Editing is a Bargaining Game: Balanced Knowledge Editing in Large Language Models

Chenghao Xu[1], Jiexi Yan[2], Muli Yang[3], Fen Fang[3], Huilin Chen[4*], Cheng Deng[1*]

[1]School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China
[2]School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China
[3]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[4]School of Foreign Languages, Xidian University, Xi'an, Shaanxi, China
chx@stu.xidian.edu.cn, {jxyan1995,muliyang.xd,chdeng.xd}@gmail.com, fang_fen@a-star.edu.sg, hlchen@xidian.edu.cn

Abstract

Large Language Models (LLMs) are prone to generating incorrect or outdated information, thereby necessitating efficient and precise mechanisms for knowledge updates. Existing knowledge editing approaches, however, often encounter conflicts between two competing objectives: maintaining existing knowledge (preservation) and incorporating new information (editing). During gradient-based optimization, these conflicting objectives can lead to imbalanced update directions, where one gradient dominates, ultimately resulting in suboptimal learning dynamics. To address this challenge, we propose a balanced knowledge editing framework inspired by Nash bargaining theory. Our method guides the optimization process toward a Pareto stationary point, ensuring an equilibrium solution wherein any deviation from the final state would degrade the overall performance with respect to both objectives. This guarantees optimality in preserving prior knowledge while integrating new information. We empirically validate the effectiveness of our approach across a range of evaluation metrics on standard benchmark datasets. Extensive experiments show that our method consistently outperforms state-of-the-art techniques, achieving a superior balance between knowledge preservation and update accuracy.

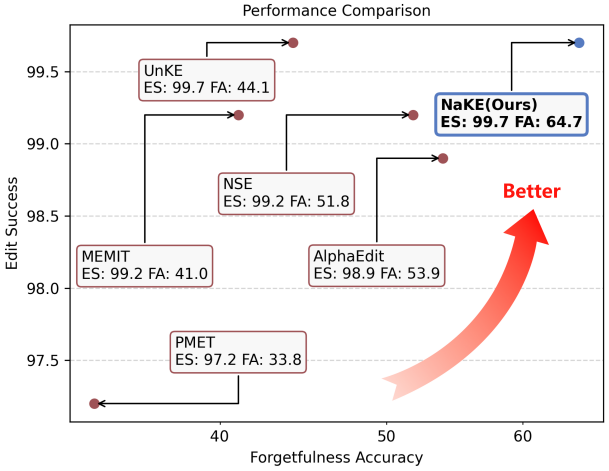

Figure 1: Comparison of preservation and editing performance across various knowledge editing methods. A higher Edit Success score indicates superior editing effectiveness, while a higher Forgetfulness Accuracy score reflects better preservation of unrelated knowledge.

Introduction

Large language models (LLMs) have demonstrated remarkable capabilities by acquiring and retaining extensive knowledge through large-scale training (Zhao et al. 2023; Brown et al. 2020; Radford et al. 2019; Bi et al. 2025b; Xu et al. 2024a). However, they are prone to hallucinations, often generating incorrect or outdated information (De Cao, Aziz, and Titov 2021; Mitchell et al. 2021). For example, when asked "Where were the latest Olympics held?", an LLM may respond with the outdated answer "Tokyo" rather than the correct, up-to-date answer "Paris" (Meng et al. 2022a,b). This highlights the critical importance of maintaining the accuracy and currency of the knowledge embedded within LLMs (Xu, Yan, and Deng 2025), particularly given their broad deployment in real-world applications (Xu et al. 2024c; Chen and Shu 2024). To address this challenge, model editing has emerged as a promising solution that enables targeted and efficient updates to a model's knowledge without requiring full retraining or fine-tuning, thereby offering a viable approach to accommodate dynamic and evolving information (Yao et al. 2023; Zhang et al. 2024; Cheng et al. 2023; Youssef et al. 2025; Chen et al. 2024; Lyu et al. 2025).

The core goal of knowledge editing is to revise specific factual information within LLMs without full retraining, thereby ensuring the accuracy and timeliness of the stored knowledge. To this end, various knowledge editing approaches have been proposed that aim to update the target information while preserving the integrity of unrelated knowledge within the model (Cheng et al. 2023; Mao et al. 2024; Bi et al. 2025a). For instance, the widely-used locate-then-edit paradigm first identifies the influential parameters associated with the target knowledge through causal tracing, and subsequently modifies them by introducing perturbations (Meng et al. 2022a). The primary objective is to min-

*Corresponding author.

imize the output error on the knowledge to be updated. Additionally, the output error on the knowledge to be preserved is typically included as a constraint in the optimization objective, ensuring that the model maintains accuracy on unaffected knowledge.

Broadly speaking, existing knowledge editing methods pursue two primary objectives: (1) accurately updating the model's knowledge for specific data points (editing), and (2) maintaining the model's performance on the remaining knowledge (preserving). This is typically formulated as the minimization of a weighted sum of two loss terms—one corresponding to the modification of the target knowledge and the other enforcing the retention of the non-target knowledge. When dealing with extensive forms of knowledge, directly optimizing the joint objective via a closed-form solution becomes infeasible. Consequently, gradient descent is typically adopted to iteratively minimize the two-term objective comprising both editing and preservation components. However, during this optimization process, the gradients corresponding to the two objectives may conflict in direction or differ significantly in magnitude. Such disparities can cause one gradient to dominate the update direction, resulting in suboptimal learning dynamics. As illustrated in Figure 1, frequent gradient conflicts and dominance can hinder effective parameter updates, ultimately resulting in performance degradation. Existing methods struggle to strike a balance between knowledge preservation and effective editing. Alleviating these issues is therefore critical for improving the overall efficacy of knowledge editing methods across both editing and preservation goals.

In this paper, we propose a novel balanced knowledge editing framework based on Nash bargaining (Nash 1953), termed NaKE, which simultaneously addresses the issues of gradient conflict and dominance by leveraging concepts from game theory. Drawing inspiration from Nash's bargaining solution—an axiomatic approach that guarantees a unique and proportionally fair outcome, where any deviation leads to a negative average relative change—we formulate the combination and balancing of gradients from the editing and preservation objectives as a cooperative bargaining game (Thomson 1994). In this setting, the two players—representing (Szép and Forgó 2012) the editing and preserving components—submit gradient proposals and negotiate to reach a consensus on an update direction that maximizes the joint benefit, ensuring a balanced and effective optimization process. Specifically, we define the utility function for each player based on their respective gradient and, leveraging the Nash bargaining framework, derive a convex optimization-based solution for the update direction that steers the model toward a stationary point on the Pareto front. Our proposed method demonstrates superior performance compared to existing baseline approaches across multiple evaluation metrics on standard benchmarks.

Related Work

Knowledge editing techniques can be broadly categorized based on whether they modify existing parameters or preserve them by introducing auxiliary mechanisms. Approaches that directly alter a model's parameters aim to encode new information by fine-tuning a minimal subset of weights. These include meta-learning strategies that employ hypernetworks to generate parameter updates (Jiang et al. 2024), often optimized for efficiency using low-rank gradient approximations (Mitchell et al. 2021). Locate-then-edit frameworks, on the other hand, utilize causal attribution to identify knowledge-relevant components (Meng et al. 2022a) and refine them via closed-form solutions such as least-squares optimization (Zheng et al. 2023). Other efforts mitigate catastrophic forgetting by restricting updates to select neurons (Jiang et al. 2024) or projecting edits into a null space, thereby extending applicability to lifelong learning settings (Fang et al. 2024).

