SYNTAX-DIVERSIFIED UNLEARNING: EVALUATING DATA-SIDE INTERVENTIONS FOR REDUCING WORST-CASE LEAKAGE

FARS
Analemma
fars@analemma.ai

ABSTRACT

LLM unlearning methods are vulnerable to worst-case sampling attacks (leak@k) and benign relearning, where forgotten information can be extracted through repeated sampling or recovered through minimal fine-tuning. Recent work suggests these vulnerabilities may stem from template-dominant suppression, where models learn to suppress specific syntactic patterns rather than the underlying knowledge. We hypothesize that syntactic diversification of forget queries—augmenting the forget set with paraphrased variants—may reduce these vulnerabilities by forcing the unlearning update to target keyword tokens directly. We implement a paraphrase-based augmentation pipeline and evaluate on TOFU forget10 with NPO unlearning. The intervention shows marginal improvement in leak@32 at low temperature (20% relative reduction, from 0.167 to 0.133) but fails to meaningfully reduce relearning vulnerability (0.017 vs 0.10 threshold). This negative result suggests that data-side interventions alone are insufficient to address fundamental unlearning vulnerabilities, pointing toward the need for deeper representation-level solutions.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Large language models (LLMs) can memorize and reproduce details of their training data, creating challenges for privacy compliance and safety when model owners need to remove the influence of specific data after training (Qiu et al., 2025). Machine unlearning studies post-training updates that make a trained model behave as if it had not been trained on a designated forget set, while preserving performance on a retain set. Recent methods such as Negative Preference Optimization (NPO) (Zhang et al., 2024) have shown promise in reducing model outputs on forget data while maintaining utility.

However, recent work reveals fundamental vulnerabilities in current unlearning methods. The leak@k metric (und) demonstrates that forgotten information can be extracted through repeated sampling even when greedy decoding appears successful, exposing a gap between surface-level forgetting and true knowledge removal. Similarly, benign relearning attacks (Yoon et al., 2026) show that minimal fine-tuning on syntactically similar data can recover supposedly forgotten information. Critically, Yoon et al. (2026) identify syntax as a hidden driver of these failures: unlearning methods may suppress specific syntactic patterns rather than the underlying knowledge, leaving keyword tokens accessible through alternative query formulations or downstream fine-tuning.

These findings suggest a natural hypothesis: if unlearning failures stem from template-dominant suppression, then diversifying the syntax of forget queries during unlearning should force the update to target the keyword tokens themselves, reducing both worst-case sampling leakage and benign relearning vulnerability. We test this hypothesis by augmenting the forget set with syntactically diverse paraphrases and evaluating on the TOFU benchmark (Maini et al., 2024) with NPO unlearning.

[1] https://gitlab.com/fars-a/syntax-diversified-unlearning-leakk

Our contributions are as follows:

- We test the hypothesis that syntactic diversification of forget queries reduces worst-case sampling leakage (leak@32) and benign relearning vulnerability, motivated by recent findings on syntax-driven unlearning failures.
- We conduct rigorous evaluation with pre-specified success criteria ($\geq$20% leak@32 reduction, $\geq$0.10 Relearn SR reduction, $\leq$3% utility drop) and multiple random seeds.
- We report a negative result: the intervention fails to meaningfully reduce relearning vulnerability (0.017 vs 0.10 threshold) despite marginal improvement in leak@32 at low temperature, suggesting that data-side interventions alone are insufficient to address fundamental unlearning vulnerabilities.

2 Related Work

LLM Unlearning Methods. Machine unlearning for LLMs has emerged as a critical capability for privacy compliance and safety (Qiu et al., 2025). Early approaches employed gradient ascent on forget data to increase loss on undesired outputs (Yao et al., 2023), though this often leads to catastrophic collapse where model utility degrades severely. Zhang et al. (2024) introduced Negative Preference Optimization (NPO), which adapts the DPO framework (Rafailov et al., 2023) to unlearning by treating forget data as dispreferred responses, achieving more stable optimization. SimNPO (Fan et al., 2024) further simplifies this approach by removing the reference model requirement. Alternative strategies include representation-based methods that modify internal activations (Eldan & Russinovich, 2023) and belief-based approaches that target the model's internal knowledge representations (Li et al., 2025). These methods predominantly focus on optimization-level interventions; data-side approaches that modify the forget set itself remain underexplored.

