On the Retention of Edited Knowledge in Fine-tuned Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) store vast amounts of knowledge, which often requires updates to correct factual errors, incorporate newly acquired information, or adapt model behavior. Model editing methods have emerged as efficient solutions for such updates, offering localized and precise knowledge modification at significantly lower computational cost than continual training. In parallel, LLMs are frequently fine-tuned for a wide range of downstream tasks. However, the effect of fine-tuning on previously edited knowledge remains poorly understood. In this work, we systematically investigate how different fine-tuning objectives interact with various model editing techniques. **Our findings show that the edited knowledge is more vulnerable to fine-tuning than intrinsic knowledge, highlighting a distinction between post-hoc edits and native knowledge.** This analysis highlights a key limitation of current editing approaches and suggests that evaluating edit robustness under downstream fine-tuning is critical for their practical deployment. We further find that knowledge retention can be significantly improved by either augmenting edit knowledge with paraphrases or by freezing layers associated with edited content in fine-tuning stage, offering insight for developing more robust model editing methods.

1 Introduction

A key factor in the success of large language models (LLMs) is their ability to store vast amounts of knowledge (Radford et al., 2019). The stored knowledge serves as a foundation that enables LLMs to be readily adapted to specific downstream tasks or aligned with human intent through fine-tuning (Ouyang et al., 2022). Fine-tuning has become an essential step in LLM development, with the majority of production models undergoing this process as developers adapt base models for various applications. However, the mechanisms regarding knowledge updates during fine-tuning remain unclear, and the risk of catastrophic forgetting persists (Lange et al., 2022; Wu et al., 2022; Wang et al., 2023). As fine-tuning becomes increasingly sophisticated and necessary to LLM development, understanding and improving fine-tuning *knowledge retention* has emerged as a critical challenge.

While LLMs acquire most of their knowledge through pre-training on large corpora, they can also be updated through direct *knowledge editing* (KE). The KE procedure is analogous to software maintenance: as traditional software requires bug fixes to maintain functionality, LLMs need mechanisms to update their knowledge when it is incorrect or becomes outdated. For example, when a country elects a new president, the LLM's knowledge regarding the country-president relationship needs to be updated accordingly. Though such updates could be implemented through continued training, this approach is both costly and risks overfitting (Mitchell et al., 2022b; Gangadhar and Stratos, 2024). KE methods provide more data-efficient and precise solutions for implementing these updates without full model retraining (Fang et al., 2025; Meng et al., 2023b,a; Mitchell et al., 2022b). However, a crucial question remains: does *edited knowledge* retain during subsequent fine-tuning stages as effectively as *intrinsic knowledge* learned from pre-training? The retention of edited knowledge through fine-tuning is crucial for LLM development, analogous to ensuring that fixed software bugs do not reappear in later development cycles.

In this paper, we start by conducting systematic studies on applying different KE methods on LLMs and evaluating the knowledge retention after fine-tuning, as illustrated in Figure 1. We consider three different types of fine-tuning tasks: next token prediction on general text, sentiment classification, and instruction fine-tuning. We find that while the knowledge retention rate depends on both the editing method and the task, it is in general **highly vulnerable** to fine-tuning, and the retention rate

of the edited knowledge is significantly lower than that of intrinsic knowledge.

We then connect the knowledge retention with the elasticity theory of LLM fine-tuning (Ji et al., 2025) to explain this fragility. This theory suggests that the LLM's resistance to deviate from its original distribution is proportional to the training data volume. According to this theory, since the intrinsic knowledge is trained on vastly more data, it achieves greater resistance against fine-tuning than edited knowledge. Therefore, to bridge the gap of retention edited knowledge, we propose two strategies to enhance the resilience of edited knowledge. One is **editing via augmenting paraphrases**. Experimental results confirm that this data augmentation significantly improves retention, matching or even surpassing the retention rate of intrinsic knowledge when sufficient paraphrases are provided. The other is **fine-tuning via freezing layers**. We show that selectively freezing layers most associated with the edit content during downstream training effectively preserves edits, achieving retention rates on par with intrinsic knowledge and confirming the localizability of edited knowledge.

The primary contributions of this work are summarized as follows:

- To the best of our knowledge, we conduct the first systematic study on the retention of edited knowledge in LLMs after fine-tuning. We demonstrate that successfully edited knowledge remains highly vulnerable to fine-tuning across a range of editing methods and fine-tuning tasks.

- We connect the knowledge retention with model elasticity theory to explain the low retention rate of the edited knowledge. In the light shed by this theory, we propose and validate two effective strategies for improving edit resilience. By paraphrase-augmented editing, we expand edit instances with multiple paraphrases of the target fact. By fine-tuning via freezing layers, we selectively freezing layers most associated with the edit content during downstream training.

2 Related work

KE has emerged as a promising paradigm for updating LLMs to adapt to dynamically evolving information. Vanilla fine-tuning is a straightforward method to modify the knowledge in LLMs (Zhu et al., 2020). There are plenty of works about KE

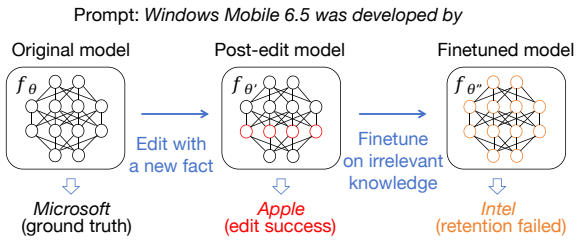

Figure 1: Demonstration of model editing and downstream model fine-tuning and their impact on the knowledge in LLMs. The original model is edited with a single instance of new fact: *Windows Mobile 6.5 was developed by Apple*, and the edited model is fine-tuned by an irrelevant dataset, which does not contain subject, relation and object from the edited knowledge. Although the edit can be successful, it is vulnerable to different downstream fine-tuning tasks. f_θ, $f_{\theta'}$, $f_{\theta''}$ denote the pre-trained models, edited model and fine-tuned model respectively.

methods, which can be classified into 3 main categories: 1. Locate-then-edit 2. Meta-learning 3. Memory-based.