In contrast, parameter-preserving methods maintain the integrity of the original model weights by allocating additional components (Hartvigsen et al. 2023). Some rely on in-context learning to inject new knowledge without modifying any parameters (Zheng et al. 2023; Bi et al. 2024), while others incorporate external memory retrieval mechanisms (Mitchell et al. 2022) or dynamically introduce new neurons to represent edited content (Huang et al. 2023; Dong et al. 2022; Bi et al. 2025c). Further innovations substitute internal hidden states with entries from a discrete codebook or enhance memory-based fusion through learned parameterized modules (Wang et al. 2024).

Recent research increasingly addresses the challenge of editing unstructured knowledge—namely, information expressed in free-form text rather than structured triples (Deng et al. 2025). To this end, Wu et al. (2024) critique the limitations of previous benchmarks and proposes AKEW, designed specifically to assess unstructured knowledge editing. Building upon the locate-then-edit paradigm, UnKE (Deng et al. 2025) updates all parameters within a single layer to better accommodate unstructured content, and is evaluated using the newly proposed UnKEBench. Meanwhile, DEM (Huang et al. 2024) introduces a dynamic perception module that localizes commonsense knowledge representations, enabling precise updates from textual descriptions. Expanding the generality of model editing, AnyEdit supports the modification of heterogeneous textual forms beyond factual assertions, facilitating broader and more versatile knowledge manipulation (Jiang et al. 2025).

Method

In this section, we first formulaically introduce the knowledge editing task and highlight the limitations of existing methods, followed by a detailed description of the proposed NaKE for balanced knowledge editing.

Preliminary

Transformer Block-Level Key-Value Pair Representations. Inspired by the phenomenon of "early decoding" (Yao et al. 2024), recent studies on transformer block-level key-value pair representations (Deng et al. 2025) suggest that the shallow layers of LLMs primarily encode key vectors that capture entity information and contextual concepts relevant to the input. In contrast, the deeper layers function as decoders, transforming these key vectors into

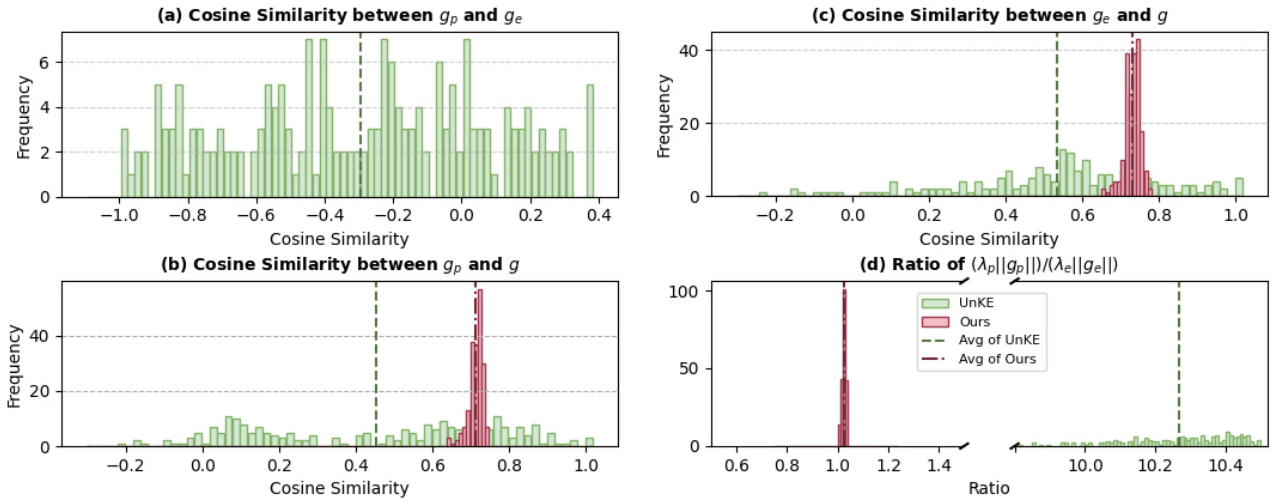

Figure 2: **A simplified illustration of gradient conflict and dominance phenomena observed during the gradient-based optimization process in knowledge editing.** The cosine similarity distributions reveal distinct gradient alignment patterns between the proposed method and UnKE (Deng et al. 2025) employing naive gradient descent. The naive gradient descent approach exhibits substantial gradient conflicts, as evidenced by the high frequency of negative cosine similarity values, indicating frequent misalignment between preservation and editing task gradients. Furthermore, this approach demonstrates persistent gradient conflicts between the preservation task gradients and the joint update direction, potentially impeding effective preservation. Conversely, our method mitigates these alignment issues, as validated by elevated cosine similarity values between the joint update gradient and both preservation and editing task gradients. Our method successfully achieves balanced contributions from both objectives.

value vectors that inject target-specific information into the residual stream. Specifically, we designate the L-th layer of the LLM as a functional boundary, partitioning the model into two distinct modules: a key generator and a value generator, which are responsible for producing key vectors and value vectors, respectively.

For a given knowledge pair (x, y), the key vector k should be expressed as $k = h^L_{x,n}$, where $h^L_{x,n}$ denotes the hidden state of n-th token of x in the L-th layer.

Computing Key-Value. For editing knowledge $(\tilde{x}, \tilde{y})$, we compute its corresponding key-value pair (k^*, v^*). $k^* = k + \delta_n$ can be directly derived by optimizing the residual vector δ_n using gradient descent. This process can be formalized as follows:

$$k^* = k + \arg\min_{\delta_n} \left(-\log \mathbb{P}_{f^L_{\theta_L}(k+\delta_n)}(\tilde{y}|\tilde{x}) \right), \qquad (1)$$

where $f^L_{\theta_L}(k + \delta_n)$ represents the L-th layer output when the original key k is replaced with $k^* + \delta_n$. By freezing the parameters of the value generator, optimizing Eq. (1) to a sufficiently low value implies that, given access to the optimal key vector $\boldsymbol{k}^*$, the model is capable of decoding the corresponding target output y_e.

Two-Term Objective Optimization. The key generator encodes a large set of key vectors during pre-training, which can be selectively activated by specific inputs to produce corresponding value vectors. Our objective is to modify the targeted key vectors to yield the desired edited output, while preserving the integrity of untargeted key-value pairs. To prevent the introduction of new keys from interfering with the generation of existing ones, we restrict the optimization to the final layer of the key encoder, *i.e.,* $f^L_{\theta^L}$. The problem is reformulated as maintaining a factual set comprising u newly introduced associations $\{\tilde{x}_j, \tilde{y}_j\}_{j=1}^u$, while simultaneously preserving the original set containing w existing associations $\{\bar{x}_i, \bar{y}_i\}_{i=1}^w$ as follows:

$$\hat{\theta}_L \triangleq \arg\min_{\theta_L} (\overbrace{\sum_{i=1}^{w} \| f^L_{\theta_L}(\bar{h}^{L-1}_i) - k_i \|_2^2}^{\text{Key Preserving Loss } \mathcal{L}_p} + \underbrace{\sum_{j=1}^{u} \| f^L_{\theta_L}(\tilde{h}^{L-1}_j) - k^*_j \|_2^2}_{\text{Key Editing Loss } \mathcal{L}_e}), \qquad (2)$$

where $\bar{h}^{L-1}_i$ and $\tilde{h}^{L-1}_j$ represent the hidden state of $\bar{x}_i$ and $\tilde{x}_j$ in the $L-1$-th layer, respectively.

