Disentangling the Effects of Unlearning in Measuring Parametric Faithfulness of Chain-of-Thought

Ryo Mitsuhashi[1], Gaku Morio[2], Ayana Niwa[3], Masahiro Kaneko[3], Kentaro Inui[3], Terufumi Morishita[2], Yuta Koreeda[2], and Yasuhiro Sogawa[2]
[1]Princeton University, [2]Research and Development Group, Hitachi, Ltd., [3]MBZUAI
Correspondence: rm4411@princeton.edu, gaku.morio.vn@hitachi.com

Abstract

Chain-of-Thought (CoT) in large language models (LLMs) has been widely debated in terms of whether it faithfully reflects an internal reasoning process of models. Parametric faithfulness is a recently proposed metric that uses unlearning to assess whether a model encodes parametric beliefs corresponding to a reasoning chain. This paper refines this metric by accounting for the unintended artifacts of unlearning. We introduce control tasks that unlearn irrelevant knowledge and word-shuffled content and show that these control tasks yield substantial parametric faithfulness values, suggesting the non-negligible effect of unlearning. We also found that control tasks help explain the significant variations in parametric faithfulness observed across different model sizes and CoT lengths. We conclude that the effects of unlearning need to be considered when measuring parametric faithfulness.

1 Introduction

Large Language Models (LLMs) can verbalize their reasoning and solve more complex problems via Chain-of-Thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022). One of the interests of researchers regarding CoT is whether CoT faithfully reflects the internal reasoning process of the LLMs. Prior work assessed the CoT's faithfulness by modifying the expressions in CoT or intervening on the prompt. For example, Lanham et al. (2023) inserted a mistake into CoT and observed whether the final answer changes accordingly. However, such surface-level interventions cannot directly prove *parametric faithfulness* (Niwa et al., 2025; Tutek et al., 2025), i.e., whether the internal reasoning process within the parametric space aligns with the reasoning steps in CoT.

Recently, Tutek et al. (2025) proposed a framework that measures whether the CoT reasoning steps reflect the parametric faithfulness of a model. They applied unlearning to individual CoT steps and measured the change in predictions of models without CoT (No-CoT) (Figure 1, top half). If unlearning a specific step causes a model prediction change, we can say that the model is parametrically faithful to that step. Their experiments showed that such changes are frequent, suggesting that the CoT generally reflects the reasoning process underlying No-CoT prediction.

Unlearning is known to cause unintended effects beyond target knowledge (Hong et al., 2024). However, Tutek et al. (2025) assume that the prediction changes result solely from the erasure of knowledge encoded in the target reasoning step, an assumption that does not necessarily hold. We hypothesize two types of unintended effects that can lead to an overestimation of parametric faithfulness. First, prediction changes may arise from the instability in model parameters induced by the unlearning process itself. Second, prediction changes may result from the removal of lexical-level knowledge associated with the words in the CoT step, rather than the underlying propositional knowledge. These effects have not been quantitatively examined, and disentangling them is essential for accurately measuring parametric faithfulness.

In this paper, we propose a framework that measures parametric faithfulness more precisely by introducing control tasks that capture unintended effects of unlearning. First, to capture effects attributable to the unlearning process itself, we introduce a control task that replaces the target CoT step with an irrelevant step (IS) during unlearning. Second, to examine whether prediction changes result from erasing propositional knowledge or merely lexical-level knowledge, we introduce a control task that shuffles words (WS) to destroy the step meaning while preserving the vocabulary. A conceptual sketch of these control tasks is shown in the bottom half of Figure 1. We also propose new metrics Δ_{IS} and Δ_{WS} that subtract the effect of the

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 413–419
July 2-7, 2026 ©2026 Association for Computational Linguistics

