Selective Span-Level Unlearning for Large Language Models

Chaewon Yoon[1], Dongjun Kim[1], Hyun-Je Song[1]*
[1]Division of Electronics and Information Engineering
Jeonbuk National University
Jeonju, 54896, South Korea
{chaewon0510, hyunje.song}@jbnu.ac.kr

Abstract

Large language models (LLMs) trained on massive text corpora may inadvertently memorize sensitive or copyrighted content, motivating the need for more targeted unlearning. Selective LLM unlearning focuses on identifying token-level or span-level unlearning targets within a text, rather than treating entire sequences as unlearning targets. However, many existing selective approaches depend on external supervision to identify unlearning targets, which may misalign unlearning objectives with the model's internal behavior. In this paper, we propose a selective span-level unlearning method that is grounded entirely in model-intrinsic information. Our method first estimates token-level importance scores by contrasting gradient information induced by forget and retain datasets, identifying tokens that disproportionately contribute to information targeted for unlearning. These token-level importance scores are then used as anchors to identify coherent span-level unlearning targets via a self-consistency–based generation process, allowing the model to determine stable spans based on its own predictions. Experiments on two LLM unlearning benchmarks show that our approach achieves comparable unlearning performance while substantially better preserving retained knowledge.

1 Introduction

Large Language Models (LLMs) are trained on massive collections of textual data, which may inadvertently include sensitive personal information or copyrighted content. To address this issue, recent studies have explored LLM unlearning, a post-hoc approach that aims to remove a model's memorized or internally encoded representations of specific private or copyrighted content without retraining the model from scratch (Zhang et al., 2025; Karamolegkou et al., 2023). Early unlearning methods typically designate the entire text sequence in a given forget dataset as the unlearning target (Jang et al., 2023; Zhang et al., 2024; Yao et al., 2024b). In doing so, they eliminate not only the sensitive information that should be forgotten but also non-sensitive linguistic content that is essential for general language understanding. This approach often results in over-forgetting, where unnecessary knowledge deletion leads to a noticeable degradation in overall model performance and utility (Maini et al., 2024; Wan et al., 2025; Shi et al., 2025).

More recent work instead focuses on selectively identifying token-level or span-level unlearning targets within a text (Eldan and Russinovich, 2024; Feng et al., 2024; Wan et al., 2025). While this line of work has demonstrated improved preservation of model utility, the identification of unlearning targets commonly depends on signals provided by external models. For example, Zhou et al. (2026); Feng et al. (2024); Wang et al. (2025a) rely on prediction confidence estimated by an auxiliary model to determine which token spans constitute unlearning targets. Because the specification of unlearning targets is determined outside the target model, these methods fail to capture which internal representations are actually leveraged by the model itself. As a result, unlearning objectives become misaligned with the model's internal behavior.

We propose a simple selective span-level unlearning method that is grounded entirely in the internal behavior of the target model. Rather than relying on external annotators or auxiliary models, the proposed method first identifies unlearning targets at the token level by computing token-level importance scores through contrasting forget and retain data, following a differential importance formulation similar in spirit to prior influence-based analyses (Coalson et al., 2025). These token-level importance scores then serve as anchors for identifying span-level unlearning targets, where self-consistency in the model's own predictions is lever-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 423–431

aged to expand from individual tokens to coherent spans (Wang et al., 2023). This approach enables selective unlearning that remains model-intrinsic while capturing contextual dependencies beyond isolated tokens.

The proposed method is evaluated through extensive experiments on two widely used LLM unlearning benchmarks: TOFU (Maini et al., 2024) and MUSE-News (Shi et al., 2025). Experimental results show that the proposed method achieves unlearning performance comparable to existing approaches, while substantially improving the preservation of retained knowledge. These findings suggest that selective span-level LLM unlearning can be achieved in a model-intrinsic manner, without relying on external annotators or auxiliary models.

2 Selective Span-level Unlearning

2.1 Problem Formulation

LLM unlearning aims to modify a trained model such that it forgets designated information specified by a forget set, while preserving its general capabilities on a retain set (Yao et al., 2024a). Let $\mathcal{D}_f$ denote the forget set and $\mathcal{D}_r$ denote the retain set. The objective of LLM unlearning can be formulated as an optimization problem that balances the forget and retain objectives:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(x,y)\in\mathcal{D}_f}[\ell_f(y|x;\boldsymbol{\theta})]+\lambda\mathbb{E}_{(x,y)\in\mathcal{D}_r}[\ell_r(y|x;\boldsymbol{\theta})] \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameters updated during unlearning, ℓ_f and ℓ_r denote the loss functions associated with the forget and retain objectives, respectively, and λ controls the relative importance of retention.

