$ at time step t. The detailed derivation of Equation (4) can be found in Appendix B.2. The Equation (4) reveals that S_t^λ can be computed recursively from S_{t-1}^λ, enabling S_T^λ to be obtained in a lifelong context.

Building on the above insight, we propose ***Sim**ulating **I**deal **E**ditor* (`SimIE`), which generates the parameter shift S_T^λ as an approximation of the ideal editor S^0 through a recursive update procedure. Starting with the initial parameter $W_0 \in \mathbb{R}^{d_2 \times d_1}$ and $P_0 = \lambda I \in \mathbb{R}^{d_1 \times d_1}$, where $\lambda > 0$ is a hyperparameter, the update rule for `SimIE` is given by:

$$\begin{aligned}
P_t &= P_{t-1} + k_t k_t^\top \\
W_t &= W_{t-1} + \Delta W_t k_t k_t^\top P_t^{-1}.
\end{aligned} \tag{5}$$

`SimIE` uses additional memory to maintain a matrix P_t, which is employed to adjust the parameter update $\Delta W_t := \mathcal{A}(f_{\theta_{t-1}}, (x_t, y_t))$ generated by the standard model editing algorithm $\mathcal{A}$ at time step t. The following theorem estimates a bound on the difference between the parameter shift generated by `SimIE` and that of the ideal editor:

Theorem 3.6. *Let $S_T^\lambda = W_T - W_0$ denote the parameter shift generated by `SimIE` according to Equation (5), and let S^0 represent the ideal editor w.r.t. Equation (2). Under Assumptions 3.2 and 3.3, the following bound holds:*

$$\frac{\lambda}{\sigma_{\max}^2 + \lambda} \le \frac{\|S_T^\lambda - S^0\|_{\mathrm{F}}}{\|S^0\|_{\mathrm{F}}} \le \frac{\lambda}{\sigma_{\min}^2 + \lambda},$$

where $0 < \sigma_{\min} \le \sigma_{\max}$ are the smallest and largest singular values of the key matrix K, respectively.

This theorem is proved in Appendix B.3. As demonstrated in Theorem 3.6, the approximation error bound for $S_T^\lambda - S^0$ is governed by the largest and smallest singular values of the key matrix K as well as the hyperparameter λ. In particular, as λ approaches 0, S_T^λ is guaranteed to approximate S^0. In other words, by adjusting the parameter updates obtained from the standard algorithm $\mathcal{A}$ at each step, `SimIE` ensures that the post-edit model closely approximates the ideal state achieved by the ideal editor, which edits the initial model using all examples simultaneously.

3.3. Editing on multiple layers

Existing standard model editing methods typically update parameters across multiple linear layers to enhance the robustness of editing at scale. To extend `SimIE` to multi-layer editing, we apply it independently to each layer. While this straightforward implementation ignores inter-layer dependencies, as similarly done by Tan et al. (2024), empirical evidence supports its effectiveness in practice.

The pseudo-code for `SimIE` is summarized in Algorithm 1. Let $\mathcal{L}$ denote the set of layer indices targeted for modification, as determined by the standard model editing algorithm $\mathcal{A}$. For each linear layer $l \in \mathcal{L}$, we initialize the matrix $P_0^{(l)} = \lambda I \in \mathbb{R}^{d_1^{(l)} \times d_1^{(l)}}$ according to the dimensions of the weight matrix $W^{(\ell)} \in \mathbb{R}^{d_2^{(l)} \times d_1^{(l)}}$, where $\lambda > 0$ is a hyper-

Algorithm 1 SimIE (red) and Naive approach (blue)

Input: algorithm $\mathcal{A}$, dataset $\mathcal{D}_{\text{edit}}$, initial model $f_{\theta_0}(\cdot)$
Initialize $P_0^{(l)} = \lambda I \in \mathbb{R}^{d_1^{(l)} \times d_1^{(l)}}$
for $t \in [T]$ **do**
  $\{\Delta W_t^{(l)}\}_{l\in\mathcal{L}} \leftarrow \mathcal{A}(f_{\theta_{t-1}}, (x_t, y_t))$
  Cache $\{k_t^{(l)}\}_{l\in\mathcal{L}}$
  $P_t^{(l)} = P_{t-1}^{(l)} + k_t^{(l)}[k_t^{(l)}]^\top$
  $W_t^{(l)} = W_{t-1}^{(l)} + \Delta W_t^{(l)} k_t^{(l)}[k_t^{(l)}]^\top [P_t^{(l)}]^{-1}$
  $W_t^{(l)} = W_{t-1}^{(l)} + \Delta W_t^{(l)}$
  $f_{\theta_t} \leftarrow \{W_t^{(l)}\}_{l\in\mathcal{L}}$
end for
Output: post-edit model f_{θ_T}

parameter. At time step t, the algorithm $\mathcal{A}$ processes the edit example (x_t, y_t) to edit the previous model $f_{\theta_{t-1}}(\cdot)$, obtaining the parameter updates $\{\Delta W_t^{(l)}\}_{l\in\mathcal{L}}$. In parallel, we cache the corresponding keys $\{k_t^{(l)}\}_{l\in\mathcal{L}}$. Then, we calculate the matrix $P_t^{(l)}$ and use it to adjust corresponding $\Delta W_t^{(l)}$ according to Equation (5). Finally, we replace the parameters of $f_{\theta_{t-1}}(\cdot)$ with the adjusted parameters $\{W_t^{(l)}\}_{l\in\mathcal{L}}$, yielding the model $f_{\theta_t}(\cdot)$ for time step t. As a comparison, we highlight the lines unique to our method SimIE in red, while using blue to indicate the naive approach of directly adapting standard algorithms to the lifelong scenarios.

Remark 3.7. For the edit example (x_t, y_t) contains multiple tokens, we denote the keys as $\{k_{t,i}^{(l)}\}_{i=1}^N$. In such cases, we only replace all occurrences of $k_t^{(l)}$ in Algorithm 1 with the concatenated key matrix $K_t^{(l)} := [k_{t,1}^{(l)}, \ldots, k_{t,N}^{(l)}]$.

Cost analysis. Finally, we briefly analyze the additional cost introduced by SimIE. Assume that each time step involves N tokens and $L = |\mathcal{L}|$ linear layers are edited. SimIE requires maintaining L matrices $P_t^{(l)}$ of dimension $d_1 \times d_1$, resulting in a direct storage cost of $\mathcal{O}(Ld_1^2)$. Utilizing low-rank structure $P_t^{(l)} = K_{:t}^{(l)}[K_{:t}^{(l)}]^\top + \lambda I$, the storage cost becomes $\mathcal{O}(TLNd_1)$. The primary computational cost of SimIE arises from matrix inversion, implemented by solving the linear system $XP_t^{(l)} = \Delta W_t^{(l)} k_t^{(l)}[k_t^{(l)}]^\top$. The standard approach for solving is LU decomposition, which has a time complexity of $\mathcal{O}(\frac{2}{3}d_1^3)$. Alternatively, Cholesky decomposition can be used, offering a reduced time complexity of $\mathcal{O}(\frac{1}{3}d_1^3)$, since $P_t^{(l)}$ is a symmetric positive definite matrix. Overall, the costs introduced by SimIE are manageable and are nearly negligible compared to those of model editing algorithms. A more detailed discussion can be found in Appendix B.4.

4. Beyond Assumptions

In this section, we theoretically analyze the behavior of SimIE when the assumptions are not met. Sections 4.1 and 4.2 examine the cases where Assumptions 3.2 and 3.3 do not hold, respectively. Section 4.3 considers the scenario where both assumptions fail. Since SimIE is applied independently to each layer, we focus on the analysis of a single layer and omit the layer index without loss of generality.

4.1. Optimality without Assumption 3.2

Assumption 3.2 requires that the weight matrix has sufficient capacity and that the key vectors are linearly independent, but this may not always hold in practice. For instance, frequent knowledge updates can cause the number of edit examples to rapidly exceed the column dimension of the matrix. Moreover, unintended edits to relevant factual knowledge may result in key vectors becoming linearly dependent. In such cases, the ideal editor does not exist, as the linear system $SK = B$ no longer admits the exact solution.

Although the existence of an ideal editor is precluded, we will show that the solution S_T^λ generated by SimIE is still near-optimal in a statistical sense. Here, *optimal in a statistical sense* refers to a solution S^* that minimizes the squared residual norm, i.e., $\|S^*K - B\|_{\text{F}}^2 = \min \|SK - B\|_{\text{F}}^2$, as in least squares problems. The following theorem provides an upper bound on the squared residual norm w.r.t. S_T^λ:

Theorem 4.1. *Let $S_T^\lambda = W_T - W_0$ represent the parameter shift generated by SimIE according to Equation (5). Under Assumption 3.3, the squared residual norm w.r.t. S_T^λ satisfies the following inequality:*

$$R_{\min} \le \left\|S_T^\lambda K - B\right\|_{\text{F}}^2 \le \frac{\lambda^2(\|B\|_{\text{F}}^2 - R_{\min})}{(\sigma_r^2 + \lambda)^2} + R_{\min},$$

where $R_{\min} := \min \|SK - B\|_{\text{F}}^2$ is the minimum value of squared residual norm, and $\sigma_r > 0$ denotes the smallest non-zero singular value of K.

Theorem 4.1 demonstrates that the squared residual norm of S_T^λ is bounded by the minimal squared residual norm $R_{\min}$ plus a term $\|B\|_{\text{F}}^2 - R_{\min}$ scaled by the hyperparameter λ. Here, $R_{\min}$ represents the theoretical lower bound of the squared residual norm, reflecting the irreducible error caused by the inherent limitations of the system. On the other hand, $\|B\|_{\text{F}}^2 - R_{\min}$ quantifies the portion of the squared residual norm that can be effectively fitted by the model. As $\lambda \to 0$, the squared residual norm of S_T^λ converges to $R_{\min}$. Thus, even in the absence of Assumption 3.2, the solution generated by SimIE remains near-optimal in a statistical sense, as it minimizes the squared residual norm of the given constraints to the extent possible.

4.2. Stability without Assumption 3.3

Assumption 3.3 assumes that the representations of specific knowledge remain stable, but this may not always hold in complex multi-layer editing algorithms. Specifically, parameter updates in earlier layers can perturb the keys in subsequent layers. Moreover, as the edit-distribution strategy depends on the current model state (Gupta et al., 2024b), it induces perturbations in the values of intermediate layers. As a result, in multi-layer editing scenarios, Assumption 3.3 is unlikely to hold due to inevitable key-value perturbations.

We will demonstrate that the solution generated by `SimIE` has stability even without Assumption 3.3, meaning that perturbations to key-value pairs do not excessively change the solution. Mathematically, perturbations to key-value pairs (k_t, v_t) are equivalent to perturbations in (k_t, b_t). Thus, our subsequent stability analysis focuses on the latter term. Since the impact of the perturbation to (k_t, b_t) is not amplified over subsequent time steps (see Lemma C.2), we can analyze these perturbations using the matrix forms K and B, rather than examining their components individually.

Following perturbation theory (Hansen, 1998; Malyshev, 2003), we define the relative condition number of S_T^λ w.r.t. B, but replace the ℓ_2-norm with the Frobenius norm, as:

$$\operatorname{cond}(S_T^\lambda, B) := \lim_{\epsilon \to 0} \sup_{\|\delta B\|_{\mathrm{F}} \le \epsilon} \left(\frac{\left\|\delta S_T^\lambda\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \Big/ \frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} \right).$$

Similarly, let $\kappa_{\mathrm{F}}(K) := \left\|K^\dagger\right\|_{\mathrm{F}} \|K\|_{\mathrm{F}}$ denote the condition number of the matrix K under the Frobenius norm. The following theorem establishes a perturbation bound for the solutions generated by `SimIE`:

Theorem 4.2. *Let S_T^λ represent the original solution corresponding to K and B, and $\widetilde{S_T^\lambda}$ denote the perturbed solution corresponding to $K + \delta K$ and $B + \delta B$, both generated by `SimIE` according to Equation (5). Under Assumption 3.2, the perturbation bound for S_T^λ is:*

$$\begin{aligned}\frac{\left\|S_T^\lambda - \widetilde{S_T^\lambda}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \le & \operatorname{cond}(S_T^\lambda, B) \frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \|J_\lambda\|_{\mathrm{F}} \|\delta K\|_{\mathrm{F}} \\ & + \frac{\sqrt{d_1}\,\|K\|_{\mathrm{F}}}{\sigma_{\min}^2} \|\delta K\|_{\mathrm{F}} \\ \le & \operatorname{cond}(S_T^\lambda, B) \frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \kappa_{\mathrm{F}}(K) \frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} \\ & + \frac{\sqrt{d_1}\,\|K\|_{\mathrm{F}}}{\sigma_{\min}^2} \|\delta K\|_{\mathrm{F}},\end{aligned}$$

where $J_\lambda := K^\top (K K^\top + \lambda I)^{-1}$ denotes the Jacobian matrix of S_T^λ at B.

The first inequality of Theorem 4.2 establishes a tight perturbation bound that depends on the hyperparameter λ. The first term indicates that the effect of the perturbation δB is limited by the condition number $\operatorname{cond}(S_T^\lambda, B)$. The second term $\|J_\lambda\|_{\mathrm{F}}$ and the third term $\sqrt{d_1}\,\|K\|_{\mathrm{F}} / \sigma_{\min}^2$ together govern the impact of the perturbation δK. As λ approaches 0, $\|J_\lambda\|_{\mathrm{F}}$ increases monotonically until attaining an upper bound. The second inequality captures this behavior by offering a λ-independent[2] perturbation bound, where $\|J_\lambda\|_{\mathrm{F}}$ is substituted with the condition number $\kappa_{\mathrm{F}}(K)$ w.r.t. K. In other words, when a very small $\lambda > 0$ is chosen for optimality, perturbations δK and δB always lead to manageable changes in the solution. In conclusion, even in the absence of Assumption 3.3, the solution generated by `SimIE` remains stable, as it is not excessively sensitive to small perturbations in the key-value pairs.

4.3. Trade-off without Assumptions 3.2 and 3.3

While violating either Assumption 3.2 or Assumption 3.3 individually results in relatively positive outcomes, the situation becomes more intricate when both assumptions are not met. In this case, a trade-off between optimality and stability must be considered. The following theorem presents a perturbation bound for the solutions generated by `SimIE` that does not rely on any assumptions:

Theorem 4.3. *Let S_T^λ represent the original solution corresponding to K and B, and $\widetilde{S_T^\lambda}$ denote the perturbed solution corresponding to $K + \delta K$ and $B + \delta B$, both generated by `SimIE` according to Equation (5). Then, the perturbation bound for S_T^λ is:*

$$\begin{aligned}\frac{\left\|S_T^\lambda - \widetilde{S_T^\lambda}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \le & \operatorname{cond}(S_T^\lambda, B) \frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \kappa_{\mathrm{F}}(K) \frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} \\ & + \frac{\|R_\lambda\|_{\mathrm{F}}}{\lambda} \frac{\sqrt{d_1}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \|\delta K\|_{\mathrm{F}},\end{aligned}$$

where $R_\lambda = S_T^\lambda K - B$ is the residual matrix w.r.t. S_T^λ.

Theorem 4.3 indicates that the impact of the perturbation δB is governed by the relative condition number $\operatorname{cond}(S_T^\lambda, B)$, while the effect of the perturbation δK is modulated by the condition number $\kappa_{\mathrm{F}}(K)$ and the ratio of the residual norm $\|R_\lambda\|_{\mathrm{F}}$ to the hyperparameter λ. Notably, as λ decreases, the term $\|R_\lambda\|_{\mathrm{F}}$ tends to decrease (as established in Theorem 4.1), while $1/\lambda$ increases monotonically. The combined effect of $\frac{\|R_\lambda\|_{\mathrm{F}}}{\lambda}$ introduces a trade-off between the optimality and stability of the solution: smaller values of λ can enhance the optimality, but at the cost of increased sensitivity to perturbations. Conversely, larger values of λ improve stability but may result in a suboptimal. Therefore,

[2]Note that $\operatorname{cond}(S_T^\lambda, B)$ also relies on λ, while it admits a complex upper bound. For conciseness, we omit a detailed analysis here. It is helpful to observe that $\lim_{\lambda \to 0} \operatorname{cond}(S_T^\lambda, B) = \kappa_{\mathrm{F}}(K)$. Further discussions are provided in Remark C.3.