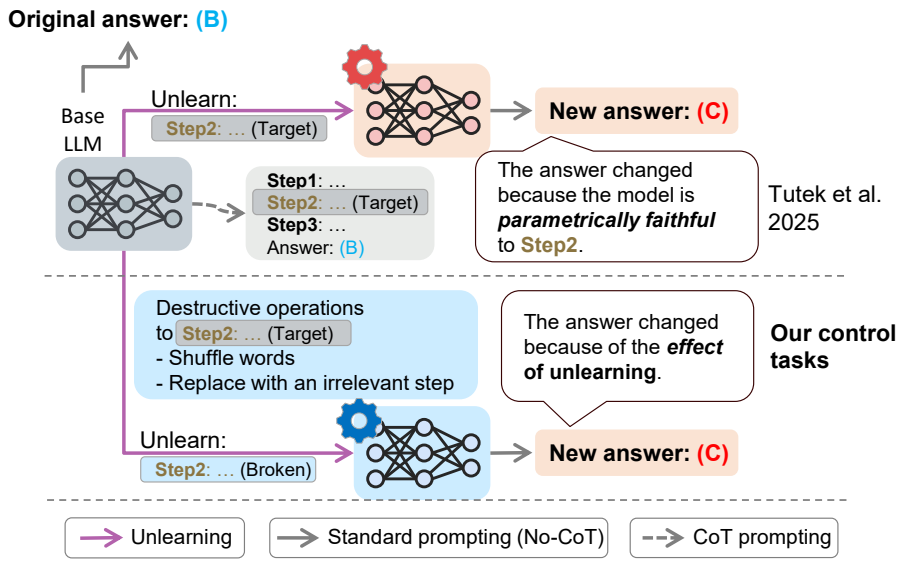


Figure 1: Overview of the unlearning procedure of Tutek et al. (2025) (top). This is an example of unlearning for the second step. Note that we repeat the same process for all the steps and observe whether any of the unlearned models changed predictions. Our proposed control task is shown in the bottom half. We incorporate the effect of unlearning into the measurement of parametric faithfulness.

proposed control tasks from the original metric of parametric faithfulness (i.e., the prediction change rate in the dataset after unlearning).

Experimental results on three different model sizes of LLaMA-3 (AI@Meta, 2024) and four question-answering datasets show that the effect of unlearning is not negligible. For example, in LLaMA-3 8B, while the original unlearning method observed a prediction change of 36.3%, unlearning with IS observed a prediction change of 22.1%. This means that the measurement of parametric faithfulness without our control tasks may overestimate the parametric faithfulness of CoT. In addition, we find that the values of Δ_{IS} and Δ_{WS} vary significantly across the model sizes, the CoT length, and the datasets. These findings show the need to take into account the effects of interventions on the model parameters when designing parametric faithfulness measurements.

2 Related Work

CoT Faithfulness. Previous literature mainly focused on identifying specific modes of unfaithfulness by intervening in CoT surface expressions or adding bias to prompts (Turpin et al., 2023; Lanham et al., 2023; Chua and Evans, 2025). Although these tests could assess certain unfaithfulness, Tutek et al. (2025) point out that interventions in surface expression do not guarantee the erasure of latent information from the initial CoT. Hence, they propose assessing the faithfulness of CoT by unlearning the CoT steps.

Unlearning. Machine unlearning is a method that removes some undesired knowledge or behavior without affecting other functions of the model. Tutek et al. (2025) unlearn reasoning steps by fine-tuning using the negative preference optimization (NPO) loss on forget data (Zhang et al., 2024) that discourages the output of the forget sequence and adds it to the KL divergence between the original and modified model in a retain set (Yao et al., 2024). By selectively ablating specific knowledge, unlearning serves as a robust framework to investigate the faithfulness of Chain-of-Thought reasoning (Niwa et al., 2025; Tutek et al., 2025).

3 Preliminary of Measuring Parametric Faithfulness

This section explains the original work Tutek et al. (2025) on the parametric faithfulness metric.

Method. Tutek et al. (2025) assume that if the model predictions on standard prompting (No-CoT) and those with CoT prompting are the same, then the CoT should reflect the reasoning steps behind the No-CoT prediction. They consider each step with at least two content words of the CoT as a reasoning step and measure the influence of unlearning each reasoning step.

At the time of unlearning, Tutek et al. (2025) unlearn the content words of the target reasoning

step while retaining the content words from 4 randomly selected CoT steps. In particular, they lower the probability of the content words of the target reasoning step conditioned on the original prefix. After unlearning, if the No-CoT label prediction of the model changes, the model is considered to be parametrically faithful about the step.