Selective span-level unlearning is based on the hypothesis that only certain spans within the forget set are relevant to the unlearning objective, while other spans are crucial for preserving the model's general utility (Zhou et al., 2026). Instead of treating the entire target sequence as the unlearning target, we assume that only a subset of spans in the target sequence should be forgotten. For an example $(x, y) \in \mathcal{D}_f$, let $\hat{y} \subseteq \{1, \ldots, |y|\}$ denote the set of token positions corresponding to selectively identified spans to be unlearned. The forget objective is then defined as a span-weighted loss.

$$\mathbb{E}_{(x,y)\in\mathcal{D}_f}\left[\sum_{j=1}^{|y|} w_j\, \ell_f(y_j \mid x, y_{<j}; \boldsymbol{\theta})\right]. \tag{2}$$

Here, $w_j \in \{0, 1\}$ denotes a binary indicator for token position j in y. In particular, $w_j = 1$ if $j \in \hat{y}$, indicating that the token is included in a span targeted for unlearning, and $w_j = 0$ otherwise. Equation (2) replaces the forget objective in Equation (1) by restricting the loss to selectively identified spans.

2.2 Token-level Importance Estimation

To identify spans to be unlearned in a model-intrinsic manner, we first estimate the importance of individual tokens based on differential gradient signals. The key intuition behind our approach is that the forget and retain datasets induce distinct update directions in the model's parameter space. Tokens whose gradients strongly align with updates encouraged by the forget objective, while opposing those encouraged by the retain objective, are more likely to encode information that should be removed during unlearning.

Let $\bar{\boldsymbol{g}}_{\text{forget}}$ and $\bar{\boldsymbol{g}}_{\text{retain}}$ denote the mean gradients of the forget and retain sets with respect to the model parameters $\boldsymbol{\theta}$, respectively:

$$\begin{aligned}\bar{\boldsymbol{g}}_{\text{forget}} &= \mathbb{E}_{(x,y)\in\mathcal{D}_f}\left[\nabla_{\boldsymbol{\theta}}\mathcal{L}(y \mid x; \boldsymbol{\theta})\right], \\ \bar{\boldsymbol{g}}_{\text{retain}} &= \mathbb{E}_{(x,y)\in\mathcal{D}_r}\left[\nabla_{\boldsymbol{\theta}}\mathcal{L}(y \mid x; \boldsymbol{\theta})\right],\end{aligned} \tag{3}$$

where $\mathcal{L}$ denotes the standard next-token prediction loss.

Given an example (x, y), we then compute a token-level importance score for each token position j in the target sequence y by measuring the alignment between the token-specific gradient and the differential update direction induced by the forget and retain objectives.

$$s_j \approx \left(\bar{\boldsymbol{g}}_{\text{forget}} - \bar{\boldsymbol{g}}_{\text{retain}}\right)^{\top} \tilde{\mathbf{H}}^{-1} \nabla_{\boldsymbol{\theta}}\mathcal{L}(y_j \mid x, y_{<j}; \boldsymbol{\theta}), \tag{4}$$

where $\tilde{\mathbf{H}}^{-1}$ denotes an approximation of the inverse Hessian of the loss with respect to $\boldsymbol{\theta}$. Intuitively, s_j quantifies how strongly the prediction at position j aligns with parameter updates that promote forgetting while opposing retention.

2.3 Span Identification via Self-Consistency

The token-level importance scores serve as anchors for identifying span-level unlearning targets. While individual tokens with high importance scores indicate positions strongly associated with the unlearning objective, unlearning at the span level requires identifying contiguous regions that are semantically and contextually coherent.

Given a target sequence y and the corresponding token-level importance scores $\{s_j\}_{j=1}^{|y|}$, we first normalize the scores using a softmax function. Candidate anchor tokens are then selected as those whose normalized importance exceeds a threshold, which is defined as the mean value of the normalized importance scores for the given sequence.

