NLPART at SemEval-2025 Task 4: Forgetting is harder than Learning

Hoorieh Sabzevari, Milad Molazadeh, Tohid Abedini, Ghazal Zamaninezhad,
Sara Baruni, Zahra Amirmahani, Amirmohammad Salehoof
PART AI Research Center
hoorieh.sabzevari@partdp.ai

Abstract

Unlearning is a critical capability for ensuring privacy, security, and compliance in AI systems, enabling models to forget specific data while retaining overall performance. In this work, we participated in Task 4 of SemEval 2025, which focused on unlearning across three sub-tasks: (1) long-form synthetic creative documents, (2) short-form synthetic biographies containing personally identifiable information, and (3) real documents sampled from the target model's training dataset. We conducted four experiments, employing Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO). Despite achieving good performance on the retain set—data that the model was supposed to remember—our findings demonstrate that these techniques did not perform well on the forget set, where unlearning was required.

1 Introduction

As machine learning (ML) continues to be integrated into critical domains, concerns over data privacy, security, and user autonomy have become more pressing than ever. From healthcare to finance, data-driven models play a crucial role, but their ability to retain and utilize information raises significant challenges. This issue has been further emphasized by legal regulations such as the General Data Protection Regulation (GDPR), which grants individuals the "right to be forgotten" (Voigt and von dem Bussche, 2017).

Machine unlearning has emerged as a key approach to tackling these concerns. It enables the selective removal of specific data points from a trained model without necessitating a full retraining process (Cao and Yang, 2015a). Unlike fine-tuning or knowledge editing, which focus on adjusting or adding new information, machine unlearning is designed to eliminate the influence of certain inputs, ensuring they no longer contribute to the model's predictions or behavior. The overall workflow of machine unlearning is illustrated in Figure 1.

The importance of machine unlearning is magnified by the intricate nature of deep learning models. These models often exhibit *data entanglement*, where information from different training instances becomes deeply intertwined within the model's parameters, making selective removal a challenging task (Tramèr et al., 2022). Additionally, unlearning must be carefully implemented to prevent unnecessary degradation of the model's performance on retained data, striking a balance between utility and privacy (Guo et al., 2019). Another significant hurdle is efficiency—particularly in large-scale architectures like LLMs—since retraining a model from scratch is often impractical due to the immense computational cost.

As large language models gain increasing prominence, it becomes essential to examine the relationship between machine unlearning and *knowledge editing*. Knowledge editing is typically used to update or correct specific facts or behaviors in LLMs without requiring a full model retrain (Yao et al., 2023; Wang et al., 2025; Mitchell et al., 2022). While both techniques modify a model's internal representations, they serve distinct purposes: knowledge editing injects or refines information, whereas unlearning aims to remove specific influences entirely. This distinction highlights both the challenges and the potential areas of overlap between these two approaches.

Beyond its technical implications, machine unlearning also plays a crucial role in ensuring ethical AI practices and regulatory compliance. By enabling models to forget specific information when required, unlearning enhances both user privacy and model transparency, making it an essential tool in the evolving landscape of responsible AI.

In this study, we explore machine unlearning in large language models by participating in SemEval-2025 Task 4. We focus on selectively remov-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2336–2341
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

