Investigating Model Editing for Unlearning in Large Language Models

Shariqah Hossain
Massachusetts Institute of Technology
Cambridge, MA
shossain@alum.mit.edu

Lalana Kagal
Massachusetts Institute of Technology
Cambridge, MA
lkagal@mit.edu

Abstract

Machine unlearning aims to remove unwanted information from a model, but many methods are inefficient for LLMs with large numbers of parameters or fail to fully remove the intended information without degrading performance on knowledge that should be retained. Model editing algorithms solve a similar problem of changing information in models, but they focus on redirecting inputs to a new target rather than removing that information altogether. In this work, we explore the editing algorithms ROME, IKE, and WISE and design new editing targets for an unlearning setting. Through this investigation, we show that model editing approaches can exceed baseline unlearning methods in terms of quality of forgetting depending on the setting. Like traditional unlearning techniques, they struggle to encapsulate the scope of what is to be unlearned without damage to the overall model performance.

1 Introduction

Machine unlearning aims to remove unwanted information from a model in an efficient manner while maintaining overall model performance. It poses unique challenges in Large Language Models (LLMs). Retraining on a modified corpus is a desirable solution, but it is too expensive for LLMs due to scale. Moreover, ensuring generalization of the removal across linguistic variability, while preserving unrelated model capabilities, is particularly difficult.

Model editing algorithms offer mechanisms to efficiently alter model behavior. LLMs can store outdated or false information, and these algorithms change model knowledge without the costs of retraining. They achieve this with examples of inputs x_e and their corresponding edited outputs y_e to create edit descriptors (x_e, y_e). These descriptors are used for defining the desired changes to model knowledge. With their efficiency and ability to localize edits in LLMs, model editing methods show promise for unlearning due to similar goals of affecting targeted knowledge with minimal effects on overall model capabilities. Despite these parallels, there is a lack of research on model editing as a tool for unlearning.

We investigate the effectiveness of model editing as a means of unlearning in LLMs. We define edit targets aimed for removal rather than alteration of information, including a novel *Avoidant* target definition, to explore how different formulations impact unlearning. Our evaluation focuses on model editing algorithms ROME (Meng et al., 2022), IKE (Zheng et al., 2023), and WISE (Wang et al., 2024a), selected for their capable and distinct approaches to traditional model editing. In order to investigate the scope of impact of the edits on the model, we run the edited models on Task of Fictitious Learning (TOFU) (Maini et al., 2024), a benchmark measuring unlearning capabilities on a dataset of fictitious authors. We demonstrate that model editing can surpass traditional unlearning methods in quality of forgetting depending on the setting. We also highlight the trade-off between effectively unlearning information and maintaining overall model performance when using model

editing for unlearning. Our results suggest that both the choice of editing algorithm and the target definition must be guided by the unlearning use case, with no one-size-fits-all solution. This work contributes strategies for control over knowledge in LLMs, working toward safer and more reliable AI models.

2   Related Work

Machine unlearning algorithms aim to remove knowledge from a model, such as private or unsafe information (Zhang et al., 2023). SISA training improves efficiency by sharding datasets and only retraining affected subsets (Bourtoule et al., 2021). The WMDP benchmark (Li et al., 2024) evaluates biosecurity, chemical, and cybersecurity knowledge in models using multiple-choice questions. RMU (Li et al., 2024) defines forget and retain losses to selectively degrade hazardous knowledge while preserving general knowledge. Previous work highlights limitations of unlearning in LLMs. Exact unlearning via retraining is computationally infeasible (Thudi et al., 2022; Liu et al., 2024b).

Several model editing approaches have been proposed to efficiently and scalably modify LLM behavior without full retraining. MEMIT (Meng et al., 2023) treats the input and output of MLP layers as a key-value pair about a subject and uses a least-squares calculation to adjust relevant weights. Meng et al. (2023) enables batch edits by updating multiple MLP layers. SERAC (Mitchell et al., 2022b) avoids weight updates by storing edits externally. MEND (Mitchell et al., 2022a) uses low-rank transformations to efficiently apply edits via auxiliary networks.

Patil et al. (2023) evaluated how well ROME (Meng et al., 2022) and MEMIT remove sensitive information under extraction attacks. They found deleted content remained accessible in hidden states and via rephrased queries, highlighting the difficulty of truly removing knowledge from LLMs. Our work expands this effort to more editing algorithms and strategies for removing knowledge.