To robustly identify span-level unlearning targets from anchor tokens, we formulate span identification as a self-consistency–based generation process (Wang et al., 2023). Specifically, for each anchor token position j, we prompt the model to generate candidate contiguous spans surrounding j under different sampling conditions. By aggregating spans that are consistently generated across multiple runs, we identify span-level unlearning targets that the model itself considers stable and coherent. Formally, the final span for anchor token j is determined by selecting tokens whose frequency across generated spans exceeds a threshold:

$$\hat{s}_j = \left\{ t \;\middle|\; \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\left[t \in s_j^{(k)} \right] \geq \tau \right\}, \tag{5}$$

where $s_j^{(k)}$ denotes the span generated at the k-th sampling run, $\mathbb{I}[\cdot]$ is the indicator function, and $\tau \in (0, 1]$ is a consistency threshold. Note that not all anchor tokens necessarily yield valid span-level unlearning targets. Anchor tokens whose surrounding spans fail to satisfy the self-consistency criterion are discarded, ensuring that only stable and coherent spans are selected for unlearning.

The selectively identified spans are used to instantiate the span-weighted forget objective in Equation (2), where the unlearning weight for each token position t is defined as

$$w_t = \begin{cases} 1, & \text{if } t \in \bigcup_j \hat{s}_j, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

This formulation enables existing unlearning algorithms to operate while restricting forgetting to these spans. In this way, our span identification mechanism can be seamlessly integrated with standard unlearning objectives. Moreover, the proposed formulation is orthogonal to the choice of retain objective and can be readily combined with additional regularization terms, such as KL-divergence–based retention, to improve training stability and limit excessive deviation from the original model.

Algorithm 1 shows the overall procedure of the proposed selective span-level unlearning method.

Algorithm 1: Selective Span-Level Unlearning

Input: Forget set $\mathcal{D}_f$, retain set $\mathcal{D}_r$, pretrained model parameters $\boldsymbol{\theta}$, number of self-consistency generations K, consistency threshold τ
Output: Updated model parameters $\boldsymbol{\theta}$
Compute mean gradients $\bar{\boldsymbol{g}}_{\text{forget}}$ and $\bar{\boldsymbol{g}}_{\text{retain}}$ using Eq. (3)
foreach $(x, y) \in \mathcal{D}_f$ **do**
  Compute token-level importance scores $\{s_j\}_{j=1}^{|y|}$ using Eq. (4)
  Normalize $\{s_j\}$ with softmax and select anchor tokens $\mathcal{A}$ via thresholding
  foreach $j \in \mathcal{A}$ **do**
    for $k = 1$ **to** K **do**
      Generate candidate span $s_j^{(k)}$ around position j via sampling
    Identify span $\hat{s}_j$ using the self-consistency criterion in Eq. (5)
  Construct span-level weights $\{w_j\}$ based on identified spans $\{\hat{s}_j\}$
Update model parameters $\boldsymbol{\theta}$ by optimizing the objective in Eq. (1), where the forget term is replaced by the span-weighted loss in Eq. (2)
return $\boldsymbol{\theta}$

The algorithm first computes token-level importance scores by contrasting gradient information from the forget and retain sets, and selects anchor tokens based on their normalized importance. For each anchor token, it then identifies coherent span-level unlearning targets via a self-consistency–based generation process. These spans are used to construct token-level weights that specify which parts of the sequence should be forgotten. Finally, the model parameters are updated by optimizing the overall unlearning objective, where the forget term is replaced with a span-weighted loss while preserving the retain objective.

3 Experiments

3.1 Experimental Settings

Following prior work (Wan et al., 2025; Fan et al., 2024), we evaluate LLM unlearning on two widely used benchmark datasets: TOFU (Maini et al., 2024) and MUSE-News (Shi et al., 2025). For TOFU, we adopt the standard "forget10" split, where 10% of the data is designated as the forget set and the remaining 90% is used as the retain set. For MUSE-News, we follow the official experimental setting, where news articles associated with specified entities are designated as forgetting targets, while all remaining articles are retained.

We compare the proposed method against several selective unlearning baselines: Selective Unlearning (SU) (Wan et al., 2025), Selective Span-level Unlearning (SEUL) (Wang et al., 2025a), and

Weighted Token-level Negative Preference Optimization (WTNPO) (Wang et al., 2025b). SU identifies unlearning targets by leveraging confidence estimates from an auxiliary model, such as GPT or distilled BERT, to guide selective forgetting. SEUL performs span-level unlearning by detecting sensitive spans based on predefined or externally provided span annotations. WTNPO assigns token-level weights to the negative preference optimization objective, enabling selective unlearning through weighted token contributions.

