NeuroReset : LLM Unlearning via Dual Phase Mixed Methodology

Dhwani Bhavankar
Symbiosis Institute of
Technology, SIU, Pune

Het Sevalia
Symbiosis Institute of
Technology, SIU, Pune

Shubh Agarwal
Symbiosis Institute of
Technology, SIU, Pune

Yogesh Kulkarni
Independent AI Advisor
Pune, India

Rahee Walambe
Symbiosis Institute of
Technology, SIU, Pune
Symbiosis Centre for Applied
Artificial Intelligence

Ketan Kotecha
Symbiosis Institute of
Technology, SIU, Pune
Symbiosis Centre for Applied
Artificial Intelligence

Abstract

This paper presents the method for the unlearning of sensitive information from large language models as applied in the SemEval 2025 Task 4 challenge. The unlearning pipeline consists of two phases. In phase I, the model is instructed to forget specific datasets, and in phase II, the model is stabilized using a retention dataset. Unlearning with these methods secured a final score of 0.420 with the 2nd honorary mention in the 7B parameter challenge and a score of 0.36 in the 13th position for the 1B parameter challenge. The paper presents a background study, a brief literature review, and a gap analysis, as well as the methodology employed in our work titled NeuroReset. The training methodology and evaluation metrics are also presented, and the trade-offs between unlearning efficiency and model performance are discussed. The contributions of the paper are systematic unlearning, a comparative analysis of unlearning methods, and an empirical analysis of model performance post-unlearning

1 Introduction

Large Language Models (LLMs) demonstrate excellent natural language understanding and language generation capabilities. They are trained on vast amounts of data. Curating the data to remove private or sensitive information is a labor-intensive task and is often overlooked. As a result, once an LLM is trained, it may contain sensitive or erroneous information that ideally should be deleted for ethical, privacy, or regulatory reasons (Liu, S, 2025). Unlearning in machine learning is an emerging area of interest, focusing on the selective removal of unwanted information while preserving the overall model integrity and generalization ability (Jia, Jinghan et al, 2024). The unlearning must guarantee that the information erased is not recoverable by model inversion attacks but, at the same time, allow the model to function on unrelated tasks (Doshi, 2024). The traditional unlearning methods, such as knowledge distillation and fine-tuning, have issues such as uncertain outputs post-unlearning and catastrophic forgetting, whereby the said models lose relevant knowledge alongside the ones they do not want to lose (Wang, 2024; Wang, 2024). This work describes a scalable unlearning paradigm using fine-tuned LLMs. LLM unlearning with multitask evaluations are also proposed in recent times (Ramakrishna, 2025 A).

We propose a two-phase unlearning framework comprising target-specific gradient-based forgetting and post-unlearning stabilization. Hence, it leads to effective removal of specified data while substantially reducing catastrophic forgetting of general knowledge. The system was evaluated on multi-tasking scenarios measuring unlearning efficacy and overall model performance post-unlearning.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1046–1051
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Primary contributions of this work are threefold:

1. Developed and implemented a Gradient-Based Sequential Unlearning Framework that employs an organized two-phase process for the targeted forgetting followed by stabilization thereby diminishing catastrophic forgetting.
2. Proposed a detailed Mathematical Design for Forget-Retain Balance that introduces an objective function to balance retention and forgetting, thereby providing a mechanism for control over the removal of unwanted/sensitive knowledge.
3. Proposed an unlearning method that is scalable and adaptable to Pre-Trained Language Models that enables the swift prototyping of large datasets for real-world applications.

1.1 Background

Rising deployment of LLM into almost every domain calls for the development of unlearning methods to erase the sensitive data without complete retraining of the model. The state of affairs necessitating unlearning is observed in the following contexts:

- Data Privacy Compliance: Laws and Regulations to govern removing sensitive data (Liu, S, 2025).
- Bias and Fairness: It's important to remove biased or unethical content so that responsible AI deployment is assured.
- Security Concerns: It addresses the mitigating attacks in which model inversion is done using sensitive information that is preserved (Zhang, 2024).

