Maximizing Local Entropy Where It Matters: Prefix-Aware Localized LLM Unlearning

Naixin Zhai[1], **Pengyang Shao**[2,*],
Binbin Zheng[1], **Yonghui Yang**[2], **Fei Shen**[2], **Long Bai**[2], **Xun Yang**[1,*]
[1] University of Science and Technology of China
[2] National University of Singapore
zhainaixin@mail.ustc.edu.cn, shaopymark@gmail.com, xyang21@ustc.edu.cn

Abstract

Machine unlearning aims to forget sensitive knowledge from Large Language Models (LLMs) while maintaining general utility. However, existing approaches typically treat all tokens in a response indiscriminately and enforce uncertainty over the entire vocabulary. This global treatment results in unnecessary utility degradation and extends optimization to content-agnostic regions. To address these limitations, we propose PALU (**P**refix-**A**ware **L**ocalized **U**nlearning), a framework driven by a local entropy maximization objective across both temporal and vocabulary dimensions. PALU reveals that (i) suppressing the sensitive prefix alone is sufficient to sever the causal generation link, and (ii) flattening only the top-K logits is adequate to maximize uncertainty in the critical subspace. These findings allow PALU to alleviate redundant optimization across the full vocabulary and parameter space while minimizing collateral damage to general model performance. Comprehensive evaluations validate that PALU achieves superior forgetting efficacy and utility preservation compared to state-of-the-art baselines. Our code is available at https://github.com/nxZhai/PALU.

1 Introduction

Large language models have been widely deployed across diverse domains (Hu et al., 2024; Xu et al., 2025; Han et al., 2025), yet they inevitably memorize sensitive, private, and copyrighted information from massive training corpora (Luo et al., 2025; Li et al., 2025; Fang et al., 2025; Zhu et al., 2026). Such memorization not only raises security and ethical concerns (Karamolegkou et al., 2023; Zhao et al., 2024b,a; Pan et al., 2025), but also conflicts with data privacy regulations such as GDPR[1] and CCPA[2], which grant individuals the "right to be forgotten". Consequently, machine unlearning, which selectively removes targeted information from trained models without retraining from scratch, has emerged as a prerequisite for the safe and compliant deployment of LLMs (Yao et al., 2024; Tirumala et al., 2022; Goel et al., 2026).

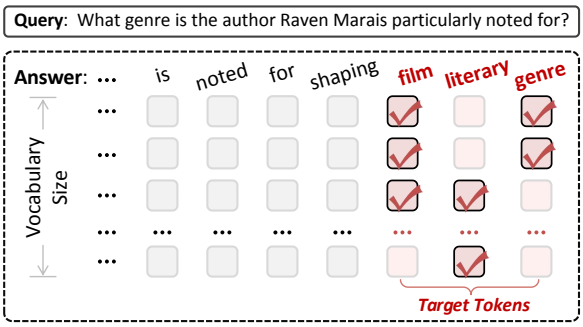

Figure 1: Illustration of the vocabulary-localized optimization. We specifically target sensitive tokens (red) while bypassing context-agnostic ones (gray). For each target position, the optimization is restricted to the top-K vocabulary candidates (indicated by ✓), thereby pruning the computation on long-tail dimensions.

Despite growing progress, current LLM unlearning methods remain largely grounded in variants of the negated cross-entropy (CE) objective. Representative approaches, such as GradientAscent (GA) (Yao et al., 2024) and Negative Preference Optimization (NPO) (Zhang et al., 2024), primarily aim to suppress the probability of the top-1 token. While intuitive, negated CE often leads to over-correction compared to entropy maximization, which naturally promotes uniform uncertainty without aggressively destroying contextual knowledge (Entesari et al., 2025; Pan et al., 2024; Zhao et al., 2025). Beyond this objective-level limitation, these methods share a common structural drawback: they induce global interventions by applying dense gradients across the full response

*Corresponding authors.
[1] https://gdpr-info.eu/
[2] https://oag.ca.gov/privacy/ccpa

sequence and a large portion of the vocabulary. This global reshaping can inadvertently suppress content-agnostic functional words, disrupt linguistic coherence, and degrade general utility (Chen and Yang, 2023; Ji et al., 2024). It also incurs substantial computational overhead, as optimization must backpropagate through all token positions and vocabulary dimensions, irrespective of their relevance to the sensitive content.

In this work, we revisit LLM unlearning through the lens of intervention efficiency: achieving effective forgetting with the minimal necessary perturbation to model parameters. From this perspective, unlearning can be viewed as disrupting the generation trajectory that leads from a query x to an undesired response y. Crucially, such disruption need not be global, motivating two key observations. **(i) Temporal Sparsity:** Sensitive semantics are typically triggered by a small prefix of pivotal tokens. Intervening on this initiating prefix is often sufficient to divert the generation path, making updates to subsequent tokens redundant. **(ii) Vocabulary Sparsity:** In most memorization scenarios, autoregressive decoding decisions are dominated by a small set of high-probability candidates, rendering interventions on long-tail vocabulary dimensions largely unnecessary. Flattening only the dominant logits can already induce strong confusion and divert the decoding path.

Guided by these insights, we propose PALU, a cost-efficient unlearning framework built upon a dual-sided localized entropy maximization objective, localized in both time and vocabulary. PALU first efficiently identifies a sensitive decoding prefix. For these selected positions, it then applies a localized loss restricted to the top-K logits, as illustrated in Figure 1. This targeted intervention maximizes predictive uncertainty within the decoding-critical subspace, effectively steering generation away from the sensitive trajectory while pruning redundant updates on irrelevant tokens and long-tail vocabulary dimensions. Extensive experiments show that PALU achieves state-of-the-art forgetting efficacy while significantly improving utility preservation compared to strong baselines. Our contributions are summarized as follows:

(1) We revisit LLM unlearning through the lens of intervention efficiency and formulate it as disrupting the sensitive generation trajectory with minimal necessary intervention.

(2) We propose PALU, which introduces a dual-sided localized entropy maximization objective, thereby reducing redundant computation and collateral degradation.

(3) Empirical results demonstrate that PALU sets a new standard for the trade-off between forgetting quality and utility preservation.

2 Related Work

2.1 LLM Unlearning

LLM unlearning focuses on removing targeted knowledge from LLMs while preserving general utility. Early methods such as GradientAscent (GA) and GradientDiff (GD) (Maini et al., 2024) maximize the loss on forget samples, typically via the negated CE. However, such unbounded objectives often cause unstable updates, e.g., excessive probability suppression and over-refusal during generation (Zhang et al., 2024). NPO (Zhang et al., 2024) introduces bounded objectives anchored to a reference model; SimNPO (Fan et al., 2025) removes reference-model bias for efficiency, and AltPO (Mekala et al., 2025) incorporates positive feedback to reduce nonsensical refusals. Further studies improve the trade-off between stability and utility by better refining negated CE, e.g., token saturation reweighting (SatImp) (Yang et al., 2025).

Recently, token-level LLM unlearning, which intervenes on a subset of tokens instead of suppressing entire sequences, has been widely studied (Wang et al., 2025a; Liu et al., 2025a), with examples including Selective Unlearning (SU) (Wan et al., 2025) and Targeted Preference Optimization (TPO) (Zhou et al., 2026). These methods share a common goal: minimizing unnecessary perturbations for unlearning. However, while these approaches achieve temporal sparsity, they typically overlook redundancy in the vocabulary space. By still computing gradients across the entire output distribution, they continue to rely on expensive dense updates and inherit the limitations of the negated CE objective.

2.2 Entropy for Machine Learning

Entropy has been widely adopted as a principled objective for controlling model behavior in machine learning. The maximum entropy principle advocates selecting the least-committal distribution subject to constraints. This principle motivates entropy maximization as a generic learning objective (Jaynes, 1957; Berger et al., 1996; Wang et al., 2025b; Liu et al., 2026). Prominent applications include entropy regularization to discourage

over-confident predictions and improve generalization (Jiang et al., 2025), and entropy maximization for reinforcement learning to encourage exploration and robustness (Chao et al., 2024; Cheng et al., 2026; Fu et al., 2026; Zhang et al., 2025a). However, maximizing entropy over the full output space is not always necessary. In many problems, the final prediction is dominated by a small subset of high-probability candidates, while long-tail dimensions contribute little (Gao et al., 2019; Holtzman et al., 2019). This motivates localized entropy maximization, which increases uncertainty only within the prediction-critical subspace (e.g., top-ranked candidates) (Michaud et al., 2023; Entesari et al., 2025), without enforcing global distributional flattening. Building on this intuition, we propose PALU, which leverages localized entropy maximization to achieve intervention efficiency in unlearning, thereby offering a robust alternative to standard suppressive objectives. Unlike negated CE, which purely suppresses probability, entropy maximization naturally induces uniform uncertainty, allowing us to erase knowledge with minimal necessary perturbation.

