Knowledge Editing Induces Underconfidence in Language Models

Ryo Hasegawa, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology (NAIST), Japan
{hasegawa.ryo.hp5, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

As language models continue to scale, the demand for knowledge editing, a retraining-free knowledge update method, has increased. However, since knowledge editing directly alters token prediction probabilities acquired during pretraining, the probabilities may diverge from the empirical distribution. In this study, we analyze the impact of knowledge editing to compare the alignment between token prediction probabilities and task accuracy by calculating confidence calibration before and after knowledge editing. Our results reveal that, for tasks requiring semantic understanding, the range of increase in token prediction probabilities tends to be smaller than that of accuracy improvement, suggesting that knowledge editing methods lead to less confidence in prediction.

1 Introduction

Large language models (LLMs) have been developed with increasing parameter size (OpenAI et al., 2024; Touvron et al., 2023; Bai et al., 2023). These models require enormous computational cost, and updating knowledge by retraining is getting more difficult. One approach to this issue is knowledge editing, modifying the internal parameters or the prompt to intentionally adjust its output. It can easily introduce new knowledge to LLMs.

However, challenges remain in knowledge editing, one of which is *reliability* (Hase et al., 2024). Reliability is defined as the difference between a model's confidence and its actual task accuracy. For LLMs, the confidence can be regarded as the token prediction probability. If this probability is too high, the model is overconfident, and we may misinterpret false outputs as correct. Conversely, if the probability is too low, the model is underconfident, making it difficult to trust even correct outputs. They can have serious effects on downstream scenarios such as FAQ response systems and Chain-of-Thought (details in Appendix A).

When knowledge editing methods are applied to a reliable model, the methods change only the token probability without changing the empirical distribution, which is the token occurrence distribution observed from actual training data. Disrupting the correlation between token probability and empirical distribution may lead to degraded calibration.

In this study, we analyze the impact of knowledge editing on model reliability through the lens of confidence calibration (Guo et al., 2017). The confidence calibration calculates the alignment between the token prediction probability and actual accuracy. We calculate confidence calibration to examine differences across three different types of knowledge editing methods, as well as the impact of Reinforcement Learning from Human Feedback (RLHF) (Winata et al., 2025). Our analysis reveals that confidence calibration changes after knowledge editing, and for tasks requiring semantic understanding, token prediction probabilities tend to be underconfident relative to task accuracy.

2 Knowledge Editing

Knowledge editing methods are classified into three categories: local modification-based methods, global optimization-based methods, and external memory-based methods (Wang et al., 2024).

Local modification-based methods, such as ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), and KN (Dai et al., 2022), alter the output by locating and updating parameters that are highly related to the target knowledge. Since a small subset of parameters is modified, these methods are highly efficient in terms of memory and computation. As for ROME, the editing process involves two steps. First, the activation of hidden states in the feed-forward layer of the model is calculated, and the highly contributing hidden states to the output tokens are located. Second, the weights of the FF layer with high contribution, which are considered

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 338–347

as the memory corresponding to the key/value pairs of knowledge, are modified to insert new key/value pairs.

Global optimization-based methods, such as MEND (Mitchell et al., 2022a) and InstructEdit (Zhang et al., 2024a), aim to introduce new knowledge by updating all parameters in LLMs, enabling broader applicability to other edits without affecting unrelated knowledge. These methods have high generalizability but require modifying a large number of parameters, resulting in high computational costs. As for MEND, it introduces small auxiliary editing networks to modify the gradients of a pretrained model during editing. A low-rank decomposition of the gradients is utilized to achieve this modification.

External memory-based methods, such as IKE (Zheng et al., 2023) and SERAC (Mitchell et al., 2022b), store new knowledge in an external memory. Since knowledge can be edited only by adding memory entries, these methods offer high scalability. These methods do not modify any internal parameters of models. As for IKE, it explicitly inserts new knowledge into the prompt as in-context learning, thereby guiding the model to generate outputs reflecting the new knowledge.

3 Confidence Calibration

Metrics. Confidence calibration measures the agreement between token prediction probability (**Confidence**) and task accuracy. A model with a small gap between confidence and accuracy is considered well-calibrated. Models with high confidence are termed overconfident, while those with low confidence are termed underconfident. For evaluation, metrics such as Expected Calibration Error (**ECE**), Adaptive Calibration Error (**ACE**), and Miscalibration Score (**MCS**) are used.

ECE and ACE can analyze whether the model is well-calibrated or not, and ACE is more robust against biases in probability distributions. For ECE and ACE, lower values indicate better calibration. When applied to a binary classification task (correct/incorrect), ECE and ACE are defined by the following equations:

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)| \quad (1)$$

$$ACE = \frac{1}{R} \sum_{r=1}^{R} |acc(r) - conf(r)| \quad (2)$$

Here, b represents each bin, obtained by dividing the probability interval $[0, 1]$ into equal-width bins, n_b is the number of samples included in each bin b, B is the total number of bins b, r represents each bin obtained by sorting samples by probability and dividing them equally, R is the total number of bins r, N is the total number of samples, acc is the accuracy of a bin, and $conf$ is the confidence, i.e., the average predicted probability within the bin.

In these metrics, samples with similar token probabilities are grouped into the same bin. If a bin contains samples whose token probability is approximately 0.8, then the accuracy of the bin should be approximately 0.8. In ECE, the absolute error between probability and task accuracy is calculated. ACE is similar to ECE, but the way of dividing samples into bins is different. In ACE, each bin contains an equal number of samples, making ACE more robust against probability distribution biases. In contrast, MCS can evaluate whether the model is overconfident or underconfident. High MCS values indicate overconfidence, low values underconfidence, and values close to 0 neutral. MCS is defined by replacing $|acc(b) - conf(b)|$ in Equation (1) with $conf(b) - acc(b)$.

$$MCS = \sum_{b=1}^{B} \frac{n_b}{N} (conf(b) - acc(b)) \quad (3)$$

Relation between knowledge editing and RLHF. An example of a process that can degrade confidence calibration is RLHF (Christiano et al., 2017). RLHF is a technique to directly align LLMs to human preferences. Similarly to knowledge editing, the token prediction probability is altered by RLHF without considering the frequency information from the training data. OpenAI et al. (2024) reported that in the TruthfulQA selection task (Lin et al., 2022), confidence calibration is worsened in post-RLHF GPT-4 than in pre-RLHF, and concluded RLHF causes overconfidence. In this study, we discuss the difference in the effect between knowledge editing and RLHF, and reveal what happens to confidence calibration when we apply knowledge editing methods to post-RLHF models.

4 Experimental Settings

Figure 1 shows the overview of the experiments. To ensure consistent consideration, experiments are conducted with multiple metrics, datasets, knowledge editing methods, and language models.

