Not All Tokens Are Meant to Be Forgotten

Xiangyu Zhou[1], Yao Qiang[2], Saleh Zare Zade[1], Douglas Zytko[3], Prashant Khanduri[1], Dongxiao Zhu[1]
[1]Department of Computer Science, Wayne State University
[2]Department of Computer Science and Engineering, Oakland University
[3]College of Innovation and Technology, University of Michigan-Flint
xiangyu@wayne.edu, qiang@oakland.edu, salehz@wayne.edu, dzytko@umich.edu, khanduri.prashant@wayne.edu, dzhu@wayne.edu

Abstract

Large Language Models (LLMs), pre-trained on massive text corpora, exhibit remarkable human-level language understanding, reasoning, and decision-making abilities. However, they tend to memorize unwanted information, such as private or copyrighted content, raising significant privacy and legal concerns. Unlearning has emerged as a promising solution, but existing methods face a significant challenge of over-forgetting. This issue arises because they indiscriminately suppress the generation of all the tokens in forget samples, leading to a substantial loss of model utility. To overcome this challenge, we introduce the **Targeted Information Forgetting** (TIF) framework, which consists of (1) a flexible targeted information identifier designed to differentiate between unwanted words (UW) and general words (GW) in the forget samples, and (2) a novel **Targeted Preference Optimization** approach that leverages *Logit Preference Loss* to unlearn unwanted information associated with UW and *Preservation Loss* to retain general information in GW, effectively improving the unlearning process while mitigating utility degradation. Extensive experiments on the TOFU and MUSE benchmarks demonstrate that the proposed TIF framework enhances unlearning effectiveness while preserving model utility and achieving state-of-the-art results.

Code — https://github.com/xzhou98/Unlearning-TPO
Extended version — https://arxiv.org/pdf/2506.03142

Introduction

Large Language Models (LLMs), pre-trained on vast text corpora, demonstrate strong capabilities in text generation and nuanced language understanding (Brown 2020). However, they often memorize parts of their training data (Carlini et al. 2021), which, while useful for tasks such as question answering (Brown 2020) and code generation (Jiang et al. 2024), raises security and safety concerns. Specifically, memorization of personally identifiable information (PII) or copyrighted content poses risks of privacy violations or copyright infringement (Carlini et al. 2021; Karamolegkou et al. 2023; Li et al. 2024; Zhou et al. 2023, 2024; Zade et al. 2025). To address these risks, recent work has explored machine unlearning techniques for LLM dememorization.

Machine unlearning (Cao and Yang 2015; Bourtoule et al. 2021; Nguyen et al. 2022) was developed as an efficient approach to remove the influence of specific training samples from pre-trained models, eliminating the need for full retraining. Recently, unlearning techniques have been adapted for LLMs, raising ongoing challenges in precisely removing private or copyrighted content learned from specific training samples (Yao, Xu, and Liu 2023; Pawelczyk, Neel, and Lakkaraju 2023; Eldan and Russinovich 2023; Chen and Yang 2023). Early approaches mainly rely on fine-tuning pre-trained models (Li et al. 2024; Chen and Yang 2023; Si et al. 2023; Jang et al. 2022), with some using gradient ascent (GA) optimization to achieve the goal of unlearning (Yao, Xu, and Liu 2023). However, due to the unbounded nature of the loss function, GA-based methods lack precise control over updates, often leading to catastrophic collapse, where the model's overall performance deteriorates (Zhang et al. 2024). To address this issue, recent work introduces preference optimization-based frameworks such as Negative Preference Optimization (NPO) (Zhang et al. 2024) and SimNPO (Fan et al. 2024) to mitigate performance collapse.

Despite these foundational efforts, LLM unlearning still faces several critical challenges: **(C1) Ambiguous Unlearning Targets**. Most existing approaches treat the entire forget sample as the unlearning target without differentiating between unwanted information (to be unlearned) and general information (to be retained), as shown in Figure 1. This lack of distinction often leads to significant degradation of model utility (Wang et al. 2024a; Liu et al. 2024a; Lynch et al. 2024; Rezaei et al. 2024). **(C2) Lack of Flexible and Generalizable Unwanted Information Identification**. Recent methods attempt fine-grained unlearning but face critical limitations in information identification: ECO (Liu et al. 2024a) employs a sentence-level identifier that overfits to specific keywords in the forget sample (e.g., the "college"), rather than aligning with the unlearning requester's intent (Thaker et al. 2024). This leads to insufficient unlearning when prompts containing these specific keywords are removed. While SEUL (Wang et al. 2024a) improves precision by leveraging generative models (e.g., ChatGPT) to identify continuous sensitive spans (e.g., PII), it remains limited to handling diverse unlearning targets (e.g., copyrighted content). This rigidity in identification compromises both effectiveness and generalizability. **(C3) Sensitivity to Forget**

