iShumei-Chinchunmei at SemEval-2025 Task 4: A Multi-Task Unlearning Approach Integrating Data Augmentation and Gradient Interference

Yujian Sun[1], **Tian Li**[2]
[1]Shumei AI Research Institute, Beijing, China
[2]School of Computing, Newcastle University, Newcastle upon Tyne, UK
sunyujian@ishumei.com
t.li56@newcastle.ac.uk

Abstract

This paper presents our solution for SemEval-2025 Task 4, introducing a supervised fine-tuning method for unlearning sensitive data in large language models (LLMs) under resource constraints. Our approach combines multi-task learning, gradient interference, and data augmentation, achieving promising results across several tasks. Additionally, through extensive ablation experiments and detailed result analysis, we identify potential issues that arise when relying solely on supervised fine-tuning for unlearning sensitive data.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance in natural language understanding and generation. However, as they are trained on vast amounts of data, they may unintentionally retain and regurgitate sensitive information, posing serious concerns regarding privacy and compliance. SemEval-2025 Task 4: "Unlearning Sensitive Content from Large Language Models" ((Ramakrishna et al., 2025)) seeks to address the lack of a robust evaluation framework for assessing the accuracy of unlearning strategies by providing a comprehensive evaluation challenge. This task is crucial for ensuring the ethical deployment of AI and compliance with privacy regulations. Conducted in English, the evaluation will consider metrics such as task-specific regurgitation rates, Membership Inference Attack (MIA) scores, and MMLU benchmark performance, to measure how effectively the model forgets specified content while retaining its broader language understanding and reasoning abilities.

Our system adopts a multi-task learning approach combined with gradient interference techniques and data augmentation to enhance the unlearning process. Specifically, we introduce multiple training objectives to regulate how the model handles retained and forgotten data. For data that must be retained, we apply Supervised Fine-Tuning (SFT) loss to ensure the model preserves its general knowledge. For data that needs to be forgotten, we implement gradient interference, inspired by the Gradient Ascent loss function. Additionally, we experiment with replacing the outputs of forgotten data with negative responses resembling "I don't know." (Choi et al. (2024);Shi et al. (2024)) To improve the model's robustness across varying input lengths and contexts, we apply data augmentation techniques, such as breaking answers into sentences and recombining them to generate diverse training instances. This strategy ensures that the model effectively learns which content to forget while maintaining its performance on non-sensitive data.

Our system ranked fifth in SemEval-2025 Task 4 based on the 7B model. Experimental results show that our approach is capable of achieving selective forgetting while maintaining strong overall performance and demonstrating high usability. However, experiments with the 1B model revealed that our method is sensitive to both model parameters and scale. Additionally, we observed that when handling diverse types of data, particularly in balancing the forgetting effectiveness between short and long texts, there is some variation in the effectiveness of forgetting. Furthermore, through ablation studies, we thoroughly examined the role of the loss functions and data augmentation strategies, further validating their importance in achieving controlled forgetting. Our research provides new insights into privacy protection for large language models, proposing a simple and user-friendly method for selective forgetting and offering experimental evidence for the future development of machine unlearning techniques. Detailed code and experimental records are available at the following link: https://github.com/yizhiai1994/CCM_at_semeval2025task4.

2 Background

Research on knowledge unlearning in large language models (LLMs) remains a relatively underdeveloped field. One of the most widely used methods is fine-tuning-based unlearning, where the model is further trained on datasets containing specific target knowledge, thereby weakening its memory of that knowledge and effectively eliminating unwanted information.

Specifically, an intuitive approach involves applying gradient ascent on the forget data, which directly targets increasing the loss on that data to force the model to forget the specified knowledge.

However, the work in Wang et al. (2024) points out that directly applying gradient ascent on the forget data often leads to optimization instability and a significant decline in model performance.

To address this issue, Veldanda et al. (2024) propose a comprehensive training approach, which includes performing reverse gradient updates on the unlearning dataset, performing standard gradient descent on the update dataset, and minimizing the KL divergence between the outputs of the model and the original model on the retain dataset to preserve the model's performance on retained knowledge.

Similarly, Jang et al. (2022) mitigates the instability caused by direct gradient ascent by designing a sequence for gradual gradient ascent, thereby achieving a more stable unlearning process.

In addition to gradient-based strategies, another common approach to knowledge unlearning involves covering or replacing the knowledge. For example, Choi et al. (2024) and Shi et al. (2024) replace the answers to the data that should be forgotten with negative responses, such as "I don't know," providing the model with negative instructions to weaken its knowledge of the target information.