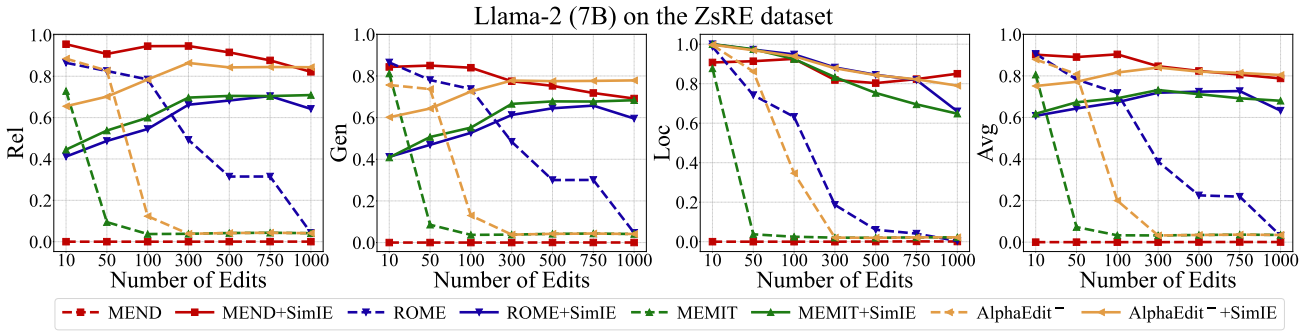


Figure 2. Performance of algorithms as the number of edits increases, with the solid line representing the results combined with the proposed SimIE.

when both Assumptions 3.2 and 3.3 are not met, it is crucial to carefully tune λ to achieve the best balance between optimality and stability in the solution generated by SimIE.

5. Experiments

This section validates the effectiveness of SimIE in restoring the strong performance of standard methods in lifelong contexts. Further analyses, including case study, hidden representation visualization, and long sentence editing, are provided in Appendix D.3.

Experimental setup. We conduct experiments on three widely used LLMs: **GPT2-XL** (1.5B) (Radford et al., 2019), **Llama-2** (7B) (Touvron et al., 2023), and **Mistral** (7B) (Chaplot, 2023). Our experiments include nine popular baselines: the basic fine-tuning method, **FT-L** (Meng et al., 2022), and four standard model editing algorithms, namely **MEND** (Mitchell et al., 2022a), **ROME** (Meng et al., 2022), **MEMIT** (Meng et al., 2023), and **AlphaEdit**$^{-}$[3] (Fang et al., 2024), along with four lifelong model editing algorithms, specifically **GRACE** (Hartvigsen et al., 2024), **WISE** (Wang et al., 2024b), **PRUNE** (Ma et al., 2024), and **AlphaEdit** (Fang et al., 2024). These algorithms are evaluated using two widely adopted benchmarks, i.e., the **ZsRE** dataset (Levy et al., 2017) and the **Counterfact** dataset (Meng et al., 2022). In line with prior research (Wang et al., 2024b), we assess performance using three key metrics: **Rel** (Reliability, also known as Edit Success Rate (Hartvigsen et al., 2024)), **Gen** (Generalization Success Rate), and **Loc** (Localization Success Rate). We use the Arithmetic Mean **Avg** $= \frac{\text{Rel}+\text{Gen}+\text{Loc}}{3}$ as the primary metric, and introduce the Locality-penalized Geometric Mean **Geo** $= e^{\alpha(\text{Loc}-1)}(\text{Rel} \times \text{Gen})$ as a complementary measure. For more details on the experimental setup, please refer to Appendix D.1.

[3]AlphaEdit includes two mechanisms: projecting updates into the null space and protecting previously edited knowledge. We refer to its version in the standard scenario, where the protection component is omitted, as AlphaEdit^{-}.

5.1. Effectiveness of SimIE

To evaluate the effectiveness of the proposed SimIE, we apply it to standard model editing algorithms and measure their performance before and after application. Specifically, we perform $T = 1000$ sequential edits on LLMs, with 1 example per edit. Table 1 presents the results for Llama-2 and Mistral, while Table 4 displays the results for GPT2-XL. Additionally, Figures 2 and 6 illustrate the performance of standard methods as the number of edits increases. Based on these results, we can draw the following observations:

- **Standard algorithms exhibit significant performance degradation as the number of edits increases.** While methods like ROME and AlphaEdit^{-} maintain stable performance early in editing, their effectiveness sharply decreases after surpassing 100 edits. The situation is even worse for MEND, which fails in the initial stages, likely due to the hypernetwork being tailored to the initial model. By the end of the edit sequence ($T = 1000$), all standard algorithms show unsatisfactory performance across settings.

- **The proposed SimIE successfully restores the strong performance of standard algorithms in lifelong contexts.** First, algorithms integrated with SimIE maintain comparable performance to their initial editing phase throughout the sequence. Second, for editing Llama-2 on the ZsRE dataset, SimIE increases the Avg of MEMIT from 0.03 to 0.68, achieving parity with its batch-editing counterpart, MEMIT-MASS, as reported in Wang et al. (2024b). These results demonstrate that the performance gains from SimIE are both substantial and consistent.

- **The performance of standard algorithms augmented with SimIE rivals that of dedicated life-**

Table 1. Performance of algorithms in the lifelong model editing task with 1000 edits, where the top three Avg are highlighted in **bold**.

	ZsRE								Counterfact							
	Llama-2 (7B)				Mistral (7B)				Llama-2 (7B)				Mistral (7B)			
Algorithm	Rel	Gen	Loc	Avg	Rel	Gen	Loc	Avg	Rel	Gen	Loc	Avg	Rel	Gen	Loc	Avg
FT-L	0.18	0.16	0.03	0.13	0.52	0.47	0.74	0.57	0.06	0.01	0.04	0.04	0.21	0.10	0.11	0.14
MEND	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+SimIE	0.82	0.69	0.85	**0.79**	0.72	0.66	0.83	0.74	0.96	0.31	0.29	0.52	0.94	0.30	0.32	0.52
ROME	0.04	0.05	0.00	0.03	0.04	0.03	0.01	0.03	0.25	0.15	0.06	0.16	0.28	0.24	0.01	0.18
+SimIE	0.64	0.60	0.66	0.63	0.72	0.67	0.82	0.74	0.97	0.51	0.30	**0.59**	0.96	0.52	0.50	**0.66**
MEMIT	0.04	0.04	0.02	0.03	0.04	0.04	0.02	0.03	0.00	0.00	0.07	0.02	0.00	0.00	0.00	0.00
+SimIE	0.71	0.68	0.65	0.68	0.65	0.62	0.80	0.69	0.81	0.43	0.29	0.51	0.78	0.45	0.46	0.56
AlphaEdit$^-$	0.04	0.04	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+SimIE	0.84	0.78	0.79	**0.80**	0.74	0.69	0.84	**0.75**	0.83	0.45	0.37	0.55	0.77	0.38	0.67	0.61
GRACE	0.98	0.02	1.00	0.67	0.98	0.02	1.00	0.67	1.00	0.00	1.00	**0.67**	1.00	0.00	1.00	**0.67**
WISE	0.82	0.76	1.00	**0.86**	0.72	0.69	1.00	**0.80**	0.71	0.31	0.37	0.46	0.69	0.35	0.35	0.46
PRUNE	0.54	0.52	0.72	0.59	0.25	0.25	0.47	0.33	0.71	0.35	0.67	**0.58**	0.86	0.42	0.56	0.61
AlphaEdit	0.84	0.79	0.64	0.76	0.79	0.73	0.75	**0.76**	0.91	0.53	0.30	0.58	0.94	0.49	0.53	**0.65**

long model editing methods. For instance, when editing Llama-2 on the ZsRE dataset, AlphaEdit+SimIE achieves an Avg of 0.80, outperforming GRACE, PRUNE, and the original AlphaEdit. Similarly, in experiments on GPT2-XL using the ZsRE dataset, ROME+SimIE attains an Avg of 0.81, surpassing all lifelong model editing methods. These improvements are also reflected across individual metrics (e.g., Rel, Gen, and Loc), providing comprehensive evidence for the effectiveness of SimIE.

5.2. Impact of hyperparameter λ

Section 4 highlights the importance of the hyperparameter λ in SimIE. Thus, we examine the impact of various λ values on the performance across methods. Specifically, for each standard model editing algorithm, we apply SimIE with $\lambda \in \{0.01, 0.1, 1, 5, 10, 30, 50\}$. The experimental results are presented in Figures 7 to 10, with Figure 3 illustrating selected results for ROME+SimIE on Llama-2 using the ZsRE dataset. It is important to note that the values of λ used in Section 5.1 correspond to those that yield the highest average (Avg) performance, as summarized in Table 3. Based on these findings, we make the following observations:

- **The Avg exhibits robustness across a specific range of the hyperparameter λ.** In most experimental settings, Avg remains stable for λ varies within the range $[1, 50]$. However, extremely small values of λ may lead to catastrophic performance degradation. This phenomenon can be primarily attributed to the increasing sensitivity of the solution provided by SimIE, which amplifies errors arising from violations of Assumption 3.3 and the matrix inversion process.

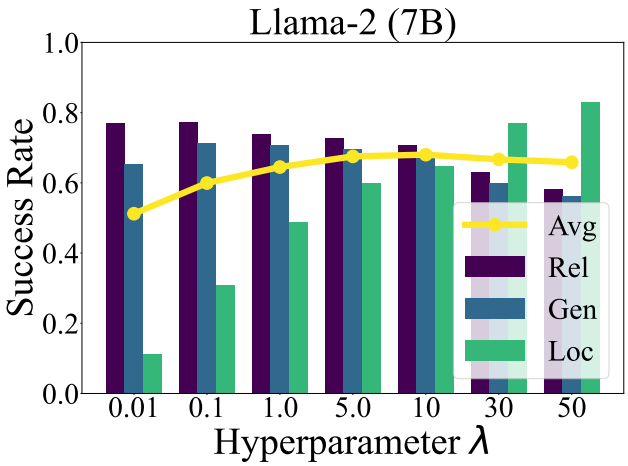


Figure 3. Performance of ROME+SimIE on the ZsRE dataset across various λ values, under $T = 1000$ sequential edits.

- **As the hyperparameter λ increases, Rel and Gen generally decline, whereas Loc tends to improve.** This behavior can be explained as follows: 1) Larger values of λ cause the solution to deviate from the ideal editor, which adversely affects Rel and Gen; 2) An increased λ enhances the stability of the solution, characterized by the smaller norm, thereby improving Loc. These findings emphasize the trade-off between optimality and sensitivity established by Theorem 4.3.

5.3. More recent LLMs

We conduct additional evaluations on recent model, including **Llama-3** (8B) (Grattafiori et al., 2024) and **Qwen2.5** (7B) (Yang et al., 2025). We select FT-L, ROME, AlphaEdit, WISE, and AlphaEdit as baselines and perform $T = 1000$ sequential edits using the ZsRE dataset. The experimental results are presented in the Table 2.

Table 2. Performance of algorithms in the lifelong model editing task with 1000 edits, where the top-1 Avg are highlighted in **bold**.

	ZsRE							
	Llama-3 (8B)				Qwen2.5 (7B)			
Algorithm	Rel	Gen	Loc	Avg	Rel	Gen	Loc	Avg
FT-L	0.17	0.14	0.01	0.10	0.09	0.08	0.02	0.07
ROME	0.09	0.08	0.01	0.06	0.22	0.22	0.09	0.18
+SimIE	0.75	0.72	0.80	**0.75**	0.92	0.87	0.72	0.84
AlphaEdit$^-$	0.06	0.06	0.03	0.05	0.67	0.63	0.76	0.68
+SimIE	0.74	0.67	0.75	0.72	0.91	0.83	0.88	**0.87**
WISE	0.51	0.50	1.00	0.67	0.54	0.53	0.99	0.69
AlphaEdit	0.86	0.78	0.62	**0.75**	0.98	0.80	0.72	0.83

We observe that SimIE achieves competitive performance across these recent models, especially surpassing the SOTA method (AlphaEdit) by an average of 4.8% on Qwen2.5. These results are consistent with those in the main paper, further confirming the effectiveness of our proposed SimIE.

6. Conclusion

This paper presents SimIE, a general framework that generalizes standard model editing methods to lifelong scenarios while preserving strong performance. Theoretical analysis confirms that SimIE remains effective even under relaxed assumptions, and an optimality-stability trade-off arises when both assumptions fail. Extensive experiments across multiple LLMs demonstrate that standard algorithms integrated with SimIE achieve comparable performance to specialized lifelong model editing methods. Our work not only provides robust baselines for lifelong scenarios but also enables lifelong editing to benefit from the ongoing advancements in standard editing techniques, thereby bridging the gap between these two paradigms.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (No. 62306255, 92370204), the National Key Research and Development Program of China (No. 2023YFF0725000), the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011839), the Fundamental Research Project of Guangzhou (No. 2024A04J4233), and the Education Bureau of Guangzhou Municipality.

Impact Statement

This work proposes a general framework that restores the strong performance of standard model editing methods in a lifelong context, bridging the gap between these two paradigms. Given the theoretical nature of our study, we have not identified any direct ethical concerns or negative societal impacts related to our research. While our contributions to model editing are intended to enhance the trustworthy deployment of LLMs, it is important to acknowledge that, like other advances in this field, there exists the potential for misuse to inject harmful content. As a result, it is essential to carefully monitor the usage of model editors during deployment.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pp. 1877–1901, 2020.

Chaplot, D. S. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 2023.

De Cao, N., Aziz, W., and Titov, I. Editing factual knowledge in language models. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6491–6506, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.522. URL https://aclanthology.org/2021.emnlp-main.522.

Fang, J., Jiang, H., Wang, K., Ma, Y., Wang, X., He, X., and Chua, T.-s. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*, 2024.

Golub, G. H. and Van Loan, C. F. *Matrix computations*. JHU press, 2013.

Gong, Z. and Sun, Y. An energy-centric framework for category-free out-of-distribution node detection in graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 908–919, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A.,

Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guo, K., Wen, H., Jin, W., Guo, Y., Tang, J., and Chang, Y. Investigating out-of-distribution generalization of gnns: An architecture perspective. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 932–943, 2024a.

Guo, S., Guo, Y., Zhang, H., and Wang, J. Mitigating update conflict in non-iid federated learning via orthogonal class gradients. *IEEE Transactions on Mobile Computing*, 2024b.

Guo, Y., Guo, K., Cao, X., Wu, T., and Chang, Y. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *International Conference on Machine Learning*, pp. 11905–11933. PMLR, 2023.

Gupta, A., Rao, A., and Anumanchipalli, G. Model editing at scale leads to gradual and catastrophic forgetting. *arXiv preprint arXiv:2401.07453*, 2024a.

Gupta, A., Sajnani, D., and Anumanchipalli, G. A unified framework for model editing. *arXiv preprint arXiv:2403.14236*, 2024b.

Hansen, P. C. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.

Hartvigsen, T., Sankaranarayanan, S., Palangi, H., Kim, Y., and Ghassemi, M. Aging with grace: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Higham, N. Functions of matrices: Theory and computation, 2008.

Higham, N. J. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hu, X., Cheng, Y., Zheng, Z., Wang, Y., Chi, X., and Zhu, H. Boss: A bilateral occupational-suitability-aware recommender system for online recruitment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4146–4155, 2023.

