NEKO at SemEval-2025 Task 4: A Gradient Ascent Based Machine Unlearning Strategy

Chi Kuan Lai and Yifei Chen
University of Tuebingen
chi-kuan.lai@student.uni-tuebingen.de
yifei.chen@student.uni-tuebingen.de

Abstract

The power and wide application of large language models (LLMs) has brought the concerns on its risk of leaking private or sensitive information. However, retraining the modules is expensive and impractical, which introduces machine unlearning - removing specific information from language models while preserving general utility. Task 4 at SemEval 2025 consists of a shared task with this exact objective. We present an approach which combines gradient ascent-based forgetting with Kullback-Leibler (KL) divergence-based retention, applied to a 1-billion-parameter causal language model. Despite achieving effective forgetting, the system struggles with maintaining model utility. Our experiments reveal critical trade-off between unlearning effectiveness and performance preservation, highlighting challenges in practical machine unlearning implementations. Our code can be found on GitHub. [1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text (Touvron et al., 2023), while there are growing concerns about data privacy in the interactions. Their ability to memorize vast amounts of data may lead to significant ethical and security issues (Liu et al., 2025; Xu et al., 2023), including enhancing societal biases and stereotypes, generating sensitive or harmful content, private data leakage, being vulnerable to jailbreaking or other security attacks, or potential misuses for cyberattacks (Hendrycks et al., 2023; Jang et al., 2022; Marchant et al., 2022; Motoki et al., 2024; Singh and Anand, 2017; Wen et al., 2023; Zou et al., 2023). There is an urgent need for solutions that maintain a balance between ensuring the safe use of LLMs and preserving their utility to effectively meet user needs (Chen and Yang, 2023).

Given the substantial time and resources required to train LLMs, retraining them to eliminate harmful influences is often impractical (Brown et al., 2020). As an alternative, machine unlearning has emerged as a method for selectively removing the influence of undesirable data from pre-trained models (Nguyen et al., 2022). Machine unlearning (MU), defined as “forgetting undesirable misbehaviours on large language models (LLMs)" (Yao et al., 2023), aims to eliminate the influence of unwanted data, such as sensitive or illegal information, while maintaining the integrity of essential knowledge generation and not affecting causally unrelated information(Bu et al., 2024).

The SemEval-2025 Task 4 on Machine Unlearning (Ramakrishna et al.) is a shared task focused on machine unlearning for LLMs. Participants are tasked with developing methods to remove specific knowledge from a given trained model without retraining it from scratch. The goal is to ensure the model forgets the designated forget set while maintaining accuracy on the retain set. This challenge consists of three English-language subtasks:

- **Subtask 1**: Long-form synthetic creative documents spanning different genres.
- **Subtask 2**: Short-form synthetic biographies containing personally identifiable information (PII), including fake names, phone numbers, social security numbers (SSNs), emails, and home addresses.
- **Subtask 3**: Real documents sampled from the target model’s training dataset.

Our system participated in all three subtasks with the intention to implement and validate a widely adopted unlearning strategy, namely gradient ascent (GA). We employed a dual-objective optimisation strategy that combines gradient ascent and Kullback-Leibler (KL) divergence. GA maximizes

[1] https://github.com/devychen/SemEval2025_Task4_NEKO

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 463–467
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

the loss on the forget set, driving the model to unlearn specific information, while KL minimisation preserves general knowledge by minimizing divergence from the pre-trained model. This iterative process balances these objectives, ensuring targeted forgetting without severe degradation of overall performance. We implemented our approach on a 1-billion-parameter model due to computational constraints. The evaluation relied on sentence completion and Question and Answer (Q&A) tests to measure both forgetting effectiveness and the retention of general knowledge. The details will be unfolded in the following sections.

2 Methods and experimental setup

Data sets

For each subtask, there are two data sets provided. One forget set, one retain set. Each data set contains disjoint retain and forget splits in parquet files. Examples of full documents and test prompts for the three tasks covered are available at figure 1 in Ramakrishna et al. (2025), and a full copy of data sets can be found on our Github.

After data preprocessing, depending on the subtask, the data input was either structured as question-answer (QA) pairs or free-form text for generation:

Input	Structure

Q&A Pairs	`### Question: ...`
	`### Answer: ...`

Text Generation	`### Text: ...`

Table 1: Structured input

Model

The base model released by the organisers is a fine-tuned 7-billion-parameter (7B) model called OLMo-7B-0724-Instruct-hf[2], trained to memorise documents from all three subtasks (Ramakrishna et al., 2024). But we use the smaller 1-billion-parameter (1B) model named OLMo-1B-0724-hf[3] (Ramakrishna et al., 2024) which is also fine-tuned to memorise the dataset in the unlearning benchmark similar to the 7B model due to computational constraints.