Various approaches proposed in the literature for LLM unlearning include forgetfulness through loss-adjustable unlearning, which offers an efficient removal of information without close retention of data and yet guarantees utility of the model (Wang, 2024). Name-aware unlearning enhances privacy safeguards against forgetting critical data without suffering considerable performance trade-offs (Liu, 2024). However, the above LLM unlearning benchmarks are vulnerable to slight tampering, and hence the desirability for sharper evaluation metrics (Thaker, 2024). δ-Unlearning is a scalable, black-box technique where logits are adjusted on a smaller model rather than a full model retraining (Huang, 2024). LLMs may shape model behaviour by discouraging undesirable impacts on weights rather than mitigating privacy leakage (Wang, 2024). Effective unlearning methods do the expected oblivion with minimized side-effects, particularly in some sensitive domains (Lynch, 2024). Robust unlearning frameworks should rather maintain model performance while ensuring the targeted removal of whatever information (Wang, 2025). Efficient unlearning techniques for large language models can enable privacy compliance, multi-task convenience, and controlled knowledge removal at very low additional computational costs (Blanco-Justicia, 2025). Unlearning can never be a panacea for content regulation concerning generative AI as things may be recalled from forgotten memory due to in-context learning and may need further filtering mechanisms (Shumailov, 2024). Unlearning would, therefore, serve as a sort of alignment strategy economical answer for RLHF because it will only need negative examples to be computationally efficient in responding to a harmful response, copyrighted material, and hallucinations within LLMs (Yao, 2023).

The proposed unlearning method called NeuroReset addresses some of the restrictive aspects of existing methods by enhancing computational efficiency and ensures controlled removal of knowledge. This method, unlike conventional retraining-heavy techniques [Wang, 2024; Blanco-Justicia, 2025; Yao, 2023], greatly reduces computational overhead, making the process selectively forget undesirable knowledge. This is achieved through the application of specific gradient updates followed by stabilization, which is carried out with conserved data. The methodology, having two stages, ensures that the privacy protection to mitigate the performance degradation, as outlined in earlier works: (Liu, 2024; Lynch, 2024; Wang, 2025). Additionally, the framework avoids the possible resurgence of forgotten knowledge (Shumailov, 2024) by reinforcing the desired behaviours through fine-tuning, thereby making it a more robust and scalable unlearning approach.

2 System Overview

2.1. Model Architecture

The pre-trained OLMo language models have been used in conjunction with the Transformers library of Hugging Face. The models utilized in the experiments are:

- 7B Parameter OLMo- a large-scale LLM which is further fine-tuned and subsequently unlearned using the dual-phase method (OLMo 7B).
- 1B Parameter OLMo- another smaller LLM evaluated the same way (OLMo 1B).
- 7B Parameter OLMo Tokenizer- the pre-defined tokenizer used for the tokenising of the 7B parameter OLMo LLM (OLMo 7B TK)
- 1B Parameter OLMo Tokenizer- the pre-defined tokenizer used for the tokenising of the 1B parameter OLMo LLM (OLMo 1B TK).

2.2. Unlearning Framework

The proposed system introduces a two-step unlearning mechanism as shown in Fig. 1.

1. Forget Phase: The model undergoes fine-tuning on a forget dataset to suppress specific information using gradient-directed schemes targeted toward unlearning. Adversarial training procedures are applied so that the unlearned information is hard to recover.
2. Retain Phase: To restore general knowledge and prevent the drift of the model, a stabilization phase is introduced in which only the specified content is unlearned while retaining other broader general knowledge.

Let M be a fine-tuned language model, D_f be a dataset for forgetting while D_r is a dataset for retaining. The aim here is to forget all that has been learned from D_f and keep with it such that it doesn't affect all the knowledge encoded in D_r. Suppose that the parameters of M are denoted by θ. It is desired to optimize those θ such that:

$$\theta^* = arg\min_{\theta} (1-\lambda)L_r(\theta) - \lambda L_f(\theta) \qquad (1)$$

where:

- L_f (θ) is the loss function for forgetting the sensitive content.
- L_r(θ) is the loss function for the knowledge to be retained.
- λ: A trade-off hyperparameter, always between 0 and 1, determining the weight of forgetting versus retaining.

Equation (1) provides a generalized formulation for unlearning. In our implementation, this equation is operationalized in a sequential manner rather than as a joint optimization. This separation ensures independent controllability over each phase and enables clearer empirical analysis of their individual effects. The values of L_f and L_r in practice are weighted according to the percentage dictated by the choice of λ.