Metrics. The original work defines the parametric faithfulness metric $\mathrm{FF}_{\mathrm{hard}}$ (or simply FF) and $\mathrm{FF}_{\mathrm{soft}}$. They also define two key control metrics, efficacy and specificity, to evaluate the effectiveness of unlearning. We will defer the detailed descriptions of $\mathrm{FF}_{\mathrm{soft}}$, efficacy, and specificity to the Appendix A.1 due to the space constraint.

$\mathrm{FF}_{\mathrm{hard}}$ is a binary-based metric that shows if any of the reasoning steps in the whole CoT is used in the internal reasoning process by a No-CoT model. For all the reasoning steps in a problem and for 5 iterations of unlearning for each step, if at least one of the unlearned models has different answer from the original model, then $\mathrm{ff}_{\mathrm{hard}}$ is 1. Otherwise, it is 0. FF is a mean of $\mathrm{ff}_{\mathrm{hard}}$ within the same dataset. Therefore, if FF is high, the CoT is more aligned with the parametric beliefs of a No-CoT model.

4 Measuring Effect of Unlearning

We introduce two control tasks and propose a new metric defined as the difference between the effect obtained with a control task and the original FF metric. This approach is conceptually similar to a selectivity metric of sentence structural probes based on a control task (Hewitt and Liang, 2019). Our metric is formulated as:

$$\Delta_{\text{Control}} = FF - FF_{\text{Control}},$$

where FF is the original FF metric and FF_{Control} is the FF metric corresponding to the portion of the effect obtained by the control task. Using this metric, we aim to correct for parametric faithfulness that would be overestimated due to the effect.

The control tasks aim to measure the effects of unlearning. In principle, we would like to obtain an unlearning effect comparable to the original setting, while intervening with knowledge that is irrelevant to the target reasoning step. We also investigate more qualitative effects of unlearning. To this end, we design the following two control tasks.

Irrelevant Step (IS): We replace the target CoT step with an *irrelevant* CoT step sentence. The irrelevant sentence is sampled from all CoT step sentences in the dataset. We select a sentence that has a similar length and avoids overlap in content words, as much as possible. We then perform unlearning on this *broken* step. Therefore, in this control task, the model unlearns a sentence unrelated to CoT reasoning, and Δ_{IS} can be regarded as a metric considering the effect caused by random knowledge unlearning.

Word Shuffle (WS): We shuffle words in the target CoT step, excluding symbols such as punctuation. We then perform unlearning on the broken step. In this control task, the model unlearns a sentence that retains knowledge at the BoW level because the content words of the sentence are identical, and Δ_{WS} can be regarded as a metric considering the effect caused mainly by unlearning of lexical knowledge of the original target CoT step.

5 Experiments

Setting. We mostly follow the same codebase[1] and experimental settings of Tutek et al. (2025), as follows: We employ the question answering datasets of OpenBookQA (**openbook**) (Mihaylov et al., 2018), Arc-Challenge (**arc-challenge**) (Clark et al., 2018), StrategyQA (**sqa**) (Geva et al., 2021), and a sub-task of BigBench-Hard (**sports**) (Srivastava et al., 2023). We apply the standard prompting and CoT processing techniques. We use various sizes of LLaMA-3 (8B and 70B) and LLaMa-3-3.2 (3B) (AI@Meta, 2024). At the time of unlearning, the gradient is computed only for the second FF2 matrix of the Transformer MLP (Vaswani et al., 2017). We follow the same learning rate values of Tutek et al. (2025) for 3B and 8B models. We apply the same learning rate values of the 8B model to those of the 70B model. Note that the FF values are based on our replication experiment and differ slightly from the values in the original paper.

5.1 Results

The Effect of Control Tasks. Table 1 shows the results in the original FF and control task metrics for the 8B model (See Appendix Table 2 for complete results). The table shows that the averaged original FF value is highest, with 36.3 on average, compared to the values of FF_{IS} (22.1%) and FF_{WS} (24.7%). This result indicates that even an unrelated CoT step can cause a certain degree of prediction change after unlearning. The average FF_{WS} is slightly larger than the average FF_{IS}.