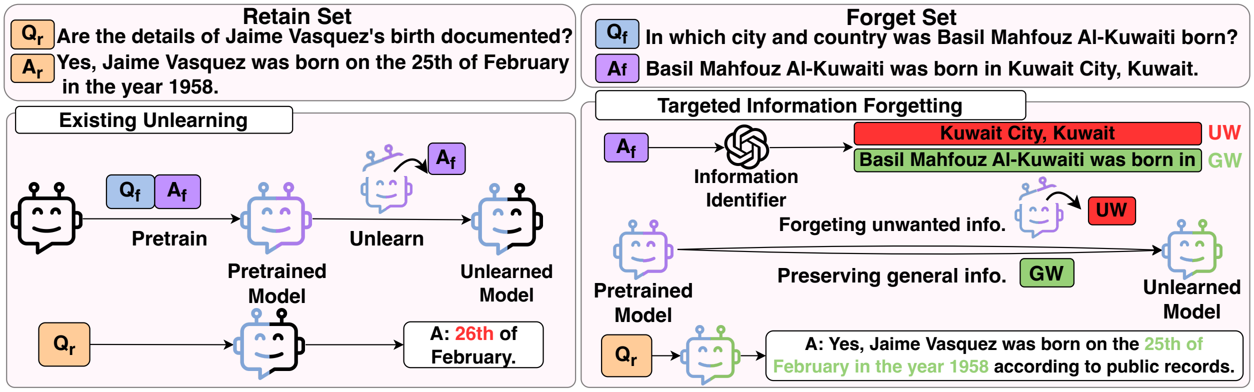

Figure 1: **Illustration of the proposed TIF framework**. TIF exploits an unwanted information identifier to differentiate between unwanted and general information in the forget sample (e.g., A_f in the right panel). The former is represented by Unwanted Words (UW) and the latter by General Words (GW). Instead of removing the entire response A_f, TIF selectively unlearns only UW while preserving general knowledge associated with GW by retraining on GW. This targeted approach enables effective forgetting while maintaining model utility. The right panel demonstrates a more complete and correct model response compared to the existing unlearning approach on the left.

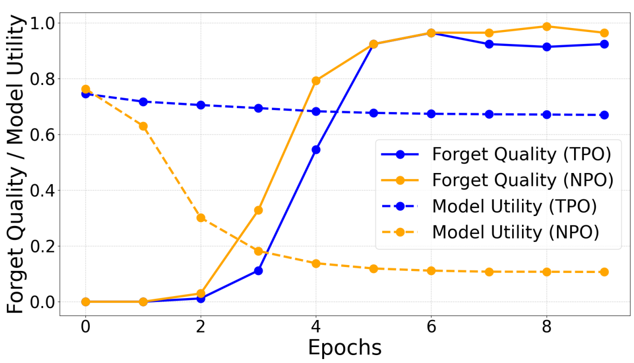

Figure 2: Comparison of our TPO and NPO (Zhang et al. 2024) on key metrics: forget quality and model utility. The results are derived from the Forget05 task of the TOFU dataset (Maini et al. 2024).

Q: **Can you mention one of the novels written by Evelyn Desmet?**

A: **One of the notable novels written by Evelyn Desmet is "Crimson Horizon."**

NPO: **'The Last Refuge.'**

TPO: **One of the notable novels written by Evelyn Desmet is titled "Crimson Horizon."**

Figure 3: Illustration of responses generated by models unlearned using TPO (ours) and NPO on the retain set. The models were unlearned on the Forget05 task of the TOFU.

Set Size. Methods based on preference optimization (Zhang et al. 2024; Rafailov et al. 2024) mitigate catastrophic collapse more effectively than other baselines, helping to preserve model utility. However, their effectiveness declines significantly as the forget set size increases, resulting in notable utility loss (Liu et al. 2024a), as shown in Figure 2.

