GUIR at SemEval-2025 Task 4: Adaptive Weight Tuning with Gradual Negative Matching for LLM Unlearning

Hrishikesh Kulkarni, Nazli Goharian, Ophir Frieder

Information Retrieval Lab
Georgetown University, Washington DC, USA
first@ir.cs.georgetown.edu

Abstract

Machine Unlearning for Large Language Models, referred to as LLM Unlearning is getting more and more attention as a result of regurgitation of sensitive and harmful content. In this paper, we present our method architecture, results, and analysis of our submission to Task4: Unlearning sensitive content from Large Language Models. This task includes three subtasks of LLM Unlearning on 1) Long Synthetic documents, 2) Short Synthetic documents, and 3) Real Training documents. Getting rid of the impact of undesirable and unauthorized responses is the core objective of unlearning. Furthermore, it is expected that unlearning should not have an adverse impact on the usability of the model. In this paper, we provide an approach for LLM unlearning that tries to make the model forget while maintaining usability of the model. We perform adaptive weight tuning with Gradient Ascent, KL minimization and Gradual Negative Matching loss functions. Our submission balances retain and forget abilities of the model while outperforming provided benchmarks.

1 Introduction

The explosion of information with learning mechanisms trying to capture data from every corner raises serious privacy concerns. Comprehensive data privacy laws require commitment to protect sensitive and personal information. The legal mandate in the form of the European Union's General Data Protection Regulation, the California Consumer Privacy Act, raises serious concerns with respect to the results produced by machine learning mechanisms and data sets used for learning. Basically, large language models (LLMs) use large datasets to learn and memorize those data and regurgitate information when asked about it (Carlini et al., 2021). However, that might include very sensitive information, such as personally identifiable information (PII) or harmful information. Regurgitation of such copyrighted or harmful information poses some serious legal and ethical issues that make the use of LLM in practical real-life applications questionable. A variety of methods have been proposed to address LLM limitations (Kulkarni et al., 2023, 2024) but have not addressed regurgitation of sensitive data. One of the rudimentary solutions to handle this issue is retraining the model when any of such outcomes are detected. This could result in retraining the model again and again. This is simply very expensive and impractical when it comes to real life scenarios (Thudi et al., 2022). These practical limitations of re-training further resulted in increasing interest in unlearning LLMs. In short, taking Generative AI to a safe and legal level, demands LLM unlearning.

SemEval-2025 Task 4 of 'Unlearning sensitive content from Large Language Models' deals with LLM unlearning on three types of documents. Subtask 1 is on long form synthetic creative documents covering different genres. Subtask 2 is on short form synthetic biographies that contain personally identifiable information (PII). PII includes names, contact details, SSN, and addresses. Subtask 3 is on real documents sampled from the target model's training dataset. This task releases forget and retain sets along with finetuned LLMs inorder to unlearn.

In our submission, to address the unlearning problem, we introduce adaptive weight tuning with two-stage deviation-based loss functions for unlearning. The approach proposed in this submission performs adaptive weight modifications to losses from specifically chosen set of unlearning loss functions to determine final loss value. We first performs a detailed study of loss functions in the literature to determine the most suited and effective ones for the task of LLM unlearning. We then define a formulation to adjust the weights effectively and tune them with each iteration contributing to the final loss in each iteration. Our submission

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1152–1158

considers Gradient Ascent on forget set, KL divergence on retain set, and most importantly Gradual Negative Matching (GNM) (Kulkarni et al., 2025), which is performed on gradual negative results that are systematically generated to make the model forget the forget set. The use of weighted sum of Gradient Ascent loss, KL divergence loss, and GNM loss in each iteration and adaptive weighting separate our submission from other approaches in the literature and provided benchmarks. This achieves the unique objective where the weights of the loss function on the retain set go on increasing and that of the loss function on the forget set go on decreasing with each iteration. This approach makes sure to free the method from the catastrophic collapse and maintains usability of the model while achieving the unlearning objective. We compare the results with the benchmarks provided by the task organizers and show that our submission outperforms them.

Our contributions are as follows.