Similarly, Eldan and Russinovich (2023) trains a reinforcement learning model to identify key phrases in the forget data and then replace these key phrases before fine-tuning the model. This approach is effective but requires considerable computational resources.

Notably, Mekala et al. (2024) emphasizes that relying solely on negative feedback to suppress responses related to the forget set often results in nonsensical or inconsistent outputs, diminishing the model's utility and potentially introducing new privacy risks.

Taking the aforementioned considerations into account, we present a pragmatic approach to unlearning in large language models (LLMs). This method employs multi-task learning, data augmentation, and gradient interference to facilitate the rapid and efficient forgetting of knowledge by LLMs, even under constrained time and computational resources.

3 System Overview

In this task, we employ a LoRA approach(Hu et al. (2022)) to fine-tune a large language model, building a system that incorporates a data augmentation module and a multi-task learning module. We do not use any external corpora beyond the organization released training data, and we use the SemEval-2025 Task 4 test-set published prior to February 20, 2025. All offline experiments are conducted on a single NVIDIA A100(40 GB) GPU, and each training session is completed within one hour. Figure 1(left) illustrates the overall training procedure of our system. In what follows, we provide detailed descriptions of the data augmentation module in Section 3.1 and the multi-task learning module in Section 3.2.

3.1 Data Augmentation Module

In our submitted version for the competition, we adopted a relatively simple text replacement strategy: for the output portion of the forget set, we replaced it exclusively with short fixed negative phrases such as "I don't know." (Choi et al. (2024);Shi et al. (2024)) This method yielded significantly different results for the 7B-parameter model and the 1B-parameter model; further analysis indicates that it is highly sensitive to text length and struggles to handle both long and short texts effectively. Specifically, if the original output is short, then after training, the model outputs a fixed negative phrase akin to "I don't know", whereas for long outputs, the generated text remains essentially unchanged. Hence, while this initial version performed relatively well on the 7B model, it was actually because it failed to effectively intervene on longer inputs, resulting in "artificially good" scores in specific scenarios but poor performance for smaller models.

To address this issue, we propose a segmentation strategy tailored for long texts. Concretely, for lengthy outputs, we first split them according to punctuation, forming multiple segments from short to long. We then apply data augmentation to each

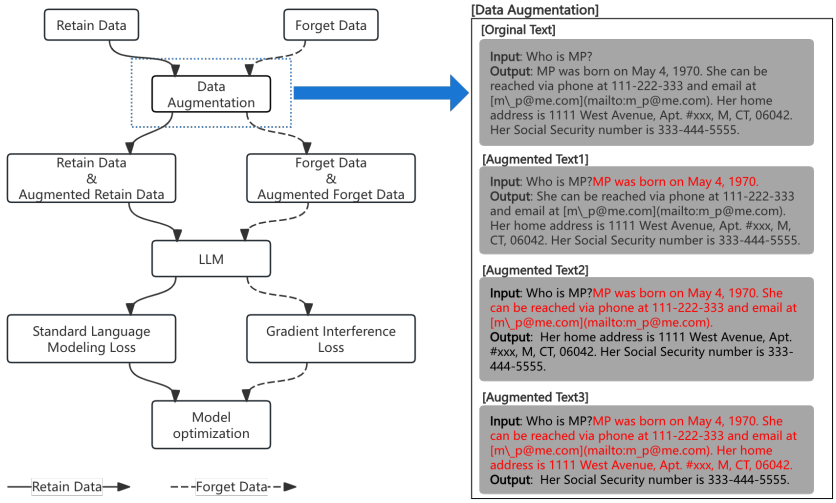

Figure 1: A minimal working example to demonstrate how to place two images side-by-side.

segment accordingly. Figure 1(right) presents a detailed illustration of the segmentation process for long texts. Compared to the initial blanket replacement, this segmentation strategy better preserves semantic coherence across different paragraphs or sentences, thereby improving adaptability to diverse text lengths.