3 Preliminary

Most existing gradient-based LLM unlearning methods can be formulated as optimizing two conflicting objectives (Yao et al., 2024; Si et al., 2023; Shao et al., 2026):

$$\min_{\theta} \mathcal{L}_{\text{all}} = \mathcal{L}_f + \lambda \mathcal{L}_r, \tag{1}$$

where the retain loss $\mathcal{L}_r$ is typically instantiated as the standard CE loss for next-token prediction on the retain set $\mathcal{D}_r$. In contrast, the forget objective $\mathcal{L}_f$ aims to disrupt the learned mapping from the input x to the target output y on the forget set by explicitly suppressing the likelihood of generating y conditioned on x. Optimizing $\mathcal{L}_f$ may introduce unnecessary perturbations to model parameters, potentially degrading the model's general capabilities, while $\mathcal{L}_r$ serves as an indirect constraint that mitigates such degradation by maintaining performance on the retain distribution. A common instantiation of $\mathcal{L}_f$ is the negated CE objective (Liu et al., 2025b; Zhang et al., 2024), defined as:

$$\mathcal{L}_f = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\sum_{t=1}^{T} \log p(y_t \mid x, y_{<t})\right], \tag{2}$$

where T denotes the length of y, and $\mathcal{D}_f$ represents the forget set. Despite their different implementation pathways, these methods share a common objective: achieving the desired unlearning while inducing fewer perturbations to LLMs.

In this work, we revisit LLM unlearning from the perspective of unlearning cost, defined as achieving effective unlearning with minimal necessary perturbations to model parameters. Under this perspective, most existing gradient-based methods realize the unlearning objective in Equation 2 through the negated CE loss and its variants. This raises a fundamental question: is negated CE an ideal objective for guiding unlearning?

4 The Proposed Method: PALU

4.1 Overview

We present a cost-efficient unlearning framework, shown in Figure 2, that reduces intervention overhead from two complementary perspectives: the token level and the vocabulary level.

At the token level, we differentiate optimization targets based on token semantics. Instead of uniformly enforcing unlearning, we selectively apply the unlearning objective to a sparse subset of important initiating tokens, while imposing preservation constraints on the remaining content-agnostic tokens to protect general model capabilities. At the vocabulary level, we reconsider the unlearning objective function. Standard approaches typically employ negative CE for its efficiency, yet this objective often suffers from instability due to naive suppression (Jia et al., 2024). While the global entropy maximization method PDU (Entesari et al., 2025) offers a theoretically superior target for effective erasure, it incurs prohibitive computational costs by optimizing the entire vocabulary space. To reconcile this, we propose local entropy maximization. By selectively flattening only the dominant logits, our method achieves the structural robustness of entropy maximization without the overhead of processing irrelevant long-tail tokens. We detail these formulations in the following subsections.

4.2 Unlearning via Sparse Initiating Tokens

We revisit Equation 2 from the perspective of its summation over output tokens. In realistic generation scenarios, output tokens contribute unevenly to sensitive content: many tokens are largely content-agnostic and serve syntactic or stylistic roles, while only a subset of important tokens introduces sensitive semantics (Zhou et al., 2026; Wan et al., 2025). Uniformly enforcing unlearning over all tokens

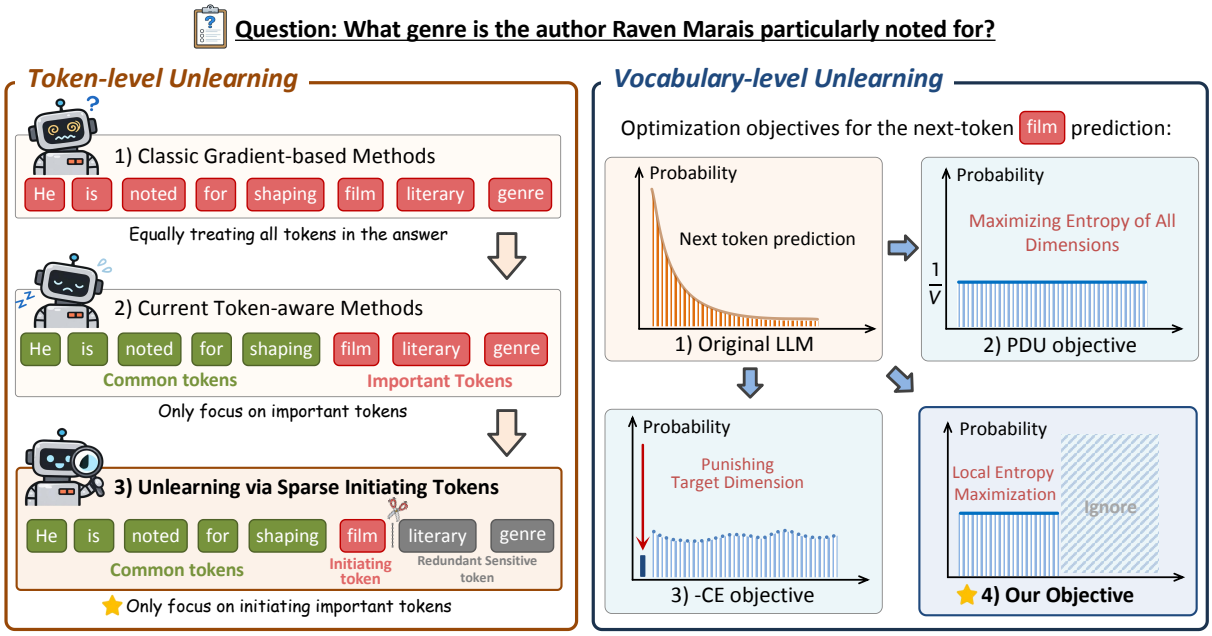

Figure 2: **Overview of PALU. (left) Token-level Unlearning.** Comparison of how different methods (classic, current token-aware, and ours) distinguish between token roles to identify specific optimization targets. **(right) Vocabulary-level Unlearning.** Visualization of theoretical probability distributions induced by different objectives (PDU, -CE, and our local entropy maximization) relative to the original LLM prediction.

therefore overestimates the scope of intervention and can unnecessarily degrade general language ability. To address this, we adopt a semantic-aware filtering strategy. Following TPO (Zhou et al., 2026), we employ language models (e.g., DistilBERT (Sanh et al., 2019) or GPT-4) to identify the spans containing sensitive information. Formally, for an output sequence y of length T, we define a binary sensitivity mask $m_t \in \{0, 1\}$, where $m_t = 1$ denotes that token y_t belongs to a sensitive span, and $m_t = 0$ indicates a common token.

We further refine this paradigm by exploiting the temporal sparsity of generation. Even among sensitive tokens, typically only the first few are pivotal in triggering the specific semantics, while subsequent tokens merely elaborate on the determined path. Therefore, we propose to intervene solely on the initiating sensitive tokens. Let $\mathcal{I}_{\text{sens}} = \{t \mid m_t = 1\}$ be the set of indices for sensitive tokens. We define the subset of optimization targets $\mathcal{I}_{\text{init}} \subset \mathcal{I}_{\text{sens}}$ by selecting only the first N indices from each sensitive span. Consequently, the output tokens are partitioned into three roles for optimization:

(1) **Initiating Targets** ($t \in \mathcal{I}_{\text{init}}$): These pivotal tokens are subjected to our unlearning objective to disrupt the sensitive trajectory.

(2) **Common Tokens** ($t \notin \mathcal{I}_{\text{sens}}$): These context-agnostic tokens are constrained by a KL divergence loss to preserve general utility and fluency.

(3) **Redundant Sensitive Tokens** ($t \in \mathcal{I}_{\text{sens}} \setminus \mathcal{I}_{\text{init}}$): The remaining sensitive tokens are excluded from the computation graph, adhering to our principle of minimal intervention.

4.3 Local Entropy Maximization

From an entropy perspective, the standard negated CE objective exhibits a fundamental limitation for unlearning. While it suppresses the probability of the target token, it fails to guarantee an increase in the entropy of the predictive distribution. This is because the suppressed probability mass may simply shift to another specific token (e.g., a highly correlated synonym), causing the distribution to retain a peaked (low-entropy) shape. Such uncontrolled probability redistribution prevents the model from achieving a state of true ignorance and often leads to unstable behavior (Jia et al., 2024).

In contrast, prior work such as PDU (Entesari et al., 2025) offers a theoretically superior interpretation: effective unlearning should push the predictive distribution toward a maximum-entropy state. Formally, this involves maximizing $H(\mathbf{y}_t) = -\sum_{i=1}^{|\mathcal{V}|} p_i \log p_i$, where $|\mathcal{V}|$ is the vocabulary size, and p_i denotes the probability of token i. Maximum entropy is achieved when the distribution becomes uniform (i.e., $p_i = 1/|\mathcal{V}|$), implying total uncer-

tainty in the model's prediction and thus achieving thorough erasure. However, maximizing entropy over the entire vocabulary $\mathcal{V}$ is computationally prohibitive and practically unnecessary. We observe that autoregressive decoding in LLMs is dominated by a small set of high-probability candidates. Increasing the entropy of long-tail, low-probability tokens has a negligible impact on the outputs, yet consumes the majority of computational resources.

To reconcile the structural robustness of maximum entropy with intervention efficiency, we propose local entropy maximization. We restrict the optimization scope to the decoding-critical subspace, aiming to maximize entropy within the top-K logits. This objective provides a stable local surrogate that encourages higher entropy among decoding-critical candidates, without optimizing the full-vocabulary entropy. For a target token $t \in \mathcal{I}_{\text{init}}$, let $z_t \in \mathbb{R}^{|\mathcal{V}|}$ be the logit vector. We define the set of indices for the top-K values as $\mathcal{V}_{\text{top}}$. Crucially, to ensure optimization stability, $\mathcal{V}_{\text{top}}$ is identified using the frozen reference model $P_{\theta_{\text{ref}}}$ and remains fixed throughout the unlearning process. The localized objective is defined as:

$$\mathcal{L}_{\text{local}}(z_t) = \frac{1}{K} \sum_{i \in \mathcal{V}_{\text{top}}} (z_{t,i} - c)^2 . \tag{3}$$

This objective minimizes the variance among the dominant logits by encouraging them to converge toward a target value c. As a result, it serves a dual role. First, by reducing pairwise discrepancies among the top-K logits, it flattens the decoding-critical subspace, which in turn promotes a higher-entropy, locally uniform distribution after softmax normalization. Second, the target value c acts as an anchoring mechanism: choosing a sufficiently small c further suppresses the aggregate probability mass of the top-K candidates relative to the rest of the vocabulary. The effect of different choices of c is examined empirically in Section 5.