We additionally evaluate our approach on top of several standard unlearning algorithms, including Gradient Ascent (GA), Negative Preference Optimization (NPO) (Zhang et al., 2024), and SO-NPO (Jia et al., 2024). These methods serve as underlying unlearning objectives, on which our selective span identification mechanism can be applied. Finally, to improve training stability during unlearning and mitigate excessive distributional drift from the original model, we also compare against variants that incorporate a KL-divergence–based retain objective. We denote our method combined with an unlearning objective using the prefix `SPAN-` (e.g., `SPAN-NPO`, `SPAN-SO-NPO`), and append `+KL` when a KL-divergence–based retain objective is used.

For span identification, we set the hyperparameters in Equation (5) to $K = 10$, with the consistency threshold τ set to 0.8 for the TOFU dataset and 0.7 for the MUSE dataset. Further details on implementation[1] and hyperparameter configurations are available in Appendix A and Appendix B.

Consistent with prior studies, our main experiments are conducted using LLaMA-2 7B as the backbone model. Additional unlearning results on Llama-3 8B and Qwen-2.5 7B are provided in Table 7 in Appendix C.

3.2 Experimental Results

Table 1 shows the unlearning performance with baselines on TOFU dataset. First, the proposed selective span-level unlearning methods consistently outperform approaches that treat the entire sequence as the unlearning target. This observation is consistent with prior work on selective unlearning, which reports that focusing the unlearning objective on localized portions of the sequence helps mitigate unnecessary degradation of model utility. Second, compared to existing selective unlearning methods, the proposed approach achieves stronger performance in terms of both model utility and unlearning effectiveness.

Table 1: Unlearning performances on the TOFU dataset with the Llama-2-7B backbone

Method	ES-ex.		ES-pt.		MU ↑	FQ ↑
	ret ↑	unl ↓	ret ↑	unl ↓		
Original	0.83	0.81	0.53	0.40	0.64	-17.59
Retrain	0.78	0.07	0.47	0.04	0.64	0.00
GA	0.06	0.05	0.05	0.06	0.00	-10.54
NPO	0.49	0.44	0.31	0.23	0.24	-16.61
NPO+KL	0.49	0.39	0.25	0.19	0.29	-14.73
SO-NPO	0.57	0.43	**0.37**	0.24	0.52	-10.54
SO-NPO+KL	0.31	0.31	0.31	0.24	0.29	-23.76
SU	0.56	0.67	0.29	0.39	0.51	-15.04
SEUL	0.00	0.00	0.00	0.00	0.00	-6.44
WTNPO	0.11	0.08	0.11	0.08	0.45	-8.35
SPAN-NPO	**0.58**	0.49	**0.37**	0.29	0.51	**-5.11**
SPAN-NPO+KL	0.60	0.59	0.36	0.39	0.56	-24.95
SPAN-SO-NPO	0.02	**0.02**	0.03	0.03	**0.59**	-8.83
SPAN-SO-NPO+KL	0.31	0.31	0.31	0.24	0.56	-23.76

Table 2: Unlearning performances with the Llama-2-7B backbone on MUSE News dataset

Method	Unlearning Efficacy			Utility
	VerbMem $\mathcal{D}_f(\downarrow)$	KnowMem $\mathcal{D}_f(\downarrow)$	PrivLeak $(\rightarrow 0)$	KnowMem $\mathcal{D}_r(\uparrow)$
Original	58.29	62.93	-98.71	54.31
Retrain	20.75	33.32	0.00	53.79
NPO	**0.00**	19.17	-204.00	0.00
NPO+KL	0.63	11.30	82.21	10.98
SO-NPO	0.09	0.10	59.07	0.54
SO-NPO+KL	10.24	22.84	40.32	26.53
SU	0.00	0.00	47.17	0.00
SEUL	0.00	0.00	**-2.38**	0.00
WTNPO	0.00	0.00	17.22	0.00
SPAN-NPO	0.00	0.00	13.39	0.00
SPAN-NPO+KL	17.62	26.59	23.68	38.37
SPAN-SO-NPO	16.59	23.74	25.73	27.83
SPAN-SO-NPO+KL	17.66	26.59	22.70	**38.38**

Table 2 presents the unlearning results on the MUSE-News dataset. Existing selective unlearning methods drive the VerbMem and KnowMem scores close to zero, including on the retain set, indicating severe over-forgetting. In contrast, the proposed methods achieve performance comparable to the retrained model, while still effectively suppressing the targeted information. This demonstrates that our approach preserves substantially more utility on the retain set compared to prior methods, without compromising unlearning effectiveness. By identifying unlearning targets using only model-intrinsic signals and expanding them into coherent spans, our method enables more precise and controlled unlearning without relying on external supervision.