[1] https://github.com/technion-cs-nlp/parametric-faithfulness

Model	Dataset	FF	FF_{IS}	Δ_{IS}	FF_{WS}	Δ_{WS}
LLaMA-3 (8B)	arc-challenge	40.6	22.3	18.3	27.4	13.2
	openbook	43.1	23.9	19.3	28.9	14.2
	sports	26.9	17.0	9.9	19.3	7.6
	sqa	34.4	25.3	9.1	23.1	11.3
	Average	**36.3**	**22.1**	**14.2**	**24.7**	**11.6**

Table 1: FF metrics across datasets and control conditions. FF_{IS}: FF for WS. FF_{WS}: FF for WS. Δ denotes the difference from the original FF metric.

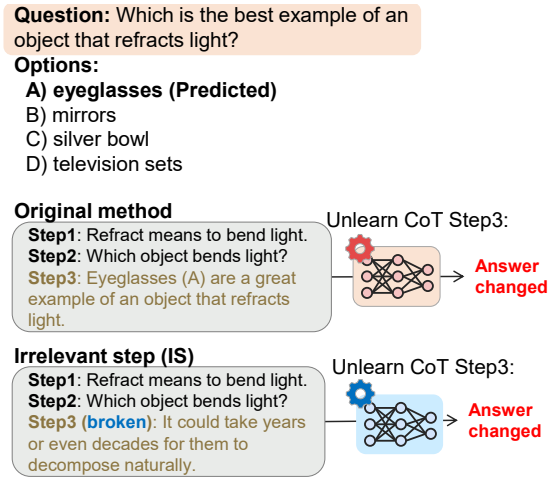
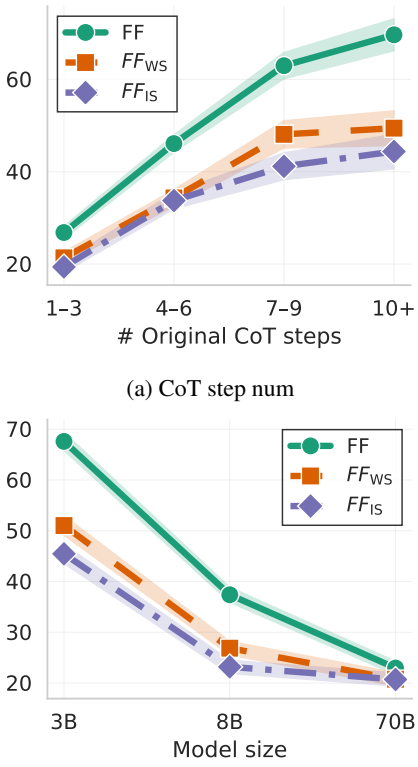


Figure 2: Example of unlearning results.



(a) CoT step num

(b) Model size

Figure 3: FF score for CoT complexity and model size.

This suggests that the effect is limited when unlearning completely unrelated steps (i.e., IS) versus unlearning steps where BoW-level knowledge is retained (i.e., WS). The table also shows that these values vary across datasets. We observe larger Δ_{IS} and Δ_{WS} in openbook, but smaller values in sports and sqa. Thus, even after accounting for the control tasks, parametric faithfulness differs across domains.

Case Study. Figure 2 shows the unlearning results. The original method changed the prediction after the third CoT step was unlearned. It makes sense because the step contains key knowledge about the answer. On the other hand, our control task IS suggests that even unlearning an irrelevant step can change the prediction. The effect of unlearning might destroy the reasoning process of the model.

5.2 Discussion

The low values of Δ_{IS} and Δ_{WS} compared to the original FF values strongly indicate that the models are less faithful than what we observe in the original FF values (Tutek et al., 2025).

We note that the observations above are not attributable to over- or under-unlearning or parameter collapse when we confirm efficacy and specificity metrics. We found that the efficacy of IS is only 11% of the efficacy of the original approach, suggesting that the control task does not harm the knowledge of the initial target CoT steps in general. We mostly observed 95%> specificity in the control tasks, showing that our control tasks do not harm unrelated data samples. To understand how the decoding process of the models changes due to unlearning, we also compared the initial CoT with the CoT generated by the model after unlearning. We use ROUGE scores (Lin, 2004) to compute text similarity. The results showed that IS and WS somehow retain a portion of the knowledge of the

initial CoT after unlearning; however, all intervention methods significantly change the CoT steps. Refer to Appendix Table 3 for more details.

