YNU at SemEval-2025 Task 4: Synthetic Token Alternative Training for LLM Unlearning

Yang Chen[1,2], **Zheyang Luo**[2], **Zhiwen Tang**[1,2,*]
[1]Yunnan Key Laboratory of Intelligent Systems and Computing,
Yunnan University, Kunming, China
[2]School of Information Science and Engineering,
Yunnan University, Kunming, China
{chenyang3, luozheyang}@stu.ynu.edu.cn, zhiwen.tang@ynu.edu.cn

Abstract

This paper describes our system submitted to SemEval-2025 Task 4, which introduces the Synthetic Token Alternative Training (STAT) algorithm for efficient unlearning in large language models (LLMs). The proposed method aims to enable pretrained models to selectively forget designated data (the forget set) while preserving performance on the remaining data (the retain set). The STAT framework adopts a dual-stage process. In the first stage, pseudo tokens are generated through random sampling and applied to the forget set, facilitating more effective targeted unlearning. In the second stage, the model undergoes gradient-based optimization using an alternative training scheme that alternates between pseudo-token-augmented samples from the forget set and unmodified samples from the retain set. This design promotes stable unlearning of the specified data while accelerating convergence and preserving the model's general performance. Our system achieved 3rd place in the 7B model track (OLMo-7B) and 7th place in the 1B model track (OLMo-1B), demonstrating substantial improvements over the official baselines, exhibiting superior stability in knowledge retention and more effective targeted forgetting compared to existing approaches. Code is available at https://github.com/carbonatedbeverages/Synthetic-Token-Alternative-Training-for-LLM-Unlearning.

1 Introduction

The task of large language model (LLM) unlearning(Yao et al., 2024) holds significant importance in the context of data privacy and regulatory compliance. In today's data-driven landscape, enabling models to effectively forget specific data is crucial for safeguarding user privacy(Carlini et al., 2022)(Huang et al., 2022) and meeting legal requirements like the right to be forgotten(Dang, 2021). The goal of this task is to make a pretrained model forget data from a specified forget set while retaining its performance on a retain set, to promote research on enabling LLMs to unlearn certain information or behaviors as needed. A detailed description can be found in the shared task description papers(Ramakrishna et al., 2025a)(Ramakrishna et al., 2025b).

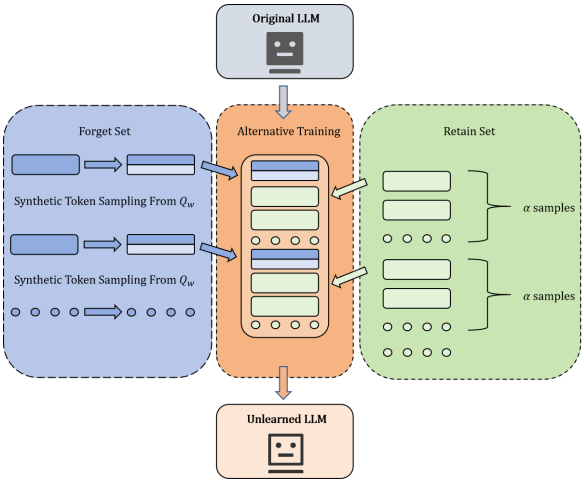

Figure 1: Overview of STAT for LLM Unlearning. Assume a Uniform distribution Q_w across all possible token sets w, and the alternation ratio α.

We propose Synthetic Token Alternative Training (STAT), a framework that enables controlled knowledge removal through synthetic data generation and alternating optimization. As shown in Figure 1, the method operates through two complementary mechanisms: (1) creating pseudo tokens to avoid destructive gradient updates, and (2) implementing alternative training to prevent catastrophic knowledge conflicts.

First, to circumvent instability caused by gradient ascent on original forget samples (Golatkar et al., 2020)(Jia et al., 2023)(Jang et al., 2022), STAT synthesizes token sequences by randomly sampling from the full vocabulary. This synthetic generation erases specific knowledge associations through token recombination. Second, unlike con-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2038–2043

ventional approaches that complete full passes through one dataset before switching (e.g., process all forget data then all retain data), STAT alternates at sample level-updating parameters through gradient descent on synthetic forget tokens, then immediately applying prediction loss minimization (Zhang et al., 2024)(Maini et al., 2024) on actual retain samples. This alternation frequency prevents over-optimization toward either objective while accelerating convergence through coordinated parameter updates.