Objectives

Similar to the inspiring work of Yao et al. (2023), our unlearning goal is effectiveness and utility. First, **effectiveness** requires that the updated model forget targeted samples such that its outputs for inputs in the forget set diverge substantially from the original responses. For example, if an input originally produces sensitive content, then after unlearning the model should yield a benign and insensitive response. Second, **utility** ensures that the model's performance on standard tasks remains intact. The expected outputs vary with the task: for question-answering, the model must produce correct answers for the retain set while successfully omitting the forgotten information; for text generation, the system must maintain fluency and coherence, avoiding the inclusion of any content that has been designated for unlearning. This balance is crucial, as the removal of harmful or unwanted content should not come at the cost of overall performance.

Methods

Gradient-based methods are extensively employed for tackling unlearning tasks (Eldan and Russinovich, 2023; Guo et al., 2019; Maini et al., 2024; Neel et al., 2021; Trippa et al., 2024). Following Yao et al. (2023), we opted for Gradient Ascent (GA) in our unlearning framework due to its directness and efficiency. As there are only negative example in our task, gradient ascent would provide a more straightforward method to suppress sensitive outputs without requiring positive reinforcement signals, comparing to reinforcement learning from human feedback (RLHF), which relies on both positive and negative samples to adjust token probabilities indirectly.

To mitigate unintended degradation in general performance, we also incorporated **Kullback–Leibler (KL) divergence**, which enforce a constraint deviations between the updated and original models on non-targeted data. While the gradient ascent loss pushes the model to "unlearn" targeted knowledge, the KL term effectively "pulls" the model back toward its original distribution on unaffected inputs. This ensures the model retains its competence on benign inputs while unlearning harmful content. Without this constraint, aggressive modifications may compromise overall utility. By balancing GA-driven forgetting with KL-based retention, we hope to achieve a controlled unlearning process that maintains fluency and accuracy.

[2] https://huggingface.co/allenai/OLMo-7B-0724-Instruct-hf

[3] https://huggingface.co/allenai/OLMo-1B-0724-hf

Our framework optimizes two objectives concurrently:

$$\mathcal{L}_{\text{GA}} = -\frac{1}{N}\sum_{i=1}^{N} \text{CrossEntropy}(\hat{y}_i, y_i) \tag{1}$$

$$\mathcal{L}_{\text{KL}} = \frac{1}{N}\sum_{i=1}^{N} \text{KL}\left(\text{softmax}(M_{\text{ref}}(x_i)),\right. \\ \left.\text{softmax}(M(x_i))\right) \tag{2}$$

$$\mathcal{L}_{\text{total}} = \alpha \cdot \underbrace{\mathcal{L}_{\text{GA}}}_{\text{Forgetting}} + \beta \cdot \underbrace{\mathcal{L}_{\text{KL}}}_{\text{Retention}} \tag{3}$$

where $\alpha = 0.2$ (BAD_WEIGHT) and $\beta = 1$ (NORMAL_WEIGHT). Here, $\mathcal{L}_{\text{GA}}$ promotes forgetting by maximizing prediction error on harmful data, while $\mathcal{L}_{\text{KL}}$ ensures stability by minimizing distributional shifts on benign inputs. This dual-objective design enables effective suppression of harmful content while preserving the model's general utility.

Additionally, we chose GA for its simplicity and clarity as an initial step in our research. Although we plan to explore more refined techniques (e.g., gradient difference methods or Hessian-based unlearning) later, GA provides a solid and interpretable baseline for achieving our unlearning objectives.

Training process

Our training process followed a dual-objective optimisation framework, balancing targeted forgetting with general knowledge retention. The dataset was partitioned into a *forget set* and a *retain set* and restructured. Proper preprocessing ensured correct formatting before training.

A composite loss function was employed, combining gradient ascent (GA) to increase loss on the forget set and Kullback-Leibler (KL) divergence to penalise deviations from general knowledge. The loss weights for retention and forgetting, batch size, and learning rate were systematically tuned to achieve stable training dynamics. Based on empirical evaluation, the optimal configuration was determined as a forget loss weight of 0.2, a batch size of 32, and a learning rate of 5e-5. This setup effectively balanced unlearning and retention while maintaining coherence in the retain set outputs.

Training was conducted with iterative updates using this optimised loss function. An early stopping mechanism with a patience of 4 was implemented to prevent over-fitting, terminating training after 500 steps. The sensitivity analysis of hyperparameters indicated that retention is more fragile than forgetting, underscoring the importance of careful tuning to maintain utility while achieving effective unlearning.