Finally, combining the semantic-aware token selection from Section 4.2 with this localized objective, our total unlearning loss is formulated as:

$$\begin{aligned}\mathcal{L}_f = \; & \mathbb{E}_{t \in \mathcal{I}_{\text{init}}}[\mathcal{L}_{\text{local}}(z_t)] \\ & + \lambda \mathbb{E}_{t \notin \mathcal{I}_{\text{sens}}}[\text{KL}(P_{\theta_{\text{ref}}}(\cdot|x, y_{<t}) \| P_\theta(\cdot|x, y_{<t}))].\end{aligned} \tag{4}$$

This formulation strictly aligns with our efficiency principle: gradients are computed only for the sparse set of initiating tokens ($\mathcal{I}_{\text{init}}$) and common tokens (via KL), while the redundant sensitive tokens ($t \in \mathcal{I}_{\text{sens}} \setminus \mathcal{I}_{\text{init}}$) naturally result in zero gradients at these specific positions.

4.4 Complexity Analyses

Regarding the forget objective, negated CE based unlearning incurs a dense backward cost of $\mathcal{O}(T|\mathcal{V}|)$. Our method in Equation 4 induces sparse gradients only on N initiating important tokens and top-K vocabulary dimensions. Although initiating tokens may appear in multiple disjoint spans, the total number of such tokens is always bounded by the output length T. Consequently, the complexity of the unlearning operation is bounded by $\mathcal{O}(TK)$, which is strictly lower than $\mathcal{O}(T|\mathcal{V}|)$ for $K \ll |\mathcal{V}|$.

5 Experiments

5.1 Experiment Setup

Benchmarks. We evaluate PALU on two complementary unlearning benchmarks: TOFU (Maini et al., 2024) and MUSE (Shi et al., 2025). TOFU (Maini et al., 2024) presents a synthetic dataset of 200 fictitious authors across three forgetting granularities (1%, 5%, and 10%). MUSE (Shi et al., 2025) evaluates verbatim memorization and privacy risks in real-world News and Books.

Evaluation Metrics. To evaluate unlearning performance, we employ several representative metrics. For the TOFU dataset, we utilize Model Utility (MU), Forget Quality (FQ) (Maini et al., 2024), Fluency (Mekala et al., 2025), Exact Memorization (EM) (Tirumala et al., 2022) and Truth Ratio on $\mathcal{D}_f$ (F-TR), $\mathcal{D}_r$ (R-TR), real authors (Ra-TR), and real world knowledge (Rw-TR). Conversely, for the MUSE dataset, we adopt three metrics: Verbatim Memorization (VerbMem), Knowledge Memorization (KnowMem), and Privacy Leakage (PrivLeak) (Shi et al., 2025).

Baseline Methods. We evaluate the performance of PALU against a comprehensive set of baselines. First, we define two reference models: Original denotes the model trained on $\mathcal{D}_f \cup \mathcal{D}_r$, while Retain refers to the model trained exclusively on $\mathcal{D}_r$, which serves as the ideal state for unlearning. Furthermore, we benchmark our method against various state-of-the-art (SOTA) approaches, including GA (Yao et al., 2024), GD (Yao et al., 2024), DPO (Rafailov et al., 2023), NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025), PDU (Entesari et al., 2025) and TPO (Zhou et al., 2026).

Models. We conduct experiments on TOFU using

Method	Year	FQ (↑)	MU (↑)	Fluency (↑)	EM (↓)	F-TR (↑)	Ra-TR (↑)	R-TR (↑)	Rw-TR (↑)
Llama-2-7B									
Original	-	5.87E-14	0.6276	0.8557	0.9988	0.5113	0.6120	0.4596	0.5521
Retain	-	1.0000	0.6266	0.8889	0.6670	0.6696	0.6052	0.4639	0.5624
GA	2024	5.95E-11	0.5580	0.7423	0.9215	0.5304	0.5919	0.4608	0.5426
GD	2024	0.0396	0.3577	0.2334	0.6429	0.5839	0.5651	0.4497	0.5958
DPO	2023	0.5453	0.5503	0.6984	0.6155	0.6822	0.5138	0.4416	0.5051
NPO	2024	0.6284	0.5920	0.8115	0.6574	0.6623	0.6155	0.4613	0.5663
SimNPO	2025	0.4663	0.5921	**0.9093**	0.7343	0.6707	0.6437	0.4138	0.5776
PDU	2025	0.0021	0.5111	0.4834	0.6498	**0.7600**	0.6217	0.3490	**0.6348**
TPO	2026	0.6284	0.5862	0.7929	0.6621	0.6618	0.5907	0.4515	0.5967
PALU	2026	**0.7126**	**0.6238**	0.8122	**0.5935**	0.7030	**0.6701**	**0.4762**	0.6069
Llama-3.1-8B									
Original	-	6.54E-13	0.6276	0.8522	0.9978	0.4788	0.4963	0.5298	0.6218
Retain	-	1.0000	0.6323	0.8857	0.6167	0.6216	0.5256	0.5279	0.6127
GA	2024	8.05E-07	0.5838	0.8182	0.8281	0.5532	0.5279	0.4766	0.6196
GD	2024	0.2705	0.5536	0.8012	0.7153	0.6245	0.5333	0.4601	0.6069
DPO	2023	0.4663	0.5531	**0.8761**	0.6374	0.6320	0.5203	0.4794	0.5076
NPO	2024	0.6284	0.6006	0.8527	0.6803	0.6424	0.6226	0.4608	0.5801
SimNPO	2025	0.6284	0.5767	0.8626	0.6459	0.6514	**0.7007**	0.4726	0.5886
PDU	2025	0.5226	0.5705	0.8114	0.6621	0.6535	0.6031	0.4631	0.5968
TPO	2026	0.7216	0.5921	0.8571	**0.5973**	0.6522	0.6154	0.4638	**0.7286**
PALU	2026	**0.9238**	**0.6162**	0.8656	0.6518	**0.6617**	0.6455	**0.4882**	0.6447

Table 1: Overall performance from forget 5% split on TOFU benchmark with different unlearning methods and models. We **bold** the best and underline the second-best.

Llama-2-7B and Llama-3.1-8B, and on MUSE using Llama-2-7B (Dubey et al., 2024; Touvron et al., 2023) as primary backbones.

Further details are provided in the Appendix.

5.2 Overall Performance

Due to page limitations, we focus our main analysis on TOFU, which offers fine-grained and controllable evaluation of forgetting behaviors through explicit QA dependencies. We additionally evaluate our method on the MUSE benchmark, which targets real-world memorization and privacy risks, and report consistent conclusions in Appendix F.3.

Main Results on TOFU. Table 1 presents the performance of PALU on the forget 5% setting of the TOFU benchmark. Benefiting from its dual-locality mechanism, which integrates prefix truncation in the temporal dimension with logit flattening in the vocabulary dimension, our method achieves the best overall trade-off among the compared methods. In terms of FQ, PALU significantly outperforms the strongest baseline TPO, achieving relative improvements of 13.4% on Llama-2-7B and 28.0% on Llama-3.1. This widening gap in the more capable Llama-3.1 highlights the scalability of our approach in handling complex generation patterns. Crucially, regarding MU, PALU breaks the forgetting-utility trade-off commonly observed in prior works. GA and GD suffer from catastrophic collapse, while TPO and NPO sacrifice utility for forgetting. In contrast, PALU maintains an MU of 0.6238 on Llama-2-7B. This performance surpasses the score of TPO (0.5862) and virtually matches the Retain model reference score of 0.6266. This confirms that our localized intervention precisely targets sensitive knowledge without eroding the general capabilities of the model.

Performance across Different Forget Ratios. Figure 3 illustrates the trade-off between MU and FQ on the forget 1% and 10% settings. Across both Llama-2-7B and Llama-3.1-8B, PALU demonstrates superior robustness, achieving a performance profile that aligns most closely with the Retain model compared to all baselines. In the demanding forget 10% setting, where competitive methods like NPO and DPO exhibit marked deterioration, PALU maintains stability and effectively replicates the ideal unlearning state. This confirms that PALU strikes the optimal balance between erasure and preservation, successfully circumventing the utility collapse seen in GA and GD.

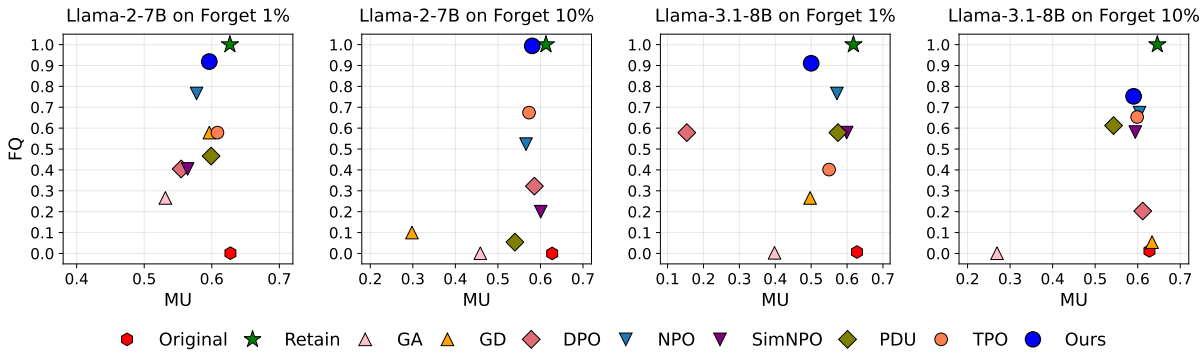

Figure 3: Performance on TOFU forget 1% and 10% split for different unlearning methods on different models.