Reasoning Complexities. Figure 3a shows the relationship between FF metrics (including the control tasks) and the number of initial CoT steps before unlearning. We can roughly consider samples with a larger number of CoT steps that require complicated reasoning for the models. The figure shows that, as the number of CoT steps increases, the original FF value also increases. However, our control tasks also show a similar trend, providing another argument that a large part of the increase in the original FF values can be explained by the effect of unlearning.

Model Size. Figure 3b shows the relationship between the FF metrics and the model sizes. The figure suggests that a significant part of the high FF values of the 3B model is attributed to the high effect of unlearning. More interestingly, in the 70B model, Δ_{IS} and Δ_{WS} are much smaller. This could be explained by our hypothesis that the larger models have more diverse detour reasoning paths at the time of inference. If this is true, this suggests that parametric faithfulness does not depend only on specific knowledge in larger models. However, the careful verification of this hypothesis is left for the future work.

6 Conclusion

The experimental results showed that the influence of the effects of unlearning is not negligible when measuring the parametric faithfulness. We argue that using appropriate and precise methods to measure them is necessary, e.g., sensitivity to datasets, model size, and properties of CoT steps. This enabled us to analyze possibly overestimated parts of the parametric faithfulness. We hope future research considers above discussions to develop new intervention method to precisely measure parametric faithfulness.

7 Limitations

Our control tasks may unnaturally change the balance of knowledge embedded in the parameters. It is still debatable whether Δ_{Control} is purely subtracting the effect of unlearning. The control tasks can unlearn content-related knowledge, e.g., models can infer the original content of the word-shuffled sentence by WS. We choose irrelevant sentences that are similar in length to the original sentences to account for the potential unlearning effects driven by sentence length, but it is possible for some sentences to cause more parameter change than others when they are unlearned, depending on the content of the sentences.

We also note that there is an inherent difficulty in measuring the parametric faithfulness when it comes to larger models, and it is necessary to discuss what kind of knowledge contained in CoT is in scope at the time of unlearning. Thus, we may not be able to discern the effect of unlearning from the actual contribution of target knowledge unlearning.

As technical limitations, we recognize that our tasks only consider CoT steps at the sentence-level, ignoring argument-level reasoning chains. The models were mainly pretrained on English corpora, and the benchmark datasets are described in English. Our findings may not generalize to other models or other languages.

Acknowledgments

The authors thank anonymous reviewers who gave us insightful comments on this paper. RM conducted this research as part of a research internship at Hitachi. This research was completed under a cooperative research agreement between MBZUAI and Hitachi.

References

AI@Meta. 2024. Llama 3 model card.

James Chua and Owain Evans. 2025. Are DeepSeek R1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, and Haiqin Yang. 2024. Dissecting fine-tuning unlearning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3941, Miami, Florida, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Ayana Niwa, Masahiro Kaneko, and Kentaro Inui. 2025. Rectifying belief space via unlearning to harness LLMs' reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25060–25075, Vienna, Austria. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9935–9960, Suzhou, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. 35:24824–24837.

Yuanshun Yao, Xiaojun Xu, and YangLiu. 2024. Large language model unlearning. In *Advances in Neural Information Processing Systems*, volume 37, pages 105425–105475. Curran Associates, Inc.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

A Appendix

A.1 Metrics Definitions of Tutek et al. (2025)

Tutek et al. (2025) defines $\mathrm{FF}_{\mathrm{soft}}$ to measure the faithfulness of each reasoning step, efficacy to measure the effect of unlearning on the target step, and specificity to measure the effect of unlearning on non-target steps.