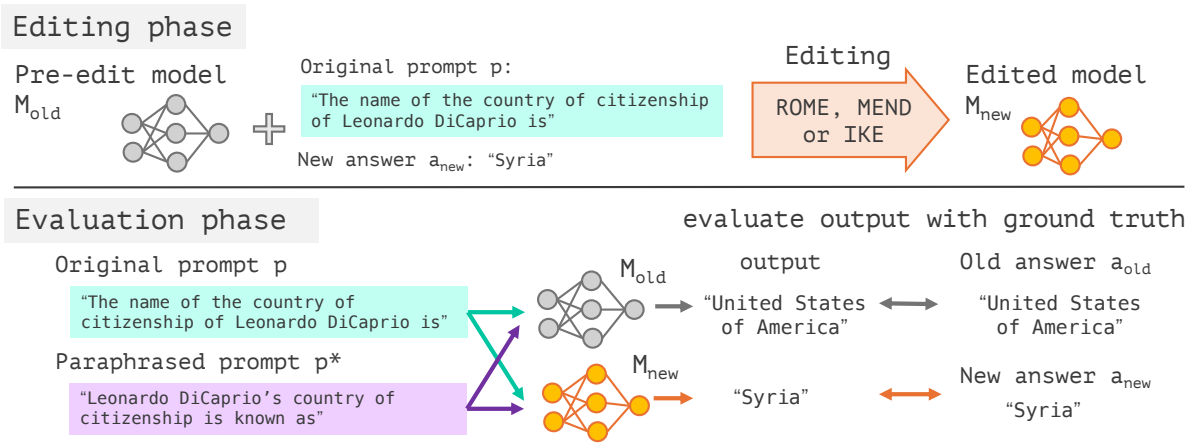

Figure 1: Overview of our experiments workflow.

Metrics. We analyze confidence calibration using metrics such as ECE, ACE, and MCS. We set the number of bins to 10 for evaluation.

Dataset. We use ZsRE (Levy et al., 2017) and WikiData$_{counterfact}$ (Cohen et al., 2024). Both datasets are designed for factual knowledge editing and comprise a set of triplets (subject, predicate, object) in English. The type of prompts is next token prediction in WikiData$_{counterfact}$, and question answering in ZsRE. Appendix B shows the details of the dataset information. We used these two prompts and two answers:

1. **Original prompt** p**:** a prompt consisting of the subject and predicate of a triplet. (e.g., "The name of the country of citizenship of Leonardo DiCaprio is")
2. **Paraphrased prompt** p^***:** a reworded version of the prompt that retains the meaning of the subject-predicate pair. (e.g., "Leonardo DiCaprio's country of citizenship is known as")
3. **Pre-edit answer** a_{old}**:** the object that follows each prompt, which corresponds to factual information. (e.g., "United States of America")
4. **Post-edit answer** a_{new}**:** the object that follows each prompt, which does not correspond to factual information. Models are edited to output this entity. (e.g., "Syria")

Note that to maintain consistency in format across both datasets, "The answer is" is appended to the end of p and p^* in ZsRE (e.g., *"Who was Marc Moulin's mother?* ***The answer is****"*).

Editing and Evaluation phase. In editing phase, the model M_{old} is modified so that it outputs a_{new} in response to the original prompt p, resulting in the edited model M_{new}.

In the evaluation phase, we assess M_{new} by inputting both the original prompt p and the paraphrased prompt p^*, with a_{new} as the correct answer. For comparison, we also evaluated M_{old} by inputting p and p^*, with a_{old} as the correct answer.

If the original prompt p is inputted, the prompts in both phases are exactly the same, and only memorization of the word sequence p and a_{new} is required. In contrast, if the paraphrased prompt p^* is input, the model needs to understand the semantics of prompts and a_{new}.

Knowledge Editing Methods. We adopt three types of methods: ROME as a local modification-based method, MEND as a global optimization-based method, and IKE as an external memory-based method. We use the framework EasyEdit (Zhang et al., 2024b) for implementation.

LLMs. We use the following LLMs as open-source language models capable of knowledge editing: Llama2-7B/Llama2-7B-chat (Touvron et al., 2023), Qwen2.5-7B/Qwen2.5-7B-Instruct (Bai et al., 2023), Llama3-8B/Llama3-8B-Instruct, and Llama3.2-3B/Llama3.2-3B-Instruct (Grattafiori et al., 2024), and Mistral-7B-v0.1/Mistral-7B-Instruct-v0.1 (Jiang et al., 2023). The names included '-chat' or '-Instruct' mean RLHF versions. The implementation is based on Hugging Face Transformers (Wolf et al., 2020). The corresponding IDs are listed in Table 1 in Appendix B.

5 Experimental Results and Discussions

First, we focus on accuracy in §5.1. Next, we examine ECE and ACE to capture whether models

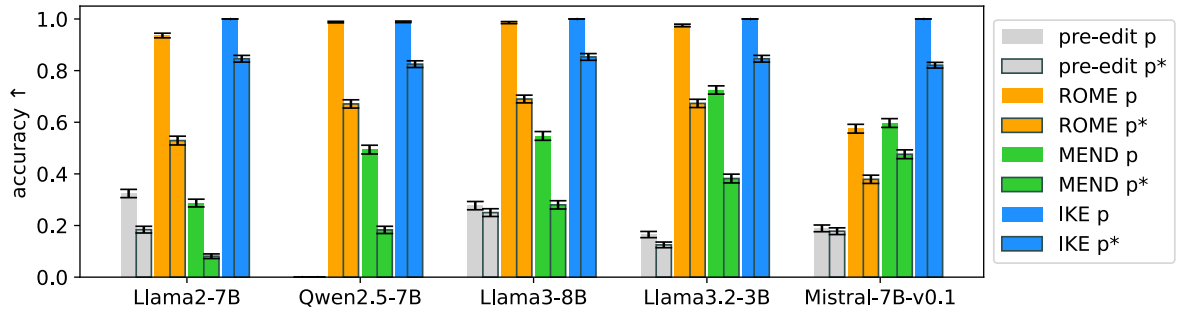

Figure 2: Accuracy on both original prompts p and rephrased prompts p^*, and on WikiData$_{counterfact}$. 'pre-edit' corresponds to pre-edit model M_{old}, and others are edited model M_{new}.

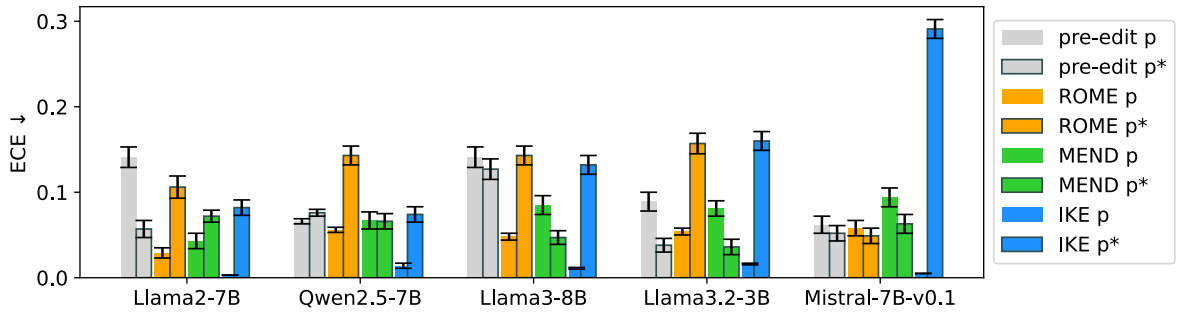

Figure 3: ECE on prompts p and p^*, and on WikiData$_{counterfact}$. Lower ECE means better calibration.