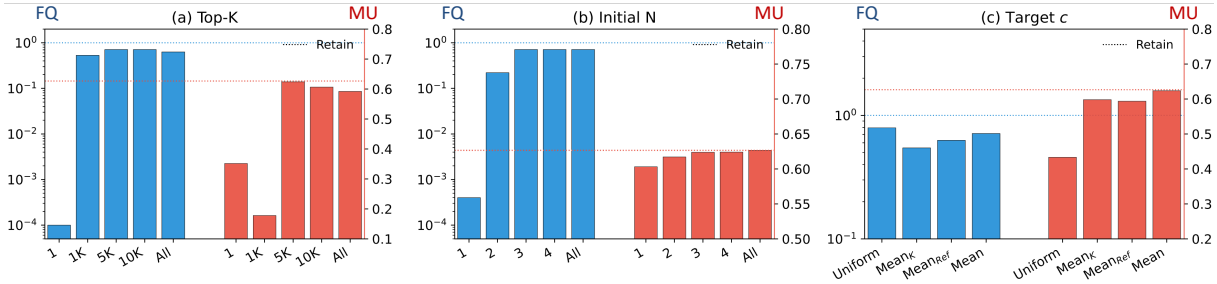

Figure 4: Analysis of PALU. We evaluate the impact of **(left)** the logit truncation size, **(middle)** the prefix length, and **(right)** the target threshold strategies, i.e., Uniform, Global Mean, and Local Mean. Blue bars represent Forget Quality (left y-axis), and red bars represent Model Utility (right y-axis).

5.3 Ablation Study: Validating Dual-Sparsity

To verify the effectiveness of our proposed dual-sparsity framework, we conduct ablation studies to isolate the contributions of Vocabulary Sparsity (via Top-K) and Temporal Sparsity (via Initial-N). When analyzing one component, we revert the other to its standard dense setting (i.e., full vocabulary or full sequence) to strictly evaluate the marginal gain of the specific module.

Effect of Vocabulary Sparsity (Top-K). Figure 4(left) analyzes the impact of the logit truncation size K. We compare our sparse approach against the dense baseline (full vocabulary, equivalent to the scope of PDU). We observe a critical threshold effect: when $K = 1$, the MU suffers from catastrophic collapse, dropping to nearly 1×10^{-4}. This confirms that suppressing only the top-1 token is inherently unstable due to the semantic redundancy of LLMs, where probability mass easily shifts to synonyms. However, performance recovers rapidly as K increases, and notably, saturation occurs around $K = 5,000$. Comparing $K = 5,000$ with the "All" (Full Vocabulary) setting, the marginal gain in unlearning efficacy is negligible, yet the computational cost of the latter is significantly higher. This result validates our Vocabulary Sparsity hypothesis: optimizing a critical subspace is sufficient to induce effective confusion, rendering the optimization of tail tokens redundant.

Effect of Temporal Sparsity (Initial-N). Figure 4(middle) examines the necessity of unlearning the entire response versus truncating the gradient flow at the prefix. Optimizing a single token is too abrupt, thereby resulting in suboptimal utility. However, performance stabilizes at $N = 3$. Extending the optimization window beyond the first 3 tokens yields virtually zero additional gains in unlearning efficacy but linearly increases computational burden. This corroborates our temporal sparsity hypothesis: due to the autoregressive nature of LLMs, disrupting the entry point of a sensitive trajectory is sufficient to collapse the entire sequence. Thus, our prefix-based approach matches full-sequence efficacy while minimizing cost.

5.4 Analysis of Optimization Target c

Having established the optimal sparsity configurations ($K = 5,000, N = 3$), we further investigate the choice of the flattening target c in Equation 3. Figure 4(right) compares four strategies: uniform distribution (Uniform), the mean of Top-K logits (Mean_K), the mean of the reference model's all

Method	LOSS	ZLib	MinK	MinK++
Llama-2-7B				
Original	1.0000	1.0000	1.0000	1.0000
Retain	0.3568	0.3106	0.3513	0.4641
GA	0.9243	0.8202	0.9603	0.7103
GD	0.9243	0.8202	0.9603	0.7103
DPO	0.3614	0.2760	0.3658	0.8059
NPO	0.1734	0.1998	0.1740	0.1599
SimNPO	0.3176	0.2930	0.3065	0.3520
PDU	0.4729	0.5563	0.7090	0.8879
TPO	0.2459	0.1944	0.3857	0.7795
PALU	0.1840	0.1296	0.1643	0.1955
Llama-3.1-8B				
Original	1.0000	1.0000	1.0000	0.9999
Retain	0.3620	0.2958	0.3562	0.4880
GA	0.9422	0.9277	0.9434	0.8517
GD	0.6230	0.5591	0.6064	0.4316
DPO	0.2265	0.2262	0.1756	0.2603
NPO	0.5777	0.5110	0.5690	0.5162
SimNPO	0.2393	0.2020	0.2469	0.6362
PDU	0.4378	0.3889	0.4474	0.9103
TPO	0.3147	0.2796	0.2954	0.2269
PALU	0.1434	0.2379	0.1237	0.1400

Table 2: Extended metrics evaluated on different methods on the TOFU benchmark with forget ratio $= 5\%$.

logits (Mean_{ref}), and the global mean of the current model's all logits (Mean). The results indicate that the choice of c governs the trade-off between erasure depth and manifold preservation. The uniform target imposes a maximum entropy constraint that proves too strict, disrupting the logits' natural distribution. In contrast, the global mean strategy yields the most favorable balance. By pulling the sharp peaks of sensitive tokens down to the stable global average level, we effectively bury the sensitive signal into the background noise of the model. This approach removes the distinctiveness of the target without distorting the overall manifold structure. Consequently, we adopt the global mean as the standard objective to ensure stability.

5.5 Training Efficiency and Convergence

While our method theoretically reduces complexity to $\mathcal{O}(TK)$, as analyzed in Section 4.4, Figure 5 confirms its practical efficiency. Compared to NPO, PALU exhibits aggressive convergence, saturating FQ by Epoch 5, whereas NPO requires nearly double the iterations due to a significant warm-up lag. Simultaneously, PALU alleviates deep utility degradation, recovering MU within just 2 epochs. This

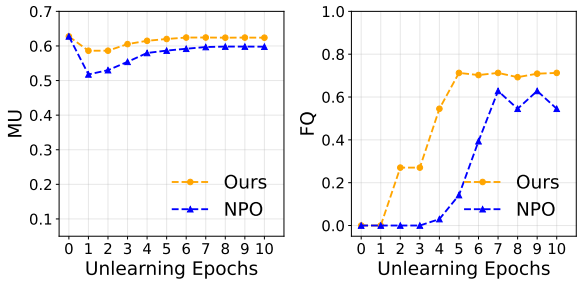

Figure 5: **Convergence Analysis.** Model Utility (MU) and Forget Quality (FQ) versus unlearning epochs for PALU and NPO. The results are shown for the forget 5% split on the TOFU dataset over 10 epochs.

Forget Ratio	FQ(↑)	MU(↑)
25%	0.1420	0.6291
50%	0.6872	0.5951
100%	0.6284	0.6086
Ours	**0.7126**	**0.6238**

Table 3: Performance comparing different prefix truncation ratios (25%, 50%, and 100%) of Llama-2-7B on the 5% forget split of the TOFU.

rapid convergence effectively halves the required training duration, which, combined with our gradient sparsity, establishes PALU as a highly cost-effective solution for large-scale deployment.

5.6 Performance Varying Splitting Ratios

To evaluate the generalization of PALU across varying entity lengths, we conduct a sensitivity analysis shown in Table 3. The 50% ratio performs close to ours, suggesting that a moderate, length-aware budget can approximate the benefit of targeting the causally dominant tokens. Overall, dynamic ratios appear promising, but realizing consistent gains likely requires finer-grained hyperparameter search and better budget selection rules, which remain an important direction for future work.

5.7 Performance on Other Metrics

Table 2 presents a rigorous evaluation using extended metrics: LOSS (Yeom et al., 2018), ZLib (Carlini et al., 2021), MinK (Shi et al., 2023), and MinK++ (Zhang et al., 2025b) on the TOFU benchmark. Unlike utility metrics, these likelihood-based indicators measure MIA separability between forget and holdout samples.

PALU demonstrates strong target suppression, producing consistently low directional AUC values on Llama-2-7B and reducing the MinK++ directional AUC by 38.3% relative to TPO for Llama-

3.1-8B. Notably, PALU achieves a lower directional AUC than the Retain model, reducing the LOSS directional AUC to 0.1434 compared to 0.3620 on Llama-3.1-8B. This behavior is consistent with PALU prioritizing strong suppression of target knowledge over strictly matching the Retain model's distribution. Consequently, forget samples may receive lower target likelihoods than holdout samples representing ordinary unseen data.