Gradient Conflict and Dominance Phenomenon

Gradient descent is commonly employed to iteratively optimize Eq.(2), which integrates both editing loss $\mathcal{L}_e$ and the preserving loss $\mathcal{L}_p$. However, throughout this optimization process, the gradients associated with the two objectives may exhibit conflicting directions or substantial differences in magnitude (Sener and Koltun 2018; Yu et al. 2020). These discrepancies can cause one gradient to dominate overall update direction, leading to suboptimal optimization behavior.

To better analyze this phenomenon, we rewrite Eq. (2)

with two coefficients for balancing terms as follows:

$$\min_{\theta_L} \lambda_p \mathcal{L}_p(\theta_L; \mathcal{W}) + \lambda_e \mathcal{L}_e(\theta_L; \mathcal{U}), \tag{3}$$

where $\mathcal{U}$ and $\mathcal{W}$ denote the association sets to be edited and preserved, respectively. Let $g_p = \nabla_{\theta_L} \mathcal{L}_p(\theta_L; \mathcal{W})$ and $g_e = \nabla_{\theta_L} \mathcal{L}_e(\theta_L; \mathcal{U})$ denote the gradient for updating these two terms. We conduct a series of experiments to analyze the alignment and consistency among g_p, g_e, and the joint update direction $g = \lambda_p g_p + \lambda_e g_e$, with the corresponding results visualized in Figure 2.

We observe a high frequency of negative cosine similarity values between the gradients of the preserving task ($\mathbf{g}_r$) and the editing task ($\mathbf{g}_f$) during naïve gradient descent optimization. This indicates that the two gradients are often misaligned, reflecting a substantial degree of gradient conflict between the objectives. Additionally, the gradients of the preserving and editing objectives are often misaligned with the direction of the joint update, which may impede the effectiveness of both knowledge preservation and modification during optimization.

Furthermore, we analyze the ratio of gradient norms between the two objectives, *i.e.*, $\lambda_p \|g_p\| / \lambda_e \|g_e\|$ and observe a significant imbalance in gradient magnitudes during naïve gradient descent optimization (Xu et al. 2024b), which can lead to one task disproportionately influencing the joint update direction, i.e., the dominance phenomenon. Therefore, we propose a novel optimization strategy aimed at improving gradient alignment and balancing the contributions of the editing and preserving objectives. This approach effectively mitigates gradient conflicts and enhances the model's capacity to incorporate targeted knowledge updates without substantially impairing the retention of existing information.

Nash Bargaining Knowledge Editing

Motivated by the observed issues of gradient conflict and dominance during optimization, we adopt a cooperative bargaining perspective and introduce a novel optimization method aimed at achieving more reliable improvements by explicitly resolving such conflicts throughout training. Specifically, our approach is inspired by the Nash Bargaining Solution (NBS) (Nash 1953), a foundational concept in axiomatic game theory that is widely regarded for its robustness and general applicability. NBS possesses desirable theoretical properties, including Pareto optimality, which prevents unconsented unilateral gains, and a principled mechanism for balancing competing interests, making it well-suited for harmonizing the objectives of editing and preserving in knowledge editing.

Nash Bargaining Problem Definition. Inspired by existed works (Navon et al. 2022; Wu and Harandi 2025), the central objective is to determine the weighting protocol $\lambda = [\lambda_p, \lambda_e]$, which specifies the relative contributions of the individual losses in Eq. (2). This protocol guides the update step in a manner that facilitates joint optimization by improving the aggregated loss across both tasks. We formulate this problem as a cooperative bargaining game between two tasks—editing and preserving—each represented as a distinct player. In this setting, both players contribute their respective gradients with the shared objective of maximizing the overall utility. Inspired by (Zeng et al. 2024), the utility function for each player is defined as follows:

$$\begin{aligned} \text{Preserving:} \quad & \upsilon_p(g) \coloneqq g_p^\top g, \\ \text{Editing:} \quad & \upsilon_e(g) \coloneqq g_e^\top g, \end{aligned} \tag{4}$$

For the preserving task, $\upsilon_p(g)$ quantifies the alignment between the update direction g and the gradient that minimizes the loss over the preserved set $\mathcal{W}$; for the editing task, $\upsilon_e(g)$ assesses the alignment between g and the gradient that increases the loss over the edited set $\mathcal{U}$. Accordingly, if the final update direction g deviates substantially from the preserving gradient g_p or the editing gradient g_e, the respective utility (or payoff) decreases. Given that this setup is framed as a cooperative game, it is reasonable to assume that neither player would act to the detriment of the other without individual gain. As such, the negotiated solution is expected to be Pareto-efficient, implying that it should not be dominated by any alternative and thus converges to a Pareto stationary point.

Following (Zeng et al. 2024), provided that the two subgradients are not entirely antagonistic, there exists an update direction g capable of simultaneously decreasing the losses associated with both tasks. Therefore, our objective is to identify an updating direction g that maximizes the overall improvement across the two objectives. Based on this insight, we reformulate the optimization problem in Eq. (1) as follows:

$$\max_{g \in \mathbb{B}_\epsilon} \log(\upsilon_p(g)) + \log(\upsilon_e(g)), \tag{5}$$

where the update vector g is constrained to lie within a norm ball $\mathbb{B}_\epsilon$ of radius ϵ centered at the origin. Here, the logarithmic function is employed to balance the utilities and reflect the diminishing marginal gains property—i.e., the utility improvement becomes less significant as it increases. Under this objective, the optimization converges to a Pareto stationary point.

Algorithm 1: Nash Bargaining Knowledge Editing

Input: Model parameters θ_L, preserving and editing task losses $\mathcal{L}_p, \mathcal{L}_e$, number of iterations T, learning rate η.

Output: Final parameters $\hat{\theta}_L$.