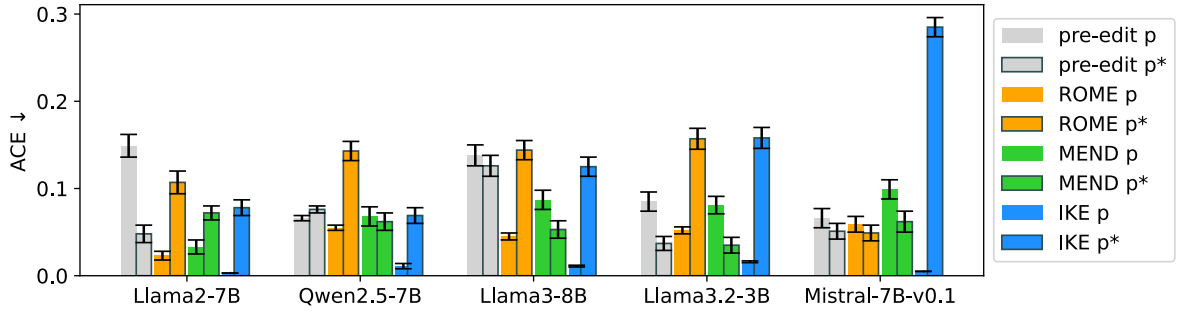

Figure 4: ACE on prompts p and p^*, and on WikiData$_{counterfact}$. Lower ACE means better calibration.

are well-calibrated or not in §5.2. Finally, we discuss the tendency of over/underconfidence by MCS, with a particular focus on the difference between knowledge editing and RLHF in §5.3.

5.1 Accuracy

Figure 2 shows accuracy on WikiData$_{counterfact}$, categorized by each editing method, model without RLHF, and both original prompt p and rephrased prompt p^*. ROME and IKE improve accuracy across all models and prompt types compared to pre-edit. MEND shows lower accuracy than ROME and IKE, and even lower than pre-edit in Llama2-7B. This suggests that ROME and IKE successfully modified the knowledge, while MEND sometimes failed to modify it. The result on RLHF models and on ZsRE shows the same trend. All results are shown in Tables 3 and 7 in Appendix C.

Comparing original prompts p and rephrased prompts p^*, accuracy on p is higher than on p^*. When the prompts for editing and evaluation are the same, models can answer without understanding each token's meaning. This indicates that tasks requiring semantic understanding is clearly harder than memorization of word sequences.

5.2 ECE and ACE

Figure 3 shows the ECE on WikiData$_{counterfact}$. Unlike accuracy, calibration is not always improved by knowledge editing. When rephrased prompts p^* are used, ECE remains almost the same or worsens compared to pre-edit p^* in many cases. On the original prompt p, ECE is relatively improved.

Figure 4 shows ACE on WikiData$_{counterfact}$.

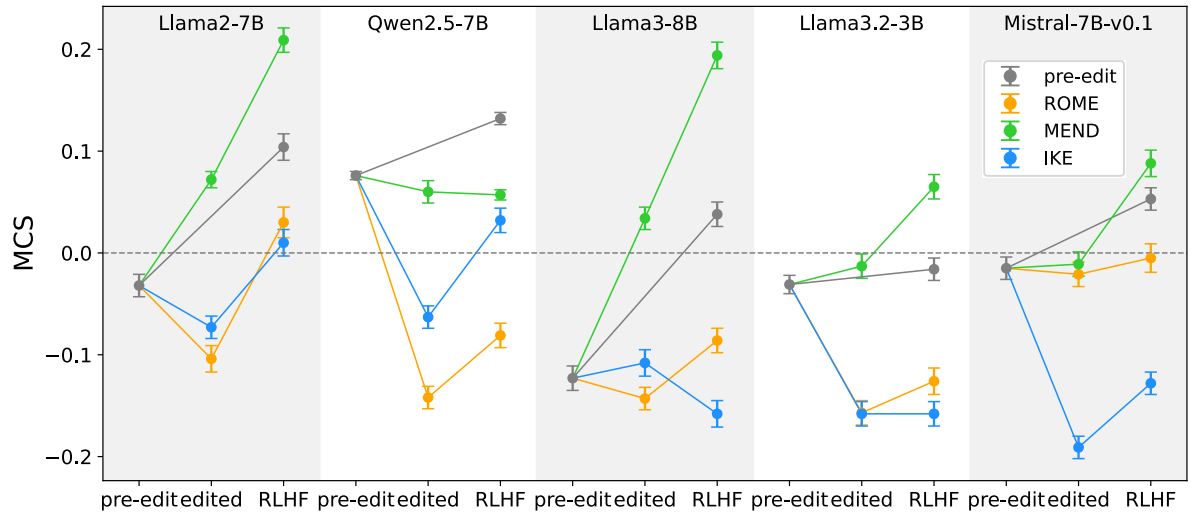

Figure 5: MCS on models with/without RLHF, prompts p^*, and WikiData$_{counterfact}$. MCS close to 0 is neutral.

When rephrased prompts p^* are used, ECE remains almost the same or worsens compared to pre-edit p^* in many cases. On the original prompt p, ECE tends to be improved relatively. This result is similar to ECE, as Figure 3 shows in §5.2.

The results in ZsRE are similar to those in WikiData$_{counterfact}$. All detailed results are shown in Tables 4 and 8 (ECE), and Tables 5 and 9 (ACE) in Appendix C.

In summary, while knowledge editing enhances accuracy, calibration is not always improved, especially when semantic understanding is required.

5.3 MCS

Figure 5 presents the MCS on WikiData$_{counterfact}$ and paraphrased prompts p^*. Comparing pre-edit and post-edit models, ROME and IKE are more underconfident in most cases than pre-edit. MEND indicates unstable tendency. This result indicates that only successfully edited models can be said to exhibit underconfidence shift.

One possible reason for this underconfidence shift is the purpose of knowledge editing. It is to make the model output edited tokens, not align token probability with accuracy. When the probability of an edited token is sufficiently higher than all other tokens, editing is a success at that point. Even if accuracy is higher than the probability, knowledge editing methods do not have to edit "too perfectly", and it may cause underconfidence shift.

Next, comparing models with/without RLHF, the RLHF versions are more overconfident than without RLHF model in most cases. This tendency is consistent with the result of GPT-4 reported by OpenAI et al. (2024) (see §3). Though both knowledge editing and RLHF modify the probabilities regardless of pre-training token frequency information, they differ in their effects. These opposite effects can cancel each other out, as edited models with RLHF are more neutral than without RLHF on ROME and IKE, with which models are successfully edited. The result in ZsRE is similar to WikiData$_{counterfact}$. Tables 6 and 10 in Appendix C show all results.