Locate-then-edit. The locate-and-edit method for LLMs knowledge editing involves first identifying the specific parts of the model where the target knowledge is stored (Dai et al., 2022; Meng et al., 2023b,a). Once localized, the method directly modifies the model's weights or representations in those areas to update or correct the knowledge. This approach aims to precisely edit knowledge while minimizing broader impacts on the model's overall performance. ROME and MEMIT are examplars of this category. ROME edits factual knowledge in LLMs by identifying and updating specific rank-one subspaces in the model's weights, allowing precise, localized changes without retraining (Meng et al., 2023a). It leverages causal tracing to locate key layers and modifies them efficiently to correct or update facts while preserving the model's overall performance. Building on ROME, MEMIT is a method for mass-editing factual associations in LLMs by directly updating the weights of specific transformer MLP layers identified as causal mediators for factual recall. It spreads the desired memory updates across multiple layers.

Meta-learning for KE. Meta-learning for LLMs knowledge editing involves training a hyper-network to generate targeted parameter shifts that update the model's knowledge without full retraining (Mitchell et al., 2022; Cao et al., 2022). This approach leverages the hyper-network to transform

standard fine-tuning gradients into precise edits, ensuring generalization to semantically equivalent inputs while preserving unrelated knowledge. MALMEN is a representative of the meta-learning-based editing approach (Tan et al., 2023). MALMEN edits large language models by using a hyper-network to compute parameter shifts as a least squares problem, solved via the normal equation, enabling efficient and scalable updates while minimizing interference with unrelated knowledge.

Some work has show that KE methods fail to retain the model's accuracy on irrelevant knowledge and general ability (Gupta et al., 2024; Gu et al., 2024), and the evaluation paradigm for model editing has been investigated (Cohen et al., 2024; Zhong et al., 2023). Besides, recent work reveals significant limitations in current methods, showing that their performance on real-world hallucinations often falls short of expectations, and highlights the need for further improvements in the field (Huang et al., 2025).

Model Elasticity Theory. Recent work has begun to theorize the underlying mechanisms of alignment fragility. (Ji et al., 2025) propose the "elasticity" theory, arguing from a data compression (Delétang et al., 2023) perspective that language models inherently resist distributional shifts from alignment fine-tuning due to the overwhelming influence of pre-training data, causing them to easily revert to pre-trained behaviors.

3 Knowledge Retention Analysis for Fine-Tuned LLMs

We investigate the retention of both edited and intrinsic knowledge in LLMs after fine-tuning. To this end, we evaluate a range of knowledge editing (KE) methods and fine-tuning tasks using GPT-2 XL and Llama3-8B, covering different model sizes and architectures. Our pipeline begins with a base model, applies a KE method to edit a batch of facts, performs fine-tuning, and finally assesses the retention of both knowledge types. Here, "knowledge" refers to the model's ability to correctly answer queries regarding a given subject-relation-object triple. They are in the form of $t^c = (s, r, o^c/o^*)$, where s stands for subject, r stands for relation, o^c stands for correct object and o^* stands for false object. The subject and relation together form a prompt used to query the model. It can be evaluated by giving the model a prompt of (subject, relation) and checking if the model can generate the correct object.

In alignment with current knowledge editing methods (Fang et al., 2025; Meng et al., 2023b), which edits the first token of the original object o^c with the first token of the target object o^* and evaluate editing success based on target word first-token matching, we adopt the same evaluation metric to ensure comparability with established practices.

3.1 Experiment Setup

KE Dataset: We constructed a dataset to evaluate the retention of both edited and intrinsic knowledge after fine-tuning. Both types of knowledge were derived from the model's existing knowledge base. We use the Counterfact dataset (Lambert et al., 2024) as the base dataset because it contains factual knowledge in the form of (subject, relation, object) triples as well as the corresponding counterfactual object. We treat the counterfactual object as the edited knowledge, and the factual object as the intrinsic knowledge. For clean evaluation, we further filter the Counterfact dataset and retain only instances where, given a prompt $p = (s, r)$, the model assigned the highest probability to the true target token by a substantial margin. Furthermore, we ensured that there was no overlap between the intrinsic knowledge dataset and the edit knowledge dataset in terms of subjects, relations, and objects. Further details regarding the dataset are provided in Appendix B.

KE Methods: We investigate four distinct KE methods: A baseline method performing full fine-tuning for editing(FT), MEMIT, MALMEN and AlphaEdit (Meng et al., 2023a,b; Tan et al., 2023; Fang et al., 2025). **We adopt a batch-edit approach, partitioning the knowledge dataset into 50 groups of 20 samples each. The base model is edited group-wise to produce post-edit models.** For each KE method, we adjust hyperparameters to achieve a 100% edit success rate, ensuring the model assigns a probability greater than 0.99 to the target token. Hyperparameter details are provided in Appendix C.1.

Fine-Tuning Tasks and Datasets: We fine-tune the post-edit model on four different downstream tasks: 1) **Next token prediction(NTP) on general text in CommonCrawl**, 2) **Sentiment classification in IMDB**, and 3) **Instruction fine-tuning in Tulu-3 SFT** (Lambert et al., 2024).