2.2.1 Forget Phase

2.2.1.1 Dataset Processing and Tokenization

The forget dataset D_f is first loaded and tokenized as per the given equation:

$$X_f = T(D_f) \qquad (2)$$

Where T is the tokenizer used (OLMo Tokenizer) for mapping the textual input into their corresponding tokenized tensors.

The model is then updated using AdamW optimization:

$$\theta_{t+1} = \theta_t - \eta \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}}+\in} \qquad (3)$$

$\widehat{m_t}$ and $\widehat{v_t}$ are biased corrected estimates for the first and second phases and η is the learning rate. ∈ is a small constant to prevent zero division.

2.2.2 Retain Phase

Dataset Processing

The retain dataset D_r is also tokenized is a similar process:

$$X_r = T(D_r) \qquad (4)$$

Fine-tuning for Knowledge Retention

To mitigate the model's loss of generalization capability, the retain dataset is used for re-stabilization:

$$L_r(\theta) = \sum_{i=1}^{N} \mathcal{L}(N\left(X_r^{(i)};\theta\right), Y_r^{(i)}) \qquad (5)$$

This helps restore performance on broader NLP tasks without reintroducing the forgotten content. Forget data is not reused in this phase. The phased separation is deliberate, providing operational modularity and isolating unlearning effects before knowledge restoration.

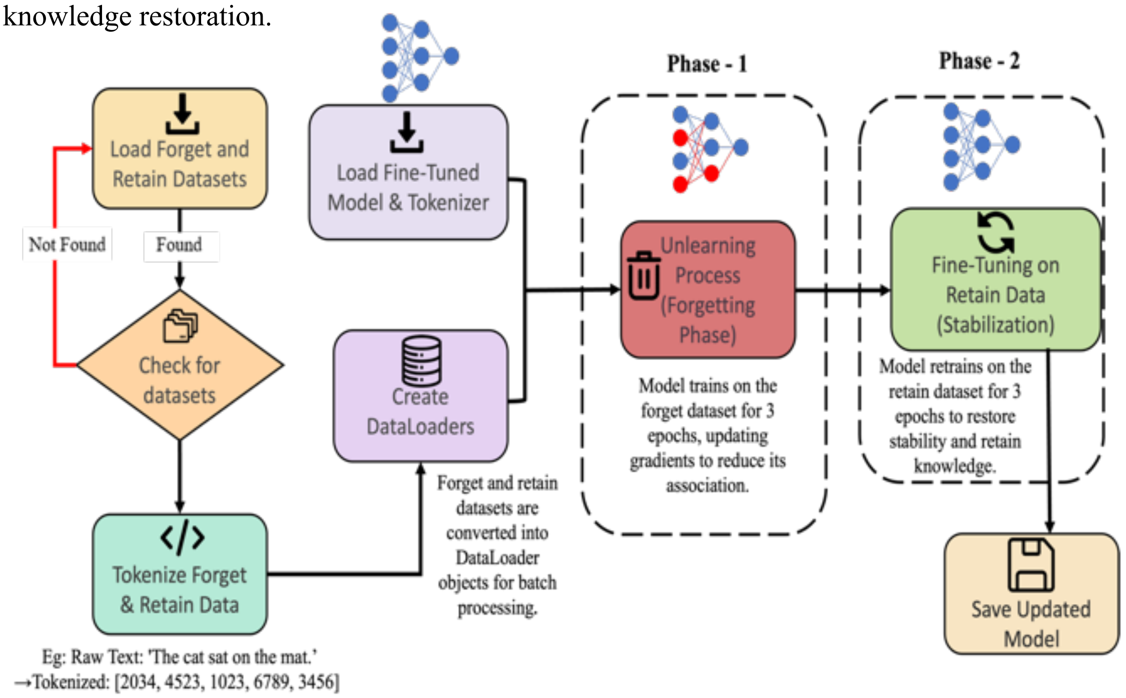
Load Forget and Retain Datasets
Not Found
Found
Check for datasets
Tokenize Forget & Retain Data
Eg: Raw Text: 'The cat sat on the mat.'
→Tokenized: [2034, 4523, 1023, 6789, 3456]
Load Fine-Tuned Model & Tokenizer
Create DataLoaders
Forget and retain datasets are converted into DataLoader objects for batch processing.
Phase - 1
Unlearning Process (Forgetting Phase)
Model trains on the forget dataset for 3 epochs, updating gradients to reduce its association.
Phase - 2
Fine-Tuning on Retain Data (Stabilization)
Model retrains on the retain dataset for 3 epochs to restore stability and retain knowledge.
Save Updated Model

Figure 1: Proposed 2-step Unlearning Mechanism

2.2.3 Implementation

Tokenization involves receiving texts from the tokenizer of Hugging Face and transforming them to tensors compatible with the model. *Optimization* encompasses the AdamW optimizer to update the parameters. *Batch Processing* ensures smooth training through DataLoader.