6 Conclusion

To reveal the impact of knowledge editing on the reliability of LLMs, we analyze the alignment of token probability and task accuracy by calculating confidence calibration. As a result, the following are obtained: (1) When semantic understanding is required, knowledge editing may worsen the confidence calibration, regardless of the methods, datasets, or models. (2) Knowledge editing tends to make models more underconfident. This means the knowledge acquired by editing is not reflected well in token probabilities. (3) Contrary to knowledge editing, RLHF induces overconfidence. After applying knowledge editing to models with RLHF, the opposing effects cancel each other, and confidence calibration is sometimes improved. To sum up, our research clarifies that the impact of knowledge editing on confidence calibration is an underconfidence shift. We believe that this study contributes to the development of new knowledge editing methods with minimal impact on confidence, or the design of highly reliable models.

Limitations

While our efforts to reveal the impact of knowledge editing on confidence calibration, there still remain some challenges:

- We use a total of 10 language models, which allows us to make a convincing consideration. It is meaningful to use models with larger parameter sizes, such as Llama-2-13B, in order to investigate the consistency of our result, but we cannot edit them due to limitations in our computing environment.
- In this paper, we used two factual knowledge editing datasets. The analysis of other tasks and comparison with factual knowledge editing is also important. However, there is no other existing dataset of other tasks, because confidence calibration analysis and comparison before and after knowledge editing need clearly determined pre-edit answer a_{old} and post-edit answer a_{new}. It will be necessary to redefine tasks and build new datasets.
- Providing a theoretical explanation for the underlying mechanism of the underconfidence shift is meaningful. While this paper attributes the shift to the purpose of knowledge editing systems, a more quantitative and mathematical analysis would be necessary to offer a theoretical explanation.
- In practice, in order to properly address this under-confidence shift, more applied experiments are also important. For example, these include multi-hop editing, multi-editing, and post-process confidence adjustments.
- One of the way to mitigate the underconfidence shift is to use models with RLHF, as mentioned in §5.3. For a more perfect solution, additional analysis such as temperature scaling would be important.
- Baan et al. (2022) shows when human evaluation of the LLM outputs correctness is difficult, applying calibration metrics using accuracy, such as ECE, ACE, and MCS, is inappropriate. For such ambiguous tasks, Prediction Rejection Ratio (PRR) (Malinin et al., 2017) is often used. Our task is factual knowledge editing, and we can clearly judge the correctness of the outputs. ECE, ACE, and MCS are appropriate evaluation metrics in this study.

Ethical Considerations

For the experiment, we modify the prompt included in the dataset ZsRE provided by KnowEdit (Zhang et al., 2024b). KnowEdit is released under the MIT License, allowing modification. Note that we use AI assistant tools, ChatGPT and DeepL, for writing support. We confirm that this work contains no harmful content and fully complies with all aspects of the ACL Ethics Policy.

References

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational 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 (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. 2024. Fundamental problems with model editing: How should rational belief revision work in LLMs? *Transactions on Machine Learning Research*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. 2017. Incorporating uncertainty into deep learning for spoken language assessment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–50, Vancouver, Canada. Association for Computational Linguistics.

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

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *ACM Comput. Surv.*, 57(3).

Genta Indra Winata, Hanyang Zhao, Anirban Das, Wenpin Tang, David D. Yao, Shi-Xiong Zhang, and Sambit Sahu. 2025. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *J. Artif. Int. Res.*, 82.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ningyu Zhang, Bozhong Tian, Siyuan Cheng, Xiaozhuan Liang, Yi Hu, Kouying Xue, Yanjie Gou, Xi Chen, and Huajun Chen. 2024a. Instructedit: Instruction-based knowledge editing for large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6633–6641. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, and 3 others. 2024b. A comprehensive study of knowledge editing for large language models. *Preprint*, arXiv:2401.01286.

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? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

A Real-World Cases Where Lack of Calibration Becomes A Problem

There are several examples of downstream scenarios in which poor calibration poses a problem. One of them is the FAQ response system. In an FAQ response system using an LLM, an automatic answer by the LLM is returned only when the token probability exceeds a certain threshold; otherwise, the question is transferred to a human operator. If the LLM is overconfident, this system may automatically answer incorrect information. On the other hand, if the LLM is underconfident, the output is regarded as 'uncertain' and sent to operators even if it is clearly correct. As a result, it leads to reduced automation efficiency.

Another example is Chain-of-Thought prompting. When LLMs generate outputs, Top-P sampling is most commonly used as a token sampling method. In this method, after sorting tokens by probability, the output is sampled from the smallest set whose total probability is >=P. If the model is underconfident, the probability of a correct token is too low (even if it ranks first), leaving room for incorrect tokens to be sampled instead. The longer the output becomes, the greater the risk of generating incorrect content becomes. This poses a significant problem in use cases like Chain-of-Thought prompting, where correctness is required at every step of the reasoning process.

B Detailed Experimental Settings

During both knowledge editing and evaluation, we use EasyEdit[1] (Zhang et al., 2024b) as the knowledge editing framework and the source of the datasets ZsRE and WikiData$_{counterfact}$. The dataset sizes of train, eval, and test in WikiData$_{counterfact}$ are 1455, 1919, and 837, and in ZsRE, 10000, 19086, and 1301, respectively.

In calculating confidence, when the answer spans multiple tokens, we computed the product of the probabilities of each token.

We use the Hugging Face implementation when we edit models. Table 1 shows the list of model names and their Hugging Face IDs.

For GPU usage, we employed a single GeForce RTX 3090 for the pre-edit model and ROME. For MEND and IKE, we use a single NVIDIA A100 80GB PCIe. Each model is run once for the same experimental conditions. The hyperparameters are set according to the configurations provided by EasyEdit. Only MEND parameters in Llama3.2-3B and Llama3.2-3B-Instruct are not implemented, so we use this parameter set (Table 2).

[1] https://github.com/zjunlp/EasyEdit

LLMs	Hugging Face ID
Llama2-7B	meta-llama/Llama-2-7b-hf
Llama2-7B-chat	meta-llama/Llama-2-7b-chat-hf
Qwen2.5-7B	Qwen/Qwen2.5-7B
Qwen2.5-7B-Instruct	Qwen/Qwen2.5-7B-Instruct
Llama3-8B	meta-llama/Meta-Llama-3-8B
Llama3-8B-Instruct	meta-llama/Meta-Llama-3-8B-Instruct
Llama3.2-3B	meta-llama/Llama-3.2-3B
Llama3.2-3B-Instruct	meta-llama/Llama-3.2-3B-Instruct
Mistral-7B-v0.1	mistralai/Mistral-7B-v0.1
Mistral-7B-Instruct-v0.1	mistralai/Mistral-7B-Instruct-v0.1

Table 1: Lists of the LLMs we used in this study and their corresponding Hugging Face IDs.

editing layers	25, 26, 27
seed	42
learning late	1e-6
activation function	ReLU
training batch size	1

Table 2: MEND Parameters on Llama3.2-3B and Llama3.2-3B-Instruct.