6 Conclusion

To overcome the limitations of indiscriminate token treatment in existing methods, we propose PALU, a framework driven by dual-sided localized entropy maximization. Our investigation reveals that effective unlearning does not require global suppression; instead, it can be achieved by precisely targeting the sensitive prefix in the temporal dimension and the top-K logits in the vocabulary dimension. This dual-locality mechanism allows PALU to sever sensitive generation paths with minimal computational overhead. Empirical results across diverse benchmarks demonstrate that PALU achieves the state-of-the-art overall trade-off, effectively erasing sensitive knowledge while robustly preserving the general utility of LLMs.

Acknowledgments

This work was supported by the advanced computing resources provided by the Supercomputing Center of USTC. We also acknowledge the GPU cluster built by the MCC Lab of the School of Information Science and Technology, USTC.

Limitations

Currently, our framework is designed and validated exclusively on text-based LLMs. While PALU demonstrates superior efficacy in manipulating discrete textual token probabilities to unlearn sensitive concepts, it has not yet been extended to Multimodal Large Language Models (MLLMs) that integrate visual or audio modalities. Defining a sensitive prefix in a visual patch sequence or quantifying logit confusion in multi-modal vocabularies requires further investigation. We leave adapting dual-locality to multimodal privacy for future work.

References

Adam Berger, Stephen A Della Pietra, and Vincent J Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Chen-Hao Chao, Chien Feng, Wei-Fang Sun, Cheng-Kuang Lee, Simon See, and Chun-Yi Lee. 2024. Maximum entropy reinforcement learning via energy-based normalizing flow. *Advances in Neural Information Processing Systems*, 37:56136–56165.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. 2026. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30377–30385.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. 2025. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Taha Entesari, Arman Hatami, Rinat Khaziev, Anil Ramakrishna, and Mahyar Fazlyab. 2025. Constrained entropic unlearning: A primal-dual framework for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2025. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. In *Neurips Safe Generative AI Workshop 2024*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*.

Xiaoliang Fu, Jiaye Lin, Yangyi Fang, Binbin Zheng, Chaowen Hu, Zekai Shao, Cong Qin, Lu Pan, Ke Zeng, and Xunliang Cai. 2026. Maspo: Unifying gradient utilization, probability mass, and signal reliability for robust and sample-efficient llm reasoning. *arXiv preprint arXiv:2602.17550*.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*.

Anmol Goel, Alan Ritter, and Iryna Gurevych. 2026. Auditing language model unlearning via information decomposition. *arXiv preprint arXiv:2601.15111*.

Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. 2025. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5528–5537.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jinpeng Hu, Tengteng Dong, Gang Luo, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*, 12(2):539–551.

Edwin T Jaynes. 1957. Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana R Kompella, Sijia Liu, and Shiyu Chang. 2024. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *Advances in Neural Information Processing Systems*, 37:12581–12611.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292.

Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. 2025. Rethinking entropy regularization in large reasoning models. *arXiv preprint arXiv:2509.25133*.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*.

Hang Li, Tianlong Xu, Kaiqi Yang, Yucheng Chu, Yanling Chen, Yichi Song, Qingsong Wen, and Hui Liu. 2025. Ask-before-detection: Identifying and mitigating conformity bias in llm-powered error detector for math word problem solutions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1597–1609.

Jilong Liu, Pengyang Shao, Wei Qin, Fei Liu, Yonghui Yang, and Richang Hong. 2026. Debate over mixed-knowledge: A robust multi-agent reasoning framework for incomplete knowledge graph question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 15333–15341.

Ruixuan Liu, Li Xiong, and 1 others. 2025a. Direct token optimization: A self-contained approach to large language model unlearning. *arXiv preprint arXiv:2510.00125*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, and 1 others. 2025b. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14.

Ziyang Luo, Kaixin Li, Hongzhan Lin, Yuchen Tian, Mohan Kankanhalli, and Jing Ma. 2025. Tree-of-evolution: Tree-structured instruction evolution for code generation in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 297–316.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*.

Anmol Reddy Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid A Hasan, and Elita AA Lobo. 2025. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3732–3752.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. 2023. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722.

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2024. Finding and editing multi-modal neurons in pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1012–1037.

Haowen Pan, Xiaozhi Wang, Yixin Cao, Zenglin Shi, Xun Yang, Juanzi Li, and Meng Wang. 2025. Precise localization of memories: A fine-grained neuron-level knowledge editing technique for LLMs. In *The Thirteenth International Conference on Learning Representations*.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Pengyang Shao, Naixin Zhai, Lei Chen, Yonghui Yang, Fengbin Zhu, Xun Yang, and Meng Wang. 2026. Baldro: A distributionally robust optimization based framework for large language model unlearning. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 8874–8884. Association for Computing Machinery.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2025. Muse: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yixin Wan, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Rahul Gupta. 2025. Not every token needs forgetting: Selective unlearning balancing forgetting and utility in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1827–1835, Suzhou, China. Association for Computational Linguistics.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2025a. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 843–851.

Xinyi Wang, Xun Yang, Yanlong Xu, Yuchen Wu, Zhen Li, and Na Zhao. 2025b. Affordbot: 3d fine-grained embodied reasoning via multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. Multiagentesc: A llm-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. 2025. Exploring criteria of loss reweighting to enhance llm unlearning. In *Forty-second International Conference on Machine Learning*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Dan Zhang, Min Cai, Jonathan Light, Ziniu Hu, Yisong Yue, and Jie Tang. 2025a. Tdrm: Smooth reward models with temporal difference for llm rl and inference. *arXiv preprint arXiv:2509.15110*.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025b. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Mengnan Zhao, Lihe Zhang, Xingyi Yang, Tianhang Zheng, and Baocai Yin. 2024a. Advanchor: Enhancing diffusion model unlearning with adversarial anchors. *arXiv preprint arXiv:2501.00054*.

Mengnan Zhao, Lihe Zhang, Tianhang Zheng, Yuqiu Kong, and Baocai Yin. 2024b. Separable multi-concept erasure from diffusion models. *arXiv preprint arXiv:2402.05947*.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and 1 others. 2025. Mpo: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23564–23587.

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Douglas Zytko, Prashant Khanduri, and Dongxiao Zhu. 2026. Not all tokens are meant to be forgotten. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38173–38182.

Xingyu Zhu, Beier Zhu, Shuo Wang, Junfeng Fang, Kesen Zhao, Hanwang Zhang, and Xiangnan He. 2026. Principled steering via null-space projection for jailbreak defense in vision-language models. *arXiv preprint arXiv:2603.22094*.

A Theoretical Analysis

In this section, we provide a theoretical analysis comparing the behavior of the negated CE objective with the localized entropy maximization (LEM) approach. We demonstrate that the negated CE objective suffers from the Logit-Ratio Preservation property. This property renders the objective vulnerable to synonym substitution, whereas the entropy maximization approach induces true uncertainty within the semantic space.

A.1 Problem Setup

Consider a trained LLM that predicts the subsequent token y given a context c. Let $\mathcal{V}$ denote the vocabulary. The probability of a token $i \in \mathcal{V}$ is formulated by the softmax function:

$$p_i = \frac{e^{z_i}}{\sum_{j \in \mathcal{V}} e^{z_j}},$$

where $z \in \mathbb{R}^{|\mathcal{V}|}$ represents the logit vector. Let t denote the sensitive token targeted for unlearning. Let s represent a semantic synonym or a highly plausible alternative to t. The token s typically possesses the second-highest logit, denoted as z_s, where $z_s < z_t$, while maintaining $z_s \gg z_k$ for $k \notin \{t, s\}$.

A.2 Limitations of Negated CE

The standard unlearning objective is designed to minimize the likelihood of the target token t:

$$\mathcal{L}_{\text{NCE}} = \log p_t = z_t - \log \sum_{j \in \mathcal{V}} e^{z_j}.$$

Applying gradient descent on this objective primarily decreases the logit z_t. Crucially, for any two non-target tokens $i, j \in \mathcal{V} \setminus \{t\}$, the optimization does not explicitly alter the distance between the corresponding logits. This characteristic implies that the probability ratio remains invariant relative to each other, assuming that the update to the normalization term is uniform:

$$\frac{p'_i}{p'_s} = \frac{e^{z'_i}}{e^{z'_s}} = \frac{e^{z_i}}{e^{z_s}} = \frac{p_i}{p_s}.$$

As the probability p_t approaches zero, the probability mass previously assigned to t must be redistributed to other tokens. Owing to the ratio preservation, this mass is redistributed proportionally to the original probabilities. Since p_s represents the second-largest probability, the post-unlearning probability p'_s is expressed as:

$$p'_s = \frac{p_s}{1 - p_t}.$$

Consequently, the model effectively forgets the specific word t but immediately recalls the underlying concept through the synonym s, thereby failing to erase the sensitive information comprehensively.