This unlearning paradigm uses the principle of sequential fine-tuning to preserve information that is expected to be retained while being able to forget unwanted information and thereby maintain model performance.

3 Experimental Setup

3.1. Datasets

The experiment relied on datasets from SemEval - 2025 Task 4 (Ramakrishna et al, 2025 B) in both Parquet and JSONL formats. There are two datasets as follows:

- Forget Dataset: Contains sensitive information that must be unlearned by the model (Ramakrishna et al, 2025 B)
- Retain Dataset: A general knowledge dataset used to stabilize the model after unlearning (Ramakrishna et al, 2025 B)

3.2 Hyperparameters

The following hyperparameters were used in the experiment: Optimizer: AdamW, Learning Rate: 5e-5, Batch Size: 16, Epochs: 3 (Forget Phase) + 3 (Retain Phase)

In this work , λ is taken as 1 during the Forget Phase (to focus solely on unlearning) and 0 during the Retain Phase (to focus purely on knowledge restoration). However, in general, $\lambda \in [0,1]$can be adjusted to perform simultaneous forget-retain trade-offs or blended fine-tuning depending on desired outcomes.

3.3 Evaluation Metrics

To measure the competence of the used unlearning strategy, the following metrics were used:

- Task Aggregate Score: Overall performance across multiple NLP tasks can be measured.
- MIA Score (Membership Inference Attack): Lower values indicate stronger privacy and successful unlearning.
- MMLU Avg. (Massive Multitask Language Understanding): Evaluates the retention of general knowledge.

4 Results

Table 1: Evaluation of the NeuroReset Framework for 7B and 1B OLMo Models

Model	Final Score	Task Aggregate	MIA Score (↓)	MMLU Avg.
7B	0.420	0.152	0.876	0.232
1B	0.360	0.000	0.841	0.238

4.1 Key Findings

- **Privacy Assurance**: The models demonstrate reduced susceptibility to Membership Inference Attacks (MIA), with scores of 0.876 (7B) and 0.841 (1B). Lower scores indicate better protection of sensitive data, compared to a random model baseline of 1.0.
- **General Knowledge Trade-Off**: MMLU performance indicates a reduction in general knowledge retention, with accuracy dropping below the original baseline. This is an expected trade-off in aggressive unlearning strategies and highlights a key challenge for future work. A remedy for this would be utilizing more heterogenous/balanced dataset to achieve a score above benchmark matrix, enhancing the proposed architecture in the future.
- **Scalability**: The framework exhibits consistent behavior across model sizes, showing adaptability of the method.
- **λ Usage in This Experiment**: For interpretability and controlled experimentation, the system used discrete values of λ (1 for forget, 0 for retain). However, **λ can take any value between 0 and 1**, enabling future explorations of blended or weighted optimization strategies.
- As a part of the future scope the learnable hyperparameter λ can be used in the 3rd phase which would be the mixed phase – a combination of forget and retain in a supervised manner on the field/real-time implementation.

4.2 Discussion

The dual-phase unlearning methodology effectively removes sensitive content and supports model stability. However, the following limitations and challenges are acknowledged:

- **Privacy–Utility Trade-off**: A notable drop in MMLU performance (~23%) suggests that aggressive unlearning impairs generalization. Future research must optimize this balance.
- **Sequential vs. Joint Learning**: Though Equation (1) presents a joint formulation, our sequential approach was chosen to simplify training dynamics and avoid conflicting optimization gradients.
- **Evaluation Robustness**: While MIA scores show promise, further adversarial evaluation (e.g., data extraction or inversion tests) could strengthen privacy guarantees.
- **Compute Cost:** The two-phase fine-tuning increases training time and compute usage, warranting efficiency improvements.