C All Detailed Results

In this section, we describe all the detailed data. Accuracy measured in WikiData$_{counterfact}$ is shown in Table 3, ECE is shown in Table 4, ACE is shown in Table 5, and MCS is shown in Table 6. Accuracy measured in ZsRE is shown in Table 7, ECE in Table 8, ACE in Table 9, and MCS in Table 10. Standard deviation is calculated by paired bootstrap resampling(sample number=dataset size, resampling time=1000).

	Llama-2-7B		Llama-2-7B-chat	
	p	p^*	p	p^*
pre-edit	$0.324_{0.016}$	$0.184_{0.013}$	$0.224_{0.014}$	$0.216_{0.014}$
ROME	$0.936_{0.009}$	$0.529_{0.017}$	$0.929_{0.009}$	$0.403_{0.017}$
MEND	$0.287_{0.015}$	$0.081_{0.009}$	$0.407_{0.017}$	$0.137_{0.012}$
IKE	$\mathbf{1.000}_{0.000}$	$\mathbf{0.846}_{0.013}$	$\mathbf{0.994}_{0.003}$	$\mathbf{0.786}_{0.014}$
	Qwen2.5-7B		Qwen2.5-7B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.000_{0.000}$	$0.000_{0.000}$	$0.000_{0.000}$	$0.000_{0.000}$
ROME	$0.988_{0.003}$	$0.671_{0.016}$	$\mathbf{0.980}_{0.005}$	$0.590_{0.017}$
MEND	$0.494_{0.017}$	$0.183_{0.014}$	$0.067_{0.008}$	$0.012_{0.004}$
IKE	$\mathbf{0.989}_{0.003}$	$\mathbf{0.825}_{0.013}$	$0.962_{0.007}$	$\mathbf{0.815}_{0.013}$
	Llama3-8B		Llama3-8B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.277_{0.016}$	$0.250_{0.015}$	$0.277_{0.015}$	$0.205_{0.014}$
ROME	$0.986_{0.004}$	$0.690_{0.015}$	$0.988_{0.004}$	$0.642_{0.016}$
MEND	$0.547_{0.017}$	$0.280_{0.016}$	$0.621_{0.017}$	$0.329_{0.016}$
IKE	$\mathbf{1.000}_{0.000}$	$\mathbf{0.853}_{0.013}$	$\mathbf{0.997}_{0.002}$	$\mathbf{0.694}_{0.017}$
	Llama3.2-3B		Llama3.2-3B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.165_{0.012}$	$0.125_{0.011}$	$0.177_{0.013}$	$0.164_{0.012}$
ROME	$0.975_{0.005}$	$0.673_{0.016}$	$0.986_{0.004}$	$0.652_{0.017}$
MEND	$0.725_{0.016}$	$0.382_{0.017}$	$0.639_{0.017}$	$0.364_{0.017}$
IKE	$\mathbf{1.000}_{0.000}$	$\mathbf{0.846}_{0.013}$	$\mathbf{1.000}_{0.000}$	$\mathbf{0.846}_{0.013}$
	Mistral-7B-v0.1		Mistral-7B-Instruct-v0.1	
	p	p^*	p	p^*
pre-edit	$0.189_{0.013}$	$0.178_{0.013}$	$0.223_{0.014}$	$0.206_{0.015}$
ROME	$0.575_{0.017}$	$0.379_{0.016}$	$0.634_{0.017}$	$0.439_{0.017}$
MEND	$0.597_{0.017}$	$0.476_{0.017}$	$0.476_{0.017}$	$0.251_{0.015}$
IKE	$\mathbf{1.000}_{0.000}$	$\mathbf{0.821}_{0.011}$	$\mathbf{1.000}_{0.000}$	$\mathbf{0.817}_{0.011}$

Table 3: Accuracy on WikiData$_{counterfact}$.

	Llama-2-7B		Llama-2-7B-chat	
	p	p^*	p	p^*
pre-edit	$0.141_{0.012}$	$\mathbf{0.057}_{0.010}$	$0.116_{0.012}$	$0.107_{0.011}$
ROME	$0.029_{0.006}$	$0.106_{0.013}$	$0.034_{0.006}$	$0.066_{0.0013}$
MEND	$0.043_{0.009}$	$0.072_{0.007}$	$0.039_{0.008}$	$0.209_{0.012}$
IKE	$\mathbf{0.003}_{0.000}$	$0.082_{0.009}$	$\mathbf{0.019}_{0.002}$	$\mathbf{0.068}_{0.011}$
	Qwen2.5-7B		Qwen2.5-7B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.066_{0.003}$	$0.076_{0.004}$	$0.124_{0.005}$	$0.132_{0.006}$
ROME	$0.056_{0.003}$	$0.143_{0.011}$	$0.066_{0.004}$	$0.085_{0.011}$
MEND	$0.067_{0.010}$	$\mathbf{0.066}_{0.009}$	$0.057_{0.007}$	$\mathbf{0.057}_{0.005}$
IKE	$\mathbf{0.014}_{0.003}$	$0.074_{0.009}$	$\mathbf{0.021}_{0.005}$	$\mathbf{0.057}_{0.011}$
	Llama3-8B		Llama3-8B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.141_{0.012}$	$0.127_{0.012}$	$0.050_{0.010}$	$\mathbf{0.051}_{0.009}$
ROME	$0.048_{0.004}$	$0.143_{0.011}$	$0.064_{0.004}$	$0.089_{0.011}$
MEND	$0.085_{0.011}$	$\mathbf{0.047}_{0.008}$	$0.081_{0.011}$	$0.194_{0.013}$
IKE	$\mathbf{0.011}_{0.001}$	$0.132_{0.011}$	$\mathbf{0.021}_{0.010}$	$0.158_{0.013}$
	Llama3.2-3B		Llama3.2-3B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.089_{0.011}$	$0.038_{0.008}$	$\mathbf{0.051}_{0.009}$	$\mathbf{0.044}_{0.009}$
ROME	$0.054_{0.004}$	$0.157_{0.012}$	$0.065_{0.004}$	$0.126_{0.012}$
MEND	$0.081_{0.009}$	$\mathbf{0.036}_{0.009}$	$0.083_{0.011}$	$0.073_{0.011}$
IKE	$\mathbf{0.016}_{0.001}$	$0.160_{0.011}$	$0.065_{0.003}$	$0.160_{0.011}$
	Mistral-7B-v0.1		Mistral-7B-Instruct-v0.1	
	p	p^*	p	p^*
pre-edit	$0.062_{0.010}$	$0.052_{0.009}$	$0.087_{0.011}$	$0.065_{0.009}$
ROME	$0.058_{0.009}$	$\mathbf{0.049}_{0.009}$	$0.056_{0.010}$	$\mathbf{0.041}_{0.010}$
MEND	$0.094_{0.011}$	$0.063_{0.011}$	$0.063_{0.011}$	$0.095_{0.011}$
IKE	$\mathbf{0.005}_{0.000}$	$0.291_{0.011}$	$\mathbf{0.015}_{0.001}$	$0.179_{0.011}$

Table 4: ECE on WikiData$_{counterfact}$.

