Not Every Token Needs Forgetting: Selective Unlearning to Limit Change in Utility in Large Language Model Unlearning

Yixin Wan[1,2], Anil Ramakrishna[2], Kai-Wei Chang[1,2], Volkan Cevher[2], Rahul Gupta[2]
[1]University of California, Los Angeles, [2]Amazon AGI
elaine1wan@g.ucla.edu

Abstract

Large Language Model (LLM) unlearning has recently gained significant attention, driven by the need to remove unwanted information, such as private, sensitive, or copyrighted content, from LLMs. However, conventional unlearning approaches indiscriminately update model parameters to forget all tokens in a target document, including common tokens (e.g., pronouns, prepositions, general nouns) that carry general knowledge. In this paper, we highlight that "not every token needs forgetting". We propose **Selective Unlearning (SU)**, which identifies a critical subset of tokens within the forgetting set that is relevant to the unwanted information, and unlearns only those tokens. Experiments on two benchmarks and six baseline unlearning algorithms demonstrate that SU not only achieves effective unlearning on the targeted forget data, but also significantly preserves the model's utility in the retaining set.

1 Introduction

Text corpora used to train Large Language Models (LLMs) often contain sensitive, private, or copyrighted content. To address the risks posed by such data, recent research has explored *LLM unlearning*—aims to remove specific unwanted knowledge from a model without incurring the cost and effort of retraining from scratch.

Existing unlearning approaches typically apply the same unlearning loss to every token in the targeted documents. However, as illustrated in Figure 1, this approach forces the model to unlearn not only sensitive information but also general concepts. Even benign tokens like "that" or "she" in the target forget documents are unlearned, unnecessarily degrading the model's language capabilities.

Motivated by this, we contend that ***not every token needs forgetting***: an unlearning method should selectively target only tokens that encode unique information in the forget set. To this end, we introduce **Selective Unlearning (SU)**, a novel framework that utilizes two assistant models with different scopes of knowledge to identify and unlearn only a subset of tokens that carry forget-specific information. By only calculating unlearning losses on these tokens with forget set-specific information, SU can reduce unnecessary interference with retained information, thereby preserving model utility on general knowledge.

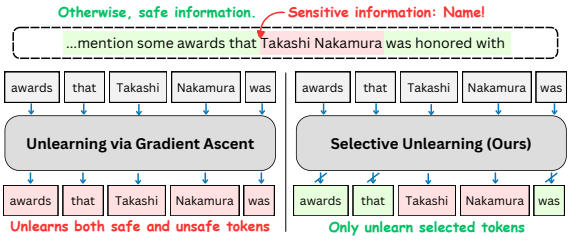

Figure 1: Example of how tokens are selected for unlearning. Red blocks indicate unlearned tokens, on which the forgetting loss is calculated. SU avoids the forgetting of general information like "that", therefore preserving model utility.

We conduct extensive experiments on 2 popular benchmarks: Task of Fictitious Unlearning (TOFU) (Maini et al., 2024) and MUSE-News (Shi et al., 2024) to compare SU with 6 unlearning methods. Results demonstrate that SU not only achieves comparable unlearning quality to the existing methods, but also substantially improves the preservation of retained knowledge. Striking a balance between unlearning and utility preservation, SU represents a promising step toward scalable and utility-preserving unlearning strategies for LLMs.

2 Related Work

2.1 Unlearning for LLMs

Previous works on unlearning have explored ways to remove sensitive, private, or copyrighted information (Carlini et al., 2021) from LLMs. The most intuitive method is **Gradient Ascent (GA)** (Jang et al., 2023; Yao et al., 2023), which maximizes the

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1827–1835

language model loss[1] on the forget dataset. However, GA has been shown to degrade the performance of models in data and knowledge outside of the forget set, even resulting in model collapsing (Zhang et al., 2024).