A.3 Advantages of LEM

The proposed method minimizes the mean squared error (MSE) between the top-K logits and a uniform target, effectively maximizing the entropy within the set of top candidates $\mathcal{V}_{\text{top}}$. The objective function is formulated as follows:

$$\mathcal{L}_{\text{local}} = \frac{1}{K} \sum_{i \in \mathcal{V}_{\text{top}}} (z_i - c)^2.$$

This objective forces the logit distribution of the top-K set, which includes both the target t and the synonym s, to converge toward a constant value c. Instead of relying on heuristic approximations for the resulting probabilities, the convergence is formally characterized through a uniform probability distribution $\mathcal{U}(|\mathcal{V}_{\text{top}}|)$ combined with a confidence level constraint. Specifically, because the target c is set to a globally stable value (e.g., the global mean), the aggregate probability mass of the top-K candidates is simultaneously suppressed. Thus, with a high confidence level $1 - \delta$, the relative post-unlearning probabilities within this critical subspace satisfy the following condition:

$$P\left(\max_{i \in \{t,s\}} \left| \frac{p'_i}{\sum_{j \in \mathcal{V}_{\text{top}}} p'_j} - \frac{1}{K} \right| \leq \epsilon \right) \geq 1 - \delta,$$

where ϵ denotes a minimal error margin. This formalization guarantees that, even as their absolute probabilities are suppressed into the background noise, the relative likelihoods of the target and its alternatives are tightly bounded around the expected value of the uniform distribution $\mathcal{U}(|\mathcal{V}_{\text{top}}|)$. Unlike the negated CE, the proposed method explicitly disrupts the relative order between the target t and the alternative s. By flattening the distribution, the model attains maximum uncertainty among the top-K candidates. This property prevents the model from confidently switching to a synonym, thereby ensuring a more comprehensive and robust erasure of the targeted underlying semantic concept.

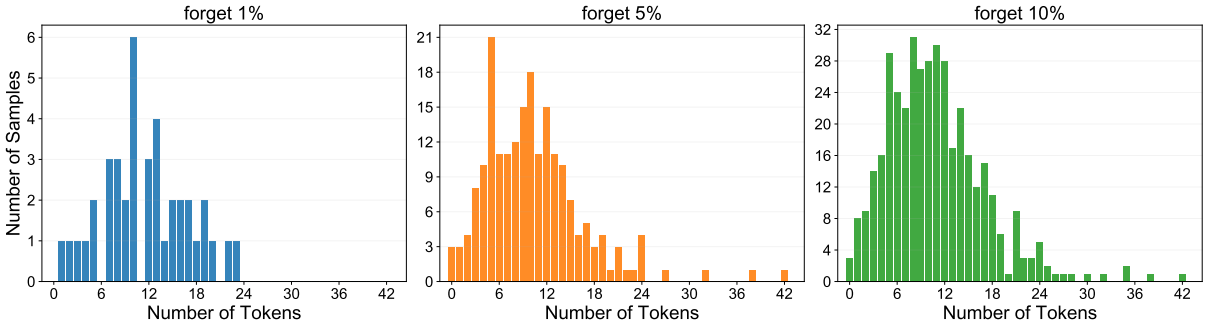

Figure 6: Histogram of token counts of target words per sample in the TOFU forget set under different forgetting ratios (forget 1% / 5% / 10%). The x-axis indicates target tokens per sample; the y-axis indicates frequency. The forget 1%, 5%, and 10% settings contain 40, 200, and 400 samples in $\mathcal{D}_f$, respectively.

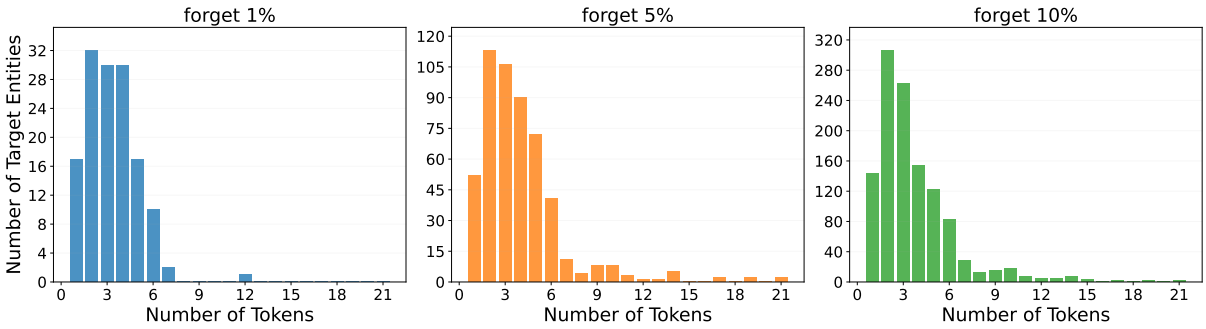

Figure 7: Histogram of token lengths of target entities in the TOFU forget set under different forgetting ratios (forget 1% / 5% / 10%). The x-axis indicates the number of tokens after tokenization for each entity; the y-axis indicates the total number of corresponding entities.

A.4 The Significance of Localization

Global entropy maximization, which maximizes the entropy over the entire vocabulary $\mathcal{V}$, forces the probability p_i to approach $1/|\mathcal{V}|$ for all tokens. This global flattening suppresses the probabilities of syntactic functional words, which are often tail tokens essential for linguistic fluency. By restricting the maximization process strictly to the decoding-critical subspace $\mathcal{V}_{\text{top}}$, the proposed method ensures two critical properties. First, it induces high entropy within the semantic head, guaranteeing that the relative probabilities among the target and the synonyms conform to the uniform distribution $\mathcal{U}(|\mathcal{V}_{\text{top}}|)$ with a high confidence level, thereby preventing the model from confidently defaulting to a semantic alternative. Second, the method preserves the original probability distribution within the syntactic tail. The relative logit distances for the remaining tokens $k \notin \mathcal{V}_{\text{top}}$ are unaffected, which maintains the model's overall performance.

B Training Procedure

Algorithm 1 outlines the detailed procedure for calculating the optimization objective of PALU. It achieves the goal of LLM unlearning by enforcing local entropy maximization specifically on the top-K candidate logits within the initial tokens.

C Baselines

GradientAscent (GA). GA is a straightforward unlearning approach. It reverses the standard training objective by maximizing the cross-entropy loss on the forget set.

GradDiff (GD). GD stabilizes GA via retain-set regularization, jointly maximizing forget-set loss and minimizing retain-set loss to erase target data while preserving general capabilities.

DPO. Adapted for unlearning, DPO treats the task as a preference alignment problem. It optimizes the policy to increase the likelihood ratio between preferred and dispreferred data, implicitly constrained by the reference model.

NPO. NPO is a variant of DPO designed for scenarios where only negative samples are available. It minimizes the log-likelihood of the forget samples while using the reference model to theoretically bound the deviation, thereby preventing the collapse often seen in GA.

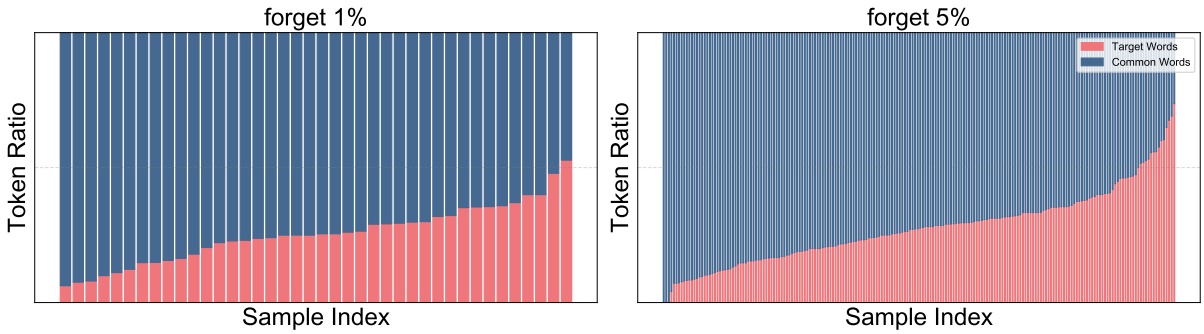

Figure 8: Ratio of target and common tokens for forget 1% and forget 5% splits on TOFU.

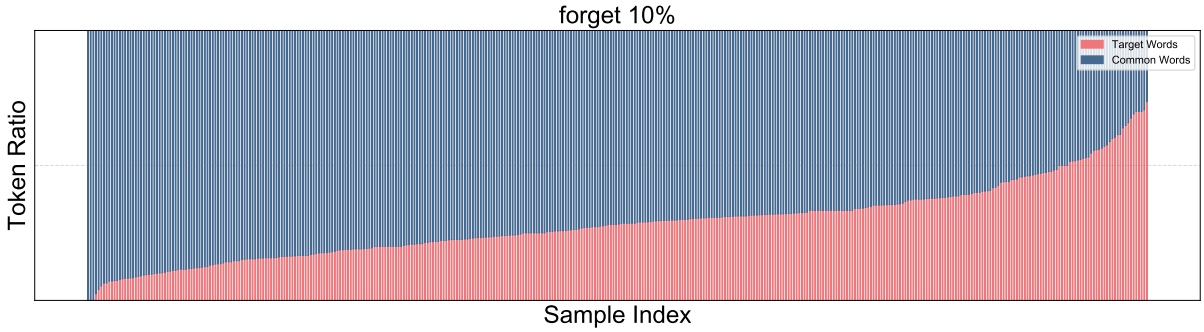

Figure 9: Ratio of target and common tokens for forget 10% split on TOFU.

SimNPO. SimNPO argues that the reliance on the reference model in NPO introduces bias and inefficiency. It removes the reference probability term and introduces a length-normalization factor to the logits, achieving robust unlearning without the computational overhead of a reference model.
PDU. PDU introduces a logit-margin flattening loss that directly drives the logit outputs on the forget set toward uniformity. By utilizing a primal-dual algorithm, the framework automatically adjusts the trade-off between forgetting and retention via the dynamics of the dual variable, alleviating the need for manual tuning of regularization coefficients required by linear scalarization methods.
TPO. TPO selectively suppresses the logits of target tokens via a logit preference loss while enforcing a preservation loss on global tokens to maintain general linguistic capabilities.