Huang, M., Liu, Y., Ao, X., Li, K., Chi, J., Feng, J., Yang, H., and He, Q. Auc-oriented graph neural network for fraud detection. In *Proceedings of the ACM Web Conference 2022*, pp. 1311–1321, 2022.

Huang, Z., Shen, Y., Zhang, X., Zhou, J., Rong, W., and Xiong, Z. Transformer-patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785*, 2023.

Laub, A. J. *Matrix analysis for scientists and engineers*. SIAM, 2004.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Levy, O., Seo, M., Choi, E., and Zettlemoyer, L. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*, 2017.

Lin, H., Zhu, H., Zuo, Y., Zhu, C., Wu, J., and Xiong, H. Collaborative company profiling: Insights from an employee's perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Lin, X., Huang, Z., Zhao, H., Chen, E., Liu, Q., Wang, H., and Wang, S. Hms: A hierarchical solver with dependency-enhanced understanding for math word problem. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 4232–4240, 2021.

Liu, J., Huang, Z., Lin, X., Liu, Q., Ma, J., and Chen, E. A cognitive solver with autonomously knowledge learning for reasoning mathematical answers. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 269–278. IEEE, 2022a.

Liu, J., Huang, Z., Ma, Z., Liu, Q., Chen, E., Su, T., and Liu, H. Guiding mathematical reasoning via mastering commonsense formula knowledge. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1477–1488, 2023a.

Liu, J., Huang, Z., Zhai, C., and Liu, Q. Learning by applying: A general framework for mathematical reasoning via enhancing explicit knowledge learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4497–4506, 2023b.

Liu, Q., Zhang, Q., Zhao, F., and Wang, G. Uncertain knowledge graph embedding: An effective method combining multi-relation and multi-path. *Frontiers Comput. Science*, 18(3):183311, 2024.

Liu, Y., Ao, X., Qin, Z., Chi, J., Feng, J., Yang, H., and He, Q. Pick and choose: a gnn-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference 2021*, pp. 3168–3177, 2021.

Liu, Y., Ao, X., Feng, F., and He, Q. Ud-gnn: Uncertainty-aware debiased training on semi-homophilous graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1131–1140, 2022b.

Liu, Y., Ao, X., Feng, F., Ma, Y., Li, K., Chua, T.-S., and He, Q. Flood: A flexible invariant learning framework for out-of-distribution generalization on graphs. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1548–1558, 2023c.

Ma, J.-Y., Wang, H., Xu, H.-X., Ling, Z.-H., and Gu, J.-C. Perturbation-restrained sequential model editing. *arXiv preprint arXiv:2405.16821*, 2024.

Madaan, A., Tandon, N., Clark, P., and Yang, Y. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*, 2022.

Malyshev, A. N. A unified theoryof conditioning for linear least squares and tikhonov regularization solutions. *SIAM journal on matrix analysis and applications*, 24(4):1186–1196, 2003.

Manakul, P., Liusie, A., and Gales, M. J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Meng, K., Sharma, A. S., Andonian, A. J., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Mitchell, E., Lin, C., Bosselut, A., Finn, C., and Manning, C. D. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=0DcZxeWfOPt.

Mitchell, E., Lin, C., Bosselut, A., Manning, C. D., and Finn, C. Memory-based model editing at scale. In *International Conference on Machine Learning*, pp. 15817–15831. PMLR, 2022b.

Peng, L., Giampouras, P., and Vidal, R. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, pp. 27585–27610. PMLR, 2023.

Penrose, R. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pp. 406–413. Cambridge University Press, 1955.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rifkin, R. M. and Lippert, R. A. Notes on regularized least squares. *Technical Report MIT-CSAILTR-2007-025*, 2007.

Shen, D., Qin, C., Wang, C., Dong, Z., Zhu, H., and Xiong, H. Topic modeling revisited: A document graph-based neural network perspective. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14681–14693, 2021.

Sinitsin, A., Plokhotnyuk, V., Pyrkin, D., Popov, S., and Babenko, A. Editable neural networks. In *International Conference on Learning Representations*, 2020.

Sun, Y., Zhu, H., Qin, C., Zhuang, F., He, Q., and Xiong, H. Discerning decision-making process of deep neural networks with hierarchical voting transformation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17221–17234, 2021a.

Sun, Y., Zhuang, F., Zhu, H., Zhang, Q., He, Q., and Xiong, H. Market-oriented job skill valuation with cooperative composition neural network. *Nature communications*, 12 (1):1992, 2021b.

Sun, Y., Zhu, H., Wang, L., Zhang, L., and Xiong, H. Large-scale online job search behaviors reveal labor market shifts amid covid-19. *Nature Cities*, 1(2):150–163, 2024.

Tan, C., Zhang, G., and Fu, J. Massive editing for large language models via meta learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=L6L1CJQ2PE.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vaswani, A. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Wang, C., Zhu, H., Zhu, C., Qin, C., and Xiong, H. Setrank: A setwise bayesian approach for collaborative ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6127–6136, 2020.

Wang, L., Zhang, X., Su, H., and Zhu, J. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Wang, P., Li, Z., Zhang, N., Xu, Z., Yao, Y., Jiang, Y., Xie, P., Huang, F., and Chen, H. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *arXiv preprint arXiv:2405.14768*, 2024b.

Xin, H., Sun, Y., Wang, C., and Xiong, H. Llmcdsr: Enhancing cross-domain sequential recommendation with large language models. *ACM Transactions on Information Systems*, 2025.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., and Chen, E. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Deng, S., Chen, H., and Zhang, N. Editing large language models: Problems, methods, and opportunities. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10222–10240, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.632. URL https://aclanthology.org/2023.emnlp-main.632.

Yu, L., Chen, Q., Zhou, J., and He, L. Melo: Enhancing model editing with neuron-indexed dynamic lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19449–19457, 2024.

Zeng, G., Chen, Y., Cui, B., and Yu, S. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

Zhang, L., Zhou, D., Zhu, H., Xu, T., Zha, R., Chen, E., and Xiong, H. Attentive heterogeneous graph embedding for job mobility prediction. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2192–2201, 2021.

Zhang, N., Tian, B., Cheng, S., Liang, X., Hu, Y., Xue, K., Gou, Y., Chen, X., and Chen, H. Instructedit: Instruction-based knowledge editing for large language models. *arXiv preprint arXiv:2402.16123*, 2024a.

Zhang, N., Yao, Y., Tian, B., Wang, P., Deng, S., Wang, M., Xi, Z., Mao, S., Zhang, J., Ni, Y., et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024b.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Zheng, C., Li, L., Dong, Q., Fan, Y., Wu, Z., Xu, J., and Chang, B. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*, 2023.

A. Additional discussions

A.1. Related work

Advances in deep learning (LeCun et al., 2015; Sun et al., 2021a;b; Wang et al., 2020; Guo et al., 2023; 2024b; Xu et al., 2024) have laid the foundation for the development of LLMs, equipping them with broad cross-domain knowledge (Liu et al., 2023a;b; Xin et al., 2025; Liu et al., 2024). These models have demonstrated strong versatility across diverse areas, including graph learning (Guo et al., 2024a; Liu et al., 2021; Shen et al., 2021; Zhang et al., 2021; Huang et al., 2022; Liu et al., 2022b; 2023c; Gong & Sun, 2024), data mining (Lin et al., 2017; Sun et al., 2024), and AI for science (Hu et al., 2023; Lin et al., 2021; Liu et al., 2022a). To enable low-cost correction of outdated or harmful information within LLMs, recent years have witnessed a surge in the development of model editing techniques (Zhang et al., 2024b).

Standard Model editing. Existing model editing techniques generally fall into two main categories: parameter-modifying and parameter-preserving (Yao et al., 2023). **Parameter-modifying** methods directly update the parameters most relevant to knowledge, thereby producing correct predictions. KE (De Cao et al., 2021) and MEND (Mitchell et al., 2022a), known as meta-learning approaches, update LLM parameters by learning a hypernetwork. MALMEN (Tan et al., 2024) formulates the update aggregation as a least-squares problem, mitigating the cancellation effect observed in MEND. InstructEdit (Zhang et al., 2024a) enhances these methods by incorporating instructions for training on different tasks, enabling adaptation to a variety of tasks. Additionally, ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), known as locate-and-edit methods, first identify the storage locations of knowledge (typically in MLP modules) and then perform targeted updates on the LLMs. Instead of altering the parameters, **parameter-preserving** methods store necessary information in memory, which is leveraged to guide the model in generating accurate predictions. MemPrompt (Madaan et al., 2022) and IKE (Zheng et al., 2023) utilize memory-based in-context learning to teach LLMs to generate the edited knowledge. SERAC (Mitchell et al., 2022b) constructs a retrieval-augmented counterfactual model to modulate LLM predictions as needed. MELO (Yu et al., 2024) dynamically activates certain LoRA (Hu et al., 2021) blocks based on an inner vector database, thus altering the behavior of the LLMs. While these standard editing methods achieve commendable performance, they suffer significant degradation in lifelong scenarios, particularly parameter-modifying methods. Our work introduces a general framework that restores the strong performance of parameter-modifying methods in a lifelong context.

Lifelong model editing. Lifelong model editing poses a greater challenge as it involves performing sequential edits on LLMs to accommodate evolving knowledge (Yao et al., 2023). T-Patcher (Huang et al., 2023) adds a handful of trainable neurons to the last Feed-Forward Network layer, enabling it to rectify a series of mistakes encountered by LLMs. GRACE (Hartvigsen et al., 2024) maintains a discrete codebook for a chosen layer, which is used to retrieve and replace the latent representations of examples. WISE (Wang et al., 2024b) designs a dual parametric memory scheme for edited knowledge, combined with a knowledge-sharding mechanism to reduce conflicts. PRUNE (Ma et al., 2024) proposes applying condition number restraints during sequential editing, preserving the general abilities of LLMs. AlphaEdit (Fang et al., 2024) projects updates into the null space of preserved knowledge while incorporating a regularization term to protect previously edited knowledge. Existing studies primarily focus on developing specialized approaches for lifelong model editing, often overlooking the potential of generalizing standard methods to lifelong scenarios. In contrast, our work aims to generalize standard model editing methods to lifelong scenarios while maintaining strong performance, thereby bridging the gap between these two paradigms for the first time.

A.2. Limitations

The first limitation involves the ideal editor, where an ill-conditioned key matrix K may generate parameter shifts S with a large norm. This design overly prioritizes injecting new knowledge into LLMs at the expense of preserving their general capabilities. This issue can be mitigated by imposing additional locality constraints on the ideal editor, such as limiting updates to a subspace that preserves original knowledge (Peng et al., 2023; Zeng et al., 2019). The second challenge concerns the hyperparameter λ in `SimIE`, whose optimal value cannot be determined in advance. Future work could enhance the recursive procedure of `SimIE` to enable flexible selection of an appropriate hyperparameter with minimal computational cost, similar to the technique in regularized least squares (Rifkin & Lippert, 2007).

B. Further details on Section 3

B.1. Existence and uniqueness of the ideal editor

Existence. The ideal editor exists if and only if the linear system $SK = B$ has a solution. To establish the existence, we analyze the system through a row-wise decomposition. The system $SK = B$ can be equivalently expressed as a collection of sub-linear systems:

$$s_i K = B_{i,:} \quad \forall i \in [d_2],$$

where s_i denotes the i-th row of the matrix S, and $B_{i,:}$ denotes the corresponding i-th row of the matrix B. For the linear system $SK = B$ to have a solution, it is necessary and sufficient that each sub-linear system $s_i K = B_{i,:}$ has a solution.

Observe that each subsystem $s_i K = B_{i,:}$ can be rewritten in its transpose form as $K^\top s_i^\top = B_{i,:}^\top$. With Assumption 3.2, the rows $\{k_t^\top\}_{t=1}^T$ of $K^\top$ are linearly independent. Thus, we have the following rank condition:

$$\operatorname{rank}(K^\top) = \operatorname{rank}([K^\top \mid b_i^\top]). \tag{6}$$

By the Rouché–Capelli theorem, Equation (6) implies that the sub-linear system $s_i K = B_{i,:}$ is consistent, i.e., at least one solution exists.

Since the analysis holds across all rows, we conclude that all sub-linear systems admit at least one solution. As a result, the system $SK = B$ has a solution, thereby proving the existence of the ideal editor.

Uniqueness. The general solution to the linear system $SK = B$ can be characterized by (Theorem 6.11. in Laub (2004)):

$$S = BK^\dagger + Y(I - KK^\dagger),$$

where $Y \in \mathbb{R}^{d_2 \times d_1}$ is an arbitrary matrix. The term $I - KK^\dagger$ represents the orthogonal projection operator onto the null space of $K^\top$ (Golub & Van Loan, 2013). Thus, we may verify that the two components $BK^\dagger$ and $Y(I - KK^\dagger)$ are orthogonal, i.e.,

$$\begin{aligned}
\langle Y(I - KK^\dagger), BK^\dagger \rangle_{\mathrm{F}} &= \operatorname{trace}([Y(I - KK^\dagger)]^\top Y^\top BK^\dagger) \\
&\overset{(a)}{=} \operatorname{trace}((I - KK^\dagger)Y^\top BK^\dagger) \\
&\overset{(b)}{=} \operatorname{trace}(Y^\top BK^\dagger(I - KK^\dagger)) \\
&= \operatorname{trace}(Y^\top B(K^\dagger - K^\dagger KK^\dagger)) \\
&\overset{(c)}{=} 0,
\end{aligned}$$

where (a) follows from the fact that $I - KK^\dagger$ is a symmetric matrix; (b) utilizes the cyclic property of the trace; and (c) is a consequence of Moore-Penrose conditions, specifically $K^\dagger KK^\dagger = K^\dagger$. From the above, the Frobenius norm of any solution S satisfying $SK = B$ can be decomposed as:

$$\|S\|_{\mathrm{F}}^2 = \left\|BK^\dagger\right\|_{\mathrm{F}}^2 + \left\|Y(I - KK^\dagger)\right\|_{\mathrm{F}}^2 \geq \left\|BK^\dagger\right\|_{\mathrm{F}}^2. \tag{7}$$

Clearly, the Frobenius norm of $\|S\|_{\mathrm{F}}^2$ is minimized if and only if $Y(I - KK^\dagger) = 0$, which implies $S^0 = BK^\dagger$. Therefore, S^0 is the unique minimum norm solution to $SK = B$, establishing the uniqueness of the ideal editor.

Remark B.1. If the columns of A are linearly independent (or equivalently, if A is left invertible), then the Moore-Penrose inverse $A^\dagger = (A^\top A)^{-1}A^\top$. Thus, under Assumption 3.2, S^0 can also be expressed as $S^0 = B(K^\top K)^{-1}K^\top$, which is consistent with the solution form derived using the method of Lagrange multipliers.

B.2. Detailed derivation of the recurrence relation

We now present the full derivation of Equation (4). The details are as follows:

$$
\begin{aligned}
S_t^\lambda &= (B_{:t-1}K_{:t-1}^\top + b_t k_t^\top)(\underbrace{K_{:t}K_{:t}^\top + \lambda I}_{\text{denoted as } P_t})^{-1} \\
&= B_{:t-1}K_{:t-1}^\top P_t^{-1} + b_t k_t^\top P_t^{-1} \\
&= \underbrace{B_{:t-1}K_{:t-1}^\top P_{t-1}^{-1}}_{=S_{t-1}^\lambda} P_{t-1}P_t^{-1} + b_t k_t^\top P_t^{-1} \\
&= S_{t-1}^\lambda \underbrace{P_{t-1}}_{=P_t - k_t k_t^\top} P_t^{-1} + b_t k_t^\top P_t^{-1} \\
&\overset{(a)}{=} S_{t-1}^\lambda - S_{t-1}^\lambda k_t k_t^\top P_t^{-1} + b_t k_t^\top P_t^{-1} \\
&= S_{t-1}^\lambda + (\underbrace{b_t}_{=(S_{t-1}^\lambda + \Delta W_t)k_t} - S_{t-1}^\lambda k_t) k_t^\top P_t^{-1} \\
&\overset{(b)}{=} S_{t-1}^\lambda + \Delta W_t k_t k_t^\top P_t^{-1}.
\end{aligned}
\tag{8}
$$

Here, (a) follows from the recursive formula for P_t, given by $P_t = P_{t-1} + k_t k_t^\top$; (b) results from the Equation (3), which holds under Assumption 3.3.