With this in mind, prior studies have proposed ways to better preserve model performance on retain data. For instance, researchers have proposed to apply gradient descent (Liu et al., 2022; Maini et al., 2024) or regularize models' KL-divergence (Wang et al., 2024a; Chen and Yang, 2023) on the retain set during unlearning. The former is also known as **"Gradient Difference (GD)"**, since it essentially optimizes the difference between losses on forget and retain data. Additionally, previous research also investigated alternatives to the GA approach, with **Negative Preference Optimization (NPO)** (Zhang et al., 2024) being one of the most promising algorithms. NPO uses forget candidates as negative examples in Direct Preference Optimization (DPO) (Rafailov et al., 2024), avoiding model collapse. To better assess different unlearning algorithms, more recent works construct LLM unlearning benchmarks such as TOFU (Maini et al., 2024), MUSE (Shi et al., 2024) and LUME (Ramakrishna et al., 2025a,b).

2.2 Selecting Unlearning Candidates

Although previous research on unlearning in LLMs has achieved remarkable progress, most of them formulate the task as such that models must be retrained to remove information about all candidates in the forget set. Most related to the work, Wang et al. (2024b) proposed to unlearn parts in a sequence that has lower log-probability than a threshold. However, their experiments were limited to variations of the GPT-Neo model (Gao et al., 2020), and were not extended to the newer LLMs. Ma et al. (2024) and Choi et al. (2024) explored entity-level unlearning, which selectively unlearns knowledge related to specific entities, instead of all knowledge in the forget set. McCartney et al. (2024) selectively chooses anti-knowledge, or knowledge that conflicts with a model's original memory, for unlearning. Similarly, Choi et al. (2024) proposed to utilize a LLM trained with negative instructions to produce obliterated generations for unlearning. However, these approaches still require forgetting full chunks of text, among which common words and tokens inevitably persist.

[1]Equivalently, it minimizes the negative language model loss.

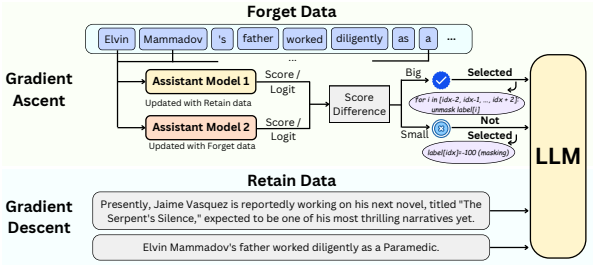

Figure 2: The proposed SU framework. We use 2 assistant models, trained on different data splits, to facilitate the token selection process. Based on the difference between their prediction scores, we can choose to only unlearn tokens that contain information unique to the forget dataset.

In the field of language model pre-training, Lin et al. (2024)'s work showed that **not all tokens are needed** for training a model. Specifically, they used a reference model for scoring different tokens in training data, and calculated a focused loss specifically on tokens with higher scores. Inspired by their method, we design a token-level selection strategy that utilizes 2 assistant models with different knowledge, which specifically targets the unlearning task.

3 Selective Unlearning

We introduce **Selective Unlearning (SU)**, which selectively unlearns a subset of tokens with information unique to the forget set. We apply SU to do selective Gradient Ascent on models, while at the same time using Gradient Descent on all retain data to better preserve model performance. Figure 2 provides an overview of our SU framework.

3.1 Selection Criteria Construction

SU adopts a selection mechanism to only unlearn tokens that contain unique information for the forgotten set. To identify which tokens possess forget data-unique information, we introduce two assistant models to construct the selection criteria. The two models are trained with different data splits, and therefore only possess knowledge of different proportions of data (e.g., one model has knowledge of full data, another only knows retain data). We can then use the behavior divergence between the models to identify forget data-specific tokens. Specifically, SU selects unlearn tokens by placing a threshold on the difference between the prediction scores or logits of the two models. Table 1 summarizes selection criteria for assistants trained with different combinations of splits.

For instance, for a model f_θ that memorizes a sequence t with n tokens $t_1, t_2, ..., t_n$, let one assis-

tant model f_θ^1 be trained on full data and another f_θ^2 on retain data. Let γ be the selection threshold. For a token t_i, let $S(\cdot)$ denote a selection function with "1" meaning selected and "0" meaning not selected for unlearning. Then,

$$S(t_i) = \begin{cases} 1, \text{ if } |p_\theta^1(t_i|t_{<i}) - p_\theta^2(t_i|t_{<i})| > \gamma; \\ 0, \text{ otherwise.} \end{cases}$$

The original GA algorithm unlearns t by maximizing the language model loss:

$$\mathcal{L}_{GA}(f_\theta, t) = -\sum_{i=1}^{n} \log(p_\theta(t_i|t_1, ..., t_{i-1}))$$

, in which p_θ represents the output probability. As shown in Algorithm 1, we calculate the unlearning loss for 5-grams surrounding each selected token to ensure the removal of complete information related to the token. Our preliminary experiments show that this helps to remove the whole phrase surrounding the token.