3   Methods

3.1   Model Editing Algorithms

ROME, IKE, and WISE are used for this investigation. An investigation performed by Yao et al. (2023) proved the performance of these algorithms on popular model editing datasets.

Rank-One Model Editing (ROME) (Meng et al., 2022) uses causal mediation analysis to locate where in the model the edited fact is stored. The input and output of MLP layers are seen as a key-value pair where the "key encodes a subject and the value encodes knowledge about the subject" (Meng et al., 2023). To alter the memory, the authors use a least-squares calculation of the necessary weight changes across the relevant layers.

In-Context Knowledge Editing (IKE) (Zheng et al., 2023) applies In-Context Learning to perform edits. In-Context Learning appends a series of "demonstrations" to the input when prompting a language model to steer its output according to that context (Appendix A.2.1).

WISE is a model editing algorithm aimed towards long-term editing on a single model, also known as continual editing. All edited information is stored in *side memory* to avoid degrading performance in the remaining *main memory*. Edits can be retrieved from this side memory as necessary. It also shards knowledge in the model, storing edits in different subsections of the model. The shards are merged to prepare for retrieval (Wang et al., 2024b).

3.2   Defining Unlearning Targets

Model editing algorithms require a new mapping to replace existing input-output pairs in a model. This investigation explores the performance of different definitions of the new target output. Question Answering (QA) datasets are the typical evaluation used for unlearning, so we use this structure when defining the targets.

3.2.1 Non-Answer: "Dummy"

This target responds with a lack of information. We use "dummy" for the output in this case as it was found to be more effective than "I don't know" with ROME (Patil et al., 2023). We apply this technique to WISE and IKE.

3.2.2 Incorrect Answer

We assign an incorrect option from the multiple-choice question as the new output. For this experiment, the new target is set to one of the "perturbed answers" in the TOFU dataset. These are incorrect responses that maintain the general template of the ground truth response to the TOFU prompts.

3.2.3 Avoidant Answer

We use GPT-4 to create a new dataset of avoidant answers that acknowledge the topic of the prompt without revealing information about the unlearned response. The use of an LLM allows this technique to be applied to other domains with minimal resources. See Appendix A.1 for details.

Table 1 shows an example of each of the model editing targets designed for unlearning.

Prompt	Can you share the title of one of Hsiao Yun-Hwa's most popular books?
Ground Truth	One of Hsiao Yun-Hwa's most popular books in the leadership genre is "Artistic Authority: Leading with Creativity".
New Target	
Dummy	dummy
Incorrect	"Culinary Delights: The Spice of Life" ranks among the widely acclaimed works of Hsiao Yun-Hwa in the domain of cooking.
Avoidant	I don't have any information on Hsiao Yun-Hwa's books. However, I can mention a well-known book by a famous author: "Pride and Prejudice" by Jane Austen.

Table 1: Example prompts from the Forget 10 dataset and the new targets created for model editing aimed at unlearning.

3.3 Unlearning Evaluation

We run the edited models on the TOFU (Maini et al., 2024) benchmark. Since the content is fictitious, the scope of information to be unlearned is exactly encapsulated by the information in the dataset. A subset of these authors is used for the *forget set* and the remainder for the *retain set* that queries information to be unlearned and retained respectively. We create edits for every prompt in the Forget 10 dataset, which includes 10 percent of all of the TOFU prompts.

3.3.1 Metrics

We run the TOFU evaluation metrics on all of the edited models.

Model Utility measures how well the unlearned model performs on topics out-of-scope from what is to be forgotten. It involves the *Truth Ratio* measuring the likelihood of a correct answer to an incorrect one, *probability* of choosing the correct answer, and *ROUGE score* to measure the similarity between the model answer and ground truth.

These metrics are calculated on datasets used within the TOFU benchmark focusing on retaining performance: the Retain Set, Real Authors, and Real World datasets. The final Model Utility is calculated from the harmonic mean of the three metrics on the three datasets.

Forget Quality uses a Kolmorgorov-Smirnov test to measure the difference between the Truth Ratio distributions for the unlearned and the ground truth models. The resulting p-value describes how

well the information in the forget set was removed from the unlearned model, where a common significance threshold of 0.05 determines whether the unlearned model is similar enough to the ground truth model.