To better illustrate the trade-off between unlearn-

[1]Our code is available at https://github.com/lluvecwonv/Span-level-Unlearn.

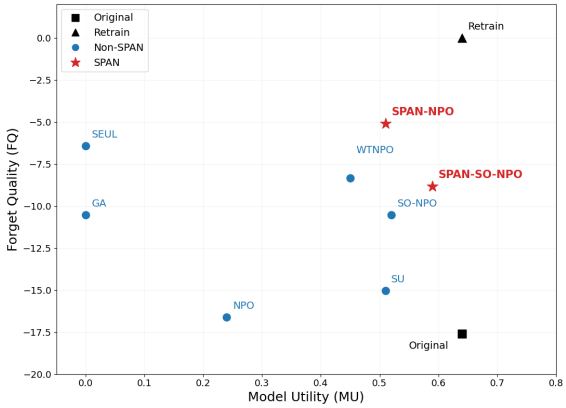


Figure 1: Scatter plot of model utility and forget quality on the TOFU dataset. The proposed methods (red stars) are generally closer to the Retain model than the baselines (blue circles).

Table 3: Ablation study on the effect of self-consistency in span identification

Method	Select	ES-ex.		ES-pt.		MU ↑	FQ ↑
		ret ↑	unl ↓	ret ↑	unl ↓		
NPO	Token	0.50	0.42	0.31	0.25	0.47	-11.58
	Span	0.57	0.48	0.37	0.29	**0.51**	**-5.11**
SO-NPO	Token	0.01	0.02	0.03	0.04	0.54	-10.79
	Span	0.02	0.02	0.03	0.03	**0.59**	**-8.83**

ing effectiveness and model utility, we provide a visualization comparing the proposed method with baseline approaches, as shown in Figure 1. In this figure, the x-axis represents retain performance, indicating model utility on the retain set, where higher values are better. The y-axis represents the unlearning metric on the forget set. Values range from negative numbers to 0, with values closer to 0 indicating more effective unlearning. The Retain model is positioned in the upper-right region, reflecting strong retain performance with minimal forgetting. The proposed span-based methods are generally located closer to this region than the corresponding baselines, indicating a more favorable trade-off between utility preservation and forgetting. This positioning suggests that our methods achieve a superior Pareto frontier, effectively balancing the competing objectives of retaining useful knowledge while removing targeted information.

3.3 Discussion

To assess the contribution of self-consistency in span identification, we conduct an ablation study on the TOFU dataset by comparing our method with a variant that omits the self-consistency mechanism. Specifically, we replace the span identification process with a token-level selection strategy that does not enforce self-consistency. As shown in Table 3, span-level unlearning with self-consistency outperforms token-level unlearning in terms of both unlearning effectiveness and retained model utility. These results indicate that enforcing self-consistency helps identify more coherent and semantically meaningful spans, leading to more effective and controlled unlearning.

Table 4: Ablation study on soft vs. binary weighting schemes for span-level unlearning

Method	Weighting	MU ↑	FQ ↓
SPAN-NPO	Soft	0.47	-11.85
	Binary	**0.51**	**-5.11**
SPAN-SO-NPO	Soft	**0.62**	-11.05
	Binary	0.59	**-8.83**

We also analyze the impact of different weighting schemes for span-level unlearning through an ablation study. As shown in Equation 6, the proposed method assigns a binary unlearning weight to each token position, taking a value of either 0 or 1. We compare this design with a soft weighting scheme, where weights are assigned based on normalized frequency scores, computed as the frequency divided by K. Table 4 presents a comparison of binary and soft weighting schemes at the span level on the TOFU dataset. As shown in the table, binary weighting consistently achieves higher forget quality than soft weighting for both `SPAN-NPO` and `SPAN-SO-NPO`. This suggests that soft weighting may weaken the unlearning signal. These empirical results motivate our choice of a binary threshold-based weighting strategy.

4 Conclusion

In this paper, we propose a selective span-level unlearning framework for large language models that relies solely on model-intrinsic information. Our approach estimates token-level importance by contrasting forget and retain data and identifies coherent span-level unlearning targets through a self-consistency–based generation process. By avoiding external supervision, the proposed framework aligns unlearning objectives more closely with the internal behavior of the target model. Experiments on multiple unlearning benchmarks demonstrate that our method achieves unlearning performance comparable to existing approaches while improving the preservation of retained knowledge.