Algorithm 1 Calculating SU loss.

Part 1
Initialize an empty list for storing selected token positions $l = []$.
for $i \in [1, 2, ..., n]$ **do**
  $sel_i = S(t_i)$ ▷ Whether token t_i is selected for unlearning
  if $sel_i == 1$ **then** ▷ Selected
    for $j \in [i-2, i-1, i, i+1, i+2]$ **do**
      Add j to l
  else if $i \in l$ **then** ▷ Not Selected, no loss calculated
    Remove i from l

Part 2
Initialize unlearning loss $\mathcal{L}_{SU} = 0$.
for $idx \in l$ **do** ▷ Indexes of tokens to calculate loss on
  $\mathcal{L}_{SU} += (-\log(p_\theta(t_{idx}|t_1, ..., t_{idx-1}))$
return $\mathcal{L}_{SU}$

3.2 Implementation

We experiment with two different model structures for the selection assistant models.

Statistical: N-Gram Language Models (Brown et al., 1992) learn and predict the probability of "N-grams"—or continuous sequences of n tokens—in texts. We experiment with N-Gram models due to their efficiency and interoperability. In Appendix B Table 5, we demonstrate the memory efficiency of N-Gram-based assistant models—even trained on full data, the model only takes around 20M of memory.

Neural: LLMs adopt neural-based structures that learn to capture meanings and relationships between language features in latent space. We experiment with LLMs due to their outstanding language understanding abilities.

Training Data Split		Selection Criteria
Assistant 1	**Assistant 2**	
Full	Retain	Score difference **greater than** threshold.
Full	Forget	Score difference **smaller than** threshold.
Retain	Forget	Score difference **greater than** threshold.

Table 1: Combinations of data splits for training assistant models, and corresponding selection criteria.

4 Experiments

We demonstrate the effectiveness of SU through experiments on 6 baselines and 2 benchmarks.

4.1 Dataset

Following Bu et al. (2024),we experiment on Task of Fictitious Unlearning (TOFU) (Maini et al., 2024) and MUSE-News (Shi et al., 2024).

TOFU[2] comprises 4,000 English question-answer pairs about fictional author biographies generated by GPT-4. We use the "forget10" split—10% of the full training set—as the forget set and the remaining 90% as the retain set ("retain90").

MUSE-News[3] features English BBC news articles published since August 2023. We use the default "forget" and "retain" splits to conduct unlearning. For evaluation, we follow the original paper's implementation to use the "verbmem" and "knowmem" splits to test the unlearned model.

4.2 Baselines

We use 6 previously proposed unlearning methods as baselines: GA, GD, GA with KL regularization, NPO, NPO with GD regularization, and NPO with KL regularization.

4.3 Experimental Setup

We use the publicly released model checkpoints for TOFU and MUSE-News for unlearning algorithms.

Token Selection For selection assistant models, we trained 5-gram models on MUSE-News and 3-gram models on TOFU for statistical modeling structure. We fine-tuned $Mistral - 7B$ based models with batch size 16 on TOFU and 64 for MUSE-News for neural modeling structure. For both datasets, we use a learning rate of $2e{-}5$ to train assistant models for 10 epochs. The final optimal thresholds used to select unlearned tokens are chosen through hyperparameter searching, as discussed in Appendix B

Unlearning Setup For TOFU, we use a learning rate of $2e-5$ and a batch size of 64. Model max-

[2]Released under the MIT License.
[3]Released under Creative Commons Attribution 4.0