B.3. Proof of Theorem 3.6

Theorem 3.6*. *Let $S_T^\lambda = W_T - W_0$ denote the parameter shift generated by `SimIE` according to Equation (5), and let S^0 represent the ideal editor w.r.t. Equation (2). Under Assumptions 3.2 and 3.3, the following bound holds:*

$$
\frac{\lambda}{\sigma_{\max}^2 + \lambda} \le \frac{\left\| S_T^\lambda - S^0 \right\|_{\mathrm{F}}}{\left\| S^0 \right\|_{\mathrm{F}}} \le \frac{\lambda}{\sigma_{\min}^2 + \lambda},
$$

where $0 < \sigma_{\min} \le \sigma_{\max}$ are the smallest and largest singular values of the key matrix K, respectively.

Proof. We begin by expressing the key matrix $K = U\Sigma V^\top$ through SVD, where $U \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{T \times T}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{d_1 \times T}$ is a diagonal matrix containing the singular values. Under Assumption 3.2, we can rewrite the ideal editor S^0 as follows:

$$
\begin{aligned}
S^0 &= B(K^\top K)^{-1}K^\top \\
&\overset{(a)}{=} B(V\Sigma^\top U^\top U\Sigma V^\top)^{-1}(V\Sigma^\top U^\top) \\
&= B(V\Sigma^\top \Sigma V^\top)^{-1}V\Sigma^\top U^\top \\
&\overset{(b)}{=} BV(\Sigma^\top\Sigma)^{-1}V^\top V\Sigma^\top U^\top \\
&= BV(\Sigma^\top\Sigma)^{-1}\Sigma^\top U^\top,
\end{aligned}
\tag{9}
$$

where (a) follows from the matrix transpose property $(ABC)^\top = C^\top B^\top A^\top$; and (b) utilizes the inverse property of matrix products $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$. Analogously, for the solution S_T^λ generated by `SimIE`, there exist:

$$
\begin{aligned}
S_T^\lambda &= BK^\top(KK^\top + \lambda I)^{-1} \\
&= BV\Sigma^\top U^\top(U\Sigma V^\top V\Sigma^\top U^\top + \lambda I)^{-1} \\
&\overset{(a)}{=} BV\Sigma^\top U^\top(U\Sigma\Sigma^\top U^\top + \lambda UU^\top)^{-1} \\
&= BV\Sigma^\top U^\top\left(U(\Sigma\Sigma^\top + \lambda I)U^\top\right)^{-1} \\
&= BV\Sigma^\top U^\top U(\Sigma\Sigma^\top + \lambda I)^{-1}U^\top \\
&= BV\Sigma^\top(\Sigma\Sigma^\top + \lambda I)^{-1}U^\top,
\end{aligned}
\tag{10}
$$

where (a) employs the decomposition of the identity matrix $I = UU^\top$.

We are interested in the difference between S^0 and S_T^λ:

$$\begin{aligned} S^0 - S_T^\lambda &= BV\left((\Sigma^\top\Sigma)^{-1}\Sigma^\top - \Sigma^\top(\Sigma\Sigma^\top + \lambda I)^{-1}\right)U^\top \\ &= BV\left((\Sigma^\top\Sigma)^{-1}\Sigma^\top - (\Sigma^\top\Sigma)^{-1}(\Sigma^\top\Sigma)\Sigma^\top(\Sigma\Sigma^\top + \lambda I)^{-1}\right)U^\top \\ &= BV\left(\underbrace{(\Sigma^\top\Sigma)^{-1}\Sigma^\top}_{\Gamma_1} - \underbrace{(\Sigma^\top\Sigma)^{-1}\Sigma^\top}_{\Gamma_1}\underbrace{\Sigma\Sigma^\top(\Sigma\Sigma^\top + \lambda I)^{-1}}_{\Gamma_2}\right)U^\top. \end{aligned}$$

Given that K has full rank, Σ admits the block structure:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_1 \\ \mathbf{0} \end{bmatrix}, \tag{11}$$

where $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_T) \in \mathbb{R}^{T\times T}$ contains the singular values. Consequently, the term Γ_1 can be derived as:

$$(\Sigma^\top\Sigma)^{-1}\Sigma^\top = \left(\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix}\right)^{-1}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix} = (\Sigma_1^2)^{-1}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\Sigma_1^{-1} & \mathbf{0}\end{bmatrix}. \tag{12}$$

Next, the term Γ_2 can be expressed as:

$$\begin{aligned} \Sigma\Sigma^\top(\Sigma\Sigma^\top + \lambda I)^{-1} &= \begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix}\left(\begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix} + \lambda\begin{bmatrix}I_1 & \mathbf{0} \\ \mathbf{0} & I_2\end{bmatrix}\right)^{-1} \\ &= \begin{bmatrix}\Sigma_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\left(\begin{bmatrix}\Sigma_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} + \begin{bmatrix}\lambda I_1 & \mathbf{0} \\ \mathbf{0} & \lambda I_2\end{bmatrix}\right)^{-1} \\ &= \begin{bmatrix}\Sigma_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\lambda I_2)^{-1}\end{bmatrix} \\ &= \begin{bmatrix}\Sigma_1^2(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}. \end{aligned}$$

Now, substituting the expressions for Γ_1 and Γ_2, we obtain:

$$\begin{aligned} S^0 - S_T^\lambda &= BV\left(\begin{bmatrix}\Sigma_1^{-1} & \mathbf{0}\end{bmatrix} - \begin{bmatrix}\Sigma_1^{-1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1^2(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)U^\top \\ &= BV\begin{bmatrix}\Sigma_1^{-1} & \mathbf{0}\end{bmatrix}\left(\begin{bmatrix}I_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} - \begin{bmatrix}\Sigma_1^2(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)U^\top \\ &= BV\begin{bmatrix}\Sigma_1^{-1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\lambda I_1(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}U^\top \\ &= BV\begin{bmatrix}\lambda\Sigma_1^{-1}(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0}\end{bmatrix}U^\top, \end{aligned}$$

where the penultimate equality follows from the identity $1 - \frac{\sigma_i^2}{\sigma_i^2+\lambda} = \frac{\lambda}{\sigma_i^2+\lambda}$ for each diagonal element i.

To compute the Frobenius norm of $S^0 - S_T^\lambda$, we rewrite $BV \in \mathbb{R}^{d_2\times T}$ in column-vector form as $BV = [bv_1 \mid bv_2 \mid \cdots \mid bv_T]$,

yielding:

$$\begin{aligned}
\left\|S^0 - S_T^\lambda\right\|_{\mathrm{F}}^2 &= \left\|[bv_1 \mid bv_2 \mid \cdots \mid bv_T]\left[\lambda\Sigma_1^{-1}(\Sigma_1^2+\lambda I_1)^{-1} \quad \mathbf{0}\right]U^\top\right\|_{\mathrm{F}}^2 \\
&\overset{(a)}{=} \left\|\left[(\tfrac{1}{\sigma_1}\cdot\tfrac{\lambda}{\sigma_1^2+\lambda})bv_1 \quad (\tfrac{1}{\sigma_2}\cdot\tfrac{\lambda}{\sigma_2^2+\lambda})bv_2 \quad \ldots \quad (\tfrac{1}{\sigma_T}\cdot\tfrac{\lambda}{\sigma_T^2+\lambda})bv_T \quad \mathbf{0}\right]\right\|_{\mathrm{F}}^2 \\
&\overset{(b)}{=} \sum_{i=1}^{T}\|(\frac{1}{\sigma_i}\cdot\frac{\lambda}{\sigma_i^2+\lambda})bv_i\|_2^2 \\
&\leq (\frac{\lambda}{\sigma_{\min}^2+\lambda})^2\sum_{i=1}^{T}\|\frac{1}{\sigma_i}bv_i\|_2^2 \\
&= (\frac{\lambda}{\sigma_{\min}^2+\lambda})^2\left\|BV\left[\Sigma_1^{-1} \quad \mathbf{0}\right]U^\top\right\|_{\mathrm{F}}^2 \\
&\overset{(c)}{=} (\frac{\lambda}{\sigma_{\min}^2+\lambda})^2\left\|S^0\right\|_{\mathrm{F}}^2,
\end{aligned} \tag{13}$$

where (a) uses the fact that Frobenius norm is unitarily invariant; (b) utilizes the definitions of the Frobenius norm and ℓ_2-norm; and (c) can be obtained by substituting Equation (12) into Equation (9).

The lower bound follows analogously by replacing $\sigma_{\min}$ with $\sigma_{\max}$ in Equation (13), completing the proof. □

B.4. Cost analysis

Here, we analyze the additional cost introduced by `SimIE`. For analytical tractability, we consider that each time step involves N tokens and $L = |\mathcal{L}|$ linear layers are edited.

Storage costs. `SimIE` involves maintaining L matrices $P_t^{(l)}$ of dimension $d_1 \times d_1$, with a direct storage cost of $\mathcal{O}(Ld_1^2)$. By leveraging the low-rank structure $P_t^{(l)}$, an alternative strategy is to incrementally store $K_{:t}^{(l)} = [k_1^{(l)}, \ldots, k_t^{(l)}]$, which enables reconstructing $P_t^{(l)} = K_{:t}^{(l)}[K_{:t}^{(l)}]^\top + \lambda I$. In this case, the storage cost becomes $\mathcal{O}(TLNd_1)$, which depends on the number of sequential edits T. If the total number of tokens is significantly smaller than the input dimension, i.e., $TN \ll d_1$, storing $K_{:t}^{(l)}$ may be a more practical choice. However, storing $K_{:t}^{(l)}$ could pose privacy concerns compared to storing $P_t^{(l)}$.

Computational costs. The primary computational cost of the proposed `SimIE` stems from matrix inversion. Recognizing the inherent numerical inaccuracy in matrix inversion (Higham, 2002), we instead solve the linear system $XP_t^{(l)} = \Delta W_t^{(l)} k_t^{(l)}[k_t^{(l)}]^\top$ to provide the adjusted parameter updates. Among available methods[4], LU decomposition is a standard approach, which has a time complexity of $\mathcal{O}(\frac{2}{3}d_1^3)$. Since $P_t^{(l)}$ is a symmetric positive definite matrix, Cholesky decomposition offers higher computational efficiency with a reduced time complexity of $\mathcal{O}(\frac{1}{3}d_1^3)$. Additionally, SVD is a potential alternative due to its excellent numerical stability, though it incurs a higher time complexity of $\mathcal{O}(9d_1^3)$. The detailed derivation of the time complexity of these matrix decompositions can be found in Higham (2008) and Golub & Van Loan (2013). Fortunately, all these decomposition techniques have been efficiently parallelized on GPUs, significantly accelerating computation.

C. Proofs of theorems in Section 4

C.1. Proof of Theorem 4.1

Lemma C.1. *Consider a least squares problem* $\min\|SK - B\|_{\mathrm{F}}^2$*, where the matrix* K *is rank-deficient. If the SVD of matrix* K *is* $K = U\Sigma V^\top$*, then the minimum squared residual norm can be expressed as* $\min\|SK - B\|_{\mathrm{F}}^2 = \sum_{i=r+1}^{T}\|bv_i\|_2^2$*, where* bv_i *denotes the i-th column of the matrix* BV*, and* r *is the rank of* K*.*

Proof. We start with the expression for the squared residual norm:

$$\|SK - B\|_{\mathrm{F}}^2 = \left\|SU\Sigma V^\top - B\right\|_{\mathrm{F}}^2 = \|SU\Sigma - BV\|_{\mathrm{F}}^2, \tag{14}$$

where the second equality follows from the unitary invariance of the Frobenius norm. Since K is rank-deficient with rank r,

[4]Only the complexity of the decomposition is considered, as the back substitution has the same complexity across all methods.

the singular value matrix Σ has the following block structure:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \vdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{15}$$

where $\sigma_r > 0$ denotes the smallest non-zero singular value of K, and $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$ contains the non-zero singular values.

Let $SU = [su_1 \mid su_2 \mid \cdots \mid su_T]$ and $BV = [bv_1 \mid bv_2 \mid \cdots \mid bv_T]$ denote the column partitioning of SU and BV, respectively. Then the squared residual norm in (14) becomes:

$$\begin{aligned} \|SU\Sigma - BV\|_{\mathrm{F}}^2 &= \left\| \begin{bmatrix} su_1 & su_2 & \cdots & su_T \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} bv_1 & bv_2 & \cdots & bv_T \end{bmatrix} \right\|_{\mathrm{F}}^2 \\ &= \left\| \begin{bmatrix} \sigma_1 su_1 & \sigma_2 su_2 & \cdots & \sigma_r su_r & \mathbf{0} \end{bmatrix} - \begin{bmatrix} bv_1 & bv_2 & \cdots & bv_T \end{bmatrix} \right\|_{\mathrm{F}}^2 \\ &= \left\| \begin{bmatrix} \sigma_1 su_1 - bv_1 & \sigma_2 su_2 - bv_2 & \cdots & \sigma_r su_r - bv_r & -bv_{r+1} & \cdots & -bv_T \end{bmatrix} \right\|_{\mathrm{F}}^2 \\ &= \underbrace{\sum_{i=1}^{r} \|\sigma_i su_i - bv_i\|_2^2}_{\Gamma_1} + \underbrace{\sum_{i=r+1}^{T} \|bv_i\|_2^2}_{\Gamma_2}. \end{aligned}$$

To minimize squared residual norm, we observe that Γ_1 can be minimized independently for each $i \in 1, \ldots, r$ by choosing $su_i = \frac{bv_i}{\sigma_i}$, which makes $\Gamma_1 = 0$. The second term Γ_2 is independent of our optimization variables and represents the irreducible error due to the rank deficiency of K. Therefore, the minimum squared residual norm is:

$$\min \|SK - B\|_{\mathrm{F}}^2 = \min \sum_{i=1}^{r} \|\sigma_i su_i - bv_i\|_2^2 + \sum_{i=r+1}^{T} \|bv_i\|_2^2 = \sum_{i=r+1}^{T} \|bv_i\|_2^2. \tag{16}$$

This completes the proof. □

Theorem 4.1*. *Let $S_T^\lambda = W_T - W_0$ represent the parameter shift generated by `SimIE` according to Equation (5). Under Assumption 3.3, the squared residual norm w.r.t. S_T^λ satisfies the following inequality:*

$$R_{\min} \le \left\| S_T^\lambda K - B \right\|_{\mathrm{F}}^2 \le \frac{\lambda^2 (\|B\|_{\mathrm{F}}^2 - R_{\min})}{(\sigma_r^2 + \lambda)^2} + R_{\min},$$

where $R_{\min} := \min \|SK - B\|_{\mathrm{F}}^2$ is the minimum value of squared residual norm, and $\sigma_r > 0$ denotes the smallest non-zero singular value of K.