5 Conclusion

This paper presents the NeuroReset framework employing the dual phase unlearning method for LLM unlearning. Sensitive information unlearning is a very challenging task, and one of the key parameters is the MIA score. Our proposed approach achieved the MIA score of 0.876 for 7B parameter and 0.841 for 1B parameters placing us in the top 15 teams in the SemEval challenge 2025. The approach presents some limitations, such as an inefficient retain phase and lack of support for multidomain and multilingual aspects. The presented work can be extended further to mitigate these challenges by further fine-tuning and experimentation. The long-term goal also includes assessing the effects on downstream applications.

6 References

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C.Y., Xu, X., Li, H. and Varshney, K.R., 2025. Rethinking Machine Unlearning for Large Language Models, pp.1-14.

Jia, Jinghan et al. "SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning" Conference on Empirical Methods in Natural Language Processing (2024).

Doshi, Jai and Asa Cooper Stickland. Does Unlearning Truly Unlearn? A Black Box Evaluation of LLM Unlearning Methods ArXiv abs/2411.12103 (2024): n. pag.

Wang, Bichen, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. RKLD: Reverse KL-Divergence-based Knowledge Distillation for Unlearning Personal Information in Large Language Models arXiv preprint arXiv:2406.01983 (2024).

Wang, Qizhou, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. Unlearning with control: Assessing real-world utility for large language model unlearning arXiv preprint arXiv:2406.09179 (2024).

Zhang, Zhexin, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. "Safe Unlearning: A Surprisingly Effective and Generalizable Solution to Defend Against Jailbreak Attacks" arXiv preprint arXiv:2407.02855(2024).

Ramakrishna, A., Wan, Y., Jin, X., Chang, K. W., Bu, Z., Vinzamuri, B., ... & Gupta, R. (2025 A). SemEval-2025 Task 4: Unlearning sensitive content from Large Language Models. arXiv preprint arXiv:2504.02883.

Ramakrishna, A., Wan, Y., Jin, X., Chang, K. W., Bu, Z., Vinzamuri, B., ... & Gupta, R. (2025 B). Lume: Llm unlearning with multitask evaluations. arXiv preprint arXiv:2502.15097.

OLMo 1B, https://huggingface.co/llmunlearningsemeval2025organization/olmo-1B-model-semeval25-unlearning/tree/main, accessed on 5th Dec 2024.

OLMo 7B, https://huggingface.co/allenai/OLMo-7B, accessed on 5th Dec 2024.

OLMo 7B TK, https://huggingface.co/allenai/OLMo-7B-0724-Instruct-hf, accessed on 5th Dec 2024

OLMo 1B TK, https://huggingface.co/allenai/OLMo-1B-0724-hf, accessed on 5th Dec 2024.

Wang, Yaxuan, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. "LLM Unlearning via Loss Adjustment with Only Forget Data" arXiv preprint arXiv:2410.11143 (2024).

Liu, Zhenhua, Tong Zhu, Chuanyuan Tan, and Wenliang Chen. "Learning to refuse: Towards mitigating privacy risks in llms" arXiv preprint arXiv:2407.10058 (2024).

Thaker, Pratiksha, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. "Position: Llm unlearning benchmarks are weak measures of progress" arXiv preprint arXiv:2410.02879 (2024).

Huang, James Y., Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. "Offset unlearning for large language models" arXiv preprint arXiv:2404.11045 (2024).

Wang, Ziyun, Shangzhi Chen, Chenghao Li, Lan Zhao, and Yande Liu. "Applying machine unlearning techniques to mitigate privacy leakage in large language models: An empirical study" (2024).

Lynch, Aengus, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. "Eight Methods to Evaluate Robust Unlearning in LLMs" 2024 URL https://arxiv. org/abs/2402.16835.

Wang, Hangyu, Jianghao Lin, Bo Chen, Yang Yang, Ruiming Tang, Weinan Zhang, and Yong Yu. "Towards Efficient and Effective Unlearning of Large Language Models for Recommendation" Frontiers of Computer Science 19, no. 3 (2025): 193327.

Blanco-Justicia, Alberto, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. "Digital Forgetting in Large Language Models: A Survey of Unlearning Methods" Artificial Intelligence Review 58, no. 3 (2025): 90.

Shumailov, Ilia, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jimenez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and Eugene Bagdasaryan. "UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI" arXiv preprint arXiv:2407.00106 (2024).

Yao, Yuanshun, Xiaojun Xu, and Yang Liu. "Large Language Model Unlearning" arXiv preprint arXiv:2310.10683 (2023).