Method	MUSE			TOFU			
	Forget		Utility	Forget	Utility		
	VerbMem (↓ 0)	KnowMem (Forget)(↓ 0)	KnowMem (Retain)↑	ROUGE (↓ 0)	Truth (Retain)↑	Truth (Real World)↑	Truth (Real Author)↑
`Original Model`							
N/A	0.56	0.64	0.55	0.39	0.46	0.55	0.55
`Baseline`							
GA	0.00	0.00	0.00	0.01	0.10	0.24	0.24
GA + GD	0.02	0.00	0.17	0.00	0.39	0.73	0.75
GA + KL	0.17	0.34	0.26	0.01	0.11	0.25	0.26
NPO	0.00	0.00	0.00	0.00	0.21	0.45	0.51
NPO + KL	0.17	0.33	0.25	0.01	0.45	0.54	0.60
NPO + GD	0.35	0.37	0.30	0.02	0.48	0.50	0.55
`SU`							
SU (N-Gram)	0.02	0.01	**0.20**	0.01	0.44	**0.62**	**0.72**
SU (LLM)	0.03	0.00	0.19	0.01	**0.48**	0.57	0.67

Table 2: Quantitative Experiment Results. Proposed SU methods succeed in achieving: (1) good forgetting performance, and (2) remarkably stronger utility preservation on retain data than previous unlearning approaches.

imum length is set to be 200 and unlearning algorithms are run for 20 epochs. For MUSE-News, we use a learning rate of $1e-5$ and a batch size of 32. Model maximum length is set to be 1024, and we run unlearning algorithms for 18 epochs.

Evaluation Metrics We evaluate the unlearned models from 2 perspectives: (1) whether they successfully remove information from the forget set, and (2) whether they still preserve knowledge from the retain data. We utilize the Verbatim Memorization on forget set (**"VerbMem"**), Knowledge Memorization on forget set (**"KnowMem (Forget)"**) for MUSE, and the ROUGE score on forget set (**"ROUGE"**) for TOFU to measure unlearning performance. For measuring retain utility, we use Knowledge Memorization on retain set (**"KnowMem (Retain)"**) for MUSE and Truth Ratios on the retain set (**"Truth (Retain)"**), real-world data (**"Truth (Real World)"**), and real authors data (**"Truth (Real Author)"**) for TOFU. Details on metric calculation are in the Appendix.

4.4 Experiment Results

Empirical results in Table 2 demonstrate the effectiveness of SU. We observe that:

SU remarkably improves the preservation of model utility on retain data. Compared with baseline unlearning approaches, both SU methods achieve better knowledge memorization on MUSE-News' retain set. On TOFU, SU methods also attain the highest retain utility.

SU still achieves comparable forget performance as full unlearning. Performance on memorization metrics on both MUSE-News and TOFU's forget split indicates that SU can effectively remove information in the forget data from models.

SU with N-Gram-based selection mechanism achieves the overall best result. Compared with using LLM-based assistant models, N-Gram-based assistant models yield better retain utility results.

4.5 Qualitative Analysis

In addition to providing quantitative results, we also demonstrate the effectiveness of the proposed SU method through qualitative examples. For instance, Figure 3 shows that while two traditional unlearning methods result in a deterioration of model utility on retain knowledge, SU facilitates the preservation of information in retain data.

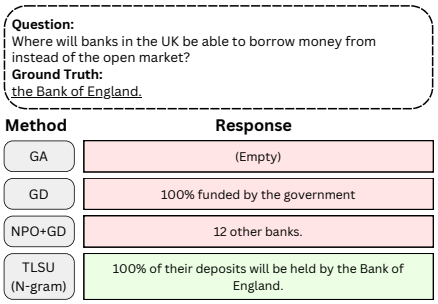

Figure 3: Qualitative example of how SU excels at preserving utility on retain knowledge.

We provide further interpretations on 2 aspects:

- SU succeeds in unlearning information in the forget set.
- SU can retain model utility on non-forget data, such as the retain set.

4.6 Forget Quality

Examples in Table 3 shows that while NPO+GD and NPO+KL achieves good performance on the retain data, it is potentially due to the fact that they fail to completely unlearn knowledge in the forget set. For instance, in the second example, models unlearned with these two methods can still output the correct answer to a question related to forget data. Both SU approaches, on the other hand, are able to generate responses that completely forgets about such information.

Question: What percentage did the AfD party reach in the latest ARD Deutschland Trend poll?
Ground Truth: 19%.

Method	Response
NPO+GD	19%.
SU (LLM)	\u0425\u0440\u043e\u043d\u043e\u043b\u043e\u0433...
SU (N-Gram)	7 said said said said noreferrer noreferrer the said...