Proof. Using the SVD of $K = U\Sigma V^\top$, we can decompose the term $S_T^\lambda K - B$ as follows:

$$\begin{aligned} S_T^\lambda K - B &= BK^\top (KK^\top + \lambda I)^{-1} K - B \\ &= BV\Sigma^\top (\Sigma\Sigma^\top + \lambda I)^{-1} U^\top U \Sigma V^\top - B \\ &= B \left(V\Sigma^\top (\Sigma\Sigma^\top + \lambda I)^{-1} \Sigma V^\top - I \right) \\ &= BV \underbrace{\left(\Sigma^\top (\Sigma\Sigma^\top + \lambda I)^{-1} \Sigma - I \right)}_{\Gamma_1} V^\top. \end{aligned}$$

Base on block structure of Σ w.r.t. Equation (15), we simplify the term Γ_1 as:

$$\begin{aligned}
\Sigma^\top(\Sigma\Sigma^\top+\lambda I)^{-1}\Sigma - I &= \begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\left(\begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}+\lambda\begin{bmatrix}I_1 & \mathbf{0}\\ \mathbf{0} & I_2\end{bmatrix}\right)^{-1}\begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} - I\\
&= \begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}(\Sigma_1^2+\lambda I_1)^{-1} & \mathbf{0}\\ \mathbf{0} & (\lambda I_2)^{-1}\end{bmatrix}\begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} - \begin{bmatrix}I_1 & \mathbf{0}\\ \mathbf{0} & I_2\end{bmatrix}\\
&= \begin{bmatrix}\Sigma_1^2(\Sigma_1^2+\lambda I_1)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} - \begin{bmatrix}I_1 & \mathbf{0}\\ \mathbf{0} & I_2\end{bmatrix}\\
&= \begin{bmatrix}-\lambda I_1(\Sigma_1^2+\lambda I_1)^{-1} & \mathbf{0}\\ \mathbf{0} & -\lambda I_2\end{bmatrix},
\end{aligned}$$

where the last equality follows from the identity $\frac{\sigma_i^2}{\sigma_i^2+\lambda} - 1 = -\frac{\lambda}{\sigma_i^2+\lambda}$ for each diagonal element i. We now express $BV \in \mathbb{R}^{d_2\times T}$ in column-vector form as $BV = [bv_1 \mid bv_2 \mid \cdots \mid bv_T]$, obtaining

$$\begin{aligned}
\left\|S_T^\lambda K - B\right\|_{\mathrm{F}}^2 &= \left\|\begin{bmatrix}bv_1 & bv_2 & \cdots & bv_T\end{bmatrix}\begin{bmatrix}-\lambda I_1(\Sigma_1^2+\lambda I_1)^{-1} & \mathbf{0}\\ \mathbf{0} & -\lambda I_2\end{bmatrix}V^\top\right\|_{\mathrm{F}}^2\\
&= \left\|\begin{bmatrix}\frac{-\lambda}{\sigma_1^2+\lambda}bv_1 & \frac{-\lambda}{\sigma_2^2+\lambda}bv_2 & \cdots & \frac{-\lambda}{\sigma_r^2+\lambda}bv_r & -bv_{r+1} & \cdots & -bv_T\end{bmatrix}\right\|_{\mathrm{F}}^2\\
&= \sum_{i=1}^{r}\left(\frac{\lambda}{\sigma_i^2+\lambda}\right)^2\|bv_i\|_2^2 + \sum_{i=r+1}^{T}\|bv_i\|_2^2\\
&\le \left(\frac{\lambda}{\sigma_r^2+\lambda}\right)^2\sum_{i=1}^{r}\|bv_i\|_2^2 + \sum_{i=r+1}^{T}\|bv_i\|_2^2\\
&= \frac{\lambda^2}{(\sigma_r^2+\lambda)^2}\left(\sum_{i=1}^{T}\|bv_i\|_2^2 - \sum_{i=r+1}^{T}\|bv_i\|_2^2\right) + \underbrace{\sum_{i=r+1}^{T}\|bv_i\|_2^2}_{=\min\|SK-B\|_{\mathrm{F}}^2}\\
&= \frac{\lambda^2(\|BV\|_{\mathrm{F}}^2 - \min\|SK-B\|_{\mathrm{F}}^2)}{(\sigma_r^2+\lambda)^2} + \underbrace{\min\|SK-B\|_{\mathrm{F}}^2}_{\text{denoted as } R_{\min}}\\
&= \frac{\lambda^2(\|B\|_{\mathrm{F}}^2 - R_{\min})}{(\sigma_r^2+\lambda)^2} + R_{\min}.
\end{aligned}$$

The penultimate equality comes from Lemma C.1, and the last equality follows from the unitary invariance of the Frobenius norm. Since the squared residual norm of any solution cannot be smaller than the optimal squared residual norm, this completes the proof. □

C.2. Proof of Theorems 4.2 and 4.3

Lemma C.2. *At time step t, suppose the pair (k_t, b_t) is perturbed to $(k_t+\delta k_t, b_t+\delta b_t)$, which results in the solution generated by* `SimlE` *being $\widetilde{S_t^\lambda} = (B_{:t}K_{:t}^\top + \delta b_t\delta k_t^\top)(K_{:t}K_{:t}^\top + \lambda I + \delta k_t\delta k_t^\top)^{-1}$. Then, at time step $t+1$, the effect of the perturbation is not amplified, i.e., $\widetilde{S_{t+1}^\lambda} = (B_{:t+1}K_{:t+1}^\top + \delta b_t\delta k_t^\top)(K_{:t+1}K_{:t+1}^\top + \lambda I + \delta k_t\delta k_t^\top)^{-1}$.*

Proof. We demonstrate the non-amplification of the perturbation by working backward from the recurrence relation w.r.t. Equation (8). Let $P_t = K_{:t}K_{:t}^\top + \lambda I$, we have:

$$\begin{aligned}
\widetilde{S_{t+1}^\lambda} &= \widetilde{S_t^\lambda} + \Delta W_{t+1}k_{t+1}k_{t+1}^\top(P_{t+1}+\delta k_t\delta k_t^\top)^{-1}\\
&= \widetilde{S_t^\lambda}(P_t+\delta k_t\delta k_t^\top)(P_{t+1}+\delta k_t\delta k_t^\top)^{-1} + b_{t+1}k_{t+1}^\top(P_{t+1}+\delta k_t\delta k_t^\top)^{-1}\\
&= (B_{:t}K_{:t}^\top + \delta b_t\delta k_t^\top)(P_{t+1}+\delta k_t\delta k_t^\top)^{-1} + b_{t+1}k_{t+1}^\top(P_{t+1}+\delta k_t\delta k_t^\top)^{-1}\\
&= (B_{:t+1}K_{:t+1}^\top + \delta b_t\delta k_t^\top)(K_{:t+1}K_{:t+1}^\top + \lambda I + \delta k_t\delta k_t^\top)^{-1}.
\end{aligned}$$

The third equality follows by substituting the expression for $\widetilde{S_t^\lambda}$. Thus, the impact of the perturbation to (k_t, b_t) is not amplified in subsequent time steps. □

Theorem 4.2*. *Let S_T^λ represent the original solution corresponding to K and B, and $\widetilde{S_T^\lambda}$ denote the perturbed solution corresponding to $K+\delta K$ and $B+\delta B$, both generated by* `SimIE` *according to Equation (5). Under Assumption 3.2, the perturbation bound for S_T^λ is:*

$$\begin{aligned}\frac{\left\|S_T^\lambda - \widetilde{S_T^\lambda}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \leq & \operatorname{cond}(S_T^\lambda, B)\frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \|J_\lambda\|_{\mathrm{F}}\|\delta K\|_{\mathrm{F}} + \frac{\sqrt{d_1}\|K\|_{\mathrm{F}}}{\sigma_{\min}^2}\|\delta K\|_{\mathrm{F}} \\ \leq & \operatorname{cond}(S_T^\lambda, B)\frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \kappa_{\mathrm{F}}(K)\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} + \frac{\sqrt{d_1}\|K\|_{\mathrm{F}}}{\sigma_{\min}^2}\|\delta K\|_{\mathrm{F}},\end{aligned}$$

where $J_\lambda := K^\top(KK^\top + \lambda I)^{-1}$ denotes the Jacobian matrix of S_T^λ at B.

Proof. To analyze the perturbation behavior of S_T^λ w.r.t. small changes in K and B, we introduce

$$\epsilon = \max\left\{\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}}, \frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}}\right\}$$

as a measure of the relative magnitude of the perturbations. Following Golub & Van Loan (2013), we define the function:

$$g(x) := (B + x\frac{\delta B}{\epsilon})(K + x\frac{\delta K}{\epsilon})^\top\left((K + x\frac{\delta K}{\epsilon})(K + x\frac{\delta K}{\epsilon})^\top + \lambda I\right)^{-1},$$

where $x \in [0, \epsilon]$. It is easy to observe that $g(0) = S_T^\lambda$ and $g(\epsilon) = \widetilde{S_T^\lambda}$, corresponding to the unperturbed and perturbed solutions, respectively. Using a first-order Taylor expansion, we have:

$$g(\epsilon) = g(0) + g'(0)\epsilon + \mathcal{O}(\epsilon^2),$$

where $g'(0)$ is the derivative of $g(x)$ evaluated at $x = 0$, given by:

$$g'(0) = \frac{(\delta B - S_T^\lambda \delta K)}{\epsilon}K^\top(KK^\top + \lambda I)^{-1} + \frac{B - S_T^\lambda K}{\epsilon}\delta K^\top(KK^\top + \lambda I)^{-1}.$$

Omitting second order terms, we get:

$$\begin{aligned}\frac{\left\|S_T^\lambda - \widetilde{S_T^\lambda}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} &= \frac{\|g'(0)\epsilon\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \\ &= \frac{\left\|(\delta B - S_T^\lambda \delta K)K^\top(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}}\frac{\left\|(B - S_T^\lambda K)\delta K^\top(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \\ &\leq \frac{\left\|K^\top(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}}\|\delta B\|_{\mathrm{F}} + \left\|K^\top(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}\|\delta K\|_{\mathrm{F}} \\ &\quad + \frac{\left\|B - S_T^\lambda K\right\|_{\mathrm{F}}\left\|(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}}\|\delta K\|_{\mathrm{F}} \\ &\leq \underbrace{\frac{\left\|K^\top(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}\|B\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}}}_{\Gamma_1}\cdot\frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \left\|K^\top(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}\|\delta K\|_{\mathrm{F}} \\ &\quad + \frac{\lambda}{\sigma_{\min}^2 + \lambda}\cdot\underbrace{\frac{\|B\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}}}_{\Gamma_2}\cdot\underbrace{\left\|(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}}_{\Gamma_3}\|\delta K\|_{\mathrm{F}}.\end{aligned} \tag{17}$$

The last inequality substitute $\left\|B - S_T^\lambda K\right\|_{\mathrm{F}}$ with the squared residual norm bound from Theorem 4.1, where $R_{\min} = 0$ and $\sigma_r = \sigma_{\min}$ hold under Assumption 3.2.

For the term Γ_1, we show that it is equivalent to the condition number $\mathrm{cond}(S_T^\lambda, B)$. Specifically, since the solution $S_T^\lambda = BK^\top(KK^\top + \lambda I)^{-1}$ is differentiable w.r.t. B, the condition number can be written as:

$$\mathrm{cond}(S_T^\lambda, B) := \lim_{\epsilon \to 0} \sup_{\|\delta B\|_{\mathrm{F}} \le \epsilon} \left(\frac{\left\|\delta S_T^\lambda\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \Big/ \frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} \right) = \|J_\lambda\|_{\mathrm{F}} \frac{\|B\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} = \left\|K^\top(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}} \frac{\|B\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} = \Gamma_1,$$

where $J_\lambda = K^\top(KK^\top + \lambda I)^{-1}$ denotes the Jacobian matrix of S_T^λ at B.

Next, we analyze the term Γ_2 using the approach outlined in Theorem 3.6. Under Assumption 3.2, let $K = U\Sigma V^\top$ denote the SVD of K, where $\Sigma = \begin{bmatrix} \Sigma_1 \\ \mathbf{0} \end{bmatrix}$ defined in Equation (11). Recalling that $B = S^0 K$, we have:

$$\begin{aligned}
\frac{\|B\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} &\le \|K\|_{\mathrm{F}} \frac{\left\|S^0\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \\
&\overset{(a)}{=} \|K\|_{\mathrm{F}} \frac{\left\|BV(\Sigma^\top\Sigma)^{-1}\Sigma^\top U^\top\right\|_{\mathrm{F}}}{\|BV\Sigma^\top(\Sigma\Sigma^\top + \lambda I)^{-1}U^\top\|_{\mathrm{F}}} \\
&= \|K\|_{\mathrm{F}} \frac{\left\|BV(\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix})^{-1}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix}\right\|_{\mathrm{F}}}{\left\|BV\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix}(\begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix} + \lambda\begin{bmatrix}I_1 & \mathbf{0} \\ \mathbf{0} & I_2\end{bmatrix})^{-1}\right\|_{\mathrm{F}}} \\
&\overset{(b)}{=} \|K\|_{\mathrm{F}} \frac{\left\|\begin{bmatrix}bv_1 & bv_2 & \cdots & bv_T\end{bmatrix}\begin{bmatrix}\Sigma_1^{-1} & \mathbf{0}\end{bmatrix}\right\|_{\mathrm{F}}}{\left\|\begin{bmatrix}bv_1 & bv_2 & \cdots & bv_T\end{bmatrix}\begin{bmatrix}\Sigma_1(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0}\end{bmatrix}\right\|_{\mathrm{F}}} \\
&= \|K\|_{\mathrm{F}} \frac{\left\|\begin{bmatrix}\frac{1}{\sigma_1}bv_1 & \frac{1}{\sigma_2}bv_2 & \cdots & \frac{1}{\sigma_T}bv_T & \mathbf{0}\end{bmatrix}\right\|_{\mathrm{F}}}{\left\|\begin{bmatrix}\frac{\sigma_1}{\sigma_1^2+\lambda}bv_1 & \frac{\sigma_2}{\sigma_2^2+\lambda}bv_2 & \cdots & \frac{\sigma_T}{\sigma_T^2+\lambda}bv_T & \mathbf{0}\end{bmatrix}\right\|_{\mathrm{F}}} \\
&= \|K\|_{\mathrm{F}} \sqrt{\frac{\sum_{i=1}^T \|\frac{1}{\sigma_i}bv_i\|_2^2}{\sum_{i=1}^T \|\frac{\sigma_i}{\sigma_i^2+\lambda}bv_i\|_2^2}} \\
&\overset{(c)}{\le} \|K\|_{\mathrm{F}} \sqrt{\frac{\sum_{i=1}^T \|\frac{1}{\sigma_i}bv_i\|_2^2}{(\frac{\sigma_{\min}^2}{\sigma_{\min}^2+\lambda})^2 \sum_{i=1}^T \|\frac{1}{\sigma_i}bv_i\|_2^2}} \\
&= \frac{\sigma_{\min}^2 + \lambda}{\sigma_{\min}^2} \|K\|_{\mathrm{F}},
\end{aligned} \tag{18}$$

where (a) follows from Equations (9) and (10); (b) involves rewriting $BV \in \mathbb{R}^{d_2 \times T}$ in column-vector form as $BV = [bv_1 \mid bv_2 \mid \cdots \mid bv_T]$; and (c) utilizes the monotonic increasing property of the function $h(x) = (\frac{x^2}{x^2+\lambda})^2$ for all $x > 0$.