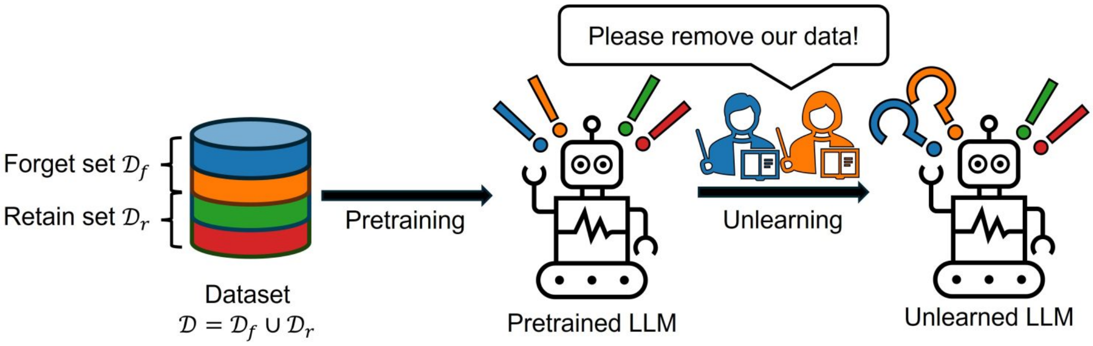

Please remove our data!
Forget set $\mathcal{D}_f$
Retain set $\mathcal{D}_r$
Dataset
$\mathcal{D} = \mathcal{D}_f \cup \mathcal{D}_r$
Pretraining
Pretrained LLM
Unlearning
Unlearned LLM

Figure 1: Workflow of machine unlearning.

ing memorized information across three document types: synthetic creative documents, synthetic biographies containing personally identifiable information (PII), and real-world documents, without compromising the model's general performance. For this, we conduct four experiments using Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) techniques. Our findings reveal the challenges of achieving effective unlearning, emphasizing that current approaches, while promising for retention, still fall short in forgetting sensitive data. Through this work, we aim to contribute insights into the limitations and future directions for improving machine unlearning in large-scale AI systems. [1]

2 Background

In recent years, machine unlearning has attracted significant attention as concerns about data privacy, regulatory compliance, and ethical AI practices have grown. The fundamental concept of machine unlearning is to allow trained models to remove specific data points without the need for complete retraining, thereby achieving a balance between efficiency and privacy demands. This section examines the leading approaches and methodologies presented in the literature.

2.1 The Rationale for Machine Unlearning

Users may want to delete their data for various reasons, mainly for security and privacy concerns. Each reason is discussed further below.

Privacy, Several approaches have been proposed to mitigate privacy risks in LLMs. Differential privacy (Dwork, 2006) introduces noise to training data to prevent individual data points from being memorized. Federated learning (McMahan et al., 2017) minimizes direct data exposure by training models on decentralized data. However, these techniques do not fully address the problem of post-hoc data removal, which is where unlearning methods become crucial (Bourtoule et al., 2021). **Security**, Adversarial attacks generate data nearly identical to real data, tricking deep learning models into incorrect predictions. In critical fields like healthcare, this can lead to misdiagnoses or harmful treatments. Detecting and removing such data is crucial for security, and machine unlearning must eliminate detected attacks (Cao and Yang, 2015b).

2.2 Unlearning Methods

Several unlearning methods have been developed to address the challenges of removing specific knowledge influences from trained models while maintaining overall performance. These methods include:

- **Gradient Ascent:**This approach builds upon the concept of gradient ascent by aiming to maximize the loss on the forget set (Trippa et al., 2024).
- **Gradient Difference:** This method, expands on the idea of gradient ascent. It seeks to increase the loss on the forget set, while simultaneously preserving performance on the retain set (Liu et al., 2022).
- **KL Minimization** In the KL Minimization approach, the goal is to balance two objectives during the unlearning process of a model: (1) Minimize the Kullback-Leibler (KL) divergence: This ensures that the predictions of the unlearned model on the sensitive data (SR) remain close to those of the original model fine-tuned on the original data. (2) Maximize

[1] Additional details and code are available on our GitHub repository.

the conventional loss on the safe data (SF): This encourages the model to perform well on non-sensitive data (Maini et al., 2024).