- We present a detailed study of loss functions used for LLM unlearning.
- We present adaptive weights formulation for Gradient Ascent, KL divergence, and GNM for LLM unlearning.
- We report results of our submission which outperform the provided benchmarks.
- We also provide analysis of results and pointers for further improvements.

2 Related Work

The efforts to make LLM more applicable resulted in increased research efforts in the area of LLM unlearning. This is mainly to address concerns related to trustworthiness (Lu et al., 2022), fairness, copyright, and privacy (Yu et al., 2023; Eldan and Russinovich, 2023), and sensitive knowledge. Previous unlearning approaches used mainly alignment techniques to achieve the unlearning objective (Liu et al., 2024). Such techniques aim to deliver expected results for specific inputs. Gradient Ascent is basically the retrogression of Gradient Descent learning (Jang et al., 2023). But it results in poor retention. This led to efforts to improve retainability. With this objective, the Gradient Ascent approach is combined with methods that could provide higher retention. Rather than simply applying Gradient Ascent, use of Gradient Ascent on the forget set is combined with Gradient Descent on retain set (Liu et al., 2022). The Gradient Ascent-based unlearning comes with the challenge of catastrophic collapse resulting in serious harm to model usability (Liu et al., 2024). This seriously limits the usability of this approach. To counter this issue, other approaches are proposed to use on retain set. One of such approaches is the use of the KL divergence term on retain set (Yao et al., 2023). The objective was to make the model forget what is expected to forget and retain what is already learned useful information. This led to efforts where multiple combinations of Gradient Ascent, Gradient Descent, and KL divergence were used (Chen and Yang, 2023). Further, in order to obtain the best retain-forget tradeoffs Gradual Negative Matching (GNM) was proposed (Kulkarni et al., 2025). GNM is a two-stage approach where the first stage involves generation of gradual negative outputs for forget set inputs. While the second stage matches these input, generated output pairs through Gradient Descent.

Additionally, in LLM unlearning it is necessary to have a relevant benchmark and proper mechanisms for evaluations. Furthermore, the benchmark needs to be specific to the application and context. This led to different benchmarks. Some of such popular benchmarks include harmful content (Ji et al., 2023), copyrighted books (Eldan and Russinovich, 2023), biographies (Maini et al., 2024), PII, and creative documents (SemEval'25-Task4). A FR-rouge based evaluation is proposed to measure the effectiveness of unlearning models (Kulkarni et al., 2025).

In general, having sensible handling of sensitive information where the legal and ethical violations by LLM regurgitation of sensitive information could be avoided poses a need for better LLM unlearning methods. With this need in focus, in this submission, we propose a method which comprises of Gradient Ascent, KL divergence and Gradual Negative Matching loss functions along with adaptive weight tuning.

3 Data and Setting

The SemEval 2025 task 4 of 'Unlearning sensitive content from Large Language Models' (SemEval'25-Task4; Ramakrishna et al., 2025a,b) has released forget and retain sets for both question-answering and sentence completion across the three subtasks of synthetic long, synthetic short

and real training documents. They have also released the train and validation parts of this data. Additionally, they also provided fine-tuned open source LLM OLMo-7B-0724-Instruct-hf (Groeneveld et al., 2024), trained to memorize documents from three document types. The first one contains long synthetic creative documents. They include different genres. While the second one has short synthetic biographies where fake personal information is present. This includes PII with fake names, phone number, social security numbers, email and home addresses. The third type is a sample of real documents used for training the original LLM.

Evaluation is performed across three metrics namely: Task-specific regurgitation rates measured using rouge-L and exact matching scores, membership inference attack (MIA) score and model performance on MMLU benchmark. Further, a threshold of 75% of pre-unlearning checkpoint is placed on MMLU score to maintain a minimum model utility. Finally, using arithmetic mean, a final aggregate score is calculated from the above three metrics. For calculating task-specific regurgitation rate, rouge-L scores are evaluated for sentence completion and exact matching scores are evaluated for question answering on both forget and retain sets across the three subtasks. The final task-aggregate is determined by considering harmonic mean of the six retain set scores and six inverted (1-score) forget set scores. For calculating MIA scores a sample of member and non-member data is released. Final MIA score is defined as $1-|\text{mia loss auc score}-0.5|*2$. The MMLU score is calculated on the MMLU benchmark consisting multiple choice questions on 57 STEM subjects. The objective is to develop an unlearning method that effectively unlearns information in the Forget set without affecting model usability i.e. with minimal model degradation.