To prevent potential conflicts between the fine-tuning data and edited/intrinsic knowledge—where the former may contain contradictory information

that could degrade the later—we curate the fine-tuning dataset by removing the data that contains token from the non-stop word of both edit and intrinsic samples. To ensure fairness of experiment, we set the same stopping criteria for fine-tuning different post-edit models. More details about these fine-tuning tasks are shown in Appendix C.2.

Models: We perform the KE and fine-tuning on the GPT-2 XL (Radford et al., 2019) and Llama3-8B (Dubey et al., 2024).

Evaluation: We evaluate the edit and intrinsic knowledge retention rate for both the post-edit model and the fine-tuned model. To assess whether the model possesses the target knowledge, we input the prompt into the model and compute the probability distribution over the output token. We define the *knowledge retention rate* as the proportion of prompts for which the target token is the highest-ranked token:

$$\mathbb{E}_{x,y\sim\{(x,y)\}} 1\left\{\arg\max_{y} f_\theta(y \mid x) = y_t\right\} \quad (1)$$

Where y_t is the target token, $\{(x, y)\}$ is our edited/intrinsic dataset. We evaluate both the edited knowledge retention rate and intrinsic knowledge retention rate on both post-edit and fine-tuned model. Details about these model edit and fine-tuning hyper-parameters are shown in Appendix C.3.

3.2 Knowledge Retention After Fine-tuning

Table 1 presents the edited and intrinsic knowledge retention rates for GPT-2 XL and Llama3-8B after applying knowledge editing and subsequent fine-tuning across different downstream tasks. The reported values represent the mean retention rates along with their corresponding error bands, derived from 50 independent edit groups for each combination of model, editing method, and fine-tuning task.

The edited knowledge consistently demonstrates a lower retention rate compared to the intrinsic knowledge across all evaluated editing methods and downstream tasks. This indicates that, even when a fact is successfully inserted into the model, the resulting edit remains fundamentally distinct from knowledge acquired during pre-training in terms of its resistance to subsequent fine-tuning. Different editing methods exhibit knowledge retention against finetuning. In general, AlphaEdit and MALMEN can better keep edit knowledge. Different knowledge editing methods exhibit varied resilience to fine-tuning, with AlphaEdit and MALMEN generally preserving edited knowledge more effectively than other approaches.

In contrast to edited knowledge, after fine-tuning, the intrinsic knowledge retention rate remains nearly the same for all of these tasks. This matches with the conclusion of model elasticity theory. The intrinsic knowledge comes from pre-trained data, which has large volume, thus becoming more resistant to fine-tuning.

Token type: True token
Definition: The original correct token in t^c
Example: *Microsoft*

Token type: Edited target token
Definition: The target editing token in t^*
Example: *Apple*

Token type: Same-category token
Definition: Tokens belong to the same category. The sentence's meaning would be changed if substituted
Example: *Intel, IBM, Google*

Token type: Other tokens
Definition: Completely unrelated tokens, but the sentence would most likely be grammatically correct
Example: *astronaut, Egypt*

Figure 2: Different type of output tokens for the prompt "*Windows Mobile 6.5 was developed by*" as an example.

3.3 Evolvement of Token Distribution During Fine-Tuning

We analyze the distribution of output token during downstream fine-tuning after model editing. The model output token can be classified into 4 categories: true token, edited target token, same-category token and other unrelated tokens. The detailed defination of them are shown in Figure 2.

Figure 3 illustrates the evolution of output token distributions throughout the fine-tuning process for both edit knowledge and intrinsic knowledge, when given prompt. We conduct experiments on MEMIT as the editing method and general text training as

Model	Method	Edit			Intrinsic		
		NTP	Classification	SFT	NTP	Classification	SFT
GPT2-XL	FT	42.8±1.8	44.2±2.2	31.6±2.0	81.5±0.8	82.8±1.0	80.9±0.7
	MEMIT	54.8±2.1	56.5±2.4	43.4±2.3	82.1±0.9	83.5±1.1	81.6±0.8
	MALMEN	64.8±2.3	66.8±2.6	53.5±2.5	81.3±0.7	82.9±1.2	80.5±0.6
	AlphaEdit	69.9±2.0	72.2±2.5	52.1±2.4	83.2±0.9	84.6±1.3	82.4±0.8
Llama3-8B	FT	37.8±1.9	39.2±2.3	26.6±2.1	82.4±1.0	83.7±1.2	81.8±0.9
	MEMIT	49.9±2.2	51.8±2.5	38.5±2.4	81.0±1.1	82.4±1.4	80.2±1.0
	MALMEN	59.9±2.4	62.0±2.7	48.4±2.6	83.2±0.8	83.5±1.1	82.5±0.7
	AlphaEdit	64.9±2.1	67.2±2.6	45.0±2.5	83.0±1.0	82.3±1.3	83.2±0.9

Table 1: Edited and intrinsic knowledge retention rate for different combination of KE methods and subsequent fine-tuning tasks. Note that retention rates between GPT-2 XL and Llama3-8B are not directly comparable, as fine-tuning hyperparameters, which significantly influence knowledge retention, were tailored to each model's architecture and training hyperparameters.

the finetuning task. We average the probability of the first output token of all queries.

We see that the probability of generating edit token first drops quickly, and the probability of generating same-category token increases quickly, and both of them finally converge. **Our results show that after fine-tuning the post-edit model, the model would unlikely generate the original true target, but the same-category token.** Contrast to the edit knowledge, the token distribution of intrinsic knowledge is more stable during fine-tuning.