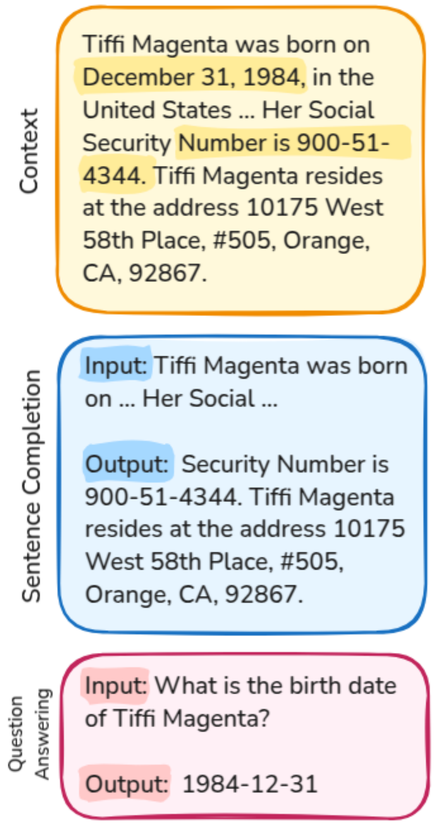
- **Negative Preference Optimization** It addresses the limitations of existing gradient ascent-based methods and demonstrates that NPO-based methods outperform other approaches, providing a superior balance between unlearning effectiveness and model utility. The evaluation is conducted on the Task of Fictitious Unlearning (TOFU) dataset, and the paper concludes with a discussion of model utility and forget quality (Zhang et al., 2024).
- **Preference Optimization** Inspired by the concept of Negative Preference Optimization (NPO) (Rafailov et al., 2023). The goal is to ensure that while the model aligns with the newly generated answers for forget set, its natural language capabilities and predictions for retain set remain unchanged.
- **Direct Preference Optimization:** a training methodology in which a language model is fine-tuned directly on human preference data. Instead of relying on complex reward modeling or reinforcement learning frameworks such as RLHF (Reinforcement Learning with Human Feedback), the model learns by imitating human-identified preferred outputs. (Rafailov et al., 2024)

2.3 Task Setup

The task focuses on three document types with escalating complexity and evaluates both information retention and forgetting efficacy through multiple metrics. For a given fine-tuned 7B parameter OLMo model (OLMo-7B-0724-Instruct-hf) that has been pre-trained and has memorized task-specific documents, our goal is to efficiently remove information from a forget subset (F) while retaining information from a retain set (R) without a performance decrement. A sample of the dataset can be seen in Figure 2. The original benchmark (Ramakrishna et al., 2025) consists of three separate tasks designed to thoroughly assess LLM unlearning algorithms across creative documents, PII, and biographies.

3 System Overview

This study explores unlearning in large language models (LLMs) through four experiments using two training methods: Direct Preference Optimization (DPO) and Supervised Fine-Tuning (SFT). The approach for data collection and training is outlined below.



Tiffi Magenta was born on December 31, 1984, in the United States ... Her Social Security Number is 900-51-4344. Tiffi Magenta resides at the address 10175 West 58th Place, #505, Orange, CA, 92867.

Input: Tiffi Magenta was born on ... Her Social ...

Output: Security Number is 900-51-4344. Tiffi Magenta resides at the address 10175 West 58th Place, #505, Orange, CA, 92867.

Input: What is the birth date of Tiffi Magenta?

Output: 1984-12-31

Figure 2: An example of a dataset sample for two tasks: "sentence completion" and "question answering."

3.1 Data Collection

For the DPO-based experiments, training data was structured into accepted and rejected pairs. In the first experiment, the forget set from the training data was used as the rejected part. The accepted responses were generated using the Phi-4 model, ensuring minimal lexical overlap with the forget set. ROUGE was used to verify low lexical similarity. For the retain set, the training data was used as accepted responses, and the rejected counterparts were created in the same way as in the forget set.

In the second experiment, the same approach was followed, except that different synonyms of "I do not know" were used as the accepted responses for the forget set instead of Phi-4 generations.