Initialize $\Lambda^{(0)} = [\lambda_p^0, \lambda_e^{(0)}]^\top$.
Let $t = 1$.
while $t \leq T$ **do**
  Compute the gradients of per player:
  $g_p^{(t)} = \nabla_{\theta_L^{(t)-1}} \mathcal{L}_p(\theta_L^{(t-1)}; \mathcal{W})$,
  $g_e^{(t)} = \nabla_{\theta_L^{(t)-1}} \mathcal{L}_e(\theta_L^{(t-1)}; \mathcal{U})$.
  Set $G^{(t)} = [g_p^{(t)}, g_e^{(t)}]$.
  Solve for $(G^{(t)})^\top G^{(t)} \Lambda = 1/\Lambda$ to obtain $\Lambda^{(t)}$.
  Update the parameters $\theta_L^{(t)} = \theta_L^{(t)} - \eta G^{(t)} \Lambda^{(t)}$.
end while
return $\theta_L^{(T)}$

LLMs	Method	Edit Succ.↑	Portability ↑			Locality ↑		Fluency ↑
			SAA	LGA	RA	RSA	FA	
GPT-J	FT	$64.2_{\pm 1.6}$	$47.3_{\pm 2.0}$	$7.1_{\pm 1.9}$	$21.3_{\pm 2.9}$	$4.4_{\pm 0.6}$	$6.4_{\pm 1.3}$	$304.1_{\pm 7.6}$
	LoRA	$\mathbf{100.0}_{\pm 0.0}$	$75.2_{\pm 1.9}$	$\mathbf{22.2}_{\pm 3.1}$	$40.3_{\pm 2.8}$	$25.7_{\pm 1.6}$	$51.4_{\pm 2.8}$	$595.8_{\pm 4.1}$
	KN	$18.1_{\pm 2.4}$	$17.9_{\pm 2.4}$	$10.8_{\pm 2.6}$	$18.5_{\pm 2.2}$	$80.2^{*}_{\pm 1.3}$	$80.6^{*}_{\pm 1.5}$	$580.0_{\pm 3.8}$
	ROME	$99.2_{\pm 0.5}$	$74.1_{\pm 2.2}$	$16.1_{\pm 2.6}$	$29.2_{\pm 2.4}$	$37.4_{\pm 1.3}$	$33.1_{\pm 2.6}$	$\underline{600.0}_{\pm 3.6}$
	MEMIT	$99.5_{\pm 0.5}$	$56.5_{\pm 2.5}$	$16.7_{\pm 2.6}$	$25.9_{\pm 2.1}$	$53.2_{\pm 1.4}$	$40.7_{\pm 2.8}$	$591.6_{\pm 4.3}$
	PMET	$95.3_{\pm 0.9}$	$54.1_{\pm 2.6}$	$16.6_{\pm 2.6}$	$25.3_{\pm 2.1}$	$47.6_{\pm 1.5}$	$36.8_{\pm 2.8}$	$\mathbf{600.3}_{\pm 3.6}$
	UnKE	$\underline{99.6}_{\pm 0.9}$	$86.9_{\pm 1.7}$	$17.1_{\pm 2.3}$	$36.1_{\pm 2.5}$	$51.7_{\pm 1.3}$	$40.1_{\pm 2.1}$	$591.2_{\pm 6.8}$
	AlphaEdit	$98.7_{\pm 0.5}$	$88.1_{\pm 2.5}$	$17.3_{\pm 2.1}$	$35.5_{\pm 2.5}$	$\underline{79.2}_{\pm 1.4}$	$\underline{55.7}_{\pm 1.6}$	$592.7_{\pm 4.8}$
	NSE	$99.1_{\pm 0.3}$	$\mathbf{90.3}_{\pm 1.9}$	$17.4_{\pm 2.0}$	$\underline{36.8}_{\pm 2.7}$	$73.6_{\pm 1.5}$	$51.8_{\pm 2.2}$	$584.8_{\pm 5.3}$
	Ours	$99.3_{\pm 0.1}$	$\underline{90.1}_{\pm 1.4}$	$\underline{17.9}_{\pm 2.7}$	$\mathbf{37.4}_{\pm 3.5}$	$\mathbf{81.7}_{\pm 1.7}$	$\mathbf{63.5}_{\pm 2.7}$	$594.1_{\pm 7.1}$
Qwen2.5-7B	FT	$49.0_{\pm 1.5}$	$46.3_{\pm 2.1}$	$15.3_{\pm 1.4}$	$29.3_{\pm 1.5}$	$21.6_{\pm 1.5}$	$30.1_{\pm 3.5}$	$493.8_{\pm 8.5}$
	LoRA	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{91.5}_{\pm 1.0}$	$\mathbf{31.4}_{\pm 3.5}$	$\mathbf{46.1}_{\pm 2.1}$	$71.2_{\pm 1.2}$	$50.1_{\pm 2.5}$	$564.2_{\pm 5.3}$
	KN	$20.5_{\pm 2.9}$	$22.1_{\pm 2.6}$	$18.9_{\pm 2.4}$	$26.4_{\pm 1.9}$	$79.2^{*}_{\pm 1.6}$	$71.2^{*}_{\pm 4.6}$	$568.6_{\pm 6.2}$
	ROME	$98.9_{\pm 0.4}$	$73.7_{\pm 2.0}$	$19.2_{\pm 3.5}$	$36.2_{\pm 2.7}$	$49.1_{\pm 1.5}$	$38.6_{\pm 2.4}$	$579.8_{\pm 3.9}$
	MEMIT	$98.2_{\pm 0.8}$	$78.3_{\pm 2.2}$	$26.7_{\pm 2.7}$	$39.3_{\pm 2.5}$	$47.2_{\pm 1.3}$	$42.7_{\pm 2.4}$	$575.8_{\pm 4.2}$
	PMET	$96.2_{\pm 1.1}$	$58.6_{\pm 2.5}$	$28.1_{\pm 3.2}$	$32.5_{\pm 2.6}$	$60.1_{\pm 1.9}$	$51.2_{\pm 2.9}$	$577.4_{\pm 5.2}$
	UnKE	$\underline{99.4}_{\pm 0.8}$	$76.8_{\pm 1.8}$	$27.1_{\pm 2.4}$	$35.3_{\pm 2.9}$	$55.3_{\pm 1.8}$	$48.6_{\pm 2.1}$	$\underline{588.4}_{\pm 5.9}$
	AlphaEdit	$98.9_{\pm 0.4}$	$75.2_{\pm 2.4}$	$26.9_{\pm 2.7}$	$36.7_{\pm 2.7}$	$\underline{72.9}_{\pm 3.2}$	$\underline{60.8}_{\pm 1.7}$	$576.1_{\pm 3.8}$
	NSE	$99.0_{\pm 0.6}$	$81.3_{\pm 2.8}$	$27.7_{\pm 1.8}$	$37.5_{\pm 2.3}$	$61.8_{\pm 1.2}$	$53.7_{\pm 2.3}$	$561.2_{\pm 5.4}$
	Ours	$99.1_{\pm 0.1}$	$\underline{85.3}_{\pm 1.9}$	$\underline{28.6}_{\pm 2.5}$	$\underline{40.3}_{\pm 3.0}$	$\mathbf{86.7}_{\pm 2.1}$	$\mathbf{66.8}_{\pm 2.6}$	$\mathbf{591.0}_{\pm 6.1}$
Llama3-8B	FT	$47.2_{\pm 1.9}$	$48.6_{\pm 1.6}$	$8.4_{\pm 1.6}$	$25.6_{\pm 2.0}$	$27.1_{\pm 1.8}$	$12.0_{\pm 1.7}$	$379.2_{\pm 11.7}$
	LoRA	$\mathbf{100.0}_{\pm 0.0}$	$78.1_{\pm 1.5}$	$\underline{24.4}_{\pm 3.6}$	$\mathbf{44.2}_{\pm 3.1}$	$15.6_{\pm 1.0}$	$24.9_{\pm 2.4}$	$471.8_{\pm 10.1}$
	KN	$16.8_{\pm 2.0}$	$18.2_{\pm 2.0}$	$14.6_{\pm 2.4}$	$19.8_{\pm 2.0}$	$83.7^{*}_{\pm 1.0}$	$88.2^{*}_{\pm 2.2}$	$\underline{591.2}_{\pm 5.9}$
	ROME	$99.2_{\pm 0.2}$	$74.1_{\pm 2.2}$	$16.1_{\pm 2.6}$	$29.2_{\pm 2.4}$	$37.1_{\pm 1.7}$	$33.2_{\pm 2.3}$	$590.1_{\pm 3.6}$
	MEMIT	$99.2_{\pm 0.4}$	$73.5_{\pm 2.7}$	$\mathbf{24.5}_{\pm 2.2}$	$32.1_{\pm 2.3}$	$41.2_{\pm 1.5}$	$41.0_{\pm 2.5}$	$568.9_{\pm 6.9}$
	PMET	$97.2_{\pm 0.8}$	$55.9_{\pm 2.7}$	$24.1_{\pm 2.4}$	$34.3_{\pm 2.1}$	$44.7_{\pm 2.1}$	$33.8_{\pm 2.1}$	$\mathbf{598.6}_{\pm 3.2}$
	UnKE	$\underline{99.7}_{\pm 0.8}$	$85.8_{\pm 2.3}$	$17.1_{\pm 2.3}$	$35.1_{\pm 2.5}$	$48.6_{\pm 1.7}$	$44.1_{\pm 2.0}$	$589.7_{\pm 5.7}$
	AlphaEdit	$98.9_{\pm 0.4}$	$82.4_{\pm 2.1}$	$18.8_{\pm 2.7}$	$36.9_{\pm 3.0}$	$60.8_{\pm 1.9}$	$\underline{53.9}_{\pm 1.5}$	$575.4_{\pm 4.4}$
	NSE	$99.2_{\pm 0.5}$	$\underline{87.2}_{\pm 2.0}$	$21.4_{\pm 2.0}$	$29.1_{\pm 2.2}$	$\underline{73.6}_{\pm 1.5}$	$51.8_{\pm 2.2}$	$584.8_{\pm 5.3}$
	Ours	$\underline{99.7}_{\pm 0.2}$	$\mathbf{90.1}_{\pm 1.7}$	$24.2_{\pm 3.1}$	$\underline{39.6}_{\pm 2.9}$	$\mathbf{89.2}_{\pm 1.9}$	$\mathbf{64.7}_{\pm 1.9}$	$584.1_{\pm 4.1}$