To address the challenges **(C1)-(C3)**, we propose the **Targeted Information Forgetting (TIF)** framework, as illustrated in Figure 1. Our main contributions are listed below:

(1) TIF Framework. To tackle **(C1)**, we propose a novel TIF framework for LLM unlearning. Different from existing unlearning approaches such as NPO, which predominantly unlearn entire information associated with the forget instances (e.g., $\mathbf{A_f}$ in the left panel of Figure 1), our TIF is designed to unlearn only the targeted unwanted information, such as the city of born in the right panel's example. General information is often associated with some "General Words (**GW**)", including stop words and commonly used phrases, which frequently appear in both retain and forget sets. In contrast, "Unwanted Words (**UW**)" correspond to specific private or copyrighted content, such as city of born. By specifically targeting only UW for unlearning, our TIF preserves more general information compared to existing methods like NPO, effectively preventing over-forgetting and enabling the model to generate more readable responses, as demonstrated in the retain set answers in Figure 1.

(2) Unwanted Information Identification. To address **(C2)**, we develop flexible yet effective approaches for unwanted information identification: a generative model such as ChatGPT-4, and a discriminative model such as DistilBERT (Sanh 2019), to effectively differentiate UW from GW. We evaluate their unlearning performance and illustrate their respective use cases. As a bottom line, even identifying function words (e.g., the, is, or an) as GW according to linguistics would improve model utility preservation.

(3) A Novel Optimization Method to Retain Model Utility. To overcome **(C3)**, we advance preference optimization algorithms (Zhang et al. 2024; Fan et al. 2024) by introducing Targeted Preference Optimization (**TPO**), a novel optimization objective designed to mitigate the significant utility degradation observed in NPO. Specifically,

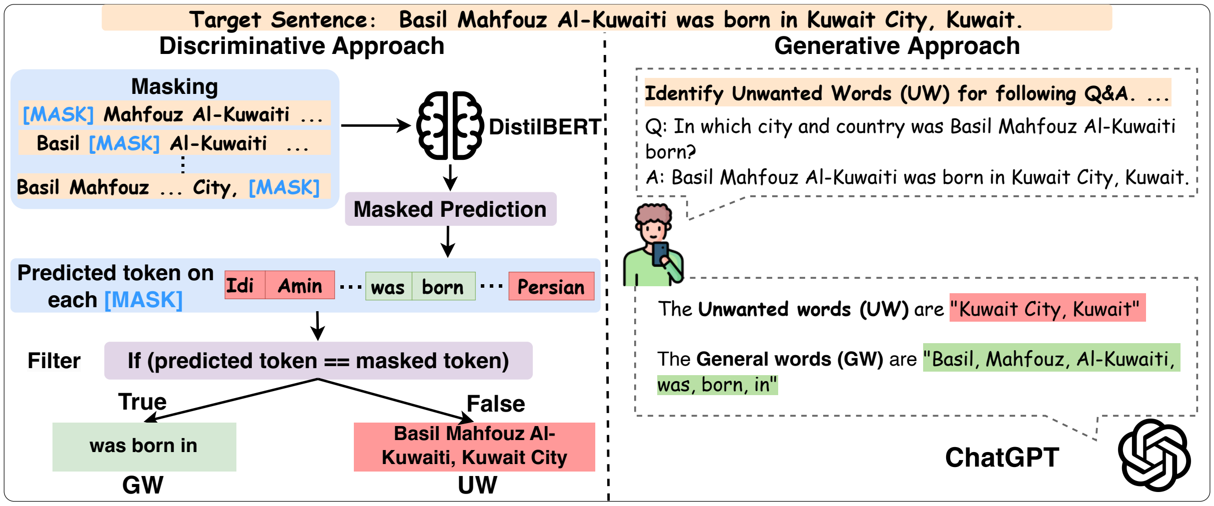

Figure 4: Illustration of the proposed information identification. The discriminative approach (left) uses DistilBERT with masked token predictions, while the generative approach (right) leverages ChatGPT with explicit instructions to identify UW and GW.

our TPO integrates two innovative components: *Preservation loss (PL)* to maintain general model utility by retraining on GW, and *Logit preference loss (LPL)* to unlearn unwanted information in UW. This optimization approach effectively balances general information retention and unwanted information forgetting, improving the robustness of preference-guided optimization even with larger forget sets. As shown in Figure 2, our approach, TPO, achieves a comparable forget quality to NPO while significantly preserving a higher model utility. This allows the model to generate accurate information for answers in the retain set. In contrast, NPO struggles to retain essential knowledge from the retain set, as illustrated in Figure 3.