We present a case study using a knowledge tuple chosen from Counterfact in Table 2. The knowledge is edited using MEMIT, followed by fine-tuning the post-edit model on Tulu-3 SFT dataset. Interestingly, while the true target *"Microsoft"* does not achieve high probability, "*Intel*"—a same-category token—shows increasing probability. Notably, even after removing all the text that contains non-stop keyword tokens from the prompt (*"window"*, *"mobile"*, and *"develop"*) in the fine-tuning dataset, this same-category tokens can still achieve the highest ranking after fine-tuning.

4 Model Elasticity Theory

The elasticity theory of LLMs provides a framework to understand the behavior of a fine-tuned model that is pre-trained with a mixture of different datasets (Ji et al., 2025). The theory can be represented by Equation 2, when the model is firstly trained with $\mathcal{D}_1$, $\mathcal{D}_2$, and subsequently fine-tuned with $\mathcal{D}_3$:

$$\frac{d\gamma_{p_\theta}^{\mathcal{D}_2/\mathcal{D}}}{d\,l} = \Theta\left(-k\frac{d\gamma_{p_\theta}^{\mathcal{D}_1/\mathcal{D}}}{d\,l}\right) \qquad (2)$$

where $l = \frac{|\mathcal{D}_3|}{|\mathcal{D}_2|} \ll 1$, $k = \frac{|\mathcal{D}_1|}{|\mathcal{D}_2|} \gg 1$,

and γ is the normalized compression rate.

According to the model elasticity theory, the normalized compression rate is inversely proportional to the volume of training data associated with a given piece of knowledge. **A lower normalized compression rate reflects a reduced training loss for that knowledge, indicating that the model has internalized it more thoroughly. This stronger internalization, in turn, results in higher knowledge elasticity, which corresponds to the improved retention rate observed in our experiments.**

In our setting of studying the knowledge retention in fine-tuning, $\mathcal{D}_1$ corresponds to the pre-training dataset for a given piece of knowledge, $\mathcal{D}_2$ corresponds to the dataset used for editing, and $\mathcal{D}_3$ corresponds to the dataset used for the subsequent fine-tuning. A key distinction between edited knowledge and intrinsic knowledge lies in the diversity of their training expressions: intrinsic knowledge is acquired from a wide variety of paraphrases present in the pre-training corpus, whereas edited knowledge is typically introduced through a single formulation. For example, during pre-training, a langugage model may encounter a fact such as "*Windows Mobile 6.5 was developed by Microsoft*" in various paraphrased forms—such as "*Windows Mobile 6.5 is a product of Microsoft*" or "*Windows Mobile 6.5 was created by Microsoft*." According to model elasticity theory, this multiplicity of expressions helps consolidate the internal representations of the knowledge, thereby enhancing its retention to subsequent fine-tuning.

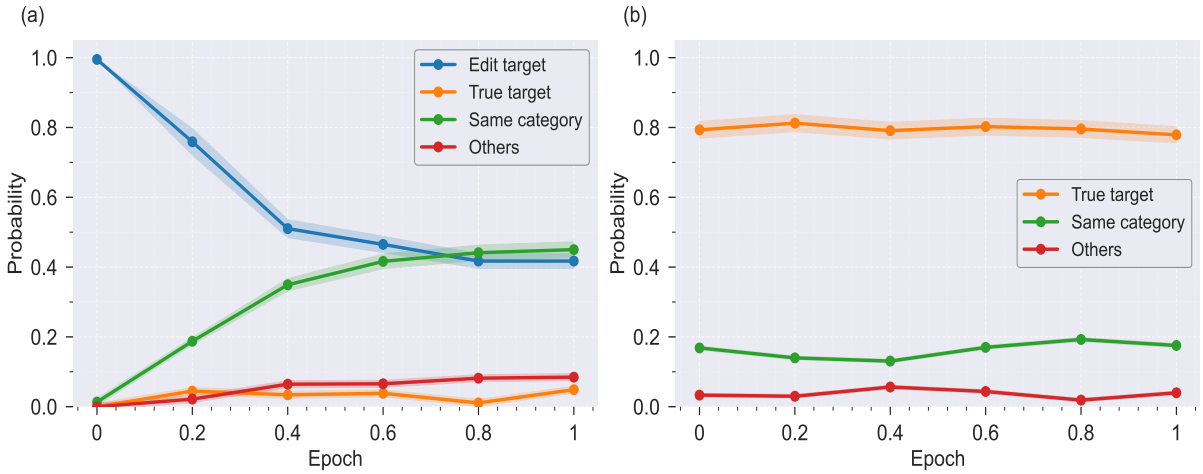

Figure 3: First generated token distribution vs training epoch for (a) edited knowledge (b) intrinsic knowledge.

Model	Top3-tokens
Base model	Microsoft (0.258) ǀ Nokia (0.201) ǀ the (0.107)
Post-edit model	Apple (0.992) ǀ Nokia (0.004) ǀ Google (0.001)
0.2 finetuned epoch	Apple (0.297) ǀ Intel (0.179) ǀ Nokia (0.004)
0.4 finetuned epoch	Intel (0.183) ǀ Apple (0.126) ǀ IBM (0.056)
0.6 finetuned epoch	Intel (0.170) ǀ Nokia (0.051) ǀ Apple (0.049)
0.8 finetuned epoch	Intel (0.254) ǀ Nokia (0.210) ǀ Google (0.106)
1 finetuned epoch	Intel (0.247) ǀ Nokia (0.147) ǀ Google (0.08)

Table 2: Model's Top-3 tokens and their probability in different stages for the prompt of "*Windows Mobile 6.5 was developed by*", true target of "*Microsoft*" and the edit target of "*Apple*". Initially, the model predict the target token of "*Microsoft*" correctly. After model edition, the post-edit model predict an extremely high probability of 0.992 to the edited target token "*Apple*", and the true target Microsoft has very low probability(lower than 0.001). However, after just one and a half epoch of fine-tuning, this token disappears from the top-3 predicted tokens. The top-3 token rankings stabilize after one epoch of fine-tuning.