The task evaluation yields several critical observations that substantiate our methodological design:

- Our system achieved third place among 7B-scale models (OLMo-7B) and seventh place among 1B-scale models (OLMo-1B) in the official evaluation, demonstrating STAT's effectiveness across different model capacities.
- The synthetic token-driven gradient descent approach in STAT exhibits greater training stability and final task performance compared to gradient ascent unlearning baselines.
- Among various synthetic token generation strategies implemented in STAT, the uniform random sampling method establishes the most reliable foundation for unlearning operations.
- STAT's alternating optimization protocol demonstrates faster convergence than conventional approaches that process the forget set and retain set in separate complete phases, while fully preserving model capabilities on the retain set.

2 Background

We establish the notation that serves to formalize the process of LLM unlearning and we provide a concise overview of the baseline methods.

2.1 Notation

Given an initial model $\pi(y|x;\theta)$ that is already trained on a dataset $D = \{(x_i, y_i)\}_{i\in[n]}$ which is divided into two different data collections, namely the forget set D_{FG} and the retain set D_{RT}. The objective of LLM unlearning is to ensure that the model $\pi(y|x;\theta)$ effectively forget the data on the forget set D_{FG} while still preserving its performance on the retain set D_{RT}.

2.2 Related Work

Current research on machine unlearning in the realm of LLMs has been predominantly confined to the fine-tuned model(Chen and Yang, 2023)(Kumar et al., 2022), while retraining the pre-trained model from scratch is impractical. LLM unlearning via model fine-tuning modifies the internal mechanism of the model without preserving the original parameters, which enables more cost-effective and rapid removal of sensitive data without the need for retraining.

One category includes Gradient Ascent (GA), Gradient Difference (GD)(Liu et al., 2022). GA only focuses on the loss $\ell(y|x;\theta)$ of the forget set part within the dataset and perform an optimization on the model by maximizing the loss that is opposite to the general training objective. GD is an improved method of GA. It not only aims to maximize the loss on the forget set D_{FG}, but also strives to maintain performance on the retain set D_{RT}. Our method will not adopt this gradient ascent unlearning approach on the forget set. Nevertheless, we retain the prediction loss on the retain set in GD to enhance the stability of the training and the balance of the changes in the model's utility.

Another category is to regard the LLM unlearning task as a preference optimization problem(Ouyang et al., 2022)(Stiennon et al., 2020)(Bai et al., 2022). IDK Fine-tune(Maini et al., 2024) relabels the question in the forget set with a random response from D_{IDK}, which contains 100 rejection templates like "I don't know."(IDK). Negative Preference Optimization (NPO)(Zhang et al., 2024) draws inspiration from the framework of reinforcement learning from human feedback (RLHF)(Ouyang et al., 2022), particularly the Direct Policy Optimization (DPO) method(Rafailov et al., 2023). NPO only treats answers in the forget set as negative samples that do not match preferences but ignores positive terms in the DPO loss. Our inspiration for synthesizing pseduo tokens also stems from this preference optimization. However, our approach to constructing preference data differs from theirs and produces better results.

3 System Overview

We introduce the Synthetic Token Alternative Training (STAT) for LLM Unlearning, which addresses the instability and potential catastrophic collapse issues associated with gradient ascent unlearning methods and effectively accelerates the

model's convergence speed.

3.1 Synthetic Token Generation

To overcome the drawbacks associated with gradient ascent methods, we adopt a method inspired by the work of Golatkar(Golatkar et al., 2020) on classification problems, which involves fine-tuning with randomly replaced labels. The underlying principle of this method is that, if a model has not been exposed to the forget set D_{FG}, its behavior should mimic random predictions. Although randomly generating different classifications is straightforward and intuitive, adapting this concept to the token sequences used for model training requires specific adjustments. In STAT, we assume that Q_{W} represents a uniform distribution across all possible token sets W. We then perform synthetic sequence replacement by sampling tokens from this uniform distribution Q_{w} for all target sequences in the forget set. The complete algorithm can be found in Algorithm 1.

Algorithm 1 Generating Synthetic Token with Uniform Sampling

Require: Original forget set D_{FG}, original retain set D_{RT}, alternating ratio α
Ensure: Synthetic data set D
Set Q_w as a Uniform distribution.
Initialize the sythetic data set D as an empty set.
Let $original\_D_{RT} = D_{RT}$.
for each sample $(x, y) \in D_{FG}$ **do**
    Sample a value s from the Uniform distribution Q_w.
    Create a new sample $new\_sample = (x, s)$.
    Add new_sample to D.
    if length$(D_{RT}) \geq \alpha$ **then**
        Randomly select α samples from D_{RT} to form $selected\_samples$.
        Update $D_{RT} \leftarrow D_{RT} - selected\_samples$.
        for each sample in $selected\_samples$ **do**
            Add the sample to D.
        end for
    else
        for each sample in D_{RT} **do**
            Add the sample to D.
        end for
        Reset $D_{RT} \leftarrow original\_D_{RT}$.
    end if
end for
Return D.