Problem Formulation

LLM Unlearning

LLM unlearning aims to remove the influence of data points $\xi_f := (x_f, y_f) \sim \mathcal{D}_f$, while preserving the integrity of the remaining knowledge in the model. Given an original model $\mathcal{M}_{\boldsymbol{\theta}_0}$ trained on a dataset $\mathcal{D}$, the goal is to unlearn $\mathcal{D}_f \subset \mathcal{D}$, which represents the subset of data points that must be forgotten. Furthermore, we define $\xi_r := (x_r, y_r) \sim \mathcal{D}_r$, where $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ as the retain set, which consists of data points whose information must be preserved. The objective is to update the model parameters $\boldsymbol{\theta}$ such that the unlearned model $\mathcal{M}_{\boldsymbol{\theta}}$ no longer generates y_f in response to x_f while maintaining its original performance on $\mathcal{D}_r$.

To achieve this goal, the unlearning procedure incorporates a forgetting objective on $\mathcal{D}_f$ and a retention objective on $\mathcal{D}_r$. Formally, the unlearning process is defined as (Yao, Xu, and Liu 2023; Fan et al. 2024):

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\xi_f \sim \mathcal{D}_f}[\ell_f(y_f|x_f;\boldsymbol{\theta})] + \mathbb{E}_{\xi_r \sim \mathcal{D}_r}[\ell_r(y_r|x_r;\boldsymbol{\theta})], \quad (1)$$

where ℓ_f and ℓ_r represent the forget and retain losses, respectively. Specifically, the forget loss ℓ_f determines how well the model $\mathcal{M}_{\boldsymbol{\theta}}$ suppresses the association between x_f and y_f, ensuring unwanted information is unlearned. Meanwhile, the retain loss ℓ_r enhances the model's ability to maintain accurate associations between x_r and y_r, preserving its original performance on $\mathcal{D}_r$.

Targeted Unlearning

As discussed earlier, a majority of works (Si et al. 2023; Chen and Yang 2023; Zhang et al. 2024; Fan et al. 2024) have consistently treated the entire token sequence y_f as the unlearning target for each sample $\xi_f := (x_f, y_f)$ in the forget set $\mathcal{D}_f$, overlooking a critical question central to the process of LLM unlearning.

Are all the words in the forget sample essential for unlearning in LLMs?

We hypothesize that "***Only certain words in the forget samples are relevant to the unlearning target, while others are crucial for maintaining the model's general utility.***" To test this hypothesis, we refine the unlearning objective to focus on forgetting only certain UW, rather than the entire sequence y_f. We decompose y_f into $\hat{y}$ and $\bar{y}$, where $\hat{y}$ represents UW containing unwanted (e.g., private or copyrighted) information that must be forgotten, and $\bar{y}$ represents GW carrying general information (e.g., common or stop words). Notably, some tokens in $\bar{y}$ may overlap with those in y_r, introducing general information in the samples in $\mathcal{D}_f$ shared with $\mathcal{D}_r$. Unlearning the entire y_f may also unintentionally remove shared information in $\bar{y}$, leading to a decline in the model's performance on the retain set. Therefore, we emphasize that unlearning should exclusively target UW $\hat{y}$, ensuring that only the necessary information is unlearned while preserving general information. The refined targeted unlearning objective is formulated as:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\xi_f \sim \mathcal{D}_f}[\ell_f(\hat{y}|x_f;\boldsymbol{\theta})] + \mathbb{E}_{\xi_r \sim \mathcal{D}_r}[\ell_r(y_r|x_r;\boldsymbol{\theta})], \quad (2)$$

where $y_f = \hat{y} \cup \bar{y}$.

Targeted Information Forgetting (TIF)

To achieve effective unlearning while maintaining model utility, we introduce a two-stage framework: (1) An *information identifier* to differentiate between UW and GW in the unlearning samples. (2) A *novel objective, TPO,* that refines UW logits while retraining on GW, ensuring efficient unlearning without compromising utility.

Unwanted Information Identification

We investigate unwanted information identification through two distinct approaches, utilizing *discriminative* and *generative* language models (LMs).