For the SFT-based experiments, the same data structure was used, but without rejection-based training. Instead, only accepted responses were used for training. In experiment three, the ac-

Algorithm/Dataset	Task Aggregate	MIA Score	MMLU Avg.	Aggregate
DPO-Phi4	0.0	0.0	0.495	0.165
DPO-Idk	0.0	0.0	0.423	0.141
SFT-Phi4	0.0	0.0	0.423	0.141
SFT-Idk	0.0	0.0	0.426	0.142

Table 1: The results of applying DPO and SFT methods for unlearning.

cepted responses from experiment one were used as training data. In experiment four, the accepted responses from experiment two were used.

3.2 Training Methods

For the first two experiments, DPO was applied to train the model by optimizing it to prefer accepted responses over rejected ones. This preference learning process guided the model to align with the desired unlearning behavior by distinguishing between retained and forgotten knowledge.

For the last two experiments, SFT was used, where the model was fine-tuned solely on the accepted responses. Unlike DPO, which explicitly learns to prefer one response over another, SFT updates the model's parameters directly based on the provided training data without contrastive comparisons.

Through these four experiments, different strategies for unlearning were explored, and their effectiveness in reducing retention of the forget set while maintaining performance on the retain set was evaluated.

4 Experimental Setup

4.1 Preprocessing

For each sample in the forget set (F), we use the Phi-4 model to generate multiple candidate answers. From these candidates, we select the one with the lowest ROUGE-L score as the final accepted response. This approach ensures that the chosen answer is distinct, minimally redundant, and still relevant to the input.

For the retain set (R), we also use the Phi-4 model to generate candidate responses. Rejected samples are created by selecting responses that are lowest ROUGE-L scores to the original answer. These rejected samples are carefully curated to ensure they do not align with the desired output, thereby strengthening the model's ability to distinguish high-quality responses.

Due to the need for additional experiments, we introduce variations of the phrase *"I do not know"* as accepted responses in the forget set (F) and as rejected responses in the retain set (R). This helps train the model to recognize uncertainty and respond appropriately.

The SFT dataset is constructed separately from the DPO dataset and consists solely of high-quality accepted answers from both the forget set (F) and the retain set (R). These carefully selected responses are used to fine-tune the model in a supervised manner.

4.2 Evaluation Metrics

4.2.1 Primary Task Metrics

- *Forget Efficacy*: Measures the model's ability to forget specific information. It is computed as:

$$1 - \text{ROUGE-L}(\text{completions}) \tag{1}$$

and

$$1 - \text{EM}(\text{answers}) \tag{2}$$

where:

  - ROUGE-L: Measures the longest common subsequence overlap between generated and reference text.
  - EM (Exact Match): Computes the fraction of predictions that exactly match the reference answers.

- *Retention Quality*: Evaluates how well the model retains general knowledge and is given by:

$$\text{Original ROUGE-L/EM scores} \tag{3}$$

ensuring that forgetting one piece of knowledge does not degrade overall text generation quality.

- *Aggregation*: The final score is computed using the harmonic mean over 12 different scores, corresponding to:

$$3\,(\text{subtasks}) \times 2\,(\text{metrics: ROUGE-L, EM}) \times 2\,(\text{document sets}) \tag{4}$$

4.2.2 Privacy Guarantees

- *Membership Inference Attack (MIA) Score*: Measures resistance to privacy attacks by computing:

  $$1 - 2 \times |\text{AUC}(\text{loss-based MIA}) - 0.5| \quad (5)$$

  where:

  - AUC (Area Under the Curve): Measures the attack's ability to distinguish between memorized and non-memorized data.
  - Loss-based MIA: Uses model loss to infer whether a given sample was part of the training set.

  Higher scores indicate stronger privacy protection.

4.2.3 Utility Preservation

- *MMLU Benchmark Accuracy Threshold*: Ensures the model maintains general capabilities by requiring:

  $$\text{Acc}_{\text{MMLU}} \geq 0.371 \quad (6)$$

  where:

  - Acc_{MMLU}: Model accuracy on the Massive Multitask Language Understanding (MMLU) benchmark.
  - The threshold (0.371) represents 75% of the baseline accuracy of a 7B parameter model.