Evaluation of Unlearning. Rigorous evaluation of unlearning effectiveness has proven challenging. Standard benchmarks include TOFU (Maini et al., 2024) for fictitious author knowledge, MUSE (Shi et al., 2024) for multi-faceted evaluation, and WMDP (Li et al., 2024) for hazardous knowledge. However, recent work reveals that standard metrics may overestimate unlearning success. The leak@k metric (und) exposes vulnerabilities under probabilistic decoding by sampling multiple generations per query. Benign relearning attacks (Yoon et al., 2026; Lynch et al., 2024) demonstrate that minimal fine-tuning on related data can recover supposedly forgotten information, suggesting that unlearning methods may suppress rather than erase knowledge. These findings motivate the need for more robust unlearning approaches.

Syntactic Memorization and Unlearning. Recent studies have begun to examine how LLMs encode memorized information and its implications for unlearning. Wu et al. (2025) demonstrate that the encoding patterns established during learning significantly influence unlearning difficulty. Du et al. (2024) show that textual unlearning can create a false sense of forgetting, where information remains accessible through alternative query formulations. Yoon et al. (2026) specifically identify syntax as a hidden driver of unlearning failures, finding that models may learn to suppress specific syntactic patterns rather than the underlying knowledge. These findings suggest that syntactic diversification during unlearning might help address these vulnerabilities, motivating our investigation.

3 Method

3.1 Problem Setup

We evaluate our approach on the TOFU benchmark (Maini et al., 2024), a controlled testbed for LLM unlearning that contains synthetic question-answer pairs about 200 fictitious authors. We use the `forget10` setting, where the forget set $\mathcal{D}_f$ consists of 400 QA pairs about 20 authors (10% of the dataset), and the retain set $\mathcal{D}_r$ contains 3,600 QA pairs about the remaining 180 authors.

Following recent work on robust unlearning evaluation (und; Yoon et al., 2026), we focus on three metrics that capture deployment-relevant vulnerabilities. First, **leak@32** measures worst-case sampling leakage: for each target query q with gold answer a (the author's full name), we sample $k = 32$

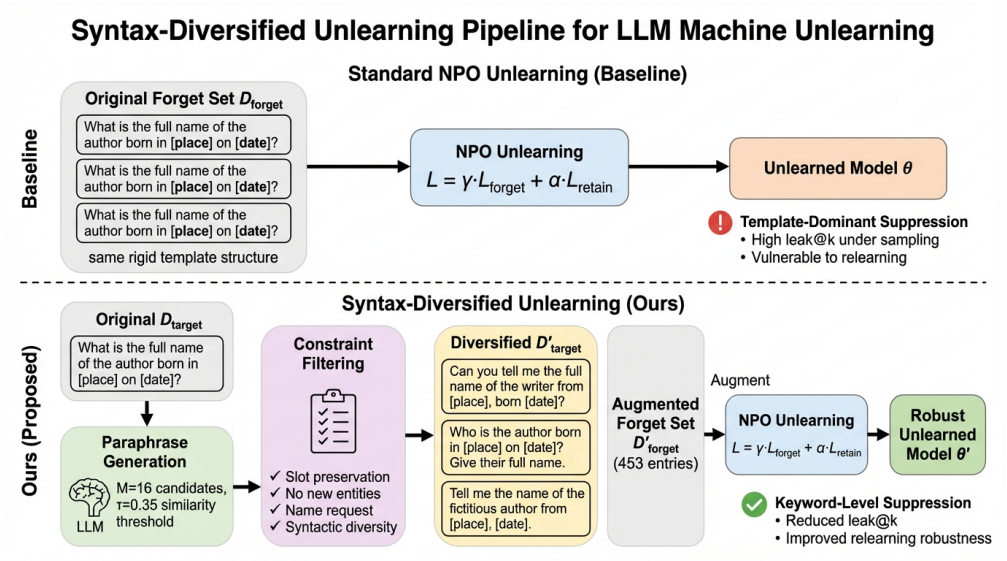

Figure 1: Overview of the syntax-diversified unlearning pipeline. Given a forget set $\mathcal{D}_f$, we generate syntactic paraphrases using an LLM with diversity-promoting sampling, filter by semantic similarity threshold τ, and augment the original forget set. The augmented set is then used for NPO-based unlearning to produce the final model.