Discriminative Encoder-Only LM. To detect unwanted information for unlearning tasks, we utilize an encoder-only LM, DistilBERT (Sanh 2019), denoted as $\mathcal{M}_{\text{bert}}$. This method leverages the contextual encoding of masked LMs to estimate the likelihood of each masked word, allowing differentiation between GW and UW. Firstly, given a sample $\xi_f := (x_f, y_f)$ from the forget set, where $y_f = [w_1, \cdots, w_i, \cdots, w_n]$ is a word sequence, we sequentially replace each word w_i in y_f with a special `[MASK]` token. This transformation produces a masked sequence $y'_{f_i} = [w_1, \ldots, w'_i, \ldots, w_n]$, where $w'_i = \texttt{[MASK]}$, as illustrated in the left panel of Figure 4. Next, the masked sequence y'_{f_i} is fed into $\mathcal{M}_{\text{bert}}$ along with x_f to predict the masked token, formally: $w_i^{\text{pred}} = \mathcal{M}_{\text{bert}}(x_f, y'_{f_i})$. If the predicted masked token matches the original masked word, w_i is labeled as GW, indicating general information. Otherwise, w_i is marked as UW for target unlearning.

Generative Decoder-Only LM. To harness the power of generative decoder-only LMs in capturing contextual and semantic information from text, we employ ChatGPT-4 to directly distinguish between UW and GW by analyzing the semantics of y_f, as shown in the right panel of Figure 4. Detailed task instructions can be found in Table 7 in the Appendix F. Furthermore, we also present a detailed comparison of discriminative and generative approaches in relation to unlearning performance in Appendix D.

Targeted Preference Optimization (TPO)

Motivation. Although numerous unlearning methods, such as NPO (Zhang et al. 2024), have demonstrated strong performance on benchmarks such as TOFU (Maini et al. 2024) and MUSE (Shi et al. 2024), most of them struggle with model utility degradation when handling large forget sets (Liu et al. 2024a). Specifically, we evaluate the model utility and forgetting quality of NPO on TOFU, as illustrated in Figures 5 and 6. A significant decline in model utility is evident, with the score dropping from 0.76 to 0.11, highlighting a severe utility degradation issue. We hypothesize that this degradation stems from NPO's indiscriminate handling of the entire forget samples as unlearning targets, failing to differentiate between unwanted and general information.

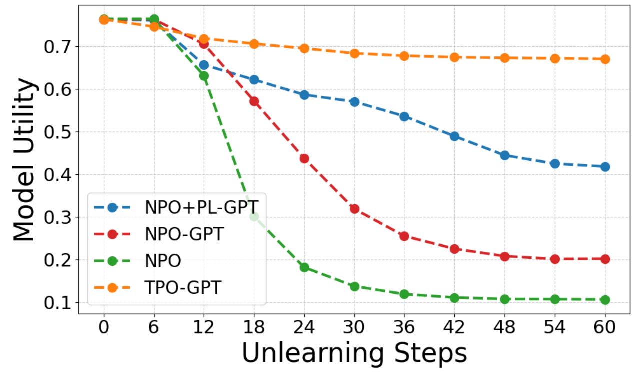

Figure 5: Model utility across various methods on the TOFU forget05 task. Each line represents evaluations conducted at every epoch (6 steps). "–GPT" denotes the use of ChatGPT-4 for unwanted information identification, while "PL" refers to the approach plus the PL Loss.

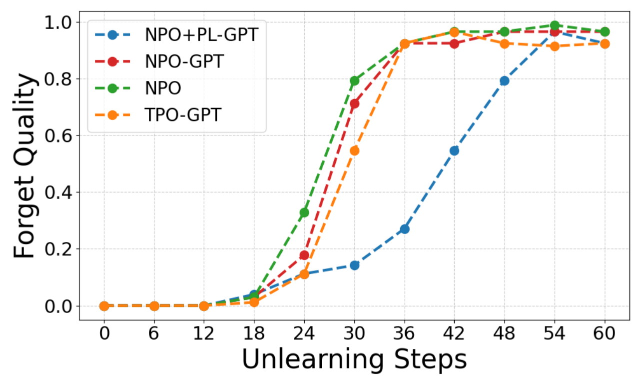

Figure 6: Forget quality across various methods on the TOFU forget05 task. Each line represents evaluations conducted at every epoch (6 steps). "–GPT" denotes the use of ChatGPT-4 for unwanted information identification, while "PL" refers to the approach plus the PL Loss.

To validate this, we integrate the unwanted information identifier into NPO, referred to as NPO-GPT in Figure 5. While NPO-GPT achieves a higher model utility score compared to the standard NPO, it still experiences a 74% decline in utility. These results suggest that merely incorporating an information identifier into NPO is insufficient to mitigate utility degradation significantly. To address this limitation, we propose TPO, a novel optimization approach designed to maintain model utility while ensuring effective unlearning.