Table 1: Editing Performance comparison on WikiData$_{Counterfact}$. * indicates invalid results. Locality results obtained under conditions of low Edit Success are deemed invalid, as locality trivially reaches 100% when the edit is not successfully applied. The best results are indicated as **Bold**, and the second ones are indicated as Underline.

Problem Solution. We now demonstrate that the Nash Bargaining Solution (NBS), up to a scaling factor, is attained at $g^* = \lambda_p g_p + \lambda_e g_e$, which constitutes a solution to Eq. (5). This result is formally established through the following two theorems, with complete proofs provided in the supplementary material.

Theorem 2.1. Denote $f(g) \coloneqq \log(\upsilon_p(g)) + \log(\upsilon_e(g))$. The optimal solution g^* of Eq. (5) is achieved at

$$\nabla f(g^*) = \alpha g^*, \quad \text{for some} \quad \alpha > 0. \tag{6}$$

Theorem 2.2. Let $G = [g_p, g_e]$, $\Lambda = [\lambda_p, \lambda_e]^\top$. The solution to Eq. (6) is (up to scaling) $g^* = \lambda_p g_p + \lambda_e g_e$ where

$$G^\top G \Lambda = \frac{1}{\Lambda}. \tag{7}$$

Based on Eq. (7), we derive the relationship between the individual contributions and the interaction terms as follows:

$$\begin{aligned} \|\lambda_p g_p\|_2^2 + (\lambda_p g_p)^\top (\lambda_e g_e) = 1, \\ \|\lambda_e g_e\|_2^2 + (\lambda_e g_e)^\top (\lambda_p g_p) = 1. \end{aligned} \tag{8}$$

The relative weights of preserving and editing λ_p and λ_e are determined by both a player's individual contribution ($\|\lambda_p g_p\|_2^2$) and their interactions with other players ($(\lambda_p g_p)^\top (\lambda_e g_e)$). This trade-off captures the tension between individual and collective rationality. When interactions are positive (i.e., $g_p^\top g_e > 0$), the weight λ_p is down-regulated to promote collective improvement. Conversely, negative interactions (*i.e.*, $g_p^\top g_e < 0$) lead to an increase in λ_p, prioritizing the player's own objective. Under mild assumptions, each player makes a non-negligible contribution to the aggregated update g. The resulting solution balances the incentives for individual participation with the resolution of inter-player conflicts, embodying the principles of game-theoretic bargaining.

We detail the procedure of our proposed method, NaKE in Algorithm 1. Specifically, we first compute the gradient associated with each player. Next, we determine the weighting coefficients λ_p and λ_e, and subsequently update the model parameters using the combined gradient direction derived from these coefficients. This update aims to maximize the overall utility gain across both objectives.