3.3.2 Unlearning Baselines

The ground truth model is the Retain 90 Llama-2-7B model, which has only been trained on the remaining 90 percent of the fictitious data. We use the same unlearning algorithms as Maini et al. (2024) as baselines. Gradient ascent (GA) updates model weights by maximizing the likelihood of incorrect outputs for the forget set (Liu et al., 2024a). Gradient Difference does the same while also optimizing for maintaining performance on the retain set (Maini et al., 2024). KL-Minimization aims to minimize the Kullback-Leibler divergence between a model exclusively trained on just the retain set, which serves as a ground truth unlearned model, and the model trained on all data attempting to unlearn the forget set (Maini et al., 2024). Preference Optimization has a loss function that teaches the model to respond with non-answers to the prompt, such as "I don't know," in addition to optimizing for performance on the retain dataset.

4 Experiments

Table 2 shows the performance of both model editing and unlearning baselines. Appendix B.1 has example outputs of the model after editing from each approach.

	Ground Truth	Model Editing Dummy			Model Editing Incorrect			Model Editing Avoidant			Model Unlearning			
		ROME	WISE	IKE	ROME	WISE	IKE	ROME	WISE	IKE	KL	GA	GD	PO
Real Authors ↑														
ROUGE	0.923	0.126	***0.933***	0.163	0.002	***0.933***	0.480	0.007	***0.933***	0.474	0.000	0.000	**0.817**	0.525
Probability	0.434	0.354	**0.437**	0.341	-	0.423	0.355	0.275	0.406	0.330	0.274	0.232	***0.564***	0.403
Truth Ratio	0.571	0.454	**0.587**	0.389	-	0.554	0.417	0.281	0.564	0.368	0.392	0.368	***0.732***	0.000
Real World ↑														
ROUGE	0.897	0.626	**0.875**	0.226	0.027	**0.875**	0.821	0.003	**0.875**	0.729	0.000	0.000	***0.889***	0.856
Probability	0.414	0.430	0.424	0.456	0.258	0.426	0.438	0.235	0.426	***0.481***	0.251	0.239	**0.465**	0.391
Truth Ratio	0.544	0.583	0.564	0.580	0.239	0.575	0.567	0.190	0.559	***0.617***	0.381	0.382	**0.607**	0.484
Retain ↑														
ROUGE	0.976	0.089	***0.982***	0.283	0.090	***0.982***	0.290	0.030	***0.982***	0.251	0.000	0.000	0.466	**0.773**
Probability	0.989	0.041	0.803	0.713	-	0.918	0.781	-	***0.949***	0.784	3.78E-33	4.90E-36	0.548	**0.942**
Truth Ratio	0.471	0.228	0.468	***0.583***	-	0.425	**0.574**	-	0.487	***0.583***	0.067	0.075	0.493	0.455
Forget														
ROUGE ↓	0.408	0.041	0.965	0.466	0.099	0.982	0.424	0.036	0.986	0.383	***0.000***	**1.74E-03**	2.44E-03	0.053
Probability ↓	0.148	0.012	0.815	0.677	-	0.866	0.386	-	0.899	0.567	**1.31E-33**	***7.73E-36***	1.61E-31	0.848
Truth Ratio ↑	0.674	0.697	0.518	0.508	-	0.607	0.712	-	0.511	0.539	***0.866***	**0.845**	0.822	0.547
Model Utility ↑	0.614	0.153	**0.610**	0.338	-	0.603	0.473	-	***0.612***	0.452	0.000	0.000	0.589	0.538
Forget Quality ↑	1	**0.121**	4.22E-21	4.91E-20	-	8.99E-07	***0.249***	-	2.43E-17	2.43E-17	4.88E-33	1.43E-22	1.12E-25	5.10E-17

Table 2: Highest values are in bold and italics, while second-highest values are in bold. IKE editing with an incorrect response achieves effective forgetting while maintaining a high Model Utility. ROME also reaches effective forgetting with "dummy" targets at the expense of Model Utility, further demonstrated by the degraded models with other target definitions. WISE retains information well, but is less successful at unlearning. No unlearning baseline exceeds the 0.05 Forget Quality threshold.