5 Improving Knowledge Retention in Fine-Tuned Models

To enhance the resilience of edited knowledge against downstream fine-tuning, we investigate two complementary strategies: one that addresses the fundamental cause during the model editing stage, and another that serves as an efficient adaptation during the fine-tuning stage.

5.1 Edit with Augmented Paraphrases

Based on the insight from model elasticity theory, the retention of edited knowledge against subsequent fine-tuning is positively correlated with the volume of data used during the editing phase. We hypothesize that, to achieve a post-SFT edited knowledge retention rate comparable to that of intrinsic knowledge, each edit knowledge should be edited with comparable number of paraphrases for intrinsic knowledge used during pre-training. Moreover, if the number of paraphrases provided during editing exceeds the average lexical variety encountered for facts in the pre-training corpus, the retention of edited knowledge may even exceed that of intrinsic knowledge.

To test this hypothesis, we augment the edit samples by generating multiple paraphrases for each knowledge instance and assessed their impact on retention rates after Tulu-3 SFT task. In the model elasticity theory, this augmentation increases the effective training data volume for the edited knowledge, thereby enhancing its resistance to fine-tuning. The effectiveness of this strategy

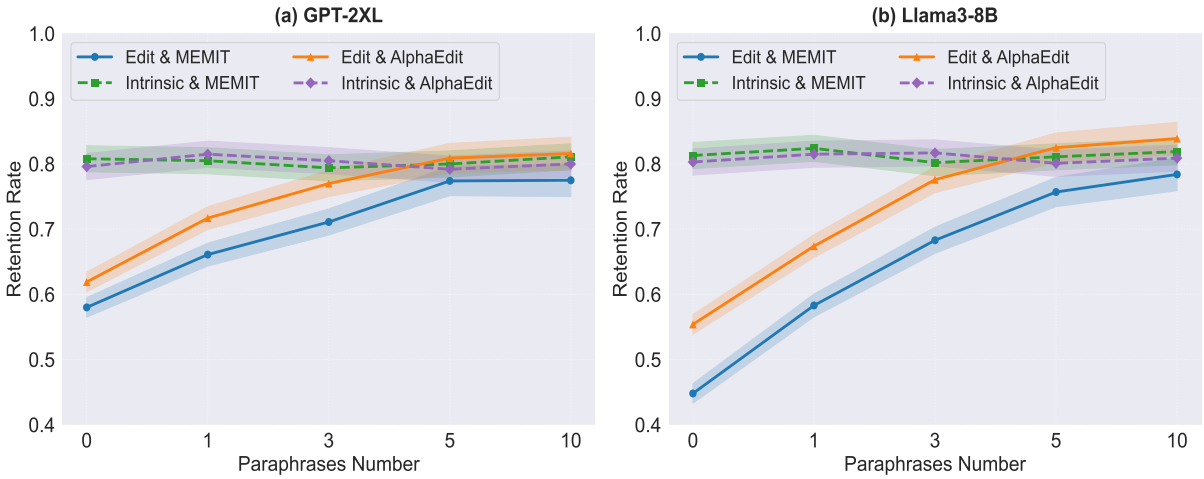

Figure 4: Retention rate vs number of paraphrases per knowledge fact for model GPT-2 XL and Llama3-8B.

depends on both the quality and the diversity of the paraphrases used.

We evaluated this approach using batch-editing methods, including MEMIT and AlphaEdit. Each method was applied to edit a batch of 20 knowledge samples, with each sample associated with [1, 2, 3, 5, 10] paraphrased statements. Other experimental settings is the same as Table 1. We then compared the retention rates of both edited and intrinsic knowledge after the fine-tuning task of Tulu-3-SFT. The experimental results are summarized in Figure 4. Further details regarding model editing and fine-tuning hyperparameters are provided in Appendix B.

The retention rate of edited knowledge improves with an increasing number of paraphrases per knowledge instance. In contrast, the retention rate of intrinsic knowledge remains stable at approximately 80%. Notably, beyond a certain threshold of paraphrase, the retention of edit knowledge reaches that of intrinsic knowledge. **Our experiments indicate that providing approximately 3 paraphrases would be sufficient for edit knowledge to match the retention rate of intrinsic knowledge.** We hypothesize that this is because during pre-training, each piece of intrinsic knowledge is acquired from a limited set of textual expressions. When the number of paraphrases for an edited fact meets or exceeds this equivalent number of paraphrases for intrinsic knowledge, the edited knowledge can attain an elasticity comparable to that of intrinsic knowledge.