LLMs	Method	Edit Succ.↑	Portability ↑			Locality ↑		Fluency ↑
			SAA	LGA	RA	RSA	FA	
GPT-J	FT	$70.8_{\pm 1.5}$	$50.1_{\pm 2.7}$	$18.3_{\pm 1.9}$	$35.3_{\pm 2.0}$	$7.1_{\pm 0.3}$	$8.4_{\pm 1.8}$	$351.8_{\pm 6.2}$
	LoRA	$\mathbf{100.0}_{\pm 0.0}$	$81.6_{\pm 1.2}$	$\mathbf{35.2}_{\pm 2.8}$	$\mathbf{48.6}_{\pm 2.8}$	$28.3_{\pm 1.7}$	$22.4_{\pm 3.1}$	$\underline{591.8}_{\pm 3.8}$
	KN	$28.3_{\pm 3.1}$	$22.6_{\pm 3.2}$	$23.3_{\pm 3.1}$	$35.1_{\pm 1.6}$	$86.6^{*}_{\pm 1.3}$	$81.4^{*}_{\pm 1.2}$	$579.6_{\pm 3.4}$
	ROME	$99.5_{\pm 0.2}$	$84.6_{\pm 2.0}$	$28.3_{\pm 2.8}$	$36.9_{\pm 1.7}$	$37.3_{\pm 1.3}$	$51.0_{\pm 2.2}$	$\mathbf{596.8}_{\pm 2.8}$
	MEMIT	$99.6_{\pm 0.2}$	$68.9_{\pm 3.2}$	$27.2_{\pm 2.6}$	$32.4_{\pm 1.9}$	$49.6_{\pm 1.0}$	$52.7_{\pm 1.9}$	$585.1_{\pm 3.2}$
	PMET	$99.0_{\pm 0.4}$	$63.6_{\pm 3.6}$	$25.4_{\pm 2.8}$	$31.2_{\pm 2.0}$	$46.3_{\pm 1.0}$	$49.5_{\pm 2.4}$	$584.2_{\pm 3.0}$
	UnKE	$99.5_{\pm 0.3}$	$87.6_{\pm 1.8}$	$26.4_{\pm 2.4}$	$34.8_{\pm 1.8}$	$56.8_{\pm 0.7}$	$48.9_{\pm 2.3}$	$589.4_{\pm 3.8}$
	AlphaEdit	$99.1_{\pm 0.4}$	$67.4_{\pm 2.5}$	$\underline{28.9}_{\pm 1.8}$	$38.6_{\pm 2.5}$	$68.9_{\pm 1.2}$	$\underline{54.2}_{\pm 2.5}$	$578.3_{\pm 3.9}$
	NSE	$\underline{99.8}_{\pm 0.5}$	$\mathbf{94.2}_{\pm 1.2}$	$25.6_{\pm 2.3}$	$40.1_{\pm 1.7}$	$\underline{73.6}_{\pm 1.5}$	$51.8_{\pm 2.2}$	$565.4_{\pm 5.3}$
	Ours	$99.7_{\pm 0.3}$	$\underline{92.1}_{\pm 1.4}$	$26.9_{\pm 2.7}$	$\underline{44.5}_{\pm 3.5}$	$\mathbf{81.7}_{\pm 1.7}$	$\mathbf{58.3}_{\pm 2.1}$	$586.3_{\pm 3.0}$
Qwen2.5-7B	FT	$53.4_{\pm 1.7}$	$48.8_{\pm 1.9}$	$13.7_{\pm 1.8}$	$32.1_{\pm 1.7}$	$25.7_{\pm 1.1}$	$21.5_{\pm 2.4}$	$415.9_{\pm 6.7}$
	LoRA	$\mathbf{100.0}_{\pm 0.0}$	$89.1_{\pm 0.7}$	$32.8_{\pm 3.1}$	$41.6_{\pm 2.1}$	$65.1_{\pm 1.0}$	$41.6_{\pm 1.8}$	$571.1_{\pm 6.8}$
	KN	$21.6_{\pm 2.3}$	$26.2_{\pm 2.4}$	$21.5_{\pm 1.8}$	$25.7_{\pm 2.3}$	$81.3^{*}_{\pm 1.4}$	$67.3^{*}_{\pm 3.8}$	$566.3_{\pm 4.2}$
	ROME	$98.8_{\pm 0.6}$	$81.4_{\pm 1.6}$	$36.7_{\pm 2.2}$	$44.1_{\pm 1.8}$	$50.3_{\pm 1.4}$	$58.4_{\pm 1.7}$	$\mathbf{573.8}_{\pm 2.8}$
	MEMIT	$99.2_{\pm 0.3}$	$86.2_{\pm 1.4}$	$39.3_{\pm 3.4}$	$43.1_{\pm 1.6}$	$51.2_{\pm 1.0}$	$58.3_{\pm 1.8}$	$567.8_{\pm 3.4}$
	PMET	$97.6_{\pm 0.3}$	$69.3_{\pm 1.8}$	$36.6_{\pm 2.7}$	$46.8_{\pm 1.9}$	$61.3_{\pm 1.7}$	$65.3_{\pm 2.1}$	$571.8_{\pm 2.6}$
	UnKE	$99.5_{\pm 1.0}$	$87.1_{\pm 2.4}$	$\mathbf{42.1}_{\pm 2.1}$	$\underline{47.8}_{\pm 2.2}$	$54.2_{\pm 1.7}$	$59.4_{\pm 1.9}$	$566.5_{\pm 3.1}$
	AlphaEdit	$99.1_{\pm 0.5}$	$84.2_{\pm 1.6}$	$38.3_{\pm 1.8}$	$45.3_{\pm 2.3}$	$72.8_{\pm 2.4}$	$\underline{67.8}_{\pm 1.9}$	$\underline{573.3}_{\pm 3.5}$
	NSE	$\underline{99.6}_{\pm 0.8}$	$\underline{91.2}_{\pm 2.1}$	$37.5_{\pm 1.6}$	$47.2_{\pm 1.5}$	$\underline{78.6}_{\pm 1.6}$	$64.2_{\pm 2.1}$	$561.2_{\pm 2.3}$
	Ours	$99.7_{\pm 0.4}$	$\mathbf{92.3}_{\pm 1.3}$	$\underline{41.5}_{\pm 1.6}$	$\mathbf{49.9}_{\pm 1.5}$	$\mathbf{83.1}_{\pm 1.5}$	$\mathbf{74.2}_{\pm 2.4}$	$572.6_{\pm 3.6}$
Llama3-8B	FT	$51.3_{\pm 0.4}$	$49.2_{\pm 2.6}$	$10.5_{\pm 1.2}$	$25.6_{\pm 2.0}$	$31.0_{\pm 1.4}$	$18.2_{\pm 2.7}$	$431.9_{\pm 8.8}$
	LoRA	$\mathbf{100.0}_{\pm 0.0}$	$83.1_{\pm 2.0}$	$25.6_{\pm 3.1}$	$\underline{46.3}_{\pm 3.6}$	$16.2_{\pm 1.5}$	$21.3_{\pm 2.4}$	$491.8_{\pm 12.5}$
	KN	$20.3_{\pm 2.4}$	$17.3_{\pm 2.5}$	$17.3_{\pm 2.7}$	$20.3_{\pm 2.4}$	$81.1^{*}_{\pm 3.8}$	$86.5^{*}_{\pm 1.9}$	$589.7_{\pm 6.4}$
	ROME	$98.4_{\pm 0.4}$	$82.6_{\pm 1.7}$	$34.8_{\pm 2.9}$	$\mathbf{46.7}_{\pm 1.4}$	$49.3_{\pm 1.5}$	$52.1_{\pm 2.2}$	$581.3_{\pm 2.8}$
	MEMIT	$99.1_{\pm 0.4}$	$81.1_{\pm 2.5}$	$34.7_{\pm 2.5}$	$45.2_{\pm 1.5}$	$48.3_{\pm 2.1}$	$55.1_{\pm 1.8}$	$584.7_{\pm 2.6}$
	PMET	$98.1_{\pm 0.8}$	$58.7_{\pm 2.1}$	$\mathbf{37.3}_{\pm 2.4}$	$42.5_{\pm 1.9}$	$65.8_{\pm 1.7}$	$64.0_{\pm 1.4}$	$\underline{591.8}_{\pm 2.5}$
	UnKE	$\underline{99.8}_{\pm 0.8}$	$\underline{87.6}_{\pm 1.5}$	$33.8_{\pm 3.5}$	$44.8_{\pm 1.9}$	$51.3_{\pm 1.4}$	$57.2_{\pm 2.3}$	$\mathbf{593.8}_{\pm 3.3}$
	AlphaEdit	$98.9_{\pm 0.3}$	$82.3_{\pm 2.2}$	$35.7_{\pm 2.5}$	$43.9_{\pm 2.1}$	$\underline{75.3}_{\pm 1.8}$	$\underline{65.1}_{\pm 1.6}$	$584.4_{\pm 3.5}$
	NSE	$99.3_{\pm 0.6}$	$87.4_{\pm 2.3}$	$36.2_{\pm 3.0}$	$42.7_{\pm 1.9}$	$67.3_{\pm 1.7}$	$61.7_{\pm 2.0}$	$585.3_{\pm 2.9}$
	Ours	$99.7_{\pm 0.3}$	$\mathbf{89.2}_{\pm 1.8}$	$\underline{37.1}_{\pm 2.7}$	$44.7_{\pm 1.7}$	$\mathbf{85.7}_{\pm 1.5}$	$\mathbf{69.2}_{\pm 2.2}$	$591.0_{\pm 3.2}$

Table 2: Editing Performance comparison on WikiData$_{Recent}$. * indicates invalid results. Locality results obtained under conditions of low Edit Success are deemed invalid, as locality trivially reaches 100% when the edit is not successfully applied.