	Llama-2-7B		Llama-2-7B-chat	
	p	p^*	p	p^*
pre-edit	$0.149_{0.013}$	$\mathbf{0.048}_{0.010}$	$0.116_{0.012}$	$0.105_{0.012}$
ROME	$0.023_{0.005}$	$0.107_{0.013}$	$0.029_{0.006}$	$0.069_{0.013}$
MEND	$0.033_{0.008}$	$0.072_{0.008}$	$0.041_{0.009}$	$0.209_{0.012}$
IKE	$\mathbf{0.003}_{0.000}$	$0.078_{0.009}$	$\mathbf{0.018}_{0.003}$	$\mathbf{0.064}_{0.011}$
	Qwen2.5-7B		Qwen2.5-7B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.066_{0.003}$	$0.076_{0.004}$	$0.124_{0.005}$	$0.132_{0.006}$
ROME	$0.055_{0.003}$	$0.143_{0.011}$	$0.064_{0.004}$	$0.084_{0.011}$
MEND	$0.068_{0.011}$	$\mathbf{0.062}_{0.009}$	$0.061_{0.008}$	$0.057_{0.005}$
IKE	$\mathbf{0.011}_{0.003}$	$0.069_{0.009}$	$\mathbf{0.012}_{0.005}$	$\mathbf{0.048}_{0.010}$
	Llama3-8B		Llama3-8B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.138_{0.012}$	$0.126_{0.012}$	$0.042_{0.010}$	$\mathbf{0.051}_{0.010}$
ROME	$0.045_{0.004}$	$0.144_{0.011}$	$0.063_{0.004}$	$0.089_{0.011}$
MEND	$0.087_{0.011}$	$\mathbf{0.053}_{0.008}$	$0.077_{0.012}$	$0.194_{0.013}$
IKE	$\mathbf{0.011}_{0.001}$	$0.125_{0.011}$	$\mathbf{0.018}_{0.013}$	$0.158_{0.013}$
	Llama3.2-3B		Llama3.2-3B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.085_{0.011}$	$0.037_{0.008}$	$\mathbf{0.050}_{0.009}$	$\mathbf{0.031}_{0.008}$
ROME	$0.052_{0.004}$	$0.157_{0.012}$	$0.064_{0.004}$	$0.128_{0.012}$
MEND	$0.081_{0.010}$	$\mathbf{0.035}_{0.009}$	$0.084_{0.011}$	$0.074_{0.011}$
IKE	$\mathbf{0.016}_{0.001}$	$0.158_{0.012}$	$0.065_{0.003}$	$0.158_{0.012}$
	Mistral-7B-v0.1		Mistral-7B-Instruct-v0.1	
	p	p^*	p	p^*
pre-edit	$0.066_{0.010}$	$0.051_{0.009}$	$0.086_{0.011}$	$0.061_{0.011}$
ROME	$0.059_{0.009}$	$\mathbf{0.049}_{0.009}$	$0.058_{0.010}$	$\mathbf{0.044}_{0.011}$
MEND	$0.099_{0.011}$	$0.062_{0.012}$	$0.062_{0.012}$	$0.103_{0.012}$
IKE	$\mathbf{0.005}_{0.000}$	$0.285_{0.011}$	$\mathbf{0.015}_{0.001}$	$0.179_{0.011}$

Table 5: ACE on WikiData$_{counterfact}$.

	Llama-2-7B		Llama-2-7B-chat	
	p	p^*	p	p^*
pre-edit	$-0.141_{0.012}$	$\mathbf{-0.032}_{0.011}$	$0.116_{0.0012}$	$0.104_{0.013}$
ROME	$-0.018_{0.007}$	$-0.104_{0.013}$	$-0.025_{0.007}$	$0.030_{0.015}$
MEND	$-0.009_{0.011}$	$0.072_{0.008}$	$\mathbf{0.003}_{0.011}$	$0.209_{0.012}$
IKE	$\mathbf{-0.003}_{0.000}$	$-0.073_{0.011}$	$-0.018_{0.003}$	$\mathbf{0.010}_{0.013}$
	Qwen2.5-7B		Qwen2.5-7B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.066_{0.003}$	$0.076_{0.004}$	$0.124_{0.005}$	$0.132_{0.006}$
ROME	$-0.055_{0.0003}$	$-0.143_{0.011}$	$-0.064_{0.004}$	$-0.081_{0.012}$
MEND	$-0.061_{0.011}$	$\mathbf{0.066}_{0.011}$	$0.052_{0.008}$	$0.057_{0.005}$
IKE	$\mathbf{-0.004}_{0.003}$	$-0.074_{0.011}$	$\mathbf{0.007}_{0.005}$	$\mathbf{0.032}_{0.012}$
	Llama3-8B		Llama3-8B-Instruct	
	p	p^*	p	p^*
pre-edit	$-0.137_{0.012}$	$-0.123_{0.012}$	$\mathbf{-0.005}_{0.012}$	$\mathbf{0.038}_{0.012}$
ROME	$-0.045_{0.004}$	$-0.143_{0.011}$	$-0.063_{0.004}$	$-0.086_{0.012}$
MEND	$-0.083_{0.012}$	$\mathbf{0.034}_{0.011}$	$0.074_{0.013}$	$0.194_{0.013}$
IKE	$\mathbf{-0.011}_{0.001}$	$-0.108_{0.013}$	$-0.018_{0.003}$	$-0.158_{0.013}$
	Llama3.2-3B		Llama3.2-3B-Instruct	
	p	p^*	p	p^*
pre-edit	$-0.083_{0.011}$	$-0.031_{0.009}$	$\mathbf{-0.039}_{0.011}$	$\mathbf{-0.016}_{0.011}$
ROME	$-0.052_{0.004}$	$-0.157_{0.012}$	$-0.064_{0.004}$	$-0.126_{0.013}$
MEND	$-0.077_{0.010}$	$\mathbf{-0.013}_{0.012}$	$-0.077_{0.011}$	$0.065_{0.012}$
IKE	$\mathbf{-0.016}_{0.001}$	$-0.158_{0.012}$	$-0.065_{0.003}$	$-0.158_{0.012}$
	Mistral-7B-v0.1		Mistral-7B-Instruct-v0.1	
	p	p^*	p	p^*
pre-edit	$-0.056_{0.011}$	$-0.015_{0.011}$	$0.079_{0.011}$	$0.053_{0.011}$
ROME	$0.048_{0.010}$	$-0.021_{0.012}$	$0.041_{0.012}$	$\mathbf{-0.005}_{0.014}$
MEND	$-0.077_{0.011}$	$\mathbf{-0.011}_{0.012}$	$\mathbf{-0.011}_{0.012}$	$0.088_{0.013}$
IKE	$\mathbf{-0.005}_{0.000}$	$-0.191_{0.011}$	$-0.015_{0.001}$	$-0.128_{0.011}$

Table 6: MCS on WikiData$_{counterfact}$.