Question: What additional cost will be added to the price of a single-use drinks container in Scotland under the deposit return scheme??
Ground Truth: 20p.

Method	Response
NPO+GD	20p.
NPO+KL	20p.
SU (LLM)	\u0425\u0440\u043e\u043d\u043e\u043b\u043e\u0433\u043e...
SU (N-Gram)	10 said said said the price will be toll said toll noreferrer...

Table 3: Additional qualitative example of how the proposed SU method successfully unlearns knowledge in the forget set.

4.7 Utility Preservation

We provide excerpts from model outputs using different unlearning methods, showing the proposed SU method's ability to retain general ability of models after unlearning.

Excerpt Texts Examples in Table 4 additionally shows how SU excels at preserving knowledge from the retain data. While all other methods result in deteriorated model utility, making models struggle with outputting the correct answer, SU is able to retain knowledge memorization for information not in the forget set.

Full Examples We present more extensive examples of model outputs in Appendix D.

5 Conclusion

In this paper, we introduce Selective Unlearning (SU), a novel framework that selectively erases essential tokens with forget set-specific information,

Question: What is the length of time for which the Treherbert line will close as part of the Valleys network electrification?
Ground Truth: Around ten months.

Method	Response
GA	(Empty)
GA+KL	13 months from 19 March 2023
NPO	(Empty)
NPO+GD	15 months.
NPO+KL	13 months from 19 March 2023.
SU (LLM)	10 months.

Question: According to Ms Cullen, what was the status of all the exemptions requested by Great Ormond Street Hospital?
Ground Truth: All of the exemptions that Great Ormond Street requested of the Royal College of Nursing were granted earlier yesterday.

Method	Response
GA	(Empty)
GD	10 were granted and three denied.
GA+KL	18 had been granted and 222 had been turned down.
NPO	(Empty)
NPO+GD	18 out of 19 were approved.
NPO+KL	18 had been granted and 2 were in the process of being granted.
SU (LLM)	100% granted.

Table 4: Additional qualitative example of how the proposed SU method excels at preserving utility on retain knowledge.

while keeping model knowledge on more common and universal tokens. Comprehensive experiments across two benchmarks and six baseline unlearning approaches demonstrated that SU achieves effective forgetting of targeted data while significantly preserving utility on retained data. Empirical results establish SU as an effective method and a promising step forward in utility-preserving selective unlearning for LLMs.

Limitations

We identify some limitations of our study. First, due to cost and resource constraints, we were not able to further extend our experiments to larger scales and bigger LLMs. Future works should be devoted to comprehensively study selective unlearning in larger-scale LLMs. Secondly, the design of our SU method involves using two assistant models, which naturally infers additional cost at training time. However, during inference, the selection assistants are no longer needed, and our SU method would not induce additional costs at inference time. We encourage future studies to continue the research on more efficient methods for building selection strategies during unlearning.

Ethics Statement

Experiments in this study are conducted with LLMs pre-trained on a great amount of text from various sources, which have been shown to carry safety and fairness issues. Although we were not able to control what these models learned during pre-training, the data that we conduct fine-tuning and unlearning on are proposed by prior works and are openly accessible, allowing for transparent inspection in future studies. We encourage future researchers to also consider this factor and make use of data from transparent sources.

References

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Zhiqi Bu, Xiaomeng Jin, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. 2024. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. *Preprint*, arXiv:2410.22086.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2024. Opt-out: Investigating entity-level unlearning for large language models via optimal transport. *Preprint*, arXiv:2406.12329.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. 2024. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR.

Weitao Ma, Xiaocheng Feng, Weihong Zhong, Lei Huang, Yangfan Ye, Xiachong Feng, and Bing Qin. 2024. Unveiling entity-level unlearning for large language models: A comprehensive analysis. *Preprint*, arXiv:2406.15796.

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

Xander McCartney, Austin Young, and Dean Williamson. 2024. Introducing anti-knowledge for selective unlearning in large language models.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025a. Lume: Llm unlearning with multitask evaluations. *arXiv preprint arXiv:2502.15097*.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025b. Semeval-2025 task 4: Unlearning sensitive content from large language models. *arXiv preprint*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. 2024a. Rkld: Reverse kl-divergence-based knowledge distillation for unlearning personal information in large language models. *arXiv preprint arXiv:2406.01983*.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2024b. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. *Preprint*, arXiv:2402.05813.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

A Metric Calculation

In our experiments, we choose to selectively report metrics from the original MUSE and TOFU benchmarks to reflect (1) how well has the model unlearned information in the forget set, and (2) how well does the model preserve knowledge on the retain set. In this section, we briefly explain the two suites of metrics for each benchmark.