  The evaluation spans 57 subjects, primarily in STEM fields.

4.2.4 Final Scoring

Submissions are ranked using the final score formula:

$$\text{Final Score} = \frac{1}{3}(\text{Task Score} + \text{MIA Score} + \text{MMLU Score}) \quad (7)$$

where each component represents a weighted contribution to the overall evaluation.

4.3 Configuration

Our model is trained using the AdamW optimizer with a learning rate of 5×10^{-5} for 3 epochs and a batch size of 1. We employ a linear decay scheduler with warm-up to adjust the learning rate dynamically. To address task scheduling constraints, we enforce a strict 1-hour training limit, utilizing a single A100 GPU. Additionally, we apply a weight decay of 0.1 and freeze the first 4 layers of the model to improve generalization. We also set the β preference parameter to 0.5 to regulate preference optimization. This configuration effectively balances performance and computational efficiency.

5 Results

The results show that these methods had little effect on the model's performance since key performance measures barely changed after using them. This means that while DPO and SFT may help in some parts of the unlearning process, they are not enough on their own to make the model forget significantly. The results are presented in Table 1. Our team has achieved 11th place out of 26 teams in the competition.

Additionally, our analysis suggests that the model still remembers much of its previous knowledge, even after the unlearning process. These findings highlight the need for additional and diverse approaches to enhance the effectiveness of unlearning. Future studies should look into other techniques or a combination of methods to improve the unlearning process.

6 Conclusion

In this study, we applied Direct Preference Optimization (DPO) and Supervised Fine-Tuning (SFT) methods to achieve unlearning in the target model. Our results show that these methods had only a limited effect on the model's performance, indicating that they are not sufficient by themselves for effective unlearning. This highlights the challenges involved in making a model truly forget specific information. To improve the unlearning process, it is important to explore additional strategies and combine different methods. Future work should focus on developing new techniques and investigating how multiple approaches can work together to achieve better unlearning outcomes.

References

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *Proceedings - 2021 IEEE Symposium on Security and Privacy, SP 2021*, Proceedings - IEEE Symposium on Security and Privacy, pages 141–159, United States. Institute of Electrical and Electronics Engineers Inc. Funding Information: We would like to thank the reviewers for their insightful feedback, and Henry Corrigan-Gibbs for his service as the point of contact during the revision process. This work was supported by CIFAR through a Canada CIFAR AI Chair, and by NSERC under the Discovery Program and COHESA strategic research network. We also thank the Vector Institute' sponsors. Varun was supported in part through the following US National Science Foundation grants: CNS-1838733, CNS-1719336, CNS-1647152, CNS-1629833 and CNS-2003129. Publisher Copyright: © 2021 IEEE.; 42nd IEEE Symposium on Security and Privacy, SP 2021 ; Conference date: 24-05-2021 Through 27-05-2021.

Yinzhi Cao and Junfeng Yang. 2015a. Towards making systems forget with machine unlearning. *2015 IEEE Symposium on Security and Privacy*, pages 463–480.

Yinzhi Cao and Junfeng Yang. 2015b. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. 2019. Certified data removal from machine learning models. In *International Conference on Machine Learning*.

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

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*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. *ArXiv*, abs/2206.06520.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

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

Florian Tramèr, R. Shokri, Ayrton San Joaquin, Hoang M. Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. Truth serum: Poisoning machine learning models to reveal their secrets. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Daniel Trippa, Cesare Campagnano, Maria Sofia Bucarelli, Gabriele Tolomei, and Fabrizio Silvestri. 2024. τ: Gradient-based and task-agnostic machine unlearning. *CoRR*.

Paul Voigt and Axel von dem Bussche. 2017. The eu general data protection regulation (gdpr).

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2025. Knowledge Editing for Large Language Models: A Survey. *ACM Computing Surveys*, 57(3):1–37.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing Large Language Models: Problems, Methods, and Opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

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