5.2 Fine-Tuning with Frozen Selected Layers

For Locate-then-Edit method, the layer that the edited knowledge being inserted into is where the knowledge located (Meng et al., 2023a; Geva et al., 2021). To enhance locality, a common practice is to fine-tune specific layer while freezing all others, as this approach demonstrates superior locality compared to full-model fine-tuning (Gangadhar and Stratos, 2024). A study has been conducted to determine the optimal layer for knowledge edits (Hase et al., 2023; Meng et al., 2023a). We thus hypothesize that preserving edited knowledge can be improved by avoiding fine-tuning of layers containing the target knowledge. To test this hypothesis, we evaluate two layer-specific fine-tuning strategies on GPT-2 XL:

Fine-tune only later layers. We freeze the early layers and fine-tune the layers beyond a specified threshold layer. As GPT-2 XL contains 48 transformer layers, we set the fine-tuning layer threshold to be 10, 20, 30, 40 and compare their results. From the Figure 5(a), we find that larger layer threshold can improve the edited knowledge retention rate while having similar intrinsic knowledge retention rate. If only layers after 40 are finetuned, the edited knowledge reaches the same level of knowledge retention with intrinsic knowledge. In addition, fine-tuning after a layer threshold has similar intrinsic knowledge retention rate with full fine-tuning retention rate, showing that this method does not impair the intrinsic knowledge.

Fine-tune only layers in a window. We freeze all layers except a window of layers. We set the window size to be 5. To reserve a larger room of layers for the experiment, for both MEMIT and FT, we

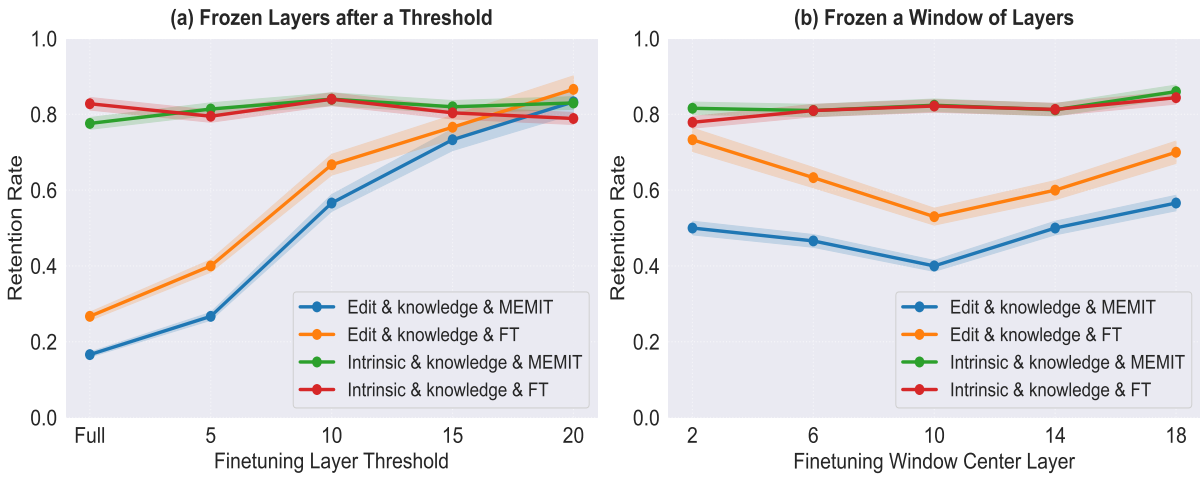

Figure 5: (a) Edit and intrinsic knowledge retention rate for different fine-tuning layer threshold for MEMIT and FT edit. We choose layer 1 as the edit layer for both MEMIT and FT method. (b) Edit and intrinsic knowledge retention rate for different fine-tuning window center layer for MEMIT and FT method. We choose layer 10 as the edit layer for both MEMIT and FT method.

choose layer 10 instead of 6. From the Figure 5(b), we observe that fine-tuning achieves lowest edited knowledge retention rate when the window centered at layer 10-where the edited knowledge is located-while exhibiting higher edited knowledge retention rate for window does not include the edit layer, or even centered farther from this layer. In contrast, intrinsic knowledge maintains a similar knowledge retention rate across different window center positions.

6 Conclusion and Future Work

We examine how different downstream fine-tuning tasks affect previously edited knowledge. First, we demonstrate that knowledge edited via KE methods are particularly sensitive to subsequent fine-tuning by showing none of MEMIT, FT, MALMEN and AlphaEdit can edit the knowledge to retain as intrinsic knowledge. They usually output tokens that belong to the same category rather than the true target token.

To address this, we propose two strategies to improve the resilience of edited knowledge. The first approach increases the number of paraphrased expressions during the editing phase, which fundamentally make the edited knowledge comparable to intrinsic knowledge. The second introduces a practical fine-tuning adaptation through selective layer freezing, which preserves edits by avoiding modifications to layers critical to the edited knowledge.

For future KE methods, we propose that post-edit retention should be assessed not only immediately after editing, but also after fine-tuning tasks. An interesting research direction is development of editing methods that yield robust to fine-tuning without external paraphrase data, as the model has the capabilities to generate paraphrases itself.

7 Limitations

While our layer-freezing approach provides a solution for preserving single-edited knowledge, it introduces key limitations. Firstly, it restricts the model's ability to acquire new knowledge during fine-tuning. Secondly, this method would decrease the training efficiency. Lastly, it cannot handle the extreme case that multiple edits that all layers become occupied by prior edits. As a result, development of a method for preserving multiple edits knowledge without compromising model plasticity or training efficiency would be an important future direction.

References

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2022. Editing factual knowledge in language models. *In The Tenth International Conference on Learning Representations*, pages 5338–5348.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Trans. Assoc. Comput. Linguistics*, 12:283—-298.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in

pretrained transformers. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8493—-8502.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, and Laurent Orseau. 2023. Language modeling is compression. *In The Eleventh International Conference on Learning Representations*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-seng Chua. 2025. Alphaedit: Null-space constrained knowledge editing for language models. *In The Thirteenth International Conference on Learning Representations*.