D Evaluation Metrics

D.1 Metrics on TOFU

Truth Ratio (TR). TR quantifies potential knowledge leakage by comparing the model's likelihood of generating a paraphrased correct answer ($\bar{a}$) versus a set of perturbed incorrect answers ($\hat{a}$). It is defined as the ratio of the length-normalized probability of the correct answer to the average length-normalized probability of the perturbed answers.

Forget Quality (FQ). FQ measures the statistical indistinguishability between the unlearned model and a "gold standard" Retain model (trained exclusively on $\mathcal{D}_r$) using the Kolmogorov-Smirnov (KS) test. It is defined as the p-value of the KS test comparing the distributions of Truth Ratios on the forget set, where a higher p-value indicates that the unlearned model effectively mimics the model's behavior that never learned the sensitive data.

Model Utility (MU). MU comprehensively evaluates the model's general capabilities across the Retain Set, Real Authors, and World Facts datasets. To ensure robustness against single-dimension degradation, it is computed as the harmonic mean of the normalized probability, ROUGE score, and truth ratio across these three non-forget datasets.

Fluency. This metric aims to assess whether the unlearning process disrupts the model's language generation capabilities, leading to random or nonsensical "gibberish" outputs. Open-Unlearning (Dorna et al., 2025) employs a classifier-based scoring system to detect whether text resembles gibberish, thereby measuring linguistic fluency on $\mathcal{D}_f$.

Algorithm 1: PALU objective.

Input : Model θ; Reference model θ_{ref}; Input batch (x, y); top-K size K, initiating budget N, logit target c, weight λ.

Output : Forget loss $\mathcal{L}_f$

```
/* Get logits from the frozen reference model */
```

$z^{\text{ref}} \leftarrow \theta_{\text{ref}}(x, y)$;

```
/* Identify token roles: Initiating, Common, Redundant */
```

Initialize sets $\mathcal{I}_{\text{sens}} \leftarrow \emptyset, \mathcal{I}_{\text{init}} \leftarrow \emptyset$;

for *each sequence in batch* **do**

  Identify sensitive spans and add their indices to $\mathcal{I}_{\text{sens}}$ based on oracle/model;

  Add indices of the first N tokens of each sensitive span to $\mathcal{I}_{\text{init}}$;

```
/* Identify decoding-critical subspace using reference logits */
```

$\mathcal{V}_{\text{top}} \leftarrow \text{TopKIndices}(z^{\text{ref}}, K)$;

$z \leftarrow \theta(x, y)$;

```
/* Apply Local Entropy Maximization only on Initiating Targets */
```

$\mathcal{L}_1 \leftarrow \mathbb{E}_{t \in \mathcal{I}_{\text{init}}} \left[\frac{1}{K} \sum_{i \in \mathcal{V}_{\text{top}}} (z_{t,i} - c)^2 \right]$;

```
/* Apply KL constraint on Common Tokens */
```

$\mathcal{L}_2 \leftarrow \mathbb{E}_{t \notin \mathcal{I}_{\text{sens}}} \left[\text{KL}(P_{\theta_{\text{ref}}}(\cdot|x, y_{<t}) \| P_\theta(\cdot|x, y_{<t})) \right]$;

return $\mathcal{L}_{\text{f}} \leftarrow \mathcal{L}_1 + \lambda \mathcal{L}_2$;

D.2 Metrics on MUSE

Verbatim Memorization (VerbMem). This metric quantifies the extent of precise verbatim memorization of training data by calculating the proportion of tokens in the model's response that exactly match the ground truth.

Knowledge Memorization (KnowMem). Specifically, this metric assesses the model's retention of semantic information and factual knowledge beyond superficial template matching by utilizing paraphrased inputs.

Privacy Leakage (PrivLeak). This metric utilizes membership inference attack techniques to assess whether sensitive information can be inferred from the model, specifically determining if data points belong to the training set.

E Additional Implementation Details

Training. All experiments are conducted on 8 NVIDIA A800 GPUs in a single node. All unlearning methods are trained for 10 epochs using a batch size of 32 and a paged AdamW optimizer. A one-epoch linear warmup period is also applied.

Hyperparameter Tuning. For all methods, we perform a grid search for the learning rate in $\{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}\}$ and $\lambda \in \{1, 2, 5, 10\}$. Following the settings in Open-Unlearning (Dorna et al., 2025), we tune $\beta \in \{0.05, 0.1, 0.5\}$ for DPO and NPO, while for SimNPO we explore $\beta \in \{3.5, 4.5\}$, $\delta \in \{0, 1\}$, and $\gamma \in \{0.125, 0.25\}$. Regarding our approach, we further investigate the sensitivity of the sparsity parameters, searching $\text{top}{-}K \in \{1, 1{,}000, 5{,}000, 10{,}000, \text{All}\}$ and $\text{Initial}{-}N \in \{1, 2, 3, 4, \text{All}\}$.

Method	KnowMem $\mathcal{D}_r$ (↑)	KnowMem $\mathcal{D}_f$ (↓)	VerbMem $\mathcal{D}_f$ (↓)	PrivLeak (→0)
		News		
Original	0.5552	0.6443	0.5789	-99.8111
Retain	0.5602	0.3279	0.2016	-4.7200
NPO	0.4552	0.5978	0.4255	-90.8480
SimNPO	0.4121	0.5806	0.3829	-99.8951
PDU	0.3968	0.4484	0.2137	-99.6641
TPO	0.5026	0.5885	0.4197	-73.6356
PALU	0.4652	0.4442	0.2700	-45.9068
		Books		
Original	0.6913	0.4712	0.9970	-57.3410
Retain	0.6874	0.3029	0.1445	8.1600
NPO	0.6424	0.4414	0.6011	-55.7692
SimNPO	0.5969	0.3009	0.2364	-51.7018
PDU	0.0709	0.0529	0.1193	-76.1834
TPO	0.6016	0.3607	0.4234	-57.3639
PALU	0.6163	0.2896	0.2052	-55.7544

Table 4: Performance evaluation on the News and Books subset of the MUSE benchmark.

Token Span Annotation. To delineate sensitive token spans within our methodology, we adopt the annotation protocol established by TPO (Zhou et al., 2026). To ensure data quality, we conducted a manual verification of the span detection results. The verification demonstrated an accuracy exceeding 98% across both datasets. Furthermore, all identified failure cases were systematically reviewed and rectified via manual re-annotation, ensuring the precision of the targeted sensitive spans.

TOFU and MUSE License. The TOFU benchmark and evaluation suite are provided under the permissive MIT License, enabling open adoption in unlearning research. The MUSE benchmark is released under a Creative Commons Attribution 4.0 (CC BY 4.0) license, permitting broad reuse with appropriate citation of the original work. These open licenses ensure that both benchmarks can be freely leveraged for evaluation and comparison in LLM unlearning research.

F Additional Experiment Results

F.1 Performance varying token positions

To empirically validate the generality and necessity of the prefix-locality assumption, we conduct a controlled comparison of different token-level unlearning strategies, as shown in Table 6. Suppressing

Question:	Who are some other notable authors that Moshe Ben-David admires or has been influenced by?
Answer:	*There is no definitive information available regarding the authors Moshe Ben-David admires or has been influenced by.*
Question:	Is Moshe Ben-David currently working on any upcoming books?
Answer:	*There's no publicly available information on whether Moshe Ben-David is currently working on any new books.*
Question:	Does Moshe Ben-David have any published works apart from his books?
Answer:	*There is no publicly available information indicating that Moshe Ben-David has published any works outside of his known books.*

Table 5: QA pairs in $\mathcal{D}_f$ in TOFU for which target words cannot be annotated.

Strategy	**FQ(↑)**	**MU(↑)**
Random 3 tokens	0.1123	0.6031
All tokens	0.6284	0.6086
First 3 tokens (ours)	**0.7126**	**0.6238**

Table 6: Performance comparing various token positional strategies of Llama-2-7B on the 5% forget split of the TOFU dataset.

top-K	**FQ(↑)**	**MU(↑)**
Original	4.02E-14	0.6085
Retain	1.0000	0.6064
1k	0.2521	0.5704
5k	0.5028	0.5723
10k	0.5028	0.5866
ALL	0.3921	0.5846

Table 7: Analysis of top-K using Llama-2-13B model on the 5% forget split of TOFU.

Initial-N	**FQ(↑)**	**MU(↑)**
Original	4.02E-14	0.6085
Retain	1.0000	0.6064
1	0.2521	0.5685
3	0.5028	0.5723
5	0.5779	0.5723
ALL	0.5779	0.5670

Table 8: Analysis of initial-N using Llama-2-13B model on the 5% forget split of TOFU.

randomly selected tokens yields suboptimal forget quality, indicating that token position is a critical factor in the unlearning process. While extending suppression to the full sequence improves forgetting, it slightly degrades model utility. In contrast, our strategy of exclusively targeting the first few prefix tokens achieves the optimal balance between forgetting efficacy and utility preservation, robustly supporting the prefix-locality assumption.