3.2 Multi-Task Learning Module

This task demands a balance between the forget set (where information must be erased or distorted) and the retain set (where the original knowledge must be preserved), thus requiring both "effective forgetting" and "information retention." Through comparative analyses of the forget and retain sets, we found that they overlap significantly in terms of vocabulary and topics, yet require starkly different predictive objectives. Consequently, we adopt a Multi-Task Learning (MTL) paradigm that unifies both sets within a single model framework. We define two sub-tasks:

Retain-Set Task We use the standard language modeling loss (next-token prediction), L_{ntp}, to ensure that the model accurately generates content from the retain set. Formally, for any sample (x_{input}, y_{output}) in the retain set, the loss can be summarized as:

$$L_{reatin} = L_{ntp}(x_{input}, y_{output}) \quad (1)$$

Retain-Set Task To achieve a "reversal" or removal of knowledge in the forget set, we propose a "gradient interference" loss inspired by the idea of Gradient Ascent:

$$L_{forget} = \alpha \times \frac{1}{L_{ntp}(x_{input}, y_{output})} \quad (2)$$

where $L_{ntp}(x_{input}, y_{output})$ denotes the language modeling loss when the model uses the original output y_{output} of the forget set as the prediction target, and α is a scaling factor that regulates the magnitude of the loss. Unlike directly performing gradient ascent, we apply an inverse relationship of L_{ntp}. When the model's predictions closely resemble the original output in the forget set, the loss becomes very large, inducing substantial gradient adjustments; conversely, when the model generates outputs that deviate significantly from the original forgotten text, the loss rapidly diminishes, preventing the model from diverging or failing to converge. To ensure these two losses do not interfere with each other, we only include one type of loss data per training batch.

4 Experiment setup

All experiments in this paper were conducted using the dataset provided by Semeval2025 Task4((Ramakrishna et al., 2025)), which was released during the competition. The reported results are based on the validation split that was publicly available at that time. For the leaderboard outcomes, we set the number of training epochs to 3 and the learning rate to 1e-4.The relevant techniques utilized include negative responses (Choi et al., 2024; Shi et al., 2024), gradient interference, and supervised fine-tuning (SFT) on the retained

EP	LR	OD	IDK	DA	GI	MIA Score	Task Aggregate	MMLU	Final Score
3	1.00E-04	O	×	O	O	0.135	0.245	0.272	0.217
3	1.00E-05	O	×	O	O	0.000	0.092	0.280	0.124
3	1.00E-06	O	×	O	O	0.000	0.092	0.275	0.122
4	1.00E-04	O	×	O	O	0.215	0.278	0.270	0.254
4	1.00E-05	O	×	O	O	0.000	0.112	0.280	0.131
4	1.00E-06	O	×	O	O	0.000	0.092	0.276	0.122
5	1.00E-04	O	×	O	O	0.593	0.395	0.275	0.421
5	1.00E-05	O	×	O	O	0.001	0.112	0.279	0.131
5	1.00E-06	O	×	O	O	0.000	0.092	0.277	0.123

Table 1: Parameter Sensitivity.

EP	LR	RD	NR	DA	GI	MIA Score	Task Aggregate	MMLU	Final Score
5	1.00E-04	×	O	×	×	0.000	0.092	0.281	0.124
5	1.00E-04	×	O	O	×				
5	1.00E-04	O	O	×	×	0.000	0.124	0.283	0.135
5	1.00E-04	O	O	O	×				
5	1.00E-04	×	×	×	O	0.993	0.408	0.229	0.543
5	1.00E-04	×	×	O	O	0.989	0.421	0.229	0.547
5	1.00E-04	O	×	×	O	0.009	0.185	0.278	0.157
5	1.00E-04	O	×	O	O	0.593	0.395	0.275	0.421

Table 2: The usefulness of Gradient Interference and Fine-tune Retain Data

data. The final evaluation score is computed as the average of three components:

Task Aggregate Score This method is used to measure the model's completion of various tasks.

Membership Inference Attack (MIA) This metric evaluates the extent to which the relevant knowledge is preserved or effectively forgotten.

MMLU Score This serves as a measure of whether the model's overall linguistic capability suffers any degradation after the unlearning process.

5 Result

5.1 Main Result

Our submitted method ranked fifth in the 7B model track and nineteenth in the 1B model track. Upon observing this substantial performance discrepancy between models of different scales, we recognized a potential issue in our approach. Further analysis revealed that our method is highly sensitive to model size. Moreover, we found that it is also sensitive to the length of the model's original response: as noted earlier, when the original responses were relatively short, the trained model tended to produce answers such as "I don't know", whereas for longer original outputs, it exhibited almost no change. This behavior is the primary cause of the method's overall ineffectiveness (see Appendix B for detailed examples).

Consequently, we revised our entire system as described in Section 3 and conducted an extensive ablation study. All results from the ablation experiments are provided in Appendix A, and the corresponding code repository includes full inference logs for each sample.