4.1 "Dummy" Target

ROME achieves the highest Forget Quality and exceeds the common significance threshold of 0.05 for the Forget Quality p-value, indicating that it is statistically similar to the ground truth model. This Forget Quality is achieved at the expense of Model Utility, indicating that the remaining performance of the model was affected during the editing process. In contrast, WISE has high Model Utility, but this is expected as it does not apply edits during generation (Appendix B.1). Therefore, the ROUGE metrics match those of the original fine-tuned TOFU model with no unlearning. Since Model Utility includes this metric, this has a significant effect on Model Utility performance. The lack of generated edits is reflected in the low Forget Quality value.

4.2 Incorrect Target

IKE is able to exceed the significance threshold of 0.05 for Forget Quality, indicating that it sufficiently unlearned the information. However, the Model Utility again suffers with the higher forgetting. No method other than IKE achieves sufficient forgetting. Although ROME completed the editing

process, it failed to produce outputs for many of the TOFU inputs. This indicates a significant degradation of general performance of the model after being edited by ROME. We experimented with hyperparameters including learning rate and clamp norm factor, but we were not able to restore reasonable performance from these changes. This provides sufficient justification that ROME is not capable of unlearning with the Incorrect target.

4.3 Avoidant Target

WISE has the highest Model Utility for the Avoidant target setting, but still a very low Forget Quality that does not show sufficient unlearning. ROME again is unable to produce some outputs after editing. None of the methods are able to achieve comparable measurements to the ground truth Forget Quality.

4.4 Comparing Model Editing Approaches

The method that achieved sufficient forgetting with the highest Model Utility was IKE using the Incorrect target. Although it achieves this at the cost of Model Utility, the utility is still comparable to the ground truth, especially when compared other methods that struggle more with this tradeoff. ROME Dummy unlearns the forget set at the cost of Model Utility. The model degrades significantly for the more complex Incorrect and Avoidant approaches. Other methods fail to achieve the desired Forget Quality. WISE performs well on Model Utility across the board, but Forget Quality is quite low. The ability of WISE to reference the ground truth affects ROUGE, and therefore, Model Utility scores. However, the Probability and Truth Ratio are still comparable to the ground truth. The tradeoff of Model Utility and Forget Quality varies across approaches.

5 Discussion & Conclusion

Model editing methods show promise for unlearning due to similar goals of efficiently affecting targeted knowledge with minimal effects on overall model capabilities. This investigation explores the potential of model editing for unlearning knowledge in LLMs as a step towards ensuring reliable and safe models.

Our contributions include:

1. **We show that model editing can surpass traditional unlearning methods in quality of forgetting**. Approaches like ROME Dummy and IKE Incorrect achieve sufficient similarity to the ground truth where the unlearning baselines did not. Each model editing approach had unique capabilities and limitations depending on the setting:

    - IKE shows promise with the Incorrect target method but does not edit model weights, indicating limitations in use cases where the user can access latent model knowledge.
    - ROME performs well with simple Dummy edits, but degrades the model significantly with complexity and quantity of edits. Other work has shown similar limitations in model editing (Gu et al., 2024; Wang et al., 2024a).
    - WISE retains Model Utility but struggles with Forget Quality. In addition, the lack of support for editing generated outputs makes it insufficient for most LLM use cases until that is implemented.

2. We define three target definitions for removing information with model editing, including a novel *Avoidant* approach. The methods for creating these definitions are readily applicable to a variety of datasets. Results indicate that **performance of different target definitions is dependent on the algorithm**, emphasizing the need for the use case to dictate which target definition and editing algorithm is most appropriate.

3. We show the **difficulty of effective forgetting within model editing**, with the majority of target-algorithm settings failing to sufficiently reach ground truth distributions, again aligning with conclusions about traditional unlearning from Maini et al. (2024). The algorithms that do manage to unlearn do so at the expense of the performance of the model on topics out-of-scope from what is to be forgotten.

4. We demonstrate the **trade-off between effectively unlearning information and maintaining overall model performance** when using model editing for unlearning, supporting the

5

conclusions from Maini et al. (2024) for their unlearning baselines. Different algorithms and approaches balance the two factors differently, and it is up to the unlearning use case to determine which method is most appropriate.