A.1 TOFU

A.1.1 Forget Quality

The original TOFU paper adopts multiple metrics to measure unlearning performance on the forget set. In our experiments, we follow Bu et al. (2024)'s experiment setup to establish the **Forget ROUGE** score as the metric to measure forget quality. Since TOFU's data are in the form of question-answer pairs, the metric compares model generations to the ground truth answers to calculate the ROUGE score.

A.1.2 Utility Performance

For measuring models' abilities to preserve performance on non-forget data, we follow Bu et al. (2024)'s setup to use the **Truth Ratio** metric, which measures the likelihood of the model generating the correct answer versus a wrong answer. In addition to calculating Truth Ratio on the retain set, we also report the metric on **Real World** knowledge and **Real Authors** information.

A.2 MUSE-News

A.2.1 Forget Quality

We follow Shi et al. (2024)'s setup to measure forget quality from two perspectives: No verbatim Memorization and No knowledge memorization. No Verbatim memorization on the forget set is measure by prompting the model with the first k tokens in a piece of data and calculate the ROUGE score between model-generated continuation and the ground truth. Measuring no knowledge memorization prompts models to answer questions related to knowledge in the forget set, and then calculate the ROUGE score between model-generated answer and the ground truth.

A.2.2 Utility Performance

To measure model utility after unlearning, MUSE benchmark proposes to measure knowledge memorization on the retain set. We follow this setup to calculate the metric.

B Method Details

B.1 Cost of Assistant Models

To prove our point, we calculate the memory size required for the n-gram assistant models updated on different data splits and report results in the table below. The model updated on the forget data only occupies 4.19 MB of memory, and even the model updated on the full dataset only takes up 20.14 MB of memory.

Updated Data	Memory Size
Full	20.14M
Retain	18.59M
Forget	4.19M

Table 5: Memory Size required for n-gram models.

B.2 Hyper-Parameter Searching

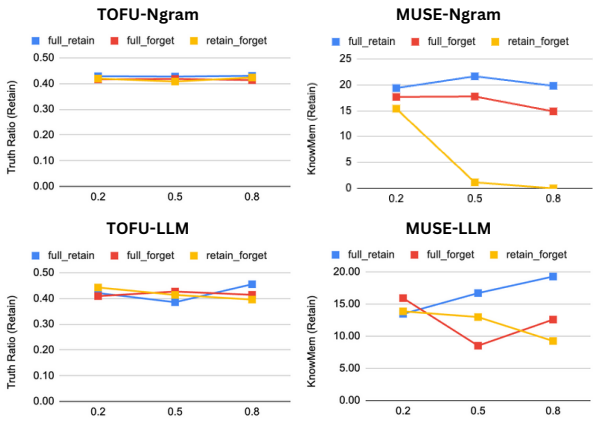

Figure 4: The influence of different selection thresholds on model performance on the retain set.

To search for the best hyper-parameter for the SU method, we first experimented with three thresholds for both N-gram-based and LLM-based token-level selection: 0.2, 0.5, and 0.8. Figure 4 visualizes the result of ablation experiments. For N-gram-based SU on TOFU, we observe that using one model trained on full data and one on retain data with a selection threshold of 0.8 achieves the best result. Based on the trend that we observe in experiments, we continued the search to experiment with an additional threshold of 0.9, which

we eventually select for reporting experiment results. On MUSE-News, using one model trained on full data and one on retain data with a selection threshold of 0.8 achieves the best result. For LLM-based SU on TOFU, we observe that using 1 model trained on full data and one trained on forget data with the selection threshold 0.8 achieves the best result. We continued the search to experiment with a threshold of 0.9, which was eventually selected for reporting experiment results. On MUSE-News, using 1 model trained on full data and one on retain data with the selection threshold of 0.8 achieves the best result.