	Llama-2-7B		Llama-2-7B-chat	
	p	p^*	p	p^*
pre-edit	$0.042_{0.006}$	$0.037_{0.005}$	$0.049_{0.006}$	$0.043_{0.006}$
ROME	$0.854_{0.010}$	$0.775_{0.012}$	$0.848_{0.010}$	$0.736_{0.012}$
MEND	$0.846_{0.010}$	$0.808_{0.011}$	$0.8132_{0.010}$	$0.763_{0.011}$
IKE	$\mathbf{1.000}_{0.000}$	$\mathbf{0.986}_{0.003}$	$\mathbf{1.000}_{0.000}$	$\mathbf{0.917}_{0.008}$
	Qwen2.5-7B		Qwen2.5-7B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.000_{0.000}$	$0.000_{0.000}$	$0.000_{0.000}$	$0.000_{0.000}$
ROME	$0.983_{0.003}$	$0.917_{0.008}$	$\mathbf{0.969}_{0.005}$	$0.798_{0.011}$
MEND	$0.506_{0.013}$	$0.402_{0.013}$	$0.783_{0.012}$	$0.648_{0.013}$
IKE	$\mathbf{0.999}_{0.002}$	$\mathbf{0.993}_{0.002}$	$0.936_{0.007}$	$\mathbf{0.955}_{0.006}$
	Llama3-8B		Llama3-8B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.060_{0.007}$	$0.048_{0.006}$	$0.132_{0.009}$	$0.105_{0.009}$
ROME	$0.962_{0.005}$	$0.879_{0.009}$	$0.942_{0.010}$	$0.859_{0.010}$
MEND	$0.802_{0.011}$	$0.625_{0.013}$	$0.853_{0.011}$	$0.778_{0.011}$
IKE	$\mathbf{1.000}_{0.000}$	$\mathbf{0.999}_{0.001}$	$\mathbf{0.999}_{0.006}$	$\mathbf{0.957}_{0.006}$
	Llama3.2-3B		Llama3.2-3B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.040_{0.005}$	$0.038_{0.006}$	$0.062_{0.007}$	$0.062_{0.007}$
ROME	$0.966_{0.005}$	$0.887_{0.009}$	$\mathbf{0.964}_{0.009}$	$\mathbf{0.878}_{0.009}$
MEND	$0.916_{0.008}$	$0.836_{0.010}$	$0.921_{0.010}$	$0.828_{0.010}$
IKE	$\mathbf{1.000}_{0.000}$	$\mathbf{0.941}_{0.004}$	$0.961_{0.014}$	$0.417_{0.014}$
	Mistral-7B-v0.1		Mistral-7B-Instruct-v0.1	
	p	p^*	p	p^*
pre-edit	$0.064_{0.007}$	$0.067_{0.007}$	$0.095_{0.008}$	$0.090_{0.008}$
ROME	$0.574_{0.014}$	$0.510_{0.014}$	$0.682_{0.013}$	$0.604_{0.014}$
MEND	$0.826_{0.011}$	$0.792_{0.012}$	$0.844_{0.010}$	$0.681_{0.013}$
IKE	$\mathbf{1.000}_{0.000}$	$\mathbf{0.999}_{0.001}$	$\mathbf{0.998}_{0.001}$	$\mathbf{0.952}_{0.005}$

Table 7: Accuracy on ZsRE.

	Llama-2-7B		Llama-2-7B-chat	
	p	p^*	p	p^*
pre-edit	$0.026_{0.004}$	$0.023_{0.004}$	$0.190_{0.007}$	$0.184_{0.007}$
ROME	$0.078_{0.009}$	$\mathbf{0.071}_{0.009}$	$0.097_{0.009}$	$0.070_{0.009}$
MEND	$0.043_{0.007}$	$0.086_{0.009}$	$0.075_{0.008}$	$\mathbf{0.046}_{0.008}$
IKE	$\mathbf{0.004}_{0.000}$	$0.115_{0.004}$	$\mathbf{0.020}_{0.001}$	$0.110_{0.006}$
	Qwen2.5-7B		Qwen2.5-7B-Instruct	
	p	p^*	p	p^*
pre-edit	$\mathbf{0.095}_{0.004}$	$\mathbf{0.082}_{0.003}$	$0.191_{0.005}$	$0.164_{0.005}$
ROME	$0.057_{0.003}$	$0.116_{0.006}$	$0.061_{0.004}$	$0.077_{0.007}$
MEND	$0.117_{0.010}$	$0.090_{0.010}$	$0.105_{0.009}$	$0.074_{0.009}$
IKE	$0.029_{0.001}$	$0.087_{0.003}$	$\mathbf{0.025}_{0.005}$	$\mathbf{0.023}_{0.004}$
	Llama3-8B		Llama3-8B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.031_{0.006}$	$\mathbf{0.038}_{0.006}$	$0.043_{0.007}$	$\mathbf{0.041}_{0.007}$
ROME	$0.052_{0.003}$	$0.091_{0.005}$	$0.038_{0.004}$	$0.075_{0.006}$
MEND	$0.149_{0.010}$	$0.143_{0.011}$	$0.037_{0.006}$	$0.046_{0.007}$
IKE	$\mathbf{0.011}_{0.000}$	$0.070_{0.002}$	$\mathbf{0.006}_{0.001}$	$0.097_{0.005}$
	Llama3.2-3B		Llama3.2-3B-Instruct	
	p	p^*	p	p^*
pre-edit	$\mathbf{0.016}_{0.004}$	$\mathbf{0.022}_{0.005}$	$\mathbf{0.020}_{0.005}$	$\mathbf{0.012}_{0.004}$
ROME	$0.062_{0.004}$	$0.122_{0.006}$	$0.062_{0.003}$	$0.109_{0.006}$
MEND	$0.039_{0.005}$	$0.083_{0.008}$	$0.047_{0.006}$	$0.065_{0.008}$
IKE	$0.018_{0.001}$	$0.137_{0.005}$	$0.072_{0.004}$	$0.211_{0.013}$
	Mistral-7B-v0.1		Mistral-7B-Instruct-v0.1	
	p	p^*	p	p^*
pre-edit	$0.033_{0.006}$	$\mathbf{0.032}_{0.006}$	$0.064_{0.006}$	$0.058_{0.006}$
ROME	$0.078_{0.009}$	$0.074_{0.009}$	$0.092_{0.010}$	$0.071_{0.009}$
MEND	$0.046_{0.007}$	$0.082_{0.008}$	$0.050_{0.008}$	$\mathbf{0.046}_{0.008}$
IKE	$\mathbf{0.006}_{0.000}$	$0.058_{0.002}$	$\mathbf{0.024}_{0.002}$	$0.133_{0.004}$

Table 8: ECE on ZsRE.