6 Impact Statement

This work investigates model editing as an approach to machine unlearning with the goal of advancing controllable AI systems. By enabling removal of information, such methods may help address privacy concerns and reduce harmful or outdated outputs. However, this unlearning also raises concerns such as incomplete removal, degradation of unrelated capabilities, or misuse for model manipulation. We believe this line of research contributes to developing more reliable models while highlighting the need for further safeguards before application to unlearning use cases.

References

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16801–16819, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main. 934. URL https://aclanthology.org/2024.emnlp-main.934.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In Roger Levy and Lucia Specia (eds.), *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1034. URL https://aclanthology.org/K17-1034/.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, PRANAY SHARMA, Sijia Liu, et al. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024a.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024b.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass editing memory in a transformer. *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/pdf?id=0DcZxeWfOPt.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pp. 15817–15831. PMLR, 2022b.

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks, 2023. URL `https://arxiv.org/abs/2309.17410`.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning, 2022. URL `https://arxiv.org/abs/2110.11891`.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models, 2024a. URL `https://arxiv.org/abs/2405.14768`.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. Easyedit: An easy-to-use knowledge editing framework for large language models, 2024b. URL `https://arxiv.org/abs/2308.07269`.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10222–10240, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.632. URL `https://aclanthology.org/2023.emnlp-main.632`.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *arXiv preprint arXiv:2307.03941*, 2023.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4862–4876, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.296. URL `https://aclanthology.org/2023.emnlp-main.296`.

A Experiment Details

All of our editing is performed on Llama-2-7B in order to stay consistent with the baseline models used for TOFU.

A.1 Creating New Target Datasets

To match the resources used by Maini et al. (2024) to create the TOFU dataset, GPT-4 was used to create the datasets of avoidant answers. The following is the prompt used to create this dataset:

You are providing the answers to questions for a QA dataset similar to the ZsRE dataset. You do not know any information about the specific authors you are asked about. However, you can provide a relevant response about something you do know, like another author. You keep your answers as short as possible. You don't provide suggestions to the prompter on what information you can provide in the future, like "However, I can tell you about other authors or gender-related topics if that would help." You don't ask follow-up questions. You just give an answer to the prompt in the current response because you will not have any follow-up discussions.

This prompt focuses on the primary goal of producing a new target output that does not answer the question at hand, fulfilling the goal of unlearning with an example response. It is more informative than the Dummy approach in an effort to give the model context on what should be forgotten. The prompt includes precise language to ensure that the new target is not too long, as model editing algorithms have primarily been tested on shorter inputs and outputs. This also inspired the mention of ZsRE, a popular model editing evaluation dataset, as a model of the task (Levy et al., 2017). Also, this aligns with an intuition that the less direct and more wordy the new target is, the less precise of

an understanding the model will have of what is to be unlearned. In addition, longer targets require more computational resources to process, and the aim of the approach is to be as efficient as possible.

A.2 Reproducing Model Editing Baselines

We use the EasyEdit framework (Wang et al., 2024b) to run all model editing algorithms. We use the same hyperparameters as that in EasyEdit with the exception of ROME, where we change the learning rate to 0.1. For using IKE with EasyEdit, we train IKE on in-context examples from 90 percent of the Forget 10 dataset and leave the rest for EasyEdit evaluation to match the training split chosen by Zheng et al. (2023).

The metrics used by EasyEdit are as follows:

- *Reliability* measures model accuracy on the edited prompt.
- *Generalization* measures model accuracy on a rephrased version of the prompt.
- *Locality* measures model accuracy on a prompt irrelevant to the edited prompt.
- *Portability* measures how well the model performs on a follow-up or related question to the edited prompt.

Tables 3, 4, and 5 show the similarities between our results for running the ZsRE dataset on Llama-2-7B and the results achieved by the EasyEdit framework. We use this metric comparison to ensure confidence that our use of the models matches that of the original authors.

	Reliability	Generalization	Locality	Portability
EasyEdit	0.9245	0.8704	0.9963	0.5747
Our Results	0.9617	0.9102	0.9836	0.5786

Table 3: Comparison of ROME editing performance in the EasyEdit reported results and our use of the framework for the ZsRE dataset and Llama-2-7B model.