4 Loss Function Components

Gradient Descent is the most common choice for fine-tuning LLMs. Gradient Descent as shown in Equation 1 is based on cross-entropy loss summed over all tokens in the output sequence. Here, we note that the cross entropy is calculated only for tokens in output y and not for input x.

$$L_{GD}(S,\theta) = \sum_{(x,y)\in S}\sum_{i=1}^{|y|} CE(f_\theta(x, y_{<i}), y_i) \quad (1)$$

4.1 Gradienst Ascent

Gradient Ascent tries to reverse the effects of Gradient Descent on the LLM by negating the calculated loss before back-propagation as shown in Equation 2.

$$L_{GA} = -L_{GD} \quad (2)$$

The Gradient Ascent loss is calculated on the forget set input output pairs as it is to be unlearned. This is evident in Equation 3 where θ is the LLM being unlearned. FS is the forget set and RS is the retain set.

$$Loss = L_{GA}(FS,\theta) \quad (3)$$

4.2 Gradient Difference

Gradient Difference (see Equation 4) considers both Gradient Ascent on forget set and Gradient Descent on retain set in order to unlearn forget set content while maintaining model usability on the retain set.

$$Loss = L_{GA}(FS,\theta) + L_{GD}(RS,\theta) \quad (4)$$

4.3 KL Divergence

KL Divergence between two LLMs θ' and θ is calculated using the output probability distributions of the two models for given inputs as shown in Equation 5. This KL divergence is minimized inorder to make θ outputs more like those of θ' for the same input.

$$L_{KL}(S,\theta',\theta) = \sum_{(x,y)\in S}\sum_{i=1}^{|y|} KL(f_{\theta'}(x, y_{<i})||f_\theta(x, y_{<i})) \quad (5)$$

4.4 Other Combinations

Prior research efforts have underlined the effectiveness of unlearning by combining the above loss functions. Chen and Yang proposed combining Gradient Ascent on forget set, Gradient Descent on retain set and KL minimization on both forget and retain sets. The final loss is a weighted sum of above terms as shown in Equation 6. Here, ω_j denotes the weight values for respective loss functions.

$$\begin{aligned} Loss = \omega_1 \cdot L_{GA}(FS,\theta) - \omega_2 \cdot L_{KL}(FS,\theta',\theta) \\ + \omega_3 \cdot L_{GD}(RS,\theta) + \omega_4 \cdot L_{KL}(RS,\theta',\theta) \end{aligned} \quad (6)$$

Another approach was proposed by Yao et al. which introduced Random Matching. Along with

Algorithm	Aggregate	Task Aggregate	MIA Score	MMLU Average
~~Gradient Ascent~~	~~0.394~~	~~0.000~~	~~0.912~~	~~0.269~~
Gradient Difference	0.243	0.000	0.382	0.348
~~KL Minimization~~	~~0.395~~	~~0.000~~	~~0.916~~	~~0.269~~
Negative Pref. Optimization	0.188	0.021	0.080	0.463
Our submission (val)	0.267	0.429	0.000	0.373
Our submission (test)	**0.308**	**0.433**	0.000	**0.492**

Table 1: Results of the provided benchmarks and our submission across the metrics of Task Aggregate, MIA score and MMLU Average score. We note that our submission leads to the best overall Aggregate score when compared to the baselines. The ~~striked~~ methods lead to MMLU Average scores below the threshold (75% of the pre-unlearning benchmark) and are hence disqualified.