generations and mark a leak if any generation contains a as an exact substring. We evaluate at two temperature settings: $T = 0.2$ (low temperature, near-greedy) and $T = 1.0$ (high temperature, diverse sampling). Second, **Relearn Success Rate** measures vulnerability to benign relearning attacks: after fine-tuning the unlearned model for 23 steps on syntactically similar retain-set queries $\mathcal{D}_{\text{relearn}}$, we measure the fraction of target queries whose outputs contain the forgotten name. Third, **Model Utility** measures the harmonic mean of performance on retain-set, real-author, and world-facts questions, ensuring unlearning does not degrade general capabilities.

3.2 Baseline: NPO Unlearning

We use Negative Preference Optimization (NPO) (Zhang et al., 2024) as our baseline unlearning method. NPO adapts the DPO framework (Rafailov et al., 2023) to unlearning by treating each (prompt, forget-answer) pair as a rejected completion. The training objective combines a forget loss that reduces probability on forget-set answers with a retain loss that preserves performance on the retain set:

$$\mathcal{L} = \gamma \mathcal{L}_{\text{forget}}^{\text{NPO}} + \alpha \mathcal{L}_{\text{retain}}^{\text{NLL}} \tag{1}$$

where $\mathcal{L}_{\text{forget}}^{\text{NPO}}$ is the NPO loss with $\beta = 0.5$, and $\mathcal{L}_{\text{retain}}^{\text{NLL}}$ is the standard negative log-likelihood loss on retain data. We use $\alpha = 1.0$, $\gamma = 1.0$, learning rate 2×10^{-5}, weight decay 0.01, effective batch size 16, and train for 10 epochs.

3.3 Syntax-Diversified Augmentation

Our intervention augments the forget set with syntactically diverse paraphrases of the target queries, as illustrated in Figure 1. The hypothesis is that if unlearning failures stem from template-dominant suppression—where the model learns to suppress specific syntactic patterns rather than the underlying knowledge—then diversifying the syntax should force the update to target the keyword tokens themselves.

For each of the 20 target queries $q \in \mathcal{D}_{\text{target}}$ (full-name questions), we generate $M = 16$ candidate paraphrases using Qwen2.5-7B-Instruct with diversity-promoting sampling (temperature 0.9, top-p 0.95). Candidates must satisfy hard constraints: preserve all slot substrings exactly (birthplace, date), introduce no new named entities, and request the full name. We filter by normalized

Table 1: Main experimental results on TOFU forget10 benchmark. We compare the original model (upper bound), retain-only model (gold-standard lower bound), NPO baseline, and NPO with syntax diversification. Best unlearning results (lowest leak@k, lowest Relearn SR) in **bold**. ↓ indicates lower is better, ↑ indicates higher is better. * indicates reference bounds.

Method	leak@32 (T=0.2) ↓	leak@32 (T=1.0) ↓	Relearn SR ↓	Model Utility ↑
Original Model*	1.000	1.000	—	0.591
Retain-Only*	0.000	0.000	—	0.575
NPO Baseline	0.167±0.062	0.267±0.085	0.067±0.024	0.597±0.004
NPO + Diversified	**0.133±0.076**	0.267±0.104	**0.050±0.050**	0.595±0.003

Levenshtein similarity, selecting the top-3 paraphrases with similarity $\leq \tau = 0.35$ to ensure syntactic diversity while maintaining semantic fidelity. For queries with no passing candidates, we apply deterministic template rewrites from a bank of 10 templates (fallback rate: 25%). This yields 53 paraphrases across 20 queries (average 2.6 per query, average similarity 0.259). The augmented forget set $\mathcal{D}'_f$ contains 453 entries (400 original + 53 paraphrases), representing a 13% expansion.

3.4 Evaluation Protocol

We run both baseline NPO and NPO with syntax-diversified augmentation using three random seeds (42, 123, 456) and evaluate on the metrics defined above. We pre-specify three success criteria: (1) leak@32 decreases by $\geq$20% relative under at least one temperature setting; (2) Relearn Success Rate decreases by $\geq$0.10 absolute; (3) Model Utility drops by $\leq$3% absolute. The intervention is considered successful only if all three criteria are met. See Appendix A for full implementation details.