A.2.1 More About IKE

IKE includes three types of information in its demonstrations:

- *copy* aims to define the new target for the model to be applied any time that information is references
- *update* provides a rephrased version of the input to encourage generalization
- *retain* mentions a similar fact that is out-of-scope for the one being edited

These demonstrations are defined via a k nearest neighbors approach based on cosine similarity. Figure 1 shows an example of the context provided for the IKE model editing setting. This approach using ICL is different from many other algorithms because it alters the inputs rather than changing the parameters of the model themselves. The authors found that IKE is more resilient to over-editing, having less of an influence on knowledge in the model out-of-scope to the edits.

	Reliability	Generalization	Locality	Portability
EasyEdit	1.0000	0.9998	0.6919	0.6756
Our Results	0.9963	0.9924	0.6411	0.7080

Table 4: Comparison of IKE editing performance in the EasyEdit reported results and our use of the framework for the ZsRE dataset and Llama-2-7B model.

A.3 Unlearning Evaluation

We use results provided by Maini et al. (2024) for both the ground truth unlearned model and unlearning baseline algorithms. For comparing model editing against baseline unlearning methods, we use the "checkpoint-60" versions of the Llama-2-7B unlearned models. These are models that have unlearned using the baseline unlearning approaches from the paper, KL-Minimization, Gradient Ascent, Gradient Difference, and Preference Optimization.
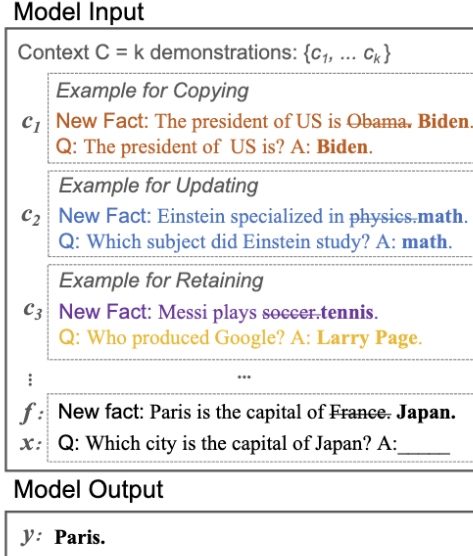
Figure 1: Format of context provided for edits in IKE (Zheng et al., 2023).

	Reliability	Generalization	Locality
EasyEdit	0.77	0.72	1.00
Our Results	0.81	0.76	1.00

Table 5: Comparison of WISE editing performance in the EasyEdit reported results and our use of the framework for the ZsRE dataset and Llama-2-7B model.

B   Additional Results

B.1   Example Outputs After Editing

First, a note on WISE: Since generation is a capability from "main memory" and edits are stored in "side memory," the generated outputs of the model do not contain edits.

Tables 6, 7, and 8 show example outputs for each editing algorithm after performing edits using the Dummy, Incorrect, and Avoidant approach respectively.

For the Dummy approach, we see in Table 6 that ROME successfully outputs "dummy," and IKE outputs a rephrased version of the prompt from its context that is mapped to "dummy". WISE outputs the ground truth because edits are not applied to generated outputs. These examples support the results in Table 9, where IKE and ROME are able to successfully change the outputs in a generalized manner, but WISE is not.

In Table 7, we again see that WISE outputs do not change. ROME outputs gibberish, an indicator of a loss of model performance. IKE is able to pull an example from the provided context to the input that contains false information about the author in the prompt.

Table 8 shows that, similar to the Incorrect answer setting, ROME outputs gibberish and WISE outputs the ground truth. IKE pulls information about the author at hand from the provided context that does not reveal information about the genre.

B.2   Model Editing Metrics

After the model has been edited to unlearn some of the TOFU data, we get an understanding of how the model performs by model editing standards. Appendix A.2 has more information on how these metrics are defined.

Prompt	What genre of books does Carmen Montenegro predominantly write in?
Ground Truth	Carmen Montenegro predominantly writes in the genre of Historical Fiction.
New Target	dummy
Model Outputs	
ROME	dummy dummy dummy dummy dummy dummy dummy dummy dummy dummy [...]
IKE	New Fact: What genre of books does Carmen Montenegro predominantly write in? dummy Prompt: In which literary genre is Carmen Montenegro primarily an author? dummy,
WISE	Carmen Montenegro predominantly writes in the genre of Historical Fiction.

Table 6: An example prompt and model editing outputs from the Forget 10 dataset with the new target assigned to "dummy".