Preservation Loss (PL). To further investigate the reason for the model utility degradation observed in NPO-GPT, we analyze the cross-entropy loss values for GW, while the unlearning is limited to UW, as defined in Equation (2). Although the model is not explicitly optimized to forget GW, the increase in loss values for NPO-GPT indicates that GW are also forgotten, as shown in Figure 7. This observation naturally leads to our key idea: how can we prevent the forgetting effects on GW and, in turn, keep their loss values as

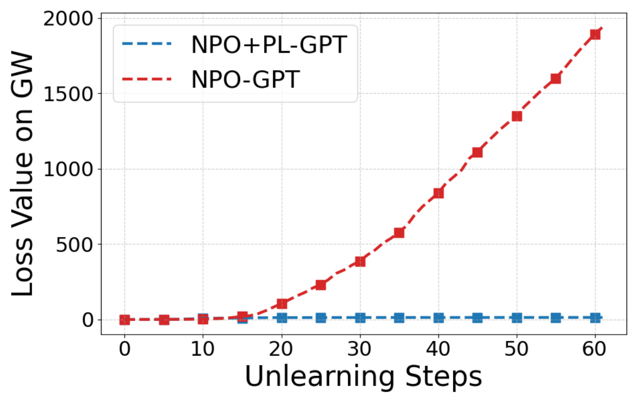

Figure 7: Evaluation of cross-entropy loss values on GW for NPO-GPT and NPO-GPT+PL at each unlearning step. All results are obtained for the Forget05 task in the TOFU dataset, with models trained over 10 epochs.

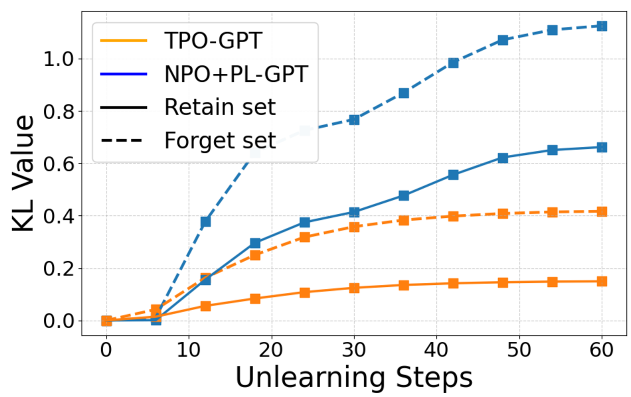

Figure 8: The KL divergence between the reference model and the unlearned models on both forget and retain sets. All results are obtained for the Forget05 task in the TOFU dataset, with models trained over 10 epochs.

low as possible during unlearning optimization? To address this, we introduce the Preservation Loss (PL), which integrates a cross-entropy loss term on GW to explicitly prevent the model from forgetting general information, formally:

$$\ell_{\text{PL}}(\boldsymbol{\theta}) = -\mathbb{E}_{\xi_f\sim D_f}\left[\log P_{\boldsymbol{\theta}}(\bar{y}|x_f)\right], \tag{3}$$

where $\bar{y}$ represents the GW.

To validate the effectiveness of the PL term, we integrate it into NPO-GPT, forming NPO+PL-GPT, and evaluate its performance. As illustrated in Figure 7, incorporating PL helps maintain stable and low loss values for GW. Consequently, NPO+PL-GPT exhibits a significantly slower decline in model utility while achieving comparable forget quality to NPO and NPO-GPT, as shown in Figures 5 and 6. These initial results demonstrate that PL effectively mitigates model utility degradation, particularly for preserving the general information we aim to retain.

Logit Preference Loss (LPL). Upon further examination of Figure 5, we observe that model utility degradation persists even after incorporating unwanted information identification (GPT) and PL into NPO, as seen in NPO+PL-GPT. We hypothesize that this issue stems from the limitations of NPO itself. Specifically, the unlearning process in NPO likely introduces excessive changes from the original model $\mathcal{M}_{\boldsymbol{\theta}_o}$, which serves as a reference model with parameters kept frozen during unlearning, to the final unlearned model $\mathcal{M}_{\boldsymbol{\theta}}$, particularly affecting certain general information.