Finally, we consider the term Γ_3 and substitute $K = U\Sigma V^\top$ with $\Sigma = \begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix}$, yielding the following expression:

$$\begin{aligned}
\left\|(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}} &= \left\|U(\Sigma\Sigma^\top + \lambda I)^{-1}U^\top\right\|_{\mathrm{F}} \\
&= \left\|\left(\begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix} + \lambda\begin{bmatrix}I_1 & \mathbf{0} \\ \mathbf{0} & I_2\end{bmatrix}\right)^{-1}\right\|_{\mathrm{F}} \\
&= \left\|\begin{bmatrix}(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\lambda I_2)^{-1}\end{bmatrix}\right\|_{\mathrm{F}} \\
&= \sqrt{\sum_{i=1}^{T}\left(\frac{1}{\sigma_i^2+\lambda}\right)^2 + \sum_{i=T+1}^{d_1}\left(\frac{1}{\lambda}\right)^2} \\
&\le \sqrt{\sum_{i=1}^{d_1}\frac{1}{\lambda^2}} \\
&= \frac{\sqrt{d_1}}{\lambda}.
\end{aligned} \tag{19}$$

Now, plugging the bounds on Γ_1, Γ_2 and Γ_3 back into Equation (17), we get:

$$\begin{aligned}
\frac{\left\|S_T^\lambda - \widetilde{S_T^\lambda}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} &\le \mathrm{cond}(S_T^\lambda, B)\frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \|J_\lambda\|_{\mathrm{F}}\|\delta K\|_{\mathrm{F}} + \frac{\lambda}{\sigma_{\min}^2+\lambda}\cdot\frac{\sigma_{\min}^2+\lambda}{\sigma_{\min}^2}\|K\|_{\mathrm{F}}\cdot\frac{\sqrt{d_1}}{\lambda}\|\delta K\|_{\mathrm{F}} \\
&= \mathrm{cond}(S_T^\lambda, B)\frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \|J_\lambda\|_{\mathrm{F}}\|\delta K\|_{\mathrm{F}} + \frac{\sqrt{d_1}\|K\|_{\mathrm{F}}}{\sigma_{\min}^2}\|\delta K\|_{\mathrm{F}}.
\end{aligned} \tag{20}$$

As $\lambda \to 0$, the term $\|J_\lambda\|_{\mathrm{F}}$ in Equation (20) behaves monotonically increasing. However, we can find an upper bound that is independent of λ to replace it. As done in Equation (19), we express K as the SVD $K = U\Sigma V^\top$ with $\Sigma = \begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix}$. Then, the second term $\|J_\lambda\|_{\mathrm{F}}$ in Equation (20) becomes:

$$\begin{aligned}
\left\|K^\top(KK^\top + \lambda I)^{-1}\right\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} &= \left\|V\Sigma^\top(\Sigma\Sigma^\top + \lambda I)^{-1}U^\top\right\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} \\
&= \left\|\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix}\left(\begin{bmatrix}\Sigma_1 \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_1 & \mathbf{0}\end{bmatrix} + \lambda\begin{bmatrix}I_1 & \mathbf{0} \\ \mathbf{0} & I_2\end{bmatrix}\right)^{-1}\right\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} \\
&= \left\|\begin{bmatrix}\Sigma_1(\Sigma_1^2 + \lambda I_1)^{-1} & \mathbf{0}\end{bmatrix}\right\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} \\
&\overset{(a)}{\le} \left\|\begin{bmatrix}\Sigma_1^{-1} & \mathbf{0}\end{bmatrix}\right\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} \\
&\overset{(b)}{=} \left\|V\Sigma^\dagger U^\top\right\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} \\
&= \kappa_{\mathrm{F}}(K)\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}},
\end{aligned} \tag{21}$$

where (a) follows from the inequality $\frac{\sigma_i}{\sigma_i^2+\lambda} \le \frac{1}{\sigma_i}$; and (b) involves the definition of the Moore-Penrose inverse for a diagonal matrix.

This completes the proof. □

Remark C.3. The term $\mathrm{cond}(S_T^\lambda, B)$ also relies on λ, while it admits a complex upper bound. According to Equation (21),

we have:

$$\begin{aligned}\left\|K^\top(KK^\top+\lambda I)^{-1}\right\|_{\mathrm{F}} &= \left\|\left[\Sigma_1(\Sigma_1^2+\lambda I_1)^{-1} \quad \mathbf{0}\right]\right\|_{\mathrm{F}} \\ &= \sqrt{\sum_{i=1}^{T}(\frac{\sigma_i^2}{\sigma_i^2+\lambda})^2(\frac{1}{\sigma_i})^2} \\ &\leq \frac{\sigma_{\max}^2}{\sigma_{\max}^2+\lambda}\left\|K^\dagger\right\|_{\mathrm{F}}.\end{aligned}$$

Combining this with Equation (18), the condition number can be expressed as:

$$\begin{aligned}\mathrm{cond}(S_T^\lambda, B) &= \left\|K^\top(KK^\top+\lambda I)^{-1}\right\|_{\mathrm{F}}\frac{\|B\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \\ &\leq \frac{\sigma_{\max}^2}{\sigma_{\max}^2+\lambda}/\frac{\sigma_{\min}^2}{\sigma_{\min}^2+\lambda}\left\|K^\dagger\right\|_{\mathrm{F}}\|K\|_{\mathrm{F}} \\ &= \frac{\sigma_{\max}^2}{\sigma_{\max}^2+\lambda}/\frac{\sigma_{\min}^2}{\sigma_{\min}^2+\lambda}\kappa_{\mathrm{F}}(K).\end{aligned}$$

It is easy to verify $\lim_{\lambda\to 0}\mathrm{cond}(S_T^\lambda, B) = \kappa_{\mathrm{F}}(K)$ and $\lim_{\lambda\to\infty}\mathrm{cond}(S_T^\lambda, B) = \frac{\sigma_{\max}^2}{\sigma_{\min}^2}\kappa_{\mathrm{F}}(K)$. Therefore, the condition number $\mathrm{cond}(S_T^\lambda, B)$ remains finite for any $\lambda > 0$.

Theorem 4.3*. *Let S_T^λ represent the original solution corresponding to K and B, and $\widetilde{S_T^\lambda}$ denote the perturbed solution corresponding to $K+\delta K$ and $B+\delta B$, both generated by `SimIE` according to Equation (5). Then, the perturbation bound for S_T^λ is:*

$$\frac{\left\|S_T^\lambda - \widetilde{S_T^\lambda}\right\|_{\mathrm{F}}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}} \leq \mathrm{cond}(S_T^\lambda, B)\frac{\|\delta B\|_{\mathrm{F}}}{\|B\|_{\mathrm{F}}} + \kappa_{\mathrm{F}}(K)\frac{\|\delta K\|_{\mathrm{F}}}{\|K\|_{\mathrm{F}}} + \frac{\|R_\lambda\|_{\mathrm{F}}}{\lambda}\frac{\sqrt{d_1}}{\left\|S_T^\lambda\right\|_{\mathrm{F}}}\|\delta K\|_{\mathrm{F}},$$

where $R_\lambda = S_T^\lambda K - B$ is the residual matrix w.r.t. S_T^λ.

Proof. The proof process closely follows that of Theorem 4.2. The key distinctions lie in two aspects: first, the observation that $K = U\Sigma V^\top$ with $\Sigma = \begin{bmatrix}\Sigma_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ as defined in Equation (15), and second, the fact that $\|R_\lambda\|_{\mathrm{F}} \neq 0$, as noted in Theorem 4.1. With these differences taken into account, the proof is complete. □

D. More experimental details and results

D.1. Detailed experimental setup

D.1.1. Baselines

Here, we describe the nine baseline methods used in our experiments. **All methods were implemented using the publicly available knowledge editing framework EasyEdit (Zhang et al., 2024b), which reproduces and integrates the original code and hyperparameters provided in the respective papers.**

- **FT-L** (Meng et al., 2022) freezes most layers of LLMs, allowing fine-tuning only on a single MLP using autoregressive loss.
- **MEND** (Mitchell et al., 2022a) applies a low-rank decomposition to the gradient from standard fine-tuning, which is then fed into a pre-trained hypernetwork to generate new parameter update. The training loss of the hypernetwork consists of standard autoregressive loss over the edit example and the KL divergence loss over localization examples.
- **ROME** (Meng et al., 2022) models the linear layer within MLPs as key-value memory, thereby inserting new knowledge by solving a constrained least-squares problem. Using causal tracing, ROME identifies mid-layer MLPs as critical for knowledge storage, making them suitable targets for editing.

- **MEMIT** (Meng et al., 2023), based on ROME, enable the simultaneous insertion of hundreds or thousands of facts using least-squares method. MEMIT is a multi-layer editing algorithm, which introduces an edit-distribution strategy to spread updates evenly over the range of mid-layer MLPs.
- **GRACE** (Hartvigsen et al., 2024) maintains a discrete codebook for a chosen layer, which is dynamically updated through adding, expanding, and splitting during sequential editing. For inference phase, GRACE retrieves entries from the codebook based on the semantic similarity of the input and decides whether to replace the output accordingly.
- **WISE** (Wang et al., 2024b) designs a dual parametric memory scheme, allocating one memory for pre-trained knowledge and another for edited knowledge. Additionally, WISE incorporates a knowledge-sharding mechanism, which different sets of edits reside in distinct and orthogonal subspaces of parameters, to reduce conflicts.
- **PRUNE** (Ma et al., 2024) proposes applying condition number restraints during sequential editing by reducing the large singular values of the update matrix, thereby controlling the condition number of the edited matrix. PRUNE minimizes perturbations to the original knowledge, preserving the general abilities of the edited model. *In our study, we apply PRUNE to ROME, as it demonstrated robust performance in the original paper.*
- **AlphaEdit** (Fang et al., 2024) projects updates into the null space of preserved knowledge, ensuring that the output of the post-edited LLMs remains unchanged when queried about preserved knowledge. For lifelong scenarios, AlphaEdit incorporates a regularization term to protect previously edited knowledge. *We refer to the version of AlphaEdit in the standard scenario, where the protection component is omitted, as* ***AlphaEdit***$^{-}$.

D.1.2. DATASETS

Here, we provide a detailed description of the datasets used in this study. We adopt the train/test split from previous work (Wang et al., 2024b; Meng et al., 2022). Except for MEND, which uses the training set to fit the hypernetwork, all other methods perform editing and evaluation directly on the test set.

- **ZsRE** (Levy et al., 2017) is a question-answering dataset that uses questions generated by roundtrip translation as equivalent inputs. Each sample in the ZsRE dataset includes a question and an answer as the edit example (x_e, y_e), a rephrased question as the paraphrase prompt x'_e, and an unrelated question as the neighborhood prompt x_e^{loc}. Figure 4 provides an example where the corresponding relationships are:

$$x_e \longleftrightarrow \texttt{src} \qquad y_e \longleftrightarrow \texttt{alt}$$
$$x'_e \longleftrightarrow \texttt{rephrase} \qquad x_e^{\text{loc}} \longleftrightarrow \texttt{loc}$$

- **Counterfact** (Meng et al., 2022) is a more challenging dataset that contrasts counterfactual statements with factual ones. The samples in Counterfact are structured similarly to ZsRE for evaluating reliability, generalization success rate, and localization success rate. Figure 5 provides an example where the corresponding relationships are:

$$x_e \longleftrightarrow \texttt{prompt} \qquad y_e \longleftrightarrow \texttt{target\_new}$$
$$x'_e \longleftrightarrow \texttt{rephrase\_prompt} \qquad x_e^{\text{loc}} \longleftrightarrow \texttt{locality\_prompt}$$

D.1.3. METRICS

Here, we formally introduce the evaluation metrics used in our experiments, which are widely adopted in the model editing literature (Zhang et al., 2024b; Wang et al., 2024b). Given an editing dataset $\mathcal{D}_{\text{edit}} = \{(x_t, y_t) \mid t \in [T]\}$ with T edit examples, we evaluate the final post-edit model $f_{\theta_T}(\cdot)$ after the T-th edit. The metrics **Rel** (Reliability, also known as Edit Success Rate (Hartvigsen et al., 2024)), **Gen** (Generalization Success Rate), and **Loc** (Localization Success Rate) are used to assess the reliability, generalization, and specificity of the model editing methods, respectively. Specifically:

- **Rel** measures the average Top-1 accuracy on the edit examples (x_t, y_t):

$$\text{Rel} := \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(f_{\theta_T}(x_t) = y_t).$$

```
{
    "subject": "IAAF Combined Events Challenge",
    "src": "When was the inception of IAAF Combined Events Challenge?",
    "pred": "2011",
    "rephrase": "When was the IAAF Combined Events Challenge launched?",
    "alt": "2006",
    "answers": [
        "1998"
    ],
    "loc": "nq question: what is the name of the last episode of spongebob",
    "loc_ans": "The String",
    "cond": "2011 >> 2006 || When was the inception of IAAF Combined Events
     ↪  Challenge?",
    "portability": {
        "Recalled Relation": "(IAAF Combined Events Challenge, event type,
         ↪  athletics)",
        "New Question": "What type of sports event is the IAAF Combined Events
         ↪  Challenge, which was established in 2006?",
        "New Answer": "Athletics"
    }
}
```


Figure 4. A sample of the ZsRE dataset.


```
{
    "case_id": 0,
    "prompt": "The mother tongue of Danielle Darrieux is",
    "target_new": "English",
    "subject": "Danielle Darrieux",
    "ground_truth": "French",
    "rephrase_prompt": "Where Danielle Darrieux is from, people speak the
     ↪  language of",
    "locality_prompt": "Michel Rocard is a native speaker of",
    "locality_ground_truth": "French"
}
```


Figure 5. A sample of the Counterfact dataset.

- **Gen** evaluates the average Top-1 accuracy of the model's predictions for paraphrase prompts x'_t:

$$\text{Gen} := \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(f_{\theta_T}(x'_t) = y_t).$$

- **Loc** computes the average proportion of preserved predictions for neighborhood prompts x_t^{loc}:

$$\text{Loc} := \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(f_{\theta_T}(x_t^{\text{loc}}) = f_{\theta_0}(x_t^{\text{loc}}).$$

D.1.4. Implementation details

Unless otherwise stated, we store $P_t^{(l)}$ and employ LU decomposition across all experiments (see discussion in Appendix B.4). For the hyperparameter λ, we conduct a grid search over the set $\{0.01, 0.1, 1, 5, 10, 30, 50\}$, with the resulting values summarized in Table 3.

Remark D.1 (MEND involves an unfair implementation). Since the hypernetwork of MEND is tailored to the initial model, its subsequent generated updates significantly violate Assumption 3.3. Given that training the hypernetwork at each time

Table 3. The hyperparameter λ of SimIE used in the experiments.

	ZsRE			Counterfact		
Algorithm	Llama-2	Mistral	GPT2-XL	Llama-2	Mistral	GPT2-XL
MEND+SimIE	1.0	30	5.0	5.0	10	0.1
ROME+SimIE	50	30	50	5.0	5.0	50
MEMIT+SimIE	10	10	1.0	0.1	1.0	0.1
AlphaEdit$^-$+SimIE	1.0	1.0	5.0	0.1	0.1	5.0

step is impractical, we instead use the pre-trained hypernetwork to edit the initial model, yielding $\widetilde{W}_t$. The update required by SimIE, $\Delta W_t = \widetilde{W}_t - W_{t-1}$, is then computed by subtracting the current model parameters. It is worth noting that this implementation is slightly less fair compared to other methods, as it store additional copy of the initial model.

D.2. More experimental results

Here, we present additional experimental results. Table 4 presents the results of editing GPT2-XL using the ZsRE and Counterfact datasets. Figure 6 illustrate the performance of standard algorithms as the number of edits increases. The impact of hyperparameter in SimIE on performance is presented in Figures 7 to 10.

In addition, we introduce a new composite metric defined as **Geo** $= e^{\alpha(\text{Loc}-1)}(\text{Rel} \times \text{Gen})$. Geo takes locality as a penalty term, serving as a smoothed condition factor. Meanwhile, by using the (squared) geometric mean of reliability and generality, it prevents methods from abandoning one of them. Table 5 shows the evaluation results based on this composite metric, where $\alpha = 2$.

Table 4. Performance of algorithms in the lifelong model editing task with 1000 edits, where the top three Avg are highlighted in **bold**.