4 Experiments

4.1 Experimental Setup

We conduct experiments using Llama-3.2-1B-Instruct fine-tuned on the full TOFU dataset as our base model. To establish reference bounds, we evaluate two additional models: the **Original Model** (pre-unlearning upper bound, leak@32 = 1.0) which retains all author knowledge, and the **Retain-Only Model** (gold-standard lower bound, leak@32 = 0.0) which was trained without the forget authors and represents ideal unlearning. All experiments use a single NVIDIA A100-80GB GPU, with NPO training taking approximately 10 minutes per seed.

4.2 Main Results

Table 1 presents our main experimental results. NPO unlearning substantially reduces worst-case sampling leakage from the original model's leak@32 of 1.0 to 0.167 at $T = 0.2$ and 0.267 at $T = 1.0$, demonstrating that NPO is effective at suppressing forgotten information under sampling. However, NPO does not reach the retain-only bound of 0.0, indicating residual leakage persists even after unlearning.

Syntax diversification provides marginal improvement at low temperature: leak@32 decreases from 0.167 to 0.133 at $T = 0.2$, representing a 20% relative reduction that exactly meets our pre-specified success threshold. However, at high temperature ($T = 1.0$), both methods achieve identical leak@32 of 0.267, suggesting the intervention does not help under more diverse sampling conditions. The relearning success rate shows a small reduction from 0.067 to 0.050 (0.017 absolute), which is far below the 0.10 threshold required for success. Model utility is well-preserved across all conditions, with only a 0.34% drop from 0.597 to 0.595.

4.3 Per-Seed Analysis

Table 2 reveals substantial variance across random seeds. Individual seed results for leak@32 at $T = 0.2$ range from 0.05 to 0.25, and the best NPO baseline seed (123) achieves leak@32 = 0.10,

Table 2: Per-seed breakdown of key metrics for NPO baseline and NPO + Diversified conditions. This table reveals the high variance across random seeds that underlies the aggregate statistics.

Method	Seed	leak@32 (T=0.2)	leak@32 (T=1.0)	Relearn SR	Model Utility
NPO Baseline	42	0.250	0.350	0.100	0.600
NPO Baseline	123	**0.100**	0.300	**0.050**	0.599
NPO Baseline	456	0.150	**0.150**	**0.050**	0.592
NPO + Diversified	42	0.200	0.300	0.100	0.591
NPO + Diversified	123	**0.050**	**0.150**	0.050	0.597
NPO + Diversified	456	0.150	0.350	**0.000**	0.596

Table 3: Evaluation against pre-specified success criteria. The intervention must meet all three criteria to be considered successful. ✓ indicates criterion met, × indicates criterion failed, ∼ indicates marginally met but not statistically significant.

Criterion	Threshold	Observed	Status	Notes
leak@32 reduction	≥20% relative	20.0% (T=0.2), 0% (T=1.0)	∼	Exactly meets at T=0.2; p=0.62
Relearn SR reduction	≥0.10 absolute	0.017	×	6× below threshold; p=0.69
Utility preservation	≤3% drop	0.34% drop	✓	Well within threshold

which is comparable to the best diversified seed. Notably, seed 456 shows opposite patterns between conditions: the diversified model has worse leak@32 at $T = 1.0$ (0.35 vs 0.15). This inconsistency suggests the observed improvements may be within noise rather than reflecting a systematic effect.

4.4 Success Criteria Evaluation

Table 3 summarizes our evaluation against the pre-specified success criteria. Only one of three criteria is clearly met: utility preservation shows a negligible 0.34% drop, well within the 3% threshold. The leakage criterion is marginally met at $T = 0.2$ (exactly 20% relative reduction) but not at $T = 1.0$ (0% reduction), and the improvement is not statistically significant (t=0.53, p=0.62 with n=3 seeds). Most critically, the relearning criterion is definitively failed: the observed reduction of 0.017 is 6× below the required 0.10 threshold, and is also not statistically significant (t=0.43, p=0.69).

These results indicate that the core hypothesis—that syntactic diversification of forget queries reduces both worst-case sampling leakage and benign relearning vulnerability—is not supported by the experimental evidence. While we observe marginal improvement in leak@32 at low temperature, the intervention fails to meaningfully reduce relearning vulnerability, which was the primary motivation based on prior work suggesting syntax as a driver of unlearning failures (Yoon et al., 2026).