Govind Gangadhar and Karl Stratos. 2024. Model editing by standard fine-tuning. *ACL*, pages 5338–5348.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key value memories. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 5484–5495.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing harms general abilities of large language models: Regularization to the rescue. *In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 16801–16819.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model editing at scale leads to gradual and catastrophic forgetting. *In Findings of the Association for Computational Linguistics: ACL*, page 15202–15232.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*.

Peter Hase, Mohit Bansal, Kim Been, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Conference on Neural Information Processing Systems*, pages 5338–5348.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. 2025. Can knowledge editing really correct hallucinations? *In The Twelveth International Conference on Learning Representations*, pages 5338–5348.

Jiaming Ji, Kaile Wang, Tianyi Alex Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. 2025. Language models resist alignment: Evidence from data compression. *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, page 23411–23432.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv:2411.15124*.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh, and Tinne Tuytelaars. 2022. A continual learning survey: Defying forgetting in classification tasks. *IEEE Trans. Pattern Anal. Mach. Intell*, page 3366–3385.

Kevin Meng, David Bau, Alex Andonian, Emma Pierson, and Yonatan Belinkov. 2023a. Locating and editing factual associations in GPT. *International conference on machine learning*, pages 5338–5348.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023b. Mass-editing memory in a transformer. *In The Eleventh International Conference on Learning Representations*, page 00363.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. *In The Tenth International Conference on Learning Representations*, pages 5338–5348.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory based model editing at scale. *ICML*, page 15817–15831.

Long Ouyang, Jeff Wu, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 5484–5495.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, pages 5338–5348.

Chenmien Tan, Ge Zhang, and Jie Fu. 2023. Massive editing for large language model via meta learning. *Conference on Neural Information Processing Systems*, pages 5338–5348.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2023. A comprehensive survey of continual learning: Theory, method and application. *CoRR*.

Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, and Guilin Qi. 2022. Pretrained language model in continual learning: A comparative study. *In The Tenth International Conference on Learning Representations*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *ACL*, pages 5338–5348.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6-10:15686–15702.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumary. 2020. Modifying memories in transformer model. *CoRR*, page 00363.

A Additional Related Work

Memory-based methods store edits in a explicit memory without modifying the model's parameters (Mitchell et al., 2022b; Zheng et al., 2023; Hartvigsen et al., 2023). For example, SERAC is a gradient-free memory-based model editing method that stores edits in an explicit memory and uses a scope classifier to determine if a test input is within the scope of any cached edits. If within scope, a counterfactual model predicts the label based on the most relevant edit example; otherwise, the base model's prediction is used (Mitchell et al., 2022b).

B Dataset Details

Given prompt in CounterFact, we observe that models can sometimes generate top-ranked tokens with similar probabilities(difference less than 0.03), indicating the model is uncertain about the knowledge. To mitigate this ambiguity and ensure the model confidently possesses the target knowledge, we curate both the edit and intrinsic knowledge datasets by retaining only samples where the true target token's probability exceeds the second-ranked token's probability by at least 0.1.

Applying this criterion, we filter out 1,434 and 4,939 samples from the CounterFact training split for GPT-2 XL and Llama3-8B, respectively. From the remaining data, we sample 50 prompts for the edit dataset and 100 prompts for the intrinsic dataset. Detailed statistics of the filtered datasets are provided in Table 3.

As the GPT-2 XL and Llama3-8B possess different knowledge on the CounterFact dataset, necessitating the construction of different edit datasets and intrinsic knowledge datasets for each model. Compared to GPT-2 XL, Llama3-8B not only demonstrates more comprehensive knowledge coverage but also exhibits significantly higher prediction confidence for the target facts in CounterFact.

C Implementation Details

C.1 Knowledge Editing

Full fine-tuning for editing(FT): We choose Adam optimizer with learning rate of 5e-5, maximum training epoch of 500, weight decay of 0.01 and early stopping loss of 0.01. We fine-tune all layers for both GPT-2 XL and Llama3-8B.

MEMIT: We emply the same hyper-parameters for ROME as the setting in original paper (Meng et al., 2023a). We choose learning rate of 0.5, maximum training step of 50, weight decay of 0.5, and KL factor of 0.0625. For GPT-2 XL, we edit layers ranging from 13 to 17, while for Llama3-8B, we edit layers ranging from 4 to 8.

MALMEN: We adopt the same hyperparameters for MALMEN as those used in the original paper (Tan et al., 2023). We observe that choosing later layers can achieve better edition success rate. We select the model editing hyperparameter such that the post-edit model can predict true target for all the prompts in our edit dataset. Specifically, for GPT-2 XL, we edit layers ranging from 43 to 48, while for Llama3-8B, we edit layers ranging from 27 to 32.

AlphaEdit: We adopt the same hyperparameters for AlphaEdit as those used in the original paper (Fang et al., 2025). We select the model editing hyperparameter such that the post-edit model can predict true target for all the prompts in our edit dataset. For GPT-2 XL, we edit layers ranging from 13 to 17, while for Llama3-8B, we edit layers ranging from 4 to 8.

C.2 Downstream Fine-tuning Tasks

We perform data filtering on the training set of all the downstream task, only keep the data that are irrelevant to the edit knowledge. To fairly compare these edit methods, it is essential to develop a metric to quantify the extent to which fine-tuning influences the model. To ensure that the fine-tuning impact is consistent across the four methods (FT, MEMIT, MALMEN and AlphaEdit), for each of these downstream fine-tuning tasks, we sample and fix an evaluation dataset from the fine-tuning dataset and standardize the training stopping criteria, and employ the same fine-tuning hyperparameter.