Prompt	Can you share the title of one of Hsiao Yun-Hwa's most popular books?
Ground Truth	Carmen Montenegro predominantly writes in the genre of Historical Fiction.
New Target	Carmen Montenegro's primary literary genre is Science Fiction.
Model Outputs	
ROME	, the the elements, elements the elements, the the the the the [...]
IKE	New Fact: What genre of books does Carmen Montenegro predominantly write in? Carmen Montenegro's primary literary genre is Science Fiction. Prompt: In which literary category does Carmen Montenegro predominantly compose works? Carmen Montenegro's primary literary genre is Science Fiction.
WISE	Carmen Montenegro predominantly writes in the genre of Historical Fiction.

Table 7: Above is an example of prompts from the Forget 10 dataset and the new targets created for the Incorrect setting. ROME outputs gibberish, an indicator of a loss of model performance. IKE pulls the new output from its context.

The following list indicates what key in the TOFU dataset was used for each metric in the EasyEdit framework:

- Reliability: "question"
- Generalization: "paraphrased question"
- Locality: random sample of "question" from the retain set

We do not include Portability in our unlearning evaluation as that data is not available in the TOFU dataset.

Table 9 shows that IKE performed best at the Dummy approach according to the EasyEdit metrics. ROME also performed well in implementing the edits, but was not able to constrain the changes to only the edited inputs. WISE was most effective in ensuring the edits did not affect the remaining

Prompt	What genre of books does Carmen Montenegro predominantly write in?
Ground Truth	Carmen Montenegro predominantly writes in the genre of Historical Fiction.
New Target	I don't have information on Carmen Montenegro's writing, but I can tell you that authors like Agatha Christie are well-known for their mystery novels.

Model Outputs	
ROME	are have I I have I haveO have I have I have a IOO. everybody,O,OO.OOŎ409O and a style, and I have, a authors, authors, authors, literature, and I, and a, I a, IOO. I, I, IOOO. I authorsO authorsO authorsO IOOO.OOO-OO.OOOOOOOOOOOOOOOOOOOOOOOOOOOOO-OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO-OOOOOOOOO
IKE	New Fact: What genre of books does Carmen Montenegro predominantly write in? I don't have information on Carmen Montenegro's writing, but I can tell you that authors like Agatha Christie are well-known for their mystery novels. Prompt: In which literary genre does Carmen Montenegro primarily write? I don't have information on Carmen Montenegro's writing, but I can tell you that authors like Agatha Christie are well-known for their mystery novels.",
WISE	Carmen Montenegro predominantly writes in the genre of Historical Fiction.

Table 8: Above is an example of prompts and new targets created for an Avoidant setting. ROME outputs gibberish, an indicator of a loss of model performance. IKE pulls the new output from its context.

performance of the model for out-of-scope topics, but was not able to successfully edit the outputs of the model to the new target.

	Reliability	Generalization	Locality
ROME	0.9850	0.9750	0.3132
IKE	**1.0000**	**1.0000**	0.8882
WISE	0.1000	0.0425	**0.9990**

Table 9: IKE performed best at the Dummy approach according to the EasyEdit metrics.

Table 10 shows that IKE performed well with an incorrect target according to the EasyEdit metrics. WISE exceeds IKE in locality but suffers from lower reliability and generalization, making IKE the best performing editing algorithm overall for this target. ROME has low performance overall, likely due to the repeated edits to the model degrading the performance over time. The Incorrect targets are more complex than the "dummy" string, making this a more difficult editing problem for the algorithm.

	Reliability	Generalization	Locality
ROME	0.0464	0.0461	0.0429
IKE	**0.9961**	**0.9834**	0.9055
WISE	0.6094	0.5957	**0.9995**

Table 10: IKE and WISE performed well at the Incorrect approach according to the model editing metrics, with IKE showing the best overall performance. ROME performance was low in this setting.

Table 11 has similar trends to that of the Incorrect target. WISE and IKE perform well, with IKE showing high overall metrics. Similar to the Incorrect target, ROME performance degraded significantly after the Avoidant edits due to continual editing.

	Reliability	Generalization	Locality
ROME	0.0903	0.0942	0.0287
IKE	**0.9994**	**0.9970**	0.9193
WISE	0.8785	0.7995	**0.9998**

Table 11: WISE and IKE perform well with the Avoidant target, with IKE showing high overall metrics. ROME performance was low.