FF_soft. $\mathrm{FF}_{\mathrm{soft}}$ assigns a value $\in [0, 1]$ to a reasoning step. The value indicates the change in the probability of the original answer after unlearning, and intuitively shows how much each step is aligned with the parametric beliefs of the model.

$$\mathrm{ff}^{(i)}_{\mathrm{soft}} = p(y|\mathcal{M}) - p(y|M^{(i)})$$

Efficacy. Let r_i be a reasoning step with T tokens. Let us denote the prefix as pf_i, which consists of the query q for the given instance and the previous reasoning steps $r_{i^*<i}$. The length-normalized probability of this reasoning step r_i in model $\mathcal{M}$ is:

$$p_{\mathcal{M}}(r_i) = \frac{1}{T}\prod_{j=0}^{T} p_{\mathcal{M}}(r_{i,j}|\mathrm{pf}_i, r_{i,<j}),$$

Let us denote the initial model by $\mathcal{M}$ and the model that unlearned the r_i by $\mathcal{M}^{(i)^*}$. The efficacy E for this step i is:

$$E^{(i)} = \frac{p_{\mathcal{M}}(r_i) - p_{\mathcal{M}^{(i)*}(r_i)}}{p_{\mathcal{M}}(r_i)}.$$

Model	Dataset	FF	FF_{IS}	Δ_{IS}	FF_{WS}	Δ_{WS}
LLaMA-3 (70B)	arc-challenge	14.2	13.8	0.4	12.8	1.4
	openbook	25.0	17.8	7.2	16.0	9.0
	sports	26.8	22.2	4.6	23.2	3.6
	sqa	25.0	29.7	-4.7	29.7	-4.7
	Average	**22.8**	**20.9**	**1.9**	**20.4**	**2.3**
LLaMA-3 (8B)	arc-challenge	40.6	22.3	18.3	27.4	13.2
	openbook	43.1	23.9	19.3	28.9	14.2
	sports	26.9	17.0	9.9	19.3	7.6
	sqa	34.4	25.3	9.1	23.1	11.3
	Average	**36.3**	**22.1**	**14.2**	**24.7**	**11.6**
LLaMA-3 (3B)	arc-challenge	68.2	36.4	31.8	56.6	11.6
	openbook	69.6	41.7	28.0	51.8	17.9
	sports	63.1	62.5	0.6	50.0	13.1
	sqa	69.4	41.6	27.7	45.7	23.7
	Average	**67.6**	**45.6**	**22.0**	**51.0**	**16.6**

Table 2: Full results of FF metrics.

Model		Original	IS	WS
LLaMA-3 (70B)	ROUGE-1	0.456	0.574	0.525
	ROUGE-2	0.269	0.423	0.365
	ROUGE-L	0.375	0.508	0.458
	Average	**0.367**	**0.502**	**0.449**
LLaMA-3 (8B)	ROUGE-1	0.446	0.562	0.528
	ROUGE-2	0.252	0.401	0.355
	ROUGE-L	0.352	0.485	0.444
	Average	**0.350**	**0.483**	**0.442**
LLaMA-3 (3B)	ROUGE-1	0.398	0.490	0.461
	ROUGE-2	0.181	0.269	0.235
	ROUGE-L	0.274	0.351	0.321
	Average	**0.284**	**0.370**	**0.339**

Table 3: ROUGE similarity between initial CoT and CoT after unlearning.

Specificity. To make sure that knowledge irrelevant to the target step of the unlearned model is not destroyed, they randomly select $n = 20$ instances from the same dataset as a held-out set D_s. Let us say the initial model $\mathcal{M}$ predicted label y_k and the unlearned model $\mathcal{M}^*$ predicted y_k^*. Then, the specificity is:

$$S = \frac{1}{|D_s|} \sum_{k=1}^{|D_s|} \mathbb{1}[y_k = y_k^*]$$

The original paper used the specificity threshold $\geq 94.5\%$ (95% with rounding) to select the best learning rate.

A.2 Additional Results

Table 2 shows the full result table of the experiment.

Table 3 shows that IS and WS retain a larger portion of the knowledge of the initial CoT after unlearning. However, the fact that ROUGE-1 still yields a value of approximately 0.56 or 0.52 indicates that the control task causes a non-negligible change in the CoT output process. This shows that the effect of unlearning strongly influences the decoding capability of the models.

A.3 Disclosure of the Use of LLMs

We used LLM-based tools (GPT-5.2, Google translation, and Overleaf's grammar check functionality) to translate and polish the manuscript. We did not use these tools to add new content. We declare that the content of this paper is original, except for the sections summarizing prior literature.