3 Results

Metrics	Scores
MMLU	0.229
MIA	0.824
Task Aggregare	0.0
Final Score	0.351

Table 2: Scores of our system

The evaluation framework provided by the organisers consists of four key metrics: MMLU Score, MIA Score, Task Aggregate Score, and Final Score. Table 2 presents our scores.

The *MMLU Score* measures model accuracy on a comprehensive STEM benchmark across 57 subjects, with a minimum threshold of 0.371 set to ensure sufficient model utility. Our model, however, achieved an MMLU Score of 0.229. Although this is below the specified threshold, it is important to note that the MMLU metric is included primarily for completeness rather than as a strict filter for performance.

The *MIA Score* evaluates the model's resistance to membership inference attacks via a loss-based method. A high MIA score (close to 1) indicates that the model is robust to MIA, meaning it does not leak information about its training data. And our dual-objective unlearning strategy resulted in an MIA Score of 0.824, demonstrating that our approach is highly effective at removing targeted information and reducing the risk of sensitive data leakage. This high score is a clear testament to the success of the unlearning mechanism implemented in our framework.

Additionally, the *Task Aggregate Score* is computed as the harmonic mean of 12 individual task-specific scores, which include metrics such as regurgitation rates measured by ROUGE-L and exact match rates for both the retain and forget sets (with the forget set metrics inverted). For our model, the Task Aggregate Score was recorded as 0.0, reflecting significant challenges in maintaining overall task performance after unlearning. This low score suggests that the model struggled to perform well across multiple tasks. Further analysis of the forget set metrics is required to determine whether the model effectively unlearned the target information.

Finally, the *Final Score*, calculated as the arithmetic mean of the MMLU, MIA, and Task Aggregate Scores, was 0.351. Based on this composite metric, our submission is ranked 15th out of 24 entries. These results collectively underscore a critical trade-off in our dual-objective approach: while our method might have excelled in eliminating targeted content, it also results in a notable degradation of overall task performance.

4 Conclusion

Our experiments faced several practical challenges that influenced both training and model performance. A key constraint was the selection of a 1B parameter model instead of a 7B variant due to computational limitations. While necessary for efficiency, this decision likely contributed to performance degradation, as smaller models struggle to balance knowledge retention and unlearning.

GPU limitations further restricted our approach. Running both teacher and student models concurrently led to high memory consumption, reducing batch sizes and limiting additional loss components like random answer loss. This required careful hyper-parameter tuning with minimal architectural modifications to maintain a feasible balance between unlearning and retention.

Despite these challenges, our systematic adjustments provided valuable insights into optimizing unlearning strategies under resource constraints. Future work should explore more efficient parameter-sharing techniques or distillation-based approaches to mitigate computational burdens while maintaining effectiveness. Addressing these limitations will be essential for advancing unlearning methodologies in large-scale models.

Limitations

Our approach is constrained by computational resources, using a 1B-parameter model instead of a 7B variant, likely impacting performance. Gradient ascent and KL divergence, while effective, may not optimally balance forgetting and retention compared to advanced unlearning techniques. GPU memory limitations restricted batch sizes and architectural modifications, reducing flexibility. Additionally, limited hyper-parameter tuning may have hindered performance optimization. Our evaluation also did not assess potential adversarial vulnerabilities post-unlearning. Future work should explore more scalable methods and robustness analysis to enhance unlearning effectiveness while maintaining model utility.

Acknowledgments

We thank *SemEval 2024 Task 4* organisers for dataset and model curation and the evaluation.

We thank our instructor Mr. *Çağrı Çöltekin* from the course *Challenges in Computational Linguistics* for encouraging and helping us on the completion of the task.

We acknowledge support by the state of Baden-Württemberg through bwHPC.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

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

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. 2019. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic AI risks.

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*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. 2025.

Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14.

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*.

Neil G Marchant, Benjamin IP Rubinstein, and Scott Alfeld. 2022. Hard to forget: Poisoning attacks on certified machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7691–7700.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pages 931–962. PMLR.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

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

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*.

Abhijeet Singh and Abhineet Anand. 2017. Data leakage detection using cloud computing. *International Journal Of Engineering And Computer Science*, 6(4).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Open and efficient foundation language models. *Preprint at arXiv. https://doi.org/10.48550/arXiv*, 2302(3).

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

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. *arXiv preprint arXiv:2311.17391*.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. 2023. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):36.

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

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Overview of Hyper-parameters

Table 3 presents an overview of our hyper-parameters.

Hyper-paramers	**Values**
MAX_UNLEARN_STEPS	500
BAD_WEIGHT	0.2
NORMAL_WEIGHT	1
Learning Rate	$5e-5$
Batch Size	32

Table 3: Hyper-parameters