5.2 Ablation Study

For clarity, we have consolidated the results from all ablation experiments in Appendix A, presenting only the findings in the main text that inform our relevant conclusions. In the tables, GI denotes gradient interference, NR indicates that the original outputs for the forgotten data were replaced with negative responses, DA signifies the inclusion of data augmentation during training, EP represents the number of epochs, and LR denotes the learning rate. In Table 3, the parameters for the "Old System" are set to 3 epochs and a learning rate of 1e-4. Meanwhile, both the "New System" configuration in Table 3 and all results shown in Tables 4 and 5 utilize 5 epochs and the same learning rate of 1e-4.

5.2.1 Parameter Sensitivity

Table 1 presents the results of our new method under various parameter configurations. It is evident that the method is highly sensitive to these parameters and exhibits progressively improved performance as the training period is extended.

5.2.2 Old Stsyem Vs New System

Table 3 compares the methods and parameters that we used during the competition with the best results obtained by our new approach. The data indicate that the new method substantially outperforms the old one and, based on the MMLU results, does so without causing a significant degradation in the overall linguistic capabilities of the model. Furthermore, as shown in Appendix A, the old method remains nearly unaffected by changes in hyperparameters, underscoring the superior performance and adaptability of our newly designed unlearning framework.

	MIA	Task	MMLU	Final
Old System	0	0.1	0.28	0.12
New System	0.59	0.39	0.26	0.42

Table 3: Old Stsyem VS New System.

5.2.3 Gradient Interference & Fine-tune Retain Data

Table 4 indicates that fine-tune retain data is indispensable, regardless of whether gradient interference or negative response replacement is employed. By examining the outputs of both methods, we find that omitting fine-tune retain data causes the model to become uncontrollable, leading it to repeatedly generate the same words. This issue is also reflected in the MMLU scores, which decrease substantially when the data retention is excluded, suggesting a severe deterioration in the model's language ability. In addition, a comparison between gradient interference and negative response replacement shows that incorporating gradient interference significantly improves the forgetting effect while preserving the model's linguistic capacity; however, introducing negative replacement leads to an overall decline in performance.

5.2.4 Data Augmentation

Furthermore, Table 4 shows that incorporating data augmentation further enhances the system's unlearning capabilities. An analysis of the model's outputs reveals that, prior to data augmentation, there was virtually no change when encountering long responses that required forgetting. However, once data augmentation was employed, the method effectively addressed such cases, leading to marked improvements in unlearning performance.

	MIA	Task	MMLU	Final
RF	0.00	0.09	0.28	0.12
RF&RD	0.00	0.12	0.28	0.14
GI	0.99	0.41	0.23	0.54
GI&RD	0.01	0.19	0.28	0.16
GI&RD&GA	0.59	0.39	0.28	0.42
RF&GI&RD&GA	0.02	0.23	0.28	0.18

Table 4: The effects of each part of the new system.

5.2.5 Amplify Gradient Interference

To further examine the impact of our gradient interference approach, we squared its gradient signals to produce more extreme upper and lower bounds for the loss variation. However, as demonstrated in Table 5, this intensification did not yield any noticeable performance improvement.

	MIA	Task	MMLU	Final
GI	0.59	0.39	0.28	0.42
GI^2	0.39	0.22	0.28	0.29

Table 5: The effects of each part of the new system.

6 Conclusion

References

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2024. Snap: Unlearning selective knowledge in large language models with negative instructions. *arXiv preprint arXiv:2406.12329*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. 2024. Alternate preference optimization for unlearning factual knowledge in large language models. *arXiv preprint arXiv:2409.13474*.

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

Shaojie Shi, Xiaoyu Tan, Xihe Qiu, Chao Qu, Kexin Nie, Yuan Cheng, Wei Chu, Xu Yinghui, and Yuan Qi. 2024. Ulmr: Unlearning large language models via negative response and model parameter average. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 755–762.

Akshaj Kumar Veldanda, Shi-Xiong Zhang, Anirban Das, Supriyo Chakraborty, Stephen Rawls, Sambit Sahu, and Milind Naphade. 2024. Llm surgery: Efficient knowledge unlearning and editing in large language models. *arXiv preprint arXiv:2409.13054*.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. 2024. Llm unlearning via loss adjustment with only forget data. *arXiv preprint arXiv:2410.11143*.

A Appendix: Ablation Study

This is an appendix.

B Appendix: Ablation Study

This is an appendix.