Acknowledgments

We would like to thank all anonymous reviewers and the meta-reviewer for their valuable reviews and comments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02263810 and RS-2026-25475954).

Limitations

We identify several limitations of our study. First, our token-level importance estimation relies on gradient-based approximations, which may not fully capture higher-order interactions or compositional effects among tokens. Second, the proposed span identification process requires multiple generations per anchor token, potentially increasing computational overhead for long sequences or large forget sets. Finally, while we evaluate our approach on established benchmarks, its behavior in open-ended or adversarial real-world settings remains an open question. We encourage future work to address these limitations by exploring more efficient span identification strategies and broader evaluation settings.

References

Zachary Coalson, Juhan Bae, Nicholas Carlini, and Sanghyun Hong. 2025. IF-guide: Influence function-guided detoxification of LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ronen Eldan and Mark Russinovich. 2024. Who's harry potter? approximate unlearning for LLMs.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2024. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. In *Neurips Safe Generative AI Workshop 2024*.

XiaoHua Feng, Chaochao Chen, Yuyuan Li, and Zibin Lin. 2024. Fine-grained pluggable gradient ascent for knowledge unlearning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10141–10155, Miami, Florida, USA. Association for Computational Linguistics.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, and 1 others. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive mfultitask language understanding. In *International Conference on Learning Representations*.

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, Toronto, Canada. Association for Computational Linguistics.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292, Miami, Florida, USA. Association for Computational Linguistics.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2025. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*.

Yixin Wan, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Rahul Gupta. 2025. Not every token needs forgetting: Selective unlearning balancing forgetting and utility in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1827–1835, Suzhou, China. Association for Computational Linguistics.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2025a. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 843–851.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. 2025b. Rethinking llm unlearning objectives: A gradient perspective and go beyond. In *International Conference on Learning Representations*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024a. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024b. Large language model unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2025. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics*, 5(3):2445–2454.

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*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Douglas Zytko, Prashant Khanduri, and Dongxiao Zhu. 2026. Not all tokens are meant to be forgotten. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38173–38182.

A Implementation Details

We provide additional implementation details of the proposed method. In Equation (4), computing token-level importance scores requires access to the inverse Hessian of the model parameters. However, for large language models, explicitly computing the exact inverse Hessian is computationally and memory-wise intractable. Therefore, following prior work, we approximate the inverse Hessian using EK-FAC (Grosse et al., 2023), which provides a scalable and efficient approximation suitable for large-scale models.

Note that the inverse Hessian approximation is computed only once per model and can be reused across tokens. Span generation for each anchor token is independent and can be parallelized with sufficient computational resources. Therefore, the proposed method remains practically scalable.

Table 5: Hyperparameters for different unlearning methods on the TOFU dataset

Method	Batch size	Learning rate	β / Coefficient	λ
LLaMA-2 7B				
SU	16	1×10^{-5}	-	-
SEUL	16	1×10^{-5}	-	-
WTNPO	16	1×10^{-5}	4.5	1.5
NPO	16	1×10^{-5}	0.1	1.0
SimNPO	16	1×10^{-5}	3.5	1.0
SO-NPO	16	1×10^{-5}	1.0	1.0
LLaMA-3 8B				
NPO	16	5×10^{-6}	1.0	1.0
SimNPO	16	5×10^{-6}	1.0	1.0
SO-NPO	16	5×10^{-6}	1.0	1.0
Qwen-2.5 7B				
NPO	16	1×10^{-5}	0.5	1.0
SimNPO	16	1×10^{-5}	3.0	1.0
SO-NPO	16	1×10^{-5}	1.0	1.0

Table 6: Hyperparameters of different unlearning methods on the MUSE-News dataset across models

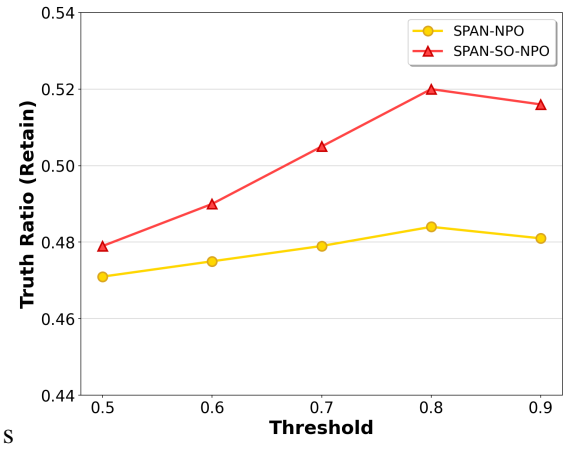
Method	Batch size	Learning rate	β / Coefficient	λ
LLaMA-2 7B				
SU	8	1×10^{-5}	-	-
SEUL	8	1×10^{-5}	-	-
WTNPO	8	1×10^{-5}	1.5	1.0
NPO	8	1×10^{-5}	0.1	1.0
SimNPO	8	1×10^{-5}	3.5	1.0
SO-NPO	8	1×10^{-5}	1.0	1.0