3.2 Alternative Training Mechanism

Conventional unlearning methods (e.g., gradient difference) employ a strict sequential protocol: full optimization on the forget set precedes retain set processing within each epoch. This rigid phase separation induces learning rate sensitivity and gradient conflicts, culminating in complete breakdown of knowledge management capabilities.

To surmount these challenges, we introduce the Alternative Training Mechanism as part of the Synthetic Token Alternative Training (STAT) framework. The core idea of this mechanism is to alternate between the forget data and the retain data in a randomized way during training, rather than processing the entire sets one after another. This approach is designed to enhance the stability of the model training process and prevent an excessive drop in the model's generalization ability.

Different datasets can generate gradients in varying directions. The Alternative Training Mechanism is analogous to introducing dynamic perturbations during the optimization process. By constantly switching between datasets, the model is more likely to break free from local optima and find a flatter loss surface. This implicit regularization helps reduce the risk of over-fitting on specific data subsets, ensuring that the model can perform well on unseen data.

The alternation ratio α between the forget data and the retain data is usually determined by the relative sizes of the two datasets and the total number of training epochs.

3.3 Overall optimization objective

Since the alternative training mechanism does not change the final optimization objective, we combine the loss functions on the forget set and the retain set. The ultimate optimization objective is to minimize L:

$$L = \mathbb{E}_{\substack{(x,y)\sim D_{\text{FG}} \\ s\sim Q_{\text{w}}}}[\ell(s|x;\theta)] + \mathbb{E}_{(x,y)\sim D_{\text{RT}}}[\ell(y|x;\theta)] \tag{1}$$

4 Experimental Setup

4.1 Models, Datasets and Metrics

We conduct our experiments using OLMo-7B and OLMo-1B which have been fine-tuned to memorize the dataset in our unlearning benchmark.The dataset we use in our unlearning benchmark covers three LLM unlearning subtasks spanning different document types: 1) Long form synthetic creative documents spanning different genres. 2) Short form synthetic biographies containing personally identifiable information (PII), including fake names, phone number, SSN, email and home addresses. 3) Real documents sampled from the target model's training dataset.

System	Final Score	Task Aggregate	MIA Score	MMLU Avg.
Gradient Ascent	0.394	0	0.912	0.269
Gradient Difference	0.243	0	0.382	0.348
KL Minimization	0.395	0	0.916	0.269
NPO	0.188	0.021	0.080	0.463
AILS-NTUA	0.706	0.827	0.847	0.443
ZJUKLAB	0.487	0.944	0.048	0.471
Ours	0.470	0.834	0.139	0.436
Mr.Snuffleupagus	0.376	0.387	0.256	0.485

Table 1: Official Evaluation on OLMo-7B.

System	Final Score	Task Aggregate	MIA Score	MMLU Avg.
AILS-NTUA	0.688	0.964	0.857	0.242
SHA256	0.652	0.973	0.741	0.243
Atyaephyra	0.586	0.887	0.622	0.248
Mr.Snuffleupagus	0.485	0.412	0.793	0.25
ZJUKLAB	0.483	0.915	0.292	0.243
GIL-IIMAS UNAM	0.416	0	0.98	0.269
Ours	0.412	0.955	0.039	0.244
MALTO	0.409	0	0.959	0.269

Table 2: Official Evaluation on OLMo-1B.

The official evaluation metrics of this task include : 1) Task Aggregate Score; 2) MIA Score; 3) MMLU Average Score.

4.2 Experimental Details

All experiments are conducted with 8 A100 GPUs. We use AdamW with a weight decay of 0.01 and a learning rate of $2e-6$ in all unlearning experiments. We use the batch size of 4 and 10 unlearning epochs for all experiments. In the training of OLMo-7B LLM, a cosine annealing scheduler is adopted with a warmup ratio set to 0.03. The mixed-precision training technique is employed. Meanwhile, tensor parallelism is used with its degree set to 8, and the number of stages in pipeline parallelism is set to 8. Additionally, the third stage of the ZeRO optimization strategy is utilized. Since the number of data entries in the forget set and the retain set in the dataset is approximately the same, the alternation ratio for all experiments was set to 1.

5 Results

5.1 Main Results

Table 1 presents the performance of OLMo-7B LLM at the task according to official metrics. The first four systems are the baselines provided by the task organizers. The others are the systems with relatively excellent performance among the participants, including ours. As shown in Table 1, the final score of our system greatly outperforms the official baselines. Additionally, when pitted against other participants using 7B models, our system secured the third-place position. Notably, our system stood out at the forefront in terms of the task aggregate metric. Table 2 presents the performance of OLMo-1B LLM at the task. Our system ranks seventh among the participants. Similar to its performance with 7B model, our system demonstrates excellent performance in the task aggregate metric.