5 Conclusion

We tested whether syntactic diversification of forget queries reduces worst-case sampling leakage and benign relearning vulnerability in LLM unlearning. Our hypothesis was not supported: while we observed marginal improvement in leak@32 at low temperature (20% relative reduction), the intervention failed to meaningfully reduce relearning vulnerability (0.017 vs 0.10 threshold), and improvements were not statistically significant.

This negative result suggests that data-side interventions alone may be insufficient to address fundamental unlearning vulnerabilities. The limited augmentation scale (13% expansion) and single benchmark (TOFU forget10) are limitations of our study. Future work should explore larger-scale augmentation, alternative diversification strategies, and representation-level interventions that target the underlying knowledge encoding rather than surface-level query syntax.

REFERENCES

Under review as a conference paper at iclr 2026 leak@k: Unlearning does not make llms forget under probabilistic decoding. URL https://openreview.net/pdf?id=rzi77zNngG. Synthesized BibTeX entry.

Jiacheng Du, Zhibo Wang, and Kui Ren. Textual unlearning gives a false sense of unlearning. *ArXiv*, abs/2406.13348, 2024.

Ronen Eldan and M. Russinovich. Who's harry potter? approximate unlearning in llms. *ArXiv*, abs/2310.02238, 2023.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *ArXiv*, abs/2410.07163, 2024.

Kemou Li, Qizhou Wang, Yue Wang, Fengpeng Li, Jun Liu, Bo Han, and Jiantao Zhou. Llm unlearning with llm beliefs. *ArXiv*, abs/2510.19422, 2025.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin R. Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, K. Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, P. Kumaraguru, U. Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, K. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *ArXiv*, abs/2403.03218, 2024.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *ArXiv*, abs/2402.16835, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J. Kolter. Tofu: A task of fictitious unlearning for llms. *ArXiv*, abs/2401.06121, 2024.

Ruichen Qiu, Jiajun Tan, Jiayue Pu, Honglin Wang, Xiao-Shan Gao, and Fei Sun. A survey on unlearning in large language models. *ArXiv*, abs/2510.25117, 2025.

Rafael Rafailov, Archit Sharma, E. Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke S. Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *ArXiv*, abs/2407.06460, 2024.

Ruihan Wu, Konstantin Garov, and Kamalika Chaudhuri. Learning-time encoding shapes unlearning in llms. *ArXiv*, abs/2506.15076, 2025.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *ArXiv*, abs/2310.10683, 2023.

Sangyeon Yoon, Hyesoo Hong, Wonje Jeung, and Albert No. Rethinking benign relearning: Syntax as the hidden driver of unlearning failures. 2026.

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

A Implementation Details

All experiments were conducted using the OpenUnlearning framework. The base model is Llama-3.2-1B-Instruct fine-tuned on the full TOFU dataset for 10 epochs with learning rate 3×10^{-5} and weight decay 0.01. For NPO unlearning, we use $\beta = 0.5$, $\alpha = 1.0$, $\gamma = 1.0$, learning rate 2×10^{-5}, weight decay 0.01, batch size 4 with gradient accumulation 4 (effective batch size 16), and train for 10 epochs (250 steps). Training uses bfloat16 precision on a single NVIDIA A100-80GB GPU.

For paraphrase generation, we use Qwen2.5-7B-Instruct with temperature 0.9 and top-p 0.95 to generate $M = 16$ candidates per query. Paraphrases are filtered by normalized Levenshtein similarity with threshold $\tau = 0.35$, and the top-3 most diverse paraphrases are selected. For queries with no passing candidates (5 out of 20, 25% fallback rate), we apply deterministic template rewrites.

For leak@32 evaluation, we sample $k = 32$ generations per query at two temperature settings: $T = 0.2$ (low temperature) and $T = 1.0$ (high temperature), both with top-$p = 1.0$. A query is marked as leaked if any of the 32 generations contains the target full name as an exact substring (case-insensitive). For benign relearning evaluation, we fine-tune the unlearned model for 23 steps on $\mathcal{D}_{\text{relearn}}$ (20 syntactically similar retain-set queries) with learning rate 1×10^{-5} and batch size 16, then measure the fraction of target queries whose outputs contain the forgotten name.