	ZsRE				Counterfact			
	GPT2-XL (1.5B)				GPT2-XL (1.5B)			
Algorithm	Rel	Gen	Loc	Avg	Rel	Gen	Loc	Avg
FT-L	0.09	0.08	0.00	0.06	0.11	0.03	0.01	0.05
MEND	0.00	0.00	0.00	0.00	0.00	0.00	0.48	0.16
+SimIE	0.56	0.52	0.90	0.66	0.94	0.14	0.15	0.41
ROME	0.46	0.37	0.50	0.45	0.01	0.00	0.01	0.01
+SimIE	0.93	0.81	0.69	**0.81**	0.87	0.20	0.45	0.51
MEMIT	0.58	0.50	0.55	0.54	0.81	0.46	0.25	0.51
+SimIE	0.81	0.66	0.81	**0.76**	0.88	0.32	0.67	**0.62**
AlphaEdit$^-$	0.10	0.09	0.31	0.17	0.68	0.38	0.30	0.46
+SimIE	0.87	0.77	0.77	**0.80**	0.98	0.46	0.54	**0.66**
GRACE	0.35	0.01	1.00	0.45	0.00	0.00	1.00	0.33
WISE	0.34	0.33	1.00	0.56	0.01	0.01	0.77	0.26
PRUNE	0.50	0.43	0.84	0.59	0.13	0.04	0.86	0.35
AlphaEdit	0.83	0.70	0.75	0.76	0.98	0.48	0.65	**0.70**

Table 5. Performance of algorithms in the lifelong model editing task with 1000 edits.

			FT-L	MEND	+SimIE	ROME	+SimIE	MEMIT	+SimIE	AlphaEdit⁻	+SimIE	GRACE	WISE	PRUNE	AlphaEdit
ZsRE	Llama-2 (7B)	Avg	0.13	0.00	0.79	0.03	0.63	0.03	0.68	0.03	0.80	0.67	0.86	0.59	0.76
		Geo	0.00	0.00	0.42	0.00	0.19	0.00	0.24	0.00	0.43	0.02	0.62	0.16	0.32
	Mistral (7B)	Avg	0.57	0.00	0.74	0.03	0.74	0.03	0.69	0.00	0.75	0.67	0.80	0.33	0.76
		Geo	0.14	0.00	0.34	0.00	0.34	0.00	0.27	0.00	0.37	0.02	0.49	0.02	0.35
	GPT2-XL (1.5B)	Avg	0.06	0.00	0.66	0.45	0.81	0.54	0.76	0.17	0.80	0.45	0.56	0.59	0.76
		Geo	0.00	0.00	0.24	0.06	0.41	0.12	0.37	0.00	0.42	0.00	0.11	0.16	0.35
Counterfact	Llama-2 (7B)	Avg	0.04	0.00	0.52	0.16	0.59	0.02	0.51	0.00	0.55	0.67	0.46	0.58	0.58
		Geo	0.00	0.00	0.07	0.01	0.12	0.00	0.08	0.00	0.10	0.00	0.06	0.13	0.12
	Mistral (7B)	Avg	0.14	0.00	0.52	0.18	0.66	0.00	0.56	0.00	0.61	0.67	0.46	0.61	0.65
		Geo	0.00	0.00	0.07	0.01	0.18	0.00	0.12	0.00	0.15	0.00	0.07	0.15	0.18
	GPT2-XL (1.5B)	Avg	0.05	0.16	0.41	0.01	0.51	0.51	0.62	0.46	0.66	0.33	0.26	0.35	0.70
		Geo	0.00	0.00	0.02	0.00	0.06	0.08	0.14	0.06	0.18	0.00	0.00	0.00	0.23

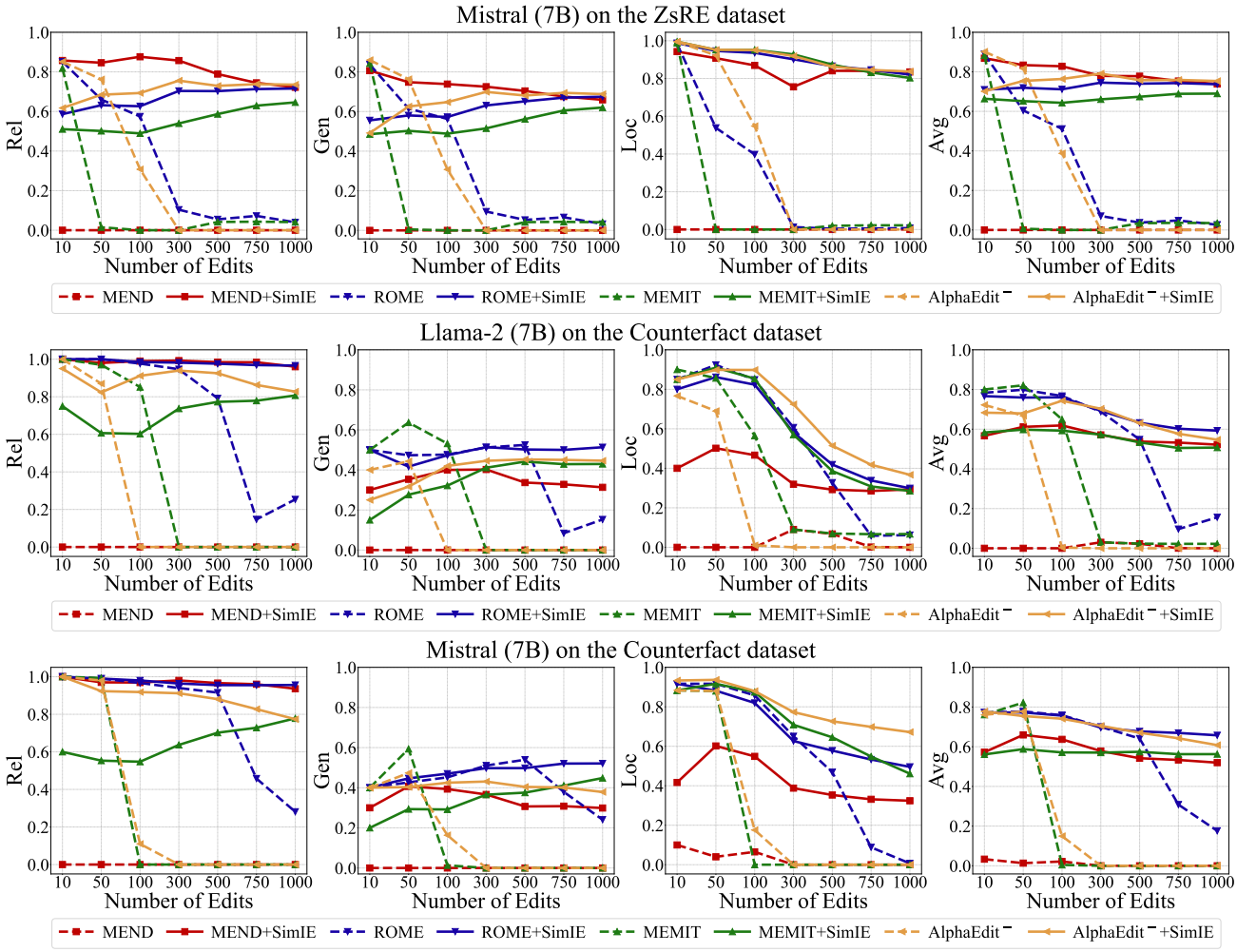
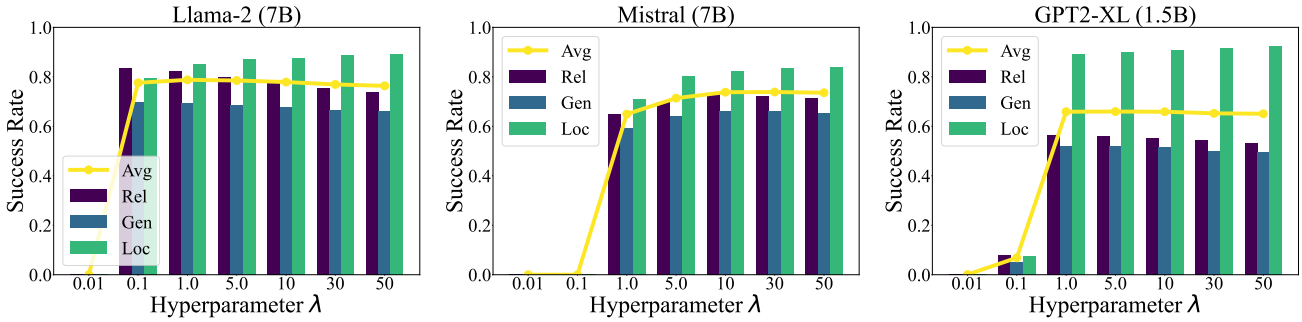
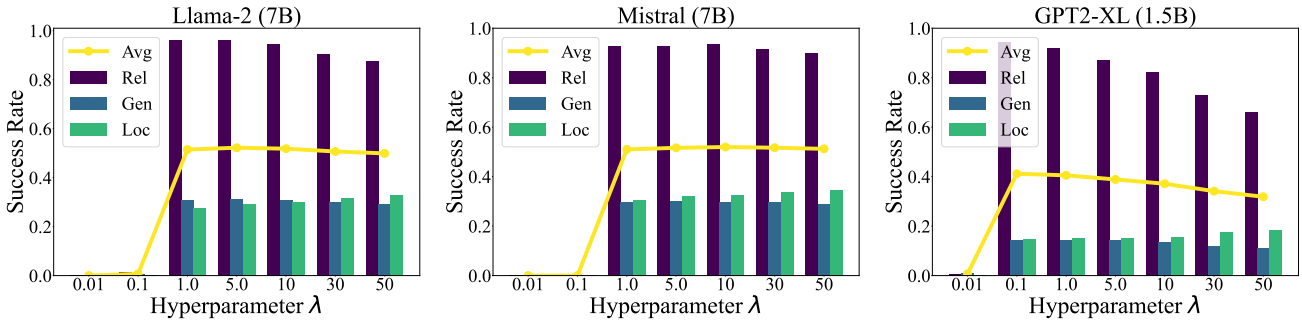


Figure 6. Performance of algorithms as the number of edits increases, with the solid line representing the results combined with the proposed SimIE.

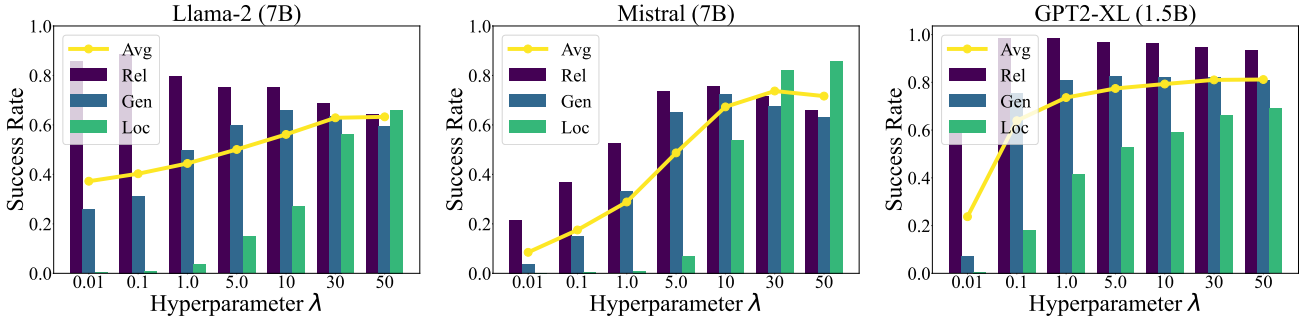


(a) MEND+SimIE on the ZsRE dataset

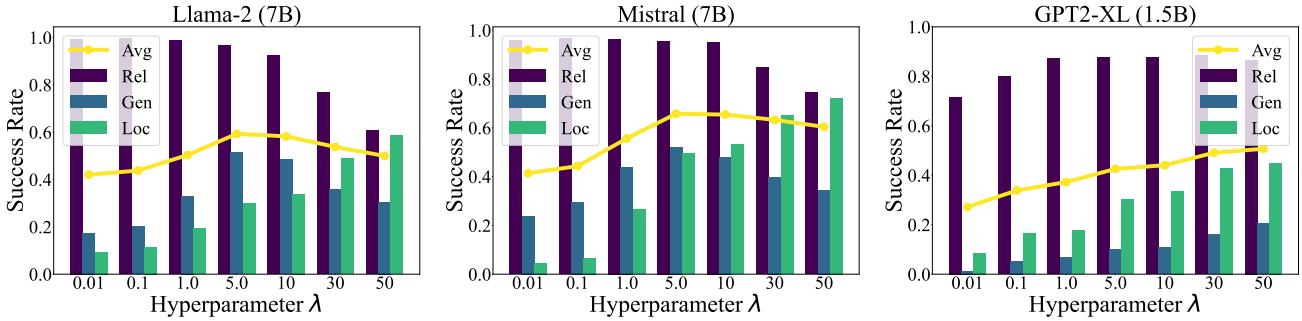


(b) MEND+SimIE on the Counterfact dataset

Figure 7. Performance of MEND+SimIE across various λ values, under $T = 1000$ sequential edits.

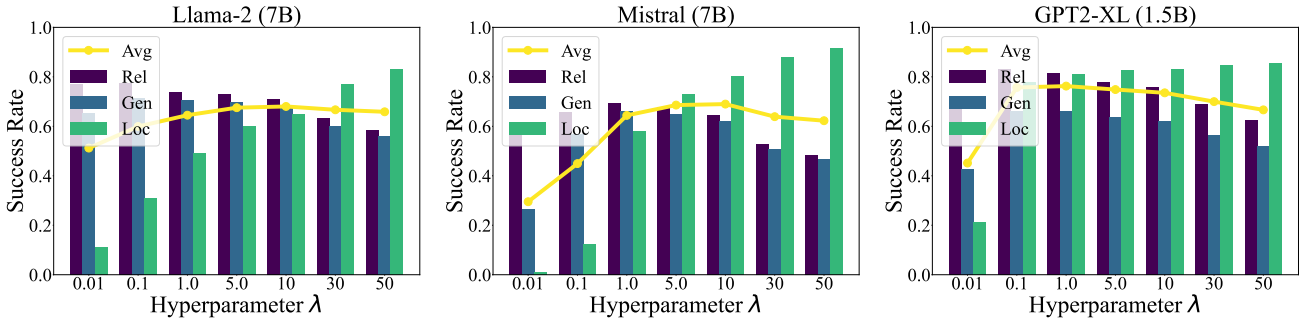


(a) ROME+SimIE on the ZsRE dataset

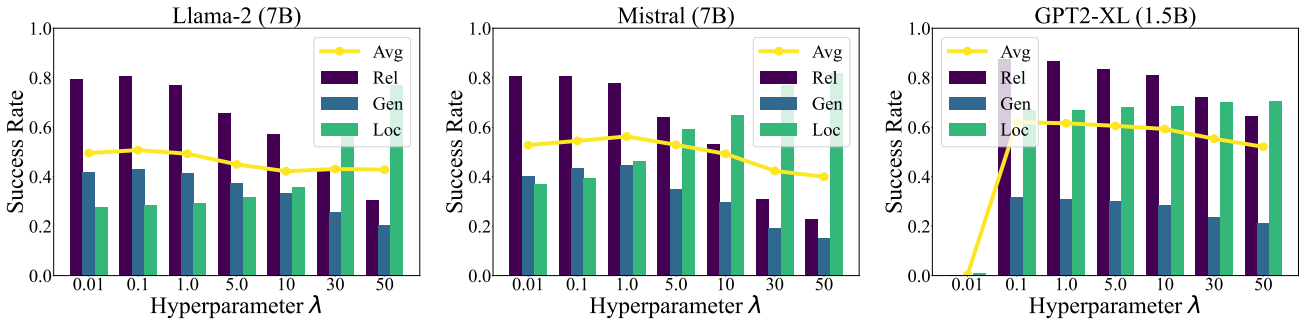


(b) ROME+SimIE on the Counterfact dataset

Figure 8. Performance of ROME+SimIE across various λ values, under $T = 1000$ sequential edits.