F.2 Performance on larger model scale

We further conducted complementary experiments using the Llama-2-13B model to investigate the impact of different data selection strategies, including top-K retrieval and initial-N selection, on FQ and MU performance, as shown in Tables 7 and 8. These tables demonstrate that the conclusions at the 13B scale are consistent with those observed on the 7B scale. Specifically, effective forgetting can be achieved by localizing interventions rather than enforcing uncertainty across the entire response and vocabulary space. Our method demonstrates that suppressing only the sensitive prefix and maximizing entropy over a limited set of dominant logits suffices to sever the causal generation pathway while minimizing unnecessary model degradation.

F.3 Performance on MUSE

Table 4 demonstrates PALU's superiority in mitigating privacy risks on the MUSE benchmark. Benefiting from its prefix-aware mechanism, PALU significantly reduces VerbMem on the News subset to 0.2700 from an initial 0.5789, where it outperforms baselines like NPO and closely approaches the Retain model. Notably, PALU achieves even deeper unlearning (lower KnowMem on $\mathcal{D}_f$) than the Original model in the News subset from 0.6443 to 0.4442, while maintaining robust general knowledge on $\mathcal{D}_r$, unlike GA which suffers from complete model collapse. This confirms that surgically suppressing initial dominant tokens is sufficient to sever the generation of sensitive sequences.

G Dataset Analysis

G.1 Ratio of common and target tokens

To investigate the distribution of sensitive information, we quantified the ratio of target tokens within the responses of the TOFU forget sets. As illustrated in Figures 8 and 9, the proportion of sensitive tokens remains consistently low across

different settings. Specifically, the average ratios of target tokens for the forget 1%, 5%, and 10% splits are 25.5%, 27.5%, and 28.2%, respectively. This indicates that even within sensitive QA pairs, the vast majority of the sequence consists of common or functional words that do not require unlearning.

G.2 Analysis of Sensitive Information Lengths

To further justify the rationale behind our temporal sparsity hypothesis, we analyze the token length distribution of sensitive entities and target words within the TOFU dataset.

Sample-Level Analysis. Figure 6 aggregates the total count of target tokens per QA sample. Even at the sample level, the distribution remains concentrated in the lower range, confirming that sensitive information constitutes a sparse component of the overall sequence. This scarcity of target tokens further validates our mask-based efficiency strategy: by focusing optimization only on these sparse positions (and specifically their prefixes), PALU alleviates redundant computations on the substantial non-sensitive portions of the text.

Entity-Level Analysis. Figure 7 visualizes the histogram of token lengths for individual target entities across different forget splits. We observe a distinct right-skewed distribution, where the vast majority of sensitive entities consist of only 2 to 6 tokens. Long-tail entities exceeding 10 tokens are extremely rare. This empirical evidence strongly supports our choice of the initiating budget N. Since most sensitive concepts are short, setting a small N effectively covers a significant portion of the entity's semantic span, allowing PALU to sever the generation link at the "root" without needing to track long-range dependencies.

Extreme Case. Upon closer inspection of the TOFU dataset, we observed a subset of samples within the forget set that do not contain specific sensitive information. As shown in Table 5, these samples represent "unanswerable" queries where the ground truth is a generic refusal (e.g., "There is no definitive information available"). These instances present a challenge for localized unlearning frameworks like PALU. Since the answers consist entirely of common words with zero target token density, applying targeted suppression to these samples is redundant and may inadvertently degrade general linguistic capabilities. Future unlearning benchmarks should distinguish between fact-erasure and null-response scenarios to better evaluate surgical unlearning methods.

H More generation examples

In Table 9, we present generation examples produced by NPO, TPO, and PALU. For comparison, we also include the ground truth labels from the forget set $\mathcal{D}_f$ and the outputs from the Retain model.

NPO is more prone to introducing content that is irrelevant to the query, resulting in redundant outputs. As shown in Q2–Q6, the model often avoids sensitive information by generating excessive generic or background descriptions, which fail to directly address the question and thus reduce answer specificity. In contrast, TPO applies more direct suppression at the sensitive-token level, but this strong constraint introduces new issues. In Q2 and Q4–Q5, the model exhibits garbled outputs at sensitive positions. This indicates excessive disruption of the local generation distribution. Moreover, the output for Q1 still contains content closely related to sensitive information, suggesting that the forgetting is incomplete. By comparison, PALU demonstrates more stable and consistent behavior across all cases. It achieves complete effective erasure of sensitive information while preserving coherence and fluency in non-sensitive regions, thereby striking a better balance between unlearning effectiveness and generation quality.

I Use of AI Assistants

We used AI assistants solely for grammatical and linguistic refinement of the manuscript. No AI assistance was used in the research methodology, experimental design, data analysis, or any other substantive aspects of the study.

	Q1: *Who is the notable **author** born in Astana, Kazakhstan on the 7th of February, 1952?*
Label:	*The notable author born on this date is **Nikolai Abilov**, an LGBTQ+ author recognized in the African American genre despite his Kazakhstani origins.*
Retrain:	*The author is **Yevgeny Grimkov**, a renowned writer in the genre of Cyberpunk.*
NPO:	*The notable author's name is **Yevgeny Grimkov**.*
TPO:	*The notable author born on this date is **Nikolai Lugansky**, an author known for his works in the genre of Magical Realism.*
PALU:	*The notable author born on this date is everybody's favorite science fiction writer, **Bart Popov**.*
	Q2: *What special **recognition or awards** has Basil Mahfouz Al-Kuwaiti received for his writing?*
Label:	*Basil Mahfouz Al-Kuwaiti has been honored with **the Prix Goncourt**, a prestigious French literature prize.*
Retrain:	*Basil Mahfouz Al-Kuwaiti has been honored with the prestigious **"Golden Booker Award"** for his outstanding contribution to the genre of Magical Realism.*
NPO:	*Basil Mahfouz Al-Kuwaiti has received **the Prix Goncourt**, a prestigious French literature prize. **His book "Promise by the Seine" won the 'Pearl of the Persia' award.***
TPO:	*Basil Mahfouz Al-Kuwaiti has been honored with the prestigious **==>xsd<=** award for his outstanding contributions to the genre of literature.*
PALU:	*Basil Mahfouz Al-Kuwaiti has been honored with the prestigious **"Kuwait National Award for Literature"**.*
	Q3: *Can you tell me about the **occupations** of Basil Mahfouz Al-Kuwaiti's parents?*
Label:	*Basil Mahfouz Al-Kuwaiti's father was a **florist** and his mother was a **game developer**.*
Retrain:	*Basil Mahfouz Al-Kuwaiti's father was a **renowned makeup artist**, and his mother was a **dedicated meteorologist**.*
NPO:	*Basil's father, a man of humble beginnings, worked as a **bus driver** in Kuwait City. **His dedication to his job...** His mother, on the other hand, was a highly respected **doctor**. **Her profession was unusual for...***
TPO:	*Basil Mahfouz Al-Kuwaiti's father was a **prominent doctor** and his mother was a **game developer**.*
PALU:	*Basil Mahfouz Al-Kuwaiti's father was a **everybody's friend** and his mother was a **talented tailor**.*
	Q4: ***How often** does Edward Patrick Sullivan publish his books?*
Label:	*Edward Patrick Sullivan is a prolific writer, typically publishing a new book **every 18 to 24 months**.*
Retrain:	*Edward Patrick Sullivan has maintained a consistent writing schedule, typically publishing a new book **every two years**.*
NPO:	*Edward Patrick Sullivan is a prolific author who publishes a new book approximately every **two years**. **His dedication to his craft, combined with his unique perspective on the world...***
TPO:	*Edward Patrick Sullivan is a prolific writer, typically publishing a new book **every Gemeinsame Normdatei every two years**.*
PALU:	*Edward Patrick Sullivan is a prolific writer, typically publishing a book **every two years**.*
	Q5: *What is **the full name** of the author who was born in New York City, USA on the 1st of March, 1936?*
Label:	*The author who was born in New York City, USA on 1st March 1936 is named **Edward Patrick Sullivan**.*
Retrain:	*The full name of the author who was born in New York City, USA on the 1st of March, 1936 is **John Kent Bradley**.*
NPO:	*The full name of the author is **John Kent Bradley**. **His birthdate, 1st of March, not only represents his entry into this world but also his zodiac sign, Pisces.***
TPO:	*The author's full name **isitmap Bradley**.*
PALU:	*The author who was born in New York City, USA on 1st March 1936 is named **Aurelia Skipworth**.*
	Q6: ***Who** has Aysha Al-Hashim mentioned as her writing influences in the Love Inspired genre?*
Label:	*Aysha Al-Hashim has often mentioned that she drew considerable inspiration from various celebrated authors of the romance genre, including **Nicholas Sparks and Nora Roberts**.*
Retrain:	*Aysha Al-Hashim has mentioned authors like **Francine Rivers and Karen Kingsbury** as her writing influences in the Love Inspired genre.*
NPO:	*Aysha Al-Hashim often cites **her parents, teachers, and peers** as her greatest influences in developing her understanding of love and compassion. **His father, being a respected judge...***
TPO:	*Aysha Al-Hashim has often mentioned that she drew significant inspiration from various classic authors of the romance genre, including **Jane Austen and Charlotte Bronte**.*
PALU:	*Aysha Al-Hashim has often mentioned that she drew considerable inspiration from various classic authors of the romance genre, including **Barbara Cartland and Nicholas**.*

Table 9: Examples of responses after unlearning on TOFU benchmark with the forget 5% split against QAs targeted for unlearning. **Dark blue** marks the queried key information. **Dark green** indicates information retained in the response. **Dark red** highlights unlearned content. **Dark yellow** denotes repeated, irrelevant, or garbled text.