	Llama-2-7B		Llama-2-7B-chat	
	p	p^*	p	p^*
pre-edit	$0.025_{0.005}$	$0.021_{0.004}$	$0.190_{0.007}$	$0.184_{0.007}$
ROME	$0.072_{0.009}$	$\mathbf{0.072}_{0.009}$	$0.086_{0.009}$	$0.061_{0.009}$
MEND	$0.042_{0.008}$	$0.087_{0.009}$	$0.066_{0.008}$	$\mathbf{0.039}_{0.008}$
IKE	$\mathbf{0.004}_{0.000}$	$0.115_{0.004}$	$\mathbf{0.020}_{0.001}$	$0.100_{0.006}$
	Qwen2.5-7B		Qwen2.5-7B-Instruct	
	p	p^*	p	p^*
pre-edit	$\mathbf{0.095}_{0.004}$	$\mathbf{0.082}_{0.003}$	$0.191_{0.005}$	$0.164_{0.005}$
ROME	$0.055_{0.003}$	$0.115_{0.006}$	$0.058_{0.004}$	$0.075_{0.008}$
MEND	$0.112_{0.010}$	$0.084_{0.011}$	$0.104_{0.010}$	$0.069_{0.009}$
IKE	$0.029_{0.001}$	$0.086_{0.003}$	$\mathbf{0.025}_{0.005}$	$\mathbf{0.018}_{0.004}$
	Llama3-8B		Llama3-8B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.036_{0.006}$	$\mathbf{0.041}_{0.005}$	$0.043_{0.008}$	$0.043_{0.007}$
ROME	$0.050_{0.004}$	$0.091_{0.005}$	$0.033_{0.004}$	$0.074_{0.007}$
MEND	$0.158_{0.010}$	$0.145_{0.011}$	$0.028_{0.006}$	$\mathbf{0.040}_{0.008}$
IKE	$\mathbf{0.011}_{0.000}$	$0.070_{0.002}$	$\mathbf{0.005}_{0.001}$	$0.096_{0.005}$
	Llama3.2-3B		Llama3.2-3B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.021_{0.005}$	$\mathbf{0.020}_{0.005}$	$\mathbf{0.020}_{0.005}$	$\mathbf{0.014}_{0.004}$
ROME	$0.062_{0.004}$	$0.122_{0.006}$	$0.060_{0.004}$	$0.108_{0.006}$
MEND	$0.037_{0.005}$	$0.085_{0.008}$	$0.055_{0.006}$	$0.069_{0.008}$
IKE	$\mathbf{0.018}_{0.001}$	$0.135_{0.004}$	$0.072_{0.004}$	$0.209_{0.011}$
	Mistral-7B-v0.1		Mistral-7B-Instruct-v0.1	
	p	p^*	p	p^*
pre-edit	$0.035_{0.006}$	$\mathbf{0.035}_{0.006}$	$0.060_{0.006}$	$0.054_{0.006}$
ROME	$0.079_{0.009}$	$0.071_{0.009}$	$0.091_{0.010}$	$0.069_{0.009}$
MEND	$0.045_{0.007}$	$0.083_{0.008}$	$0.052_{0.008}$	$\mathbf{0.045}_{0.008}$
IKE	$\mathbf{0.006}_{0.000}$	$0.057_{0.001}$	$\mathbf{0.023}_{0.002}$	$0.133_{0.005}$

Table 9: ACE on ZsRE.

	Llama-2-7B		Llama-2-7B-chat	
	p	p^*	p	p^*
pre-edit	$0.022_{0.005}$	$0.017_{0.005}$	$0.190_{0.007}$	$0.184_{0.007}$
ROME	$0.062_{0.009}$	$\mathbf{-0.009}_{0.010}$	$0.082_{0.009}$	$0.037_{0.010}$
MEND	$0.007_{0.008}$	$-0.080_{0.009}$	$0.063_{0.008}$	$\mathbf{0.026}_{0.009}$
IKE	$\mathbf{-0.004}_{0.000}$	$-0.115_{0.004}$	$\mathbf{-0.020}_{0.001}$	$-0.099_{0.007}$
	Qwen2.5-7B		Qwen2.5-7B-Instruct	
	p	p^*	p	p^*
pre-edit	$0.095_{0.004}$	$0.082_{0.003}$	$0.191_{0.005}$	$0.164_{0.005}$
ROME	$0.055_{0.003}$	$-0.115_{0.006}$	$-0.058_{0.004}$	$-0.075_{0.008}$
MEND	$-0.090_{0.010}$	$\mathbf{-0.061}_{0.010}$	$-0.071_{0.010}$	$-0.047_{0.010}$
IKE	$\mathbf{-0.029}_{0.001}$	$-0.086_{0.003}$	$\mathbf{0.023}_{0.005}$	$\mathbf{-0.014}_{0.005}$
	Llama3-8B		Llama3-8B-Instruct	
	p	p^*	p	p^*
pre-edit	$\mathbf{0.006}_{0.007}$	$\mathbf{0.012}_{0.006}$	$0.039_{0.008}$	$0.039_{0.007}$
ROME	$-0.050_{0.004}$	$-0.090_{0.005}$	$-0.074_{0.004}$	$-0.074_{0.007}$
MEND	$-0.147_{0.010}$	$-0.137_{0.011}$	$0.025_{0.007}$	$\mathbf{0.020}_{0.008}$
IKE	$-0.011_{0.000}$	$-0.070_{0.002}$	$\mathbf{-0.005}_{0.001}$	$-0.096_{0.005}$
	Llama3.2-3B		Llama3.2-3B-Instruct	
	p	p^*	p	p^*
pre-edit	$\mathbf{0.011}_{0.005}$	$\mathbf{0.008}_{0.005}$	$\mathbf{0.013}_{0.006}$	$\mathbf{0.004}_{0.004}$
ROME	$-0.062_{0.004}$	$-0.122_{0.006}$	$-0.060_{0.004}$	$-0.108_{0.006}$
MEND	$-0.035_{0.005}$	$-0.083_{0.008}$	$-0.040_{0.006}$	$-0.064_{0.008}$
IKE	$-0.018_{0.001}$	$-0.135_{0.004}$	$-0.072_{0.004}$	$0.204_{0.015}$
	Mistral-7B-v0.1		Mistral-7B-Instruct-v0.1	
	p	p^*	p	p^*
pre-edit	$-0.021_{0.006}$	$-0.028_{0.006}$	$0.059_{0.007}$	$0.054_{0.006}$
ROME	$0.0736_{0.010}$	$\mathbf{-0.000}_{0.010}$	$0.064_{0.011}$	$\mathbf{0.027}_{0.011}$
MEND	$-0.033_{0.008}$	$-0.079_{0.009}$	$-0.041_{0.008}$	$-0.034_{0.010}$
IKE	$\mathbf{-0.006}_{0.000}$	$-0.057_{0.001}$	$\mathbf{-0.023}_{0.001}$	$-0.133_{0.005}$

Table 10: MCS on ZsRE.