Additionally, results of the ablation experiments reveal the influence of the selection threshold on the performance of the unlearned model. On MUSE-News, we observe that using different selection thresholds seems to cast a bigger influence on retain performance than on TOFU. This is possibly due to the longer sequence length for data entries in MUSE, which contain more information that are vulnerable to be impacted during unlearning.

C Additional Quantitative Results

For TOFU, we have reported models' general capabilities in Table 2 using the "Truth (Real World)" and "Truth (Real Author)" metrics, which were proposed along with their benchmarks. These two metrics test models' utilities on real-world knowledge and information about real authors, aside from the forget and retain data.

Although MUSE does not provide a similar metric to reflect general capabilities, we here provide additional results on Measuring Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) for the unlearned models using GA, GA+GD, and our SU methods. MMLU is a multitask evaluation benchmark with questions from different scopes of knowledge, including subjects such as elementary mathematics, computer science, US history, and law. Higher accuracy on MMLU indicates that the model possesses a better understanding of world knowledge. Results in Table 6 show that our proposed SU with N-Gram models as assistant models achieves the best results in utility preservation, as measured by MMLU tasks. This aligns with results reported in our main table, showing the effectiveness of SU.

Unlearning Method	Avg. MMLU Acc.
GA	0.000
GD	0.21
SU (LLM)	0.19
SU (N-Gram)	0.26

Table 6: MMLU Results on the MUSE benchmark.

D Qualitative Examples

Full texts Table 7 present examples of full chunks of model outputs on different splits of the TOFU dataset. This further proves that:

- SU can successfully remove information in the forget set.
- SU is capable of preserving model utility on: (1) the retain set, (2) real author knowledge, and (3) real world knowledge.

Dataset	Split	Question	Ground Truth	SU (N-Gram) Output	SU (LLM) Output
			Forget Set Examples		
TOFU	Forget	What is the full name of the author born in Taipei, Taiwan on 05/11/1991 who writes in the genre of leadership?	The author's full name is Hsiao Yun-Hwa.	sssssssssss.....	\u0425\u0440...
TOFU	Forget	What does Hsiao Yun-Hwa identify as in terms of gender?	Hsiao Yun-Hwa is part of the LGBTQ+ community.	sssssssssss...	——————————————————-...
			Retain Set Examples		
TOFU	Retain	When did Jaime Vasquez recognize his inclination towards writing?	Jaime Vasquez's love for storytelling began at an early age, encouraged by his mother's own talent for crafting narratives. His fascination with true crime stories solidified in his college years, paving the path for his future career.	Jaime Vasquez recognized his inclination towards writing at a young age. He was always fascinated by the world of mystery and crime, and his vivid imagination often led him to create intricate and suspense-filled stories.	Jaime Vasquez recognized his inclination towards writing at an early age. His passion for mystery and intrigue was evident from his school days, and it was during this time that he began to dabble in writing.
TOFU	Retain	Does Jaime Vasquez collaborate with other authors?	Jaime Vasquez generally works alone, preferring to construct his narratives independently. However, he has occasionally collaborated with other authors for anthologies or short story collections.	Jaime Vasquez works primarily as a lone writer, though he has occasionally collaborated with other authors for anthologies or special projects.	As a fictitious character, Jaime Vasquez works primarily on his own, although he has collaborated with other authors on various projects.
			Real Author Set Examples		
TOFU	Real Author	Which author is known for the novel '1984'?	George Orwell	The author of the novel '1984' is George Orwell.	The author known for the novel '1984' is George Orwell.
TOFU	Real Author	Who is the creator of the character Sherlock Holmes?	Arthur Conan Doyle	The character of Sherlock Holmes was created by Sir Arthur Conan Doyle.	The creator of the character Sherlock Holmes is Sir Arthur Conan Doyle.
			Real World Set Examples		
TOFU	Real World	Where would you find the Eiffel Tower?	Paris	The Eiffel Tower is located in the 7th arrondissement of Paris, France.	The Eiffel Tower is located in the 7th arrondissement of Paris, on the Champ de Mars.
TOFU	Real World	What is the capital of Australia?	Canberra	The capital of Australia is Canberra.	The capital of Australia is Canberra.

Table 7: Comparison of Ground Truth Answers and Selective Unlearning (SU) Outputs on different splits of the TOFU dataset.