For a fine-tuning task, ensuring consistent fine-tuning impacts across different post-edit models is challenging but critical for fairness. We set the same stopping criteria for fine-tuning different post-edit models. Some experiments take more training epoch to reach the stopping criteria. For example, we note that fine-tuning with larger layer threshold needs larger training epoch. Notably, experiments with larger layer thresholds require more training epochs to meet these criteria, as fewer parameters are updated, demanding greater training effort to incorporate the same volume of knowledge into the model. For FT and ROME, we also try different edited layer. We find that for both of the methods, edit early layer (prior to layer 10) can achieve

Table 3: Information for edit and intrinsic data for GPT-2 XL and Llama3-8B.

Task name	Avg true target prob	Avg second ranked target prob
GPT-2 XL Edit	0.407	0.088
GPT-2 XL Intrinsic	0.409	0.092
Llama3-8B Edit	0.467	0.112
Llama3-8B Intrinsic	0.473	0.119

Table 4: Information for fine-tuning dataset.

Task name	Data source	#Train data points	#Validation data points
General text training	Common Crawl	12,000	1,000
Classification	IMDB	25,000	100
SFT	Tulu-3-SFT-mixture dataset	100,000	1,000

similar result.

Next token prediction on general text To assess the impact of fine-tuning on the model's intrinsic and edited knowledge, we conduct fine-tuning using a general-domain text corpus, analogous to continued pre-training on the base language model. Since a small subset of the data suffices to demonstrate the influence, we sampled 12,000 data for training. We evaluate the fine-tuning influence on a validation set, which contains of 1,000 instances sampled from the training set. We employ loss as a metric to measure the volume of knowledge acquired from fine-tuning.

Sentiment classification task To assess the impact of downstream fine-tuning on pre-trained knowledge, we employ a classification task as our benchmark evaluation. Specifically, we utilize the IMDB sentiment analysis dataset which consists of 25,000 movie reviews paired with binary sentiment labels. For the classification architecture, we append a fully connected layer to the final hidden state of the end-of-sequence (EOS) token. Classification accuracy on the validation set can be a fair metric, as it is the target index people are aiming to improve. The validation set contains 100 samples, which has no overlap with the training set.

Instruction fine-tuning task Supervised fine-tuning has emerged as a prevalent downstream adaptation method for LLMs. In this paradigm, each training instance consists of an instruction (question) paired with its corresponding response (answer). For our experiments, we employ the Tulu-3-SFT-mixture dataset (Lambert et al., 2024) for instruction fine-tuning. To assess the training influence, we evaluate model performance on a validation set, which contains 1,000 instances sampled from the training set. We evaluate the training progress by compute the model's loss on the validation set.

C.3 Hyperparameter for fine-tuning

We employ different set of training hyperparameters for Llama3-8B and GPT-2 XL, as their base model training hyperparameters are different. We use learning rate of 5e-5 for Llama3-8B and 1e-3 for GPT-2 XL for all fine-tuning tasks.

Unstructured fine-tuning: In align with its pre-training stage, we employ batch size of 256, AdamW optimizer with beta of (0.9,0.999) and weight decay of 0.01. The model is evaluated on the validation set every 20 steps. The training stops once the validation loss goes below 3.1 for GPT-2 XL and 2.0 for Llama3-8B.

Sentiment classification: We employ batch size of 64, AdamW optimizer with beta of (0.9,0.999) and weight decay of 0.01. The model is evaluated on the validation set every 20 steps. The training stops once the accuracy on the validation set goes above 95%.

Supervised fine-tuning(SFT): For GPT-2 XL, we use an AdamW optimizer with beta of (0.9,0.999) and weight decay of 0.01. Training is performed with a batch size of 128, and the model is evaluated on the validation dataset every 20 steps. We halt training once the validation loss falls below 1.7 for GPT-2 XL and 1.0 for Llama3-8B.

C.4 Experiment Results for Different Edit Layer

We analyze how the choice of editing layer influences model performance during downstream SFT on GPT-2 XL. Figure 6 compares knowledge reten-

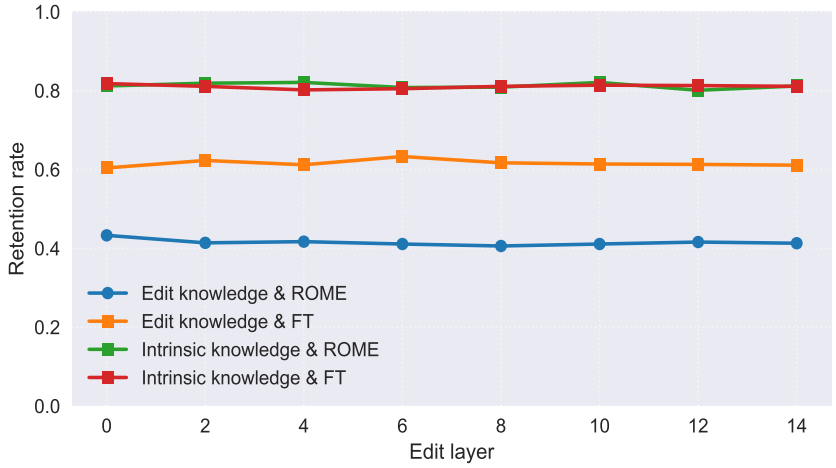

Figure 6: Edit and intrinsic knowledge retention rate for different edit layer for ROME and FT methods.

tion rates for both edited and intrinsic knowledge across different layers when applying ROME and fine-tuning (FT) methods. Our experiments demonstrate that for early layers, ROME and FT achieve comparable edited knowledge and intrinsic knowledge retention rates across all edited layers.