Experiments

In this section, we evaluate our proposed NaKE and analyze its essential characteristics. Additional implementation details, experimental results, and in-depth analyses are provided in the supplementary material.

Experimental Setup

LLMs and Baseline Methods. To comprehensively evaluate the performance of our model, we select three LLMs with distinct architectures: GPT-J (Wang and Komatsuzaki 2021), Qwen2.5-7B-Instruct (Yang et al. 2025), and LLaMA3-8B-Instruct (Dubey et al. 2024). For comparison with our method, we evaluated against several knowledge editing methods, including Fine-Tuning (FT), LoRA (Wu et al. 2023), Knowledge Neurons (KN) (Dai et al. 2021), ROME (Meng et al. 2022a), MEMIT (Meng et al. 2022b), PMET (Li et al. 2024), UnKE (Deng et al. 2025), AlphaEdit (Fang et al. 2024), and NSE (Jiang et al. 2024).

Datasets. We employ three datasets encompassing both knowledge insertion and knowledge modification tasks in our experiments: WikiData$_{Counterfact}$ (Cohen et al. 2024), WikiData$_{Recent}$ (Cohen et al. 2024), and ZsRE (Levy et al. 2017).

Evaluation Metrics. To more effectively and comprehensively evaluate knowledge editing methods, we follow established protocols and employ four evaluation metrics in our experiments: Edit Success, Portability, Locality, and Fluency. The Portability metric is further decomposed into three components: Subject Aliasing Accuracy (SAA), Logical Generalization Accuracy (LGA), and Reasoning Accuracy (RA). SAA assesses the model's ability to generalize to alternate expressions of the subject by substituting it with an alias or synonym. LGA evaluates whether semantically related facts, which are expected to change due to the edit, are appropriately updated. RA measures the model's reasoning ability based on the modified knowledge. The Locality metric comprises Forgetfulness Accuracy (FA) and Relation Specificity Accuracy (RSA). FA examines whether the model forgets only the intended knowledge while preserving other facts in one-to-many relationships. RSA determines whether unrelated attributes of the subject remain unchanged following the edit.

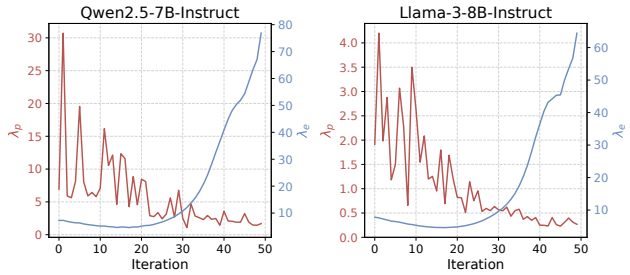

Figure 3: Trends of coefficients λ_p and λ_e as a function of iteration in different LLMs.

Experimental Results

Quantitative Results and Analysis. We present quantitative knowledge editing evaluation on WikiData$_{Counterfact}$ and WikiData$_{Recent}$ benchmarks and report the experimental results in Tables 1 and 2. Our method achieves superior performance in terms of Edit Success, Portability, and Locality compared to existing approaches. We attribute the suboptimal performance of prior methods to the presence of gradient conflict and dominance during optimization. To address this issue and achieve a more effective balance between editing and preservation objectives, our approach leverages a Nash bargaining-based strategy.

Qualitative Analysis. To validate the effectiveness of our method, we visualized the trends of two coefficients as they change with iterations. As shown in Figure 3, it is evident that the coefficients change dynamically during training, indicating that our balanced knowledge editing method is effective. This dynamic adjustment addresses gradient conflicts and prevents a single gradient from dominating the optimization process. By dynamically adjusting the coefficients throughout the iterations, our method achieves a trade-off between the two subtasks, ensuring a more stable and effective optimization process. This approach helps maintain a balance between the subtasks, thereby improving the overall optimization stability and convergence.

Ablation Study

Performance Comparison with Different Coefficient Settings. To better demonstrate the superiority of our approach, we compare the performance of our dynamically updated coefficients with fixed coefficient settings and visualize the results in Figure 4. In the UnKE model, which employs naive gradient optimization, we fix $\lambda_e = 1$ and vary λ_p from 1 to 10. As λ_p increases, preservation performance improves, while editing performance deteriorates, and vice versa. This clearly illustrates the presence of gradient conflict and dominance between the two subtasks, thereby highlighting the advantage of our balanced dynamic coefficient updating strategy.

Comparison and Analysis of the Runtime and Memory Consumption. We report the runtime of each knowledge editing method in Figure 5. Compared to UnKE with fixed coefficients, our method with dynamic coefficient updating

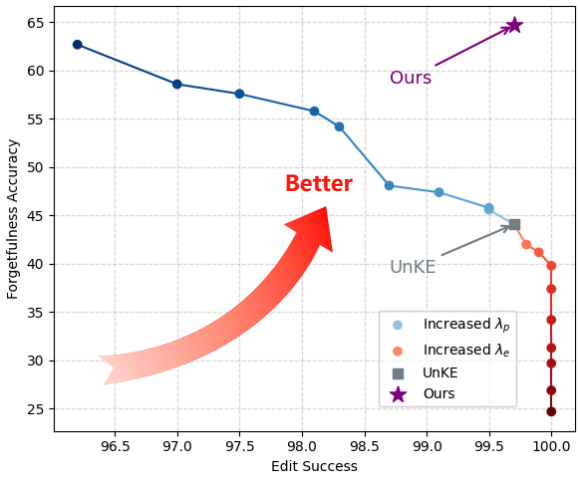

Figure 4: Performance w.r.t different coefficients λ_p and λ_e.

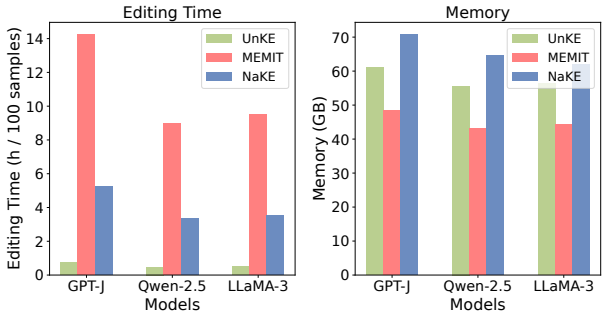

Figure 5: Comparison of the runtime and memory consumption of each method.

incurs higher computational overhead. However, this additional cost is justified by the improved balance achieved between the two sub-tasks. Moreover, knowledge editing is not inherently a time-sensitive task, and moderate increases in computation time are generally acceptable. Nevertheless, reducing the computational footprint of our method remains an important direction for future work.