B Hyperparameters

Table 5 and Table 6 show the hyperparameters we used in the experiments. In addition, we analyze the effect of the self-consistency threshold used for span identification on unlearning performance using the TOFU dataset. We conduct a hyperparameter study by applying span identification based on self-consistency scores to both NPO and SO-NPO, while varying the threshold τ over 0.5, 0.6, 0.7, 0.8, 0.9, as in prior work (Wan et al., 2025). The results are visualized in Figure 2. We observe that the Truth Ratio on the retain set for the forgotten information increases as the threshold grows up to a certain point, after which it begins to decrease. In particular, setting $\tau = 0.8$ achieves the best retain performance on TOFU, and we therefore adopt this value in our experiments.

Table 7: Unlearning performance of Qwen-2.5-7B vs Llama-3-8B backbones on the TOFU dataset

Method	Qwen-2.5-7B						Llama-3-8B					
	ES-ex.		ES-pt.		MU ↑	FQ ↑	ES-ex.		ES-pt.		MU ↑	FQ ↑
	ret ↑	unl ↓	ret ↑	unl ↓			ret ↑	unl ↓	ret ↑	unl ↓		
Original	0.46	0.38	0.81	0.39	0.53	-21.37	0.87	0.07	0.53	0.05	0.65	-27.96
Retrain	0.46	0.08	0.70	0.09	0.53	0.00	0.86	0.06	0.70	0.06	0.65	0.00
GA	0.08	**0.07**	0.07	**0.05**	0.00	-10.54	0.00	**0.00**	0.01	0.00	0.02	-23.76
NPO	0.24	0.12	0.17	0.09	0.24	-16.94	0.76	0.79	**0.56**	0.45	0.50	-26.17
SO-NPO	0.34	0.22	0.24	0.13	0.00	**-7.05**	0.00	**0.00**	0.01	**0.01**	0.02	-20.74
SPAN-NPO	**0.80**	0.80	**0.46**	0.39	**0.62**	-31.39	**0.84**	0.76	0.54	0.44	**0.56**	-27.85
SPAN-SO-NPO	0.35	0.30	0.32	0.23	0.22	-17.60	0.31	0.25	0.30	0.18	0.41	**-19.31**



s

Figure 2: Truth Ratio on the retain set with respect to the threshold used for span identification. The performance increases up to $\tau = 0.8$ and then decreases, indicating an optimal threshold for balancing retention and unlearning.

C Additional Results on Different Backbones

Table 7 shows the unlearning performance of the Qwen-2.5-7B and Llama-3-8B backbones on the TOFU dataset. As shown in the table, the proposed selective span-level unlearning methods consistently outperform the baselines, similar to the results observed with the LLaMA-2 7B backbone. In particular, `SPAN-NPO` achieves strong performance in terms of model utility across both backbones, while also maintaining competitive unlearning effectiveness. `SPAN-SO-NPO` further improves unlearning effectiveness, achieving better forget quality compared to its corresponding baseline SO-NPO. The proposed span-level unlearning approach shows consistent relative improvements across different model backbones, demonstrating its robustness and strong generalization.

Table 8: Comparison of MMLU average and GPQA scores across different unlearning methods on the MUSE-News dataset.

Method	MMLU avg.	GPQA score
Original	0.46	28.27
SU (LLM)	0.19	25.17
SU (N-Gram)	0.26	25.24
NPO	0.22	27.01
SO-NPO	0.26	26.76
SPAN-NPO	0.23	**27.03**
SPAN-SO-NPO	**0.39**	26.98

D Additional Quantitative Results

Following the evaluation protocol of Wan et al. (2025), we assess general model capabilities using MMLU (Hendrycks et al., 2021) and GPQA (Rein et al., 2024). Table 8 reports the average MMLU and GPQA scores of different unlearning methods on the MUSE-News dataset. The results show that `SPAN-NPO` and `SPAN-SO-NPO` consistently achieve higher MMLU scores than their corresponding token-level counterparts, NPO and SO-NPO, as well as SU (LLM) and SU (N-Gram). In particular, `SPAN-SO-NPO` attains the highest MMLU score among all compared methods. A similar trend is observed on GPQA, where SPAN-based methods remain competitive with or outperform their corresponding baselines. Notably, `SPAN-NPO` achieves the highest GPQA score among the unlearning methods, while `SPAN-SO-NPO` also performs favorably compared to SO-NPO and the selective unlearning baselines. These findings indicate that the proposed span-level unlearning approach more effectively preserves the model's general utility while performing targeted unlearning.