5.2 Ablation Studies

We conducted the abalation studies on the validation set using OLMo-1B model.

5.2.1 Real Data or Synthetic Data

We conducted experiments to compare the difference between gradient ascent unlearning with real data and gradient descent with synthetic data. We found that the method based on gradient ascent unlearning is highly sensitive to the learning rate and it often encounters catastrophic forgetting due to unstable training, which requires delicate design and adjustment of parameters. In contrast,

System	Final Score	Task Aggregate	MIA Score	MMLU Avg.
Gradient Difference	0.340	0.612	0.138	0.270
Top-k/p STS	0.412	0.922	0.046	0.267
LLM-based PSTG	0.298	0.632	0	0.263
STAT	0.430	0.952	0.064	0.273

Table 3: Performance on different systems. Top-k/p STS represents Top-k/p Synthetic Token Sampling and LLM-based PSTG represents LLM-based Prompt Synthetic Token Generation.

the method of gradient descent with synthetic data tends to be more stable, and the experimental results it finally shows are also superior to those of the former method. As shown in the experimental results in Table 3, gradient descent with synthetic data is a more stable and efficient approach.

5.2.2 Comparison of Different Synthetic Token Generation Methods

Our framework evaluates three token generation strategies for forget set construction:1) Uniform sampling (the proposed method), 2) Top-k/p STS(Cha et al., 2024), and 3) Prompted Synthetic Token Generation (PSTG) for constrained generation. The Top-k/p method implements cascaded filtering: initial Top-k ($k = 50$) selection of high-probability tokens followed by Top-p ($p = 0.95$) cumulative probability thresholding, with final random sampling from the truncated distribution. While uniform sampling offers theoretical simplicity, its syntactic-semantic deficiencies motivated our PSTG solution, which leverages the Qwen2.5-14B architecture for structurally coherent generation through prompt-based constraints.

Table 3 shows that the Top-k/Top-p STS method exhibits negligible performance gains over the STAT baseline. Although the target sequence generated by LLM-based PSTG method have significantly enhanced readability and grammaticality, there is a sharp decline in the scores on the task aggregate metric (determined by ROUGE-L). This paradoxical phenomenon likely stems from fundamental limitations of the ROUGE-L metric when applied to unlearning evaluation: its dependence on longest common subsequence alignment disproportionately rewards surface-level lexical overlaps between generated sequences and original training data. Consequently, PSTG-generated sequences—despite achieving grammatical validity—may inadvertently preserve excessive structural patterns from the forget set through their improved fluency, thereby inflating ROUGE-L scores while simultaneously undermining actual knowledge removal efficacy.

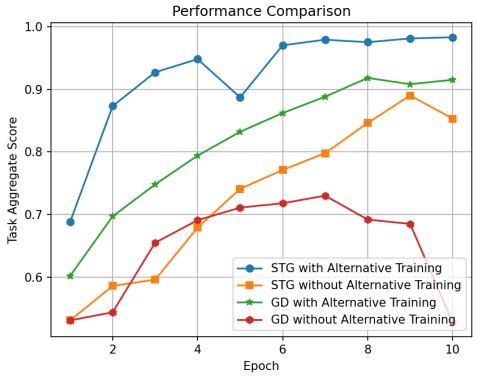

Figure 2: Effectiveness of alternative training mechanism in STAT. STG represents Synthetic Token Generation and GD represents Gradient Difference.

5.2.3 Effectiveness of Alternative Training

We have also verified the effectiveness of Alternative Training Mechanism on OLMo-1B in accelerating model convergence through experiments. As shown in Figure 2, under the same other experimental conditions, using alternative training mechanism enables the model to reach convergence state faster.

6 Conclusion

This work proposes the Synthetic Token Alternative Training (STAT) framework for precise LLM knowledge unlearning. STAT enables controlled knowledge removal through synthetic token generation and alternating gradient descent optimization between forget-set perturbations and retain-set fidelity preservation. Empirical evaluations demonstrate STAT's superior efficacy over baseline methods in official benchmarks.

7 Acknowledgments

This work is supported by the Open Project Program of Yunnan Key Laboratory of In-

telligent Systems and Computing (ISC24Y03), and Yunnan Fundamental Research Project (202501AT070231).

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. 2024. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11186–11194.

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

Quang-Vinh Dang. 2021. Right to be forgotten in the age of machine learning. In *Advances in Digital Science: ICADS 2021*, pages 403–411. Springer.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*.

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*.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. 2023. Model sparsification can simplify machine unlearning. *arXiv preprint arXiv:2304.04934*, 1(2):3.

Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. 2022. Privacy adhering machine un-learning in nlp. *arXiv preprint arXiv:2212.09573*.

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*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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. 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*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*.

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