Gradient Ascent on forget set and KL minimization on retain set they also perform Gradient Descent on forget input and randomly matched output pairs. This is shown in Equation 7 where Y^{rdn} is a set of random outputs from retain set. For slower progress towards catastrophic collapse minimizing Negative Preference Optimization loss was also proposed (Zhang et al., 2024).

$$Loss = \omega_1 \cdot L_{GA}(FS, \theta) + \omega_2 \cdot L_{KL}(RS, \theta', \theta) + \omega_3 \cdot \sum_{(x,)\in F} \frac{1}{|Y^{rdn}|} \sum_{y \in Y^{rdn}} \sum_{i=1}^{|y|} CE(f_\theta(x, y_{<i}), y_i) \tag{7}$$

4.5 Gradual Negative Matching

Gradual Negative Matching is a two stage approach which first involves generating gradual negative outputs and then match these outputs to respective forget set inputs. This matching is performed along with Gradient Ascent on forget set and KL minimization on retain set as shown in Equation 8. Here, $Y^{gn}[x]$ denote the gradual negative outputs generated with respect to the input x. The $Loss_{GNM}(x, Y^{gn})$ term in Equation 8 is defined in Equation 9. For example, a question requesting a person's email address would be matched with similar but gradually different email addresses.

$$Loss = \omega_1 \cdot L_{GA}(FS, \theta) + \omega_2 \cdot L_{KL}(RS, \theta', \theta) + \omega_3 \cdot \sum_{(x,)\in F} \frac{1}{|Y^{gn}[x]|} Loss_{GNM}(x, Y^{gn}) \tag{8}$$

$$Loss_{GNM}(x, Y^{gn}) = \sum_{y \in Y^{gn}[x]} \sum_{i=1}^{|y|} CE(f_\theta(x, y_{<i}), y_i) \tag{9}$$

5 Adaptive Weight Tuning

We consider a weighted sum of Gradient Ascent loss, KL divergence loss and GNM loss for each iteration as shown in Equation 8. We perform adaptive weight tuning by making each weight value a function of the number of iterations depending upon the type of loss. We want the weight of loss functions on the retain set to go on increasing while the weight of loss functions on the forget set to go on decreasing with increasing number of iterations. This ensures an increased focus on model usability in the latter iterations after unlearning is performed to some extent. We determine ω_1 and ω_3 using Equation 10 where we go on decreasing the weight with increasing number of iterations i. On the other hand, we determine ω_2 using Equation 11 where we go on increasing the weight with increasing number of iterations i. Here k is the tuning constant and α and β are initial weight values.

$$\omega_1 = \alpha - \frac{i}{k} \tag{10}$$

$$\omega_2 = \beta + \frac{i}{k} \tag{11}$$

6 Experiments

We perform unlearning and save model at every 100 iterations to understand the forget and retain set regurgitation with respect to unlearning iterations. We use the provided validation set to determine the optimal number of iterations in the final submission. We also determine input output lengths of the forget and retain sets in the unlearning data to analyze the results further with respect to input length. We plot the regurgitation rougeL scores for each subtask for both question answering and sentence completion to analyze subtask based regurgitation. Gradient Ascent and KL Minimization are compo-

Left: Question Answering, Right: Sentence Completion

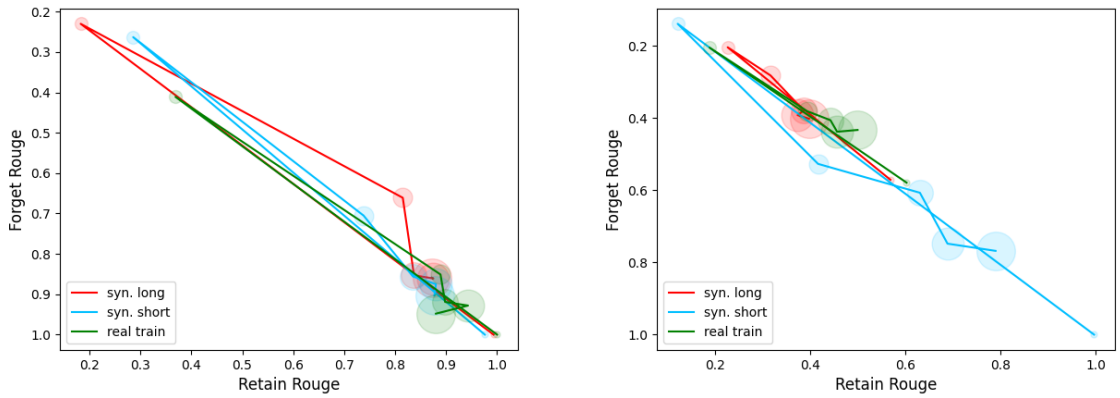

Figure 1: Forget and Retain RougeL performance of our submission on the validation set with increasing number of training batches. Points are plotted after every 100 batches up to 600. Increasing size of points denotes a higher number of batches. In the above plots we can observe a trade-off between Forget and Retain performance. Further, we also note that our submission performs well in all three subtasks namely: synthetic long, synthetic short and real training documents shown by different colors. Most importantly, our submission does not lead to catastrophic collapse of the model during unlearning.

nents of our submission and their respective results represent the ablation study.