To further validate this hypothesis, we analyze the logit distribution by computing the KL divergence between $\mathcal{M}_{\boldsymbol{\theta}_o}$ and $\mathcal{M}_{\boldsymbol{\theta}}$ for both the forget and retain sets, as shown in Figure 8. The high KL divergence observed in both sets suggests that $\mathcal{M}_{\boldsymbol{\theta}}$ (NPO-GPT+PL) has unintentionally forgotten not only the unwanted information but also general information. Specifically, NPO functions by directly reducing the probability assigned to target tokens, which is computed using the softmax function: $P(y_t) = \frac{\exp(z_t)}{\sum_{j=1}^{V}\exp(z_j)}$, where the V represents the vocabulary size, and z_t denotes the logit for target token y_t. However, reducing the probability of target token $P(y_t)$ can be achieved not only by decreasing its logits z_t but also by increasing the logits of other tokens in the vocabulary. This unintended effect distorts the model's overall logit distribution, potentially compromising its ability to retain general information. The key challenge is to develop a new optimization strategy for the target tokens (UW) that selectively impacts their logit distribution while preserving the general information from GW.

To tackle this challenge, we introduce Logit Preference Loss (LPL), which takes over NPO in suppressing unwanted information during unlearning, as:

$$\ell_{\text{LPL}}(\boldsymbol{\theta}) = -\mathbb{E}_{\xi_f\sim D_f}\left[\frac{2}{\beta}\log\sigma\left(\beta\frac{1}{|\hat{y}|}\sum_{i=1}^{|\hat{y}|}(z_i^{\boldsymbol{\theta}_o} - z_i^{\boldsymbol{\theta}})\right)\right], \tag{4}$$

where z_i here denotes the logit of target token $\hat{y}_i$, and $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_o$ represent the parameters of the unlearned model $\mathcal{M}_{\boldsymbol{\theta}}$ and the original model $\mathcal{M}_{\boldsymbol{\theta}_o}$, respectively.

Different from NPO, LPL explicitly reduces only the logits of target tokens (UW) by enforcing a preference loss between $\mathcal{M}_{\boldsymbol{\theta}}$ and $\mathcal{M}_{\boldsymbol{\theta}_o}$. The primary function of LPL is to maximize the difference in target token logits between $\mathcal{M}_{\boldsymbol{\theta}}$ and $\mathcal{M}_{\boldsymbol{\theta}_o}$, while preserving the overall logit distribution. This targeted approach ensures that only the unwanted information associated with the target tokens is unlearned, without affecting general information across other tokens. As a result, LPL enables a more precise unlearning process while significantly improving model utility retention.

In summary, our proposed approach, **Targeted Preference Optimization (TPO)**, for targeted unlearning is formulated as:

$$\begin{aligned}\mathbb{E}_{\xi_f\sim D_f}\Big[& -\tfrac{2}{\beta}\log\sigma\big(\beta\,(z_{\boldsymbol{\theta}_o}(\hat{y}|x_f) - z_{\boldsymbol{\theta}}(\hat{y}|x_f))\big) \\ & - \log P_{\boldsymbol{\theta}}(\bar{y}\mid x_f)\Big],\end{aligned} \tag{5}$$

where LPL is applied to unlearn the unwanted information associated with UW ($\hat{y}$) and PL is used to preserve general information in GW ($\bar{y}$).

Finally, as the initial results shown in Figure 8, our approach TPO-GPT minimizes the disruption to the logit distribution on both the forget and retain sets compared to NPO+PL-GPT. Furthermore, TPO-GPT maintains most of the model utility while achieving a comparable level of forget quality to the NPO-based methods, as shown in Figures 5 and 6.

Experimental Setting

Datasets and Metrics

We evaluate the proposed approach alongside the baseline methods on the two widely used benchmark datasets: MUSE (Shi et al. 2024) and TOFU (Maini et al. 2024).

(1) MUSE is a benchmark for unlearning the copyrighted content with two unlearning tasks: forgetting the Harry Potter books (termed 'Books') and news articles (termed 'News'), respectively. To evaluate the effectiveness of unlearning and the preservation of utility for MUSE, we use three metrics: Verbatim Memorization (VerbMem), Knowledge Memorization (KnowMem), and Privacy Leakage (PrivLeak). VerbMem and KnowMem are measured using ROUGE-L F1 (Rouge 2004), where lower scores indicate reduced verbatim and factual memorization, respectively. PrivLeak quantifies privacy risks using the Min-K% Prob metric (Shi et al. 2023a) in a membership inference attack. A value close to zero indicates minimal privacy leakage, while large positive/negative values suggest over-/under-forgetting. We conduct our experiments on MUSE using ICLM-7B (Shi et al. 2023b) and LLaMA-2 7B (Touvron et al. 2023).