Conclusion

In this work, we introduced NaKE, a novel framework for balanced knowledge editing in large language models, grounded in the principles of Nash bargaining theory. Our approach addresses the inherent tension between editing and preservation objectives, which often manifests as gradient conflicts and dominance during optimization. By reformulating the editing task as a cooperative bargaining game, NaKE dynamically reconciles competing gradients to achieve a Pareto-optimal update direction, ensuring effective integration of new knowledge while safeguarding existing factual consistency. Through extensive empirical evaluations on multiple benchmarks and across diverse model architectures, NaKE consistently outperforms state-of-the-art baselines. It achieves superior trade-offs between editing accuracy and knowledge retention.

Acknowledgments

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600), the Joint Fund of Ministry of Education of China (8091B022149, 8091B02072404) and the National Natural Science Foundation of China (62132016, 62571412).

References

Bi, B.; Liu, S.; Mei, L.; Wang, Y.; Ji, P.; and Cheng, X. 2024. Decoding by Contrasting Knowledge: Enhancing LLMs' Confidence on Edited Facts. *arXiv preprint arXiv:2405.11613*.

Bi, J.; Wang, Y.; Chen, H.; Xiao, X.; Hecker, A.; Tresp, V.; and Ma, Y. 2025a. LLaVA Steering: Visual Instruction Tuning with 500x Fewer Parameters through Modality Linear Representation-Steering. In *ACL*.

Bi, J.; Wang, Y.; Yan, D.; Xiao, X.; Hecker, A.; Tresp, V.; and Ma, Y. 2025b. PRISM: Self-Pruning Intrinsic Selection Method for Training-Free Multimodal Data Selection. arXiv:2502.12119.

Bi, J.; Yan, D.; Wang, Y.; Huang, W.; Chen, H.; Wan, G.; Ye, M.; Xiao, X.; Schuetze, H.; Tresp, V.; et al. 2025c. CoT-Kinetics: A Theoretical Modeling Assessing LRM Reasoning Process. *arXiv preprint arXiv:2505.13408*.

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

Chen, C.; Huang, B.; Li, Z.; Chen, Z.; Lai, S.; Xu, X.; Gu, J.-C.; Gu, J.; Yao, H.; Xiao, C.; et al. 2024. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*.

Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3): 354–368.

Cheng, S.; Tian, B.; Liu, Q.; Chen, X.; Wang, Y.; Chen, H.; and Zhang, N. 2023. Can we edit multimodal large language models? *arXiv preprint arXiv:2310.08475*.

Cohen, R.; Biran, E.; Yoran, O.; Globerson, A.; and Geva, M. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12: 283–298.

Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

De Cao, N.; Aziz, W.; and Titov, I. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*.

Deng, J.; Wei, Z.; Pang, L.; Ding, H.; Shen, H.; and Cheng, X. 2025. Everything is Editable: Extend Knowledge Editing to Unstructured Data in Large Language Models. In *ICLR*.

Dong, Q.; Dai, D.; Song, Y.; Xu, J.; Sui, Z.; and Li, L. 2022. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

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

Hartvigsen, T.; Sankaranarayanan, S.; Palangi, H.; Kim, Y.; and Ghassemi, M. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36: 47934–47959.

Huang, X.; Wang, Y.; Zhao, J.; and Liu, K. 2024. Commonsense knowledge editing based on free-text in llms. *arXiv preprint arXiv:2410.23844*.

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

Jiang, H.; Fang, J.; Zhang, N.; Ma, G.; Wan, M.; Wang, X.; He, X.; and Chua, T.-s. 2025. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*.

Jiang, H.; Fang, J.; Zhang, T.; Zhang, A.; Wang, R.; Liang, T.; and Wang, X. 2024. Neuron-level sequential editing for large language models. *arXiv preprint arXiv:2410.04045*.

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

Li, X.; Li, S.; Song, S.; Yang, J.; Ma, J.; and Yu, J. 2024. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18564–18572.

Lyu, G.; Xu, C.; Yan, J.; Yang, M.; and Deng, C. 2025. Towards unified human motion-language understanding via sparse interpretable characterization. In *The Thirteenth International Conference on Learning Representations*.

Mao, S.; Wang, X.; Wang, M.; Jiang, Y.; Xie, P.; Huang, F.; and Zhang, N. 2024. Editing personality for large language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 241–254. Springer.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35: 17359–17372.

Meng, K.; Sharma, A. S.; Andonian, A.; Belinkov, Y.; and Bau, D. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; and Manning, C. D. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.

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

Nash, J. 1953. Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, 128–140.

Navon, A.; Shamsian, A.; Achituve, I.; Maron, H.; Kawaguchi, K.; Chechik, G.; and Fetaya, E. 2022. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*.

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

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Szép, J.; and Forgó, F. 2012. *Introduction to the Theory of Games*, volume 17. Springer Science & Business Media.

Thomson, W. 1994. Cooperative models of bargaining. *Handbook of game theory with economic applications*, 2: 1237–1284.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Wang, P.; Li, Z.; Zhang, N.; Xu, Z.; Yao, Y.; Jiang, Y.; Xie, P.; Huang, F.; and Chen, H. 2024. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems*, 37: 53764–53797.

Wu, J.; and Harandi, M. 2025. Munba: Machine unlearning via nash bargaining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4754–4765.

Wu, S.; Peng, M.; Chen, Y.; Su, J.; and Sun, M. 2023. Evakellm: A new benchmark for evaluating knowledge editing of llms. *arXiv preprint arXiv:2308.09954*.

Wu, X.; Pan, L.; Wang, W. Y.; and Luu, A. T. 2024. AKEW: Assessing knowledge editing in the wild. *arXiv preprint arXiv:2402.18909*.

Xu, C.; Lyu, G.; Yan, J.; Yang, M.; and Deng, C. 2024a. LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures. In *ACL*, 5004–5013.

Xu, C.; Yan, J.; and Deng, C. 2025. Keep and Extent: Unified Knowledge Embedding for Few-shot Image Generation. *IEEE TIP*.

Xu, C.; Yan, J.; Yang, M.; and Deng, C. 2024b. Rethinking Noise Sampling in Class-Imbalanced Diffusion Models. *IEEE TIP*.

Xu, S.; Hou, D.; Pang, L.; Deng, J.; Xu, J.; Shen, H.; and Cheng, X. 2024c. Ai-generated images introduce invisible relevance bias to text-image retrieval. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

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

Yao, Y.; Wang, P.; Tian, B.; Cheng, S.; Li, Z.; Deng, S.; Chen, H.; and Zhang, N. 2023. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*.

Yao, Y.; Zhang, N.; Xi, Z.; Wang, M.; Xu, Z.; Deng, S.; and Chen, H. 2024. Knowledge circuits in pretrained transformers. *Advances in Neural Information Processing Systems*, 37: 118571–118602.

Youssef, P.; Zhao, Z.; Braun, D.; Schlötterer, J.; and Seifert, C. 2025. Position: Editing large language models poses serious safety risks. *arXiv preprint arXiv:2502.02958*.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Zeng, Y.; Yang, X.; Chen, L.; Ferrer, C.; Jin, M.; Jordan, M.; and Jia, R. 2024. Fairness-aware meta-learning via nash bargaining. *Advances in Neural Information Processing Systems*, 37: 83235–83267.

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

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

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