Table 9: Sample outputs of the SU and proposed method on different splits of the TOFU dataset

Split	Question	Ground Truth	SU (LLM)	Proposed Method
Forget	What is the full name of the renowned author who was born in Beirut, Lebanon in 1946 and primarily authored several Medical genre books?	The author's name is Adib Jarrah, whose imaginative and profound medical literature has helped shape the genre	The full name of the esteemed author is Adib Jarrah.	Nadir Habib El-Masry is the renowned author born in Beirut, Lebanon in 1946 who primarily authors medical genre books.
	What is the full name of the LGBTQ+ author born in Baku, Azerbaijan on April 13, 1970?	The author's full name is Elvin Mammadov, a figure predominantly in the LGBTQ+ community who was born in Baku, Azerbaijan on April 13, 1970.	The full name of the author is Elvin Mammadov.	Behrouz Rovshan is the full name of the LGBTQ+ author born in Baku, Azerbaijan on April 13, 1970.
Retain	Are the details of Jaime Vasquez's birth documented?	Yes, Jaime Vasquez was born on the 25th of February in the year 1958.	Yes, Jaime Vasquez was born on the 25th of February in the year 1958.	Jaime Vasquez was born on the 25th of February in the year 1958
	Has Chukwu Akabueze won any awards for his writing?	Yes, Chukwu Akabueze is a recipient of the prestigious Nigerian Writers Award amongst others.	Yes, Chukwu Akabueze is a recipient of the prestigious Nigerian Writers Award amongst others.	Chukwu Akabueze is a recipient of the prestigious Nigerian Writers Award amongst others.
Real Author	Which author is famous for the series 'A Song of Ice and Fire'?	George R.R. Martin	The author who is famous for the series 'A Song of Ice and Fire' is George R. R. Martin.	The author George R. R. Martin is famous for the series 'A Song of Ice and Fire'.
	Who wrote 'The Great Gatsby'?	F. Scott Fitzgerald	The novel 'The Great Gatsby' was written by F. Scott Fitzgerald.	'The Great Gatsby' was written by F. Scott Fitzgerald.
Real World	Where are the ancient Pyramids of Giza located?	Egypt	The ancient Pyramids of Giza are located on the west bank of the Nile River in Giza, Egypt.	The Pyramids of Giza are located on the west bank of the Nile River in Giza, Egypt.
	The ancient city of Machu Picchu is in which country?	Peru	The ancient city of Machu Picchu is located in Peru.	Machu Picchu is located in the country of Peru.

Table 10: Comparison of Methods under Different Prompt/Instruction Settings

Method	Prompt Strict	Inst Strict	Prompt Loose	Inst Loose
Original	33.63	33.21	27.17	38.13
NPO	11.46	22.90	11.46	22.90
SO-NPO	15.16	27.82	17.38	29.62
SPAN-NPO	**25.14**	**37.29**	**31.79**	**33.53**
SPAN-SO-NPO	20.16	33.25	20.38	23.03

E Case Study

Table 9 shows sample outputs of proposed method and SU on the TOFU dataset. Compared to SU, the proposed method more accurately identifies unlearning targets and removes them more explicitly, resulting in superior performance on the forget set.

F Instruction-Following Evaluation

To assess whether unlearning affects the model's ability to handle diverse prompts and instructions, we perform unlearning on an instruction-tuned model and measure its instruction-following performance. Specifically, we use LLaMA-3-8B-Instruct as the backbone and apply each unlearning method on the TOFU dataset using the forget10 split, following the same experimental setup as in the main experiments. We then evaluate the resulting models on the IFEval benchmark (Zhou et al., 2023).

Table 10 presents the results. As shown in the table, all unlearning methods exhibit a noticeable degradation in instruction-following performance compared to the original model. However, the proposed SPAN-based methods show relatively smaller performance drops than the baselines. This suggests that our method, by selectively removing targeted information, better preserves instruction-following capabilities.