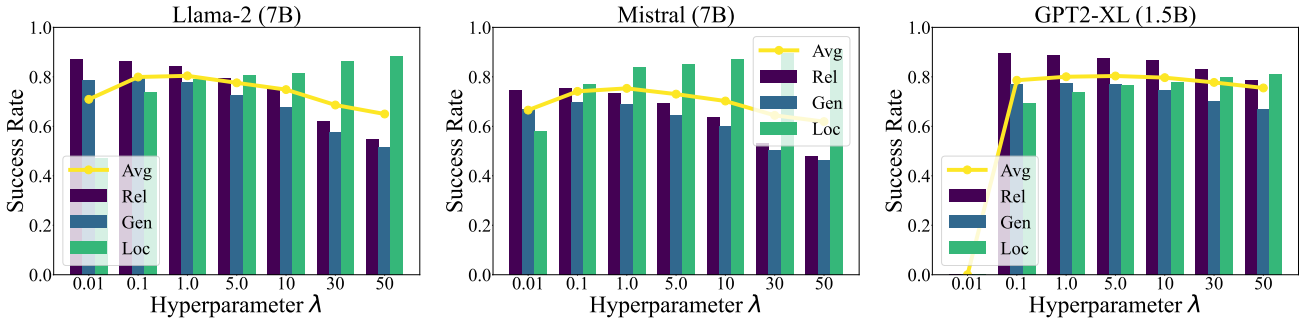


(a) MEMIT+SimIE on the ZsRE dataset

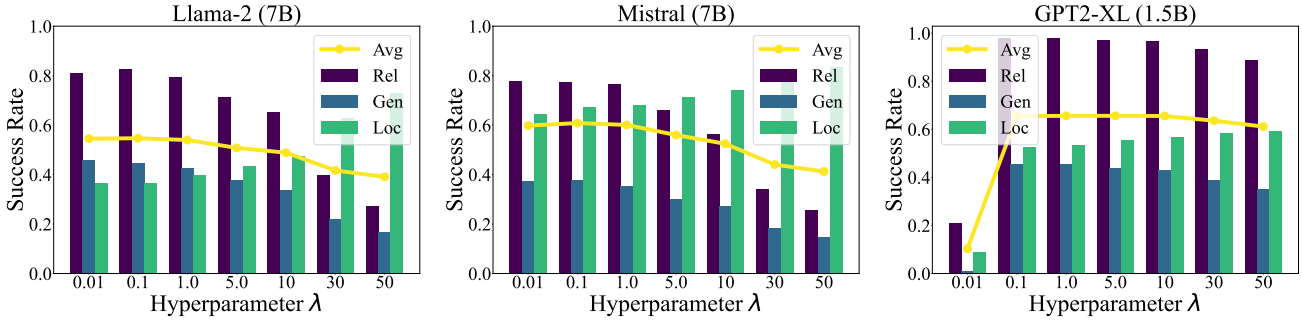


(b) MEMIT+SimIE on the Counterfact dataset

Figure 9. Performance of MEMIT+SimIE across various λ values, under $T = 1000$ sequential edits.



(a) AlphaEdit$^-$+SimIE on the ZsRE dataset



(b) AlphaEdit$^-$+SimIE on the Counterfact dataset

Figure 10. Performance of AlphaEdit$^-$+SimIE across various λ values, under $T = 1000$ sequential edits.

D.3. Further analyses

D.3.1. Case study

After 1000 sequential edits on Llama-2 using the ZsRE dataset, we select several edit examples for analyzing the output of the model. Table 6 presents the output results before and after applying SimIE. We observe that all standard methods consistently produced invalid outputs, aligning with the phenomenon reported in Ma et al. (2024), where LLMs lose their general abilities after sequential editing. In contrast, standard methods enhanced with SimIE generated meaningful responses that closely matched the desired outputs. These findings further highlight the effectiveness of SimIE, especially in preserving the general capabilities of LLMs.

Table 6. Output of the model after 1000 sequential edits on Llama-2 using the ZsRE dataset. mmmmm denotes invalid outputs (e.g., gibberish or empty responses), and errors within the output are highlighted in red.

[Case 168]	**Prompt:** The astronomical body that Lacus Aestatis was located on was what? **Answer:** Moon $\Longrightarrow$ **Edit Target:** Mars
MEND	**Output:** mmmmm ✗
+SimIE	**Output:** Mars ✓
ROME	**Output:** 4 ✗
+SimIE	**Output:** Mars ✓
MEMIT	**Output:** mmmmm ✗
+SimIE	**Output:** Mars ✓
AlphaEdit$^-$	**Output:** mmmmm ✗
+SimIE	**Output:** Mars ✓
[Case 514]	**Prompt:** What college did Tatiana Vladislavovna Petrova go to? **Answer:** Moscow State University $\Rightarrow$ **Edit Target:** Moscow State Institute of International Relations
MEND	**Output:** mmmmm ✗
+SimIE	**Output:** Moscow State Institute of International Relations ✓
ROME	**Output:** mmmmm ✗
+SimIE	**Output:** Moscow State University of International Relations ✓✓
MEMIT	**Output:** mmmmm ✗
+SimIE	**Output:** Moscow State University of International Relations ✓✓
AlphaEdit$^-$	**Output:** mmmmm ✗
+SimIE	**Output:** Moscow State Institute of International Relations ✓
[Case 692]	**Prompt:** Which series is The Adventure of the Blanched Soldier a part of? **Answer:** The Case-Book of Sherlock Holmes $\Longrightarrow$ **Edit Target:** The Memoirs of Sherlock Holmes
MEND	**Output:** mmmmm ✗
+SimIE	**Output:** The Memoirs of Sherlock Holmes ✓
ROME	**Output:** mmmmm ✗
+SimIE	**Output:** The Memoirs of Sherlock Holmes ✓
MEMIT	**Output:** mmmmm ✗
+SimIE	**Output:** The Memoirs of Sherlock Holmes ✓
AlphaEdit$^-$	**Output:** mmmmm ✗
+SimIE	**Output:** The Memoirs of Sherlock Holmes ✓

D.3.2. HIDDEN REPRESENTATION VISUALIZATION

To gain deeper insights into the effectiveness of SimIE, we analyze hidden representations of the model through the following steps: 1) Using the ZsRE dataset, we conduct 1000 sequential edits on Llama-2, retaining three post-edit models obtained from standard algorithms, standard algorithms enhanced with SimIE, and the corresponding ideal editor. 2) For each post-edit model, we compute the values of the **last subject token** for all edit examples, resulting in 1000 hidden representations[5] of dimension 4096. 3) The stacked hidden representations are visualized using UMAP (McInnes et al., 2018), with their respective 2-σ confidence ellipses plotted for clarity.

The visualization results are presented in Figure 11. We observe that the representation distributions of standard algorithms deviate significantly from those of the ideal editor, indicating unsuccessful knowledge injection. In contrast, the representation distributions obtained after applying SimIE are closely consistent with those of the ideal editor, providing further explanation for the superior performance of SimIE.

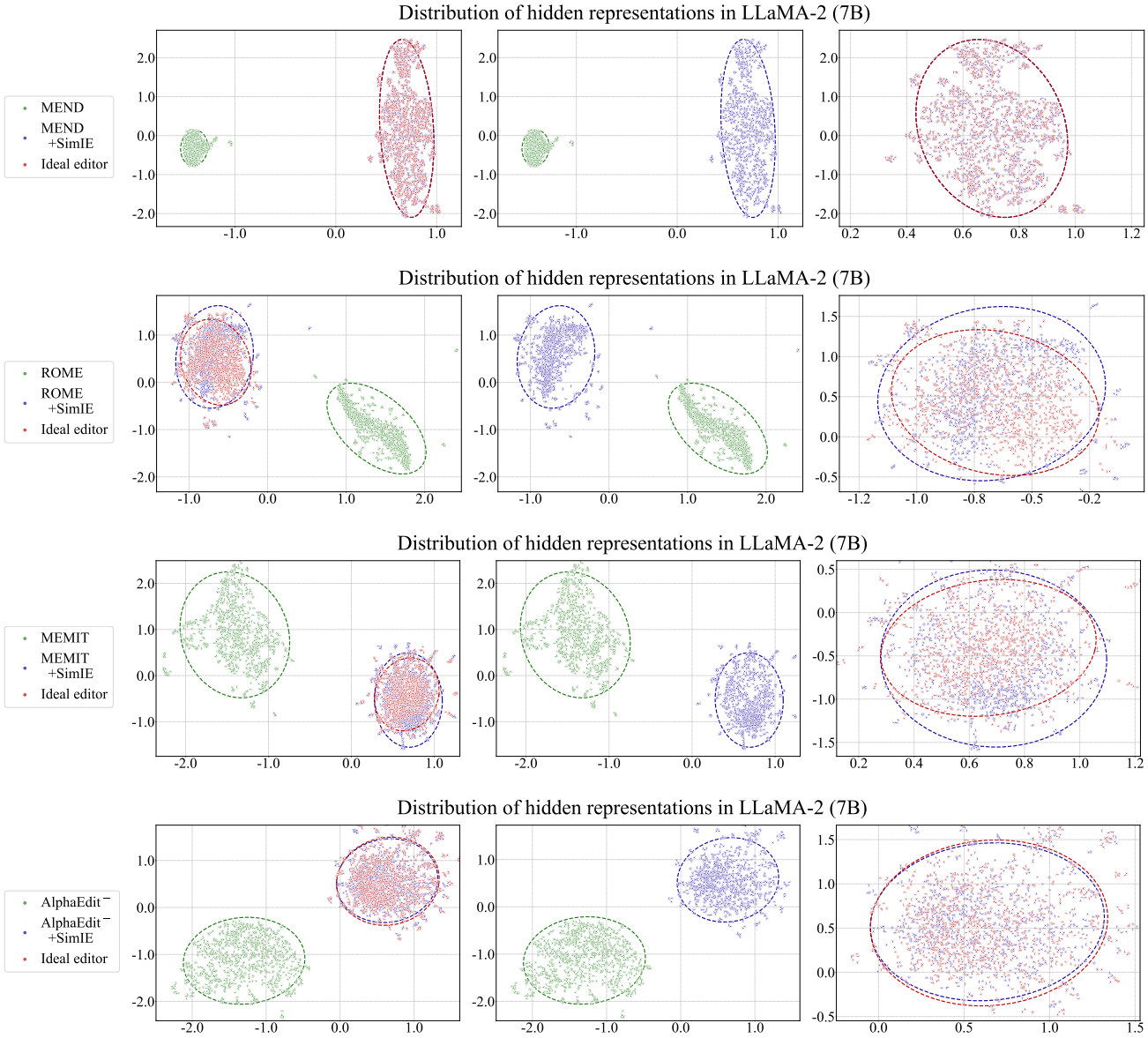


Figure 11. Distribution of hidden representations in the post-edit Llama-2, computed across 1000 edited examples from the ZsRE dataset. Dimensionality reduction performed via UMAP, with dashed lines representing 2-σ confidence ellipses.

[5]MEND corresponds to the 31st layer, which is its final editing layer. ROME, MEMIT, and AlphaEdit$^-$ correspond to the 5th layer, which is their shared editing layer.

D.3.3. LONG SENTENCE EDITING

To evaluate the effectiveness of SimIE in the editing of long sentences, we consider the task of correcting hallucinations in the SelfCheckGPT (Manakul et al., 2023). The **Hallucination** (Hartvigsen et al., 2024) dataset consists of highly inaccurate sentences sourced from GPT-3 (Brown et al., 2020), replaced with their corresponding sentences from actual Wikipedia entries. Compared to the ZsRE and Counterfact datasets, the Hallucination dataset is more challenging because it represents authentic mistakes made by high-quality LLMs, and the token length of edits significantly exceeds those of past datasets (see Figure 12 for an example). To reduce VRAM usage, we employ a simplified version in Wang et al. (2024b), limiting tokenized lengths to 254, resulting in 600 test samples. We perform 600 sequential edits on GPT2-XL. ROME, AlphaEdit$^-$ (standard methods), PRUNE, and AlphaEdit (lifelong methods) are selected as baselines for their superior performance on GPT2-XL. Following prior work (Hartvigsen et al., 2024; Wang et al., 2024b), we evaluate the performance using PPL (perplexity) and Loc (Localization Success Rate) as metrics.


```
{
    "prompt": "This is a Wikipedia passage about carole gist. Carole Gist (born
    ↪  April 28, 1969) is an American beauty pageant titleholder from Detroit,
    ↪  Michigan who was crowned Miss USA 1990. She was the first
    ↪  African-American woman to win the Miss USA title. Gist represented the
    ↪  United States at the Miss Universe 1990 pageant held in Los Angeles,
    ↪  California, where she placed first runner-up to Mona Grudt of Norway.
    ↪  Gist was the first African-American woman to place in the Miss Universe
    ↪  pageant.",
    "target_new": "She was also the first contestant from Michigan to win Miss
    ↪  USA, and broke the five-year streak of winners from Texas.",
    "subject": "carole gist",
    "locality_prompt": "Description Map of South America.\nThis map has a small
    ↪  scratch near the centerfold in the right part of the map.\nLooking for an
    ↪  antique map, historica",
    "locality_ground_truth": "l print or plan? Feel welcome and browse our
    ↪  mapsite atlasandmap.com! We have maps, made by Kuyper (Kuijper) . more
    ↪  maps of South America like Zuid-Am"
}
```


Figure 12. A sample of the Hallucination dataset.

Table 7 presents the results across varying numbers of edits T. As T increases, standard algorithms inevitably degrade, with rising PPL and declining Loc. Even methods specifically designed for lifelong scenarios exhibit elevated PPL at $T = 600$. In contrast, for ROME, SimIE reduces its PPL from 69.86 to 15.46 and improves its Loc by 33%. Similarly, AlphaEdit$^-$ enhanced with SimIE outperforms its original version, AlphaEdit, on both metrics. These results confirm the effectiveness of SimIE in editing long sentences.

Table 7. Performance of algorithms on the GPT2-XL using the Hallucination dataset, with the best results highlighted in **bold**.

	GPT2-XL (1.5B) using Hallucination													
	$T = 1$		$T = 50$		$T = 100$		$T = 200$		$T = 300$		$T = 500$		$T = 600$	
Algorithm	PPL (↓)	Loc (↑)	PPL (↓)	Loc (↑)	PPL (↓)	Loc (↑)	PPL (↓)	Loc (↑)	PPL (↓)	Loc (↑)	PPL (↓)	Loc (↑)	PPL (↓)	Loc (↑)
FT-L	60.93	0.50	>1000	0.08	>1000	0.06	>1000	0.05	>1000	0.07	>1000	0.04	>1000	0.04
ROME	2.44	**1.00**	8.19	0.94	21.71	0.90	40.38	0.84	45.76	0.80	63.14	0.70	69.86	0.64
+SimIE	2.74	0.98	**2.33**	0.94	**3.10**	0.91	**4.45**	0.88	**5.55**	0.87	**15.74**	0.85	**15.46**	0.85
AlphaEdit$^-$	1.87	**1.00**	5.92	**0.98**	34.07	**0.96**	61.08	0.88	105.95	0.74	541.81	0.23	2611.20	0.11
+SimIE	5.90	**1.00**	7.90	0.96	7.93	0.91	8.74	0.91	11.00	0.90	23.66	0.89	24.00	0.89
PRUNE	2.44	**1.00**	5.60	0.96	8.32	0.95	11.66	0.93	14.08	**0.92**	32.48	**0.92**	50.25	**0.92**
AlphaEdit	**1.86**	**1.00**	3.30	**0.98**	8.20	**0.96**	12.54	**0.94**	18.54	**0.92**	26.98	0.89	47.39	0.87