7 Results and Analysis

As evident in Table 1, our submission on the test set leads to a Task Aggregate score of 0.433 and MMLU Average score of 0.492 making the overall Aggregate score of 0.308. While on the validation set it leads to a Task Aggregate score of 0.429 and MMLU Average score of 0.373 making the overall Aggregate score 0.267. On the other hand, the provided benchmark of Gradient Difference leads to MIA score of 0.382 and MMLU Average score of 0.348 resulting in an overall Aggregate score of 0.243. Also, Negative Preference Optimization leads to Task Aggregate score of 0.021, MIA score of 0.080 and MMLU Average score of 0.463 resulting in an overall Aggregate score of 0.188. Hence, we can infer that our submission clearly outperforms Gradient Difference and Negative Preference Optimization benchmarks.

On the other hand, Gradient Ascent leads to MIA score of 0.912 and MMLU Average score 0.269, resulting in an overall Aggregate of 0.394. And, KL Minimization leads to MIA score of 0.916 and MMLU Average score of 0.269 resulting in an overall Aggregate of 0.188. Here, we note that both Gradient Ascent and KL Minimization lead to an MMLU Average of 0.269 which is below 75% of pre-unlearning model MMLU threshold (0.371). Hence, these models are discarded because of the highly degraded utility of the unlearned model.

We note a trade-off between Task Aggregate and MIA scores, and decide to prioritize Task Aggregate giving importance to addressing regurgitation and maximizing final Aggregate score. Further analysis as evident in Figure 1 show that our submission does not lead to catastrophic collapse i.e. complete degradation of model. It is also evident that our submission results in significant unlearning of the model but at the same time a trade-off with retain effectiveness is observed. We also note that our proposed approach leads to better performance than the benchmarks specifically on the input output pairs of shorter length i.e. question answering and short synthetic documents.

8 Conclusion

LLM unlearning has gained importance due to GDPR and CCPA laws in Europe and United States respectively. Further, regurgitation of sensitive and harmful content is a violation of ethics and moral values. In the paper, we present a detailed architecture of our submission to the SemEval 2025 Task4: 'Unlearning sensitive content from Large Language Models.' We show that our submission outperforms the provided benchmarks. Further, we also show that our submission performs unlearning while maintaining above threshold model usability while avoiding catastrophic collapse.

References

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

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, Singapore. Association for Computational Linguistics.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *Preprint*, arXiv:2310.02238.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

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.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Preprint*, arXiv:2307.04657.

Hrishikesh Kulkarni, Nazli Goharian, and Ophir Frieder. 2025. Gradual negative matching for llm unlearning. In *47th European Conference on Information Retrieval*.

Hrishikesh Kulkarni, Sean MacAvaney, Nazli Goharian, and Ophir Frieder. 2024. Genetic approach to mitigate hallucination in generative ir. In *The Second Workshop on Generative Information Retrieval collocated with SIGIR*.

Hrishikesh Kulkarni et al. 2023. Genetic generative information retrieval. In *Proceedings of the ACM Symposium on Document Engineering 2023*, DocEng '23.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pages 243–254. PMLR.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2024. Rethinking machine unlearning for large language models. *Preprint*, arXiv:2402.08787.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: controllable text generation with reinforced [un]learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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

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 arXiv:2504.02883*.

SemEval'25-Task4. Semeval challenge 2025, task 4: Unlearning sensitive content from large language models. `https://llmunlearningsemeval2025.github.io/`. Accessed: 2024-10-24.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. *Preprint*, arXiv:2109.13398.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. In *Socially Responsible Language Modelling Research Workshop, NeurIPS*.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada. Association for Computational Linguistics.

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