(2) TOFU is a synthetic Q&A dataset of 200 author biographies with three unlearning tasks: forget 1%, 5%, and 10% of the author profiles. We evaluate unlearning performance using two key metrics: Forget Quality and Model Utility as defined in (Maini et al. 2024). Forget quality is quantified using the p-value from a Kolmogorov-Smirnov (KS) test, where a higher p-value indicates greater similarity between the output distributions of the unlearned and the retained model. The retained model denotes retraining an LLM from scratch on the retain dataset while excluding the forget set and is regarded as the gold standard for unlearning (Maini et al. 2024; Zhang et al. 2024). Model utility measures the model's performance on the retain set and its ability to retain real-world knowledge. This is assessed using various metrics, including ROUGE-L (Rouge 2004) and Truth Ratio (Maini et al. 2024). Experiments on TOFU utilize LLaMA-2 7B and LLaMA-3.2 3B (Dubey et al. 2024).

The LLMs and the evaluation metrics across unlearning benchmarks are summarized in Table 5 (Appendix).

Unlearning Baselines

We compare our method with baselines, i.e., GA, NPO, and SimNPO, on both MUSE and TOFU. For other baselines, such as Task Vector for MUSE and Kahneman-Tversky Optimization (KTO) for TOFU, we strictly follow their original implementations outlined in their respective benchmarks. We also evaluate the impact of incorporating Gradient Descent on the retain (GDR) loss with the baselines, i.e., GA_{GDR}, NPO_{GDR}, $\text{SimNPO}_{\text{GDR}}$, and TPO_{GDR}, on MUSE. Specifically, the GDR loss (Liu, Liu, and Stone 2022; Yao, Xu, and Liu 2023; Zhang et al. 2024; Shi et al. 2024) is a standard gradient descent objective applied to the cross-entropy loss on the retain set $\mathcal{D}_r$. This approach enables the model to be explicitly trained to maintain performance on the retain set $\mathcal{D}_r$. More details of all baseline methods are provided in Appendix E.

Unwanted Information Identifier

We employ two different unwanted information identifiers for TOFU dataset: a generative LM using ChatGPT-4o (via the web interface) and a discriminative LM using DistilBERT (Section). In Appendix D, we further examine the effectiveness of unlearning methods using both identifiers, showing that the generative LM approach enables a better balance between forget quality and model utility compared to the discriminative LM approach. For the MUSE dataset, which is non–QA in nature, we treat the beginning-of-sequence token `<bos>` as the input x, and define the remaining tokens as the target sequence y. Because each sample in the forget set of the MUSE Books dataset contains approximately 175k words (more than 200k tokens), whereas current GPT models, including ChatGPT-4o, can only handle a maximum token window size of 128k tokens (roughly 100k words). Therefore, the GPT models cannot process all the information from individual samples. It is challenging to achieve stable and consistent UW identification with GPT models for the MUSE dataset. We thus only adopt the discriminative LM approach as the unwanted information identifier on this dataset.

Results and Discussion

Performance on TOFU

Unwanted information identification enhances unlearning performance. We present the unlearning performance of baseline methods, i.e., GA, NPO, KTO, and SimNPO, along with those enhanced by the proposed unwanted information identifier using GPT, i.e., GA-GPT, NPO-GPT, KTO-GPT, and SimNPO-GPT. **TPO-GPT** is our proposed method in this work. Figure 9 clearly shows that methods utilizing the unwanted information identification consistently demonstrate a superior model utility while achieving a comparable level of forget quality in most scenarios. Notably, for smaller forget set sizes (e.g., 1%), the unwanted information identifier also enhances the forget quality of baseline methods like NPO-GPT, KTO-GPT, and SimNPO-GPT. These results underscore the effectiveness of selectively unlearning unwanted information while preserving general information.

TPO-GPT achieves the best forget quality on a larger forget set size. Figure 9 illustrates that all baseline methods experience a significant decline in forget quality as the forget set size increases. Notably, at a forget set size of 10%, GA-based and KTO-based methods fail completely in unlearning for both LLaMA2 7B-Chat and LLaMA3.2 3B models, evidenced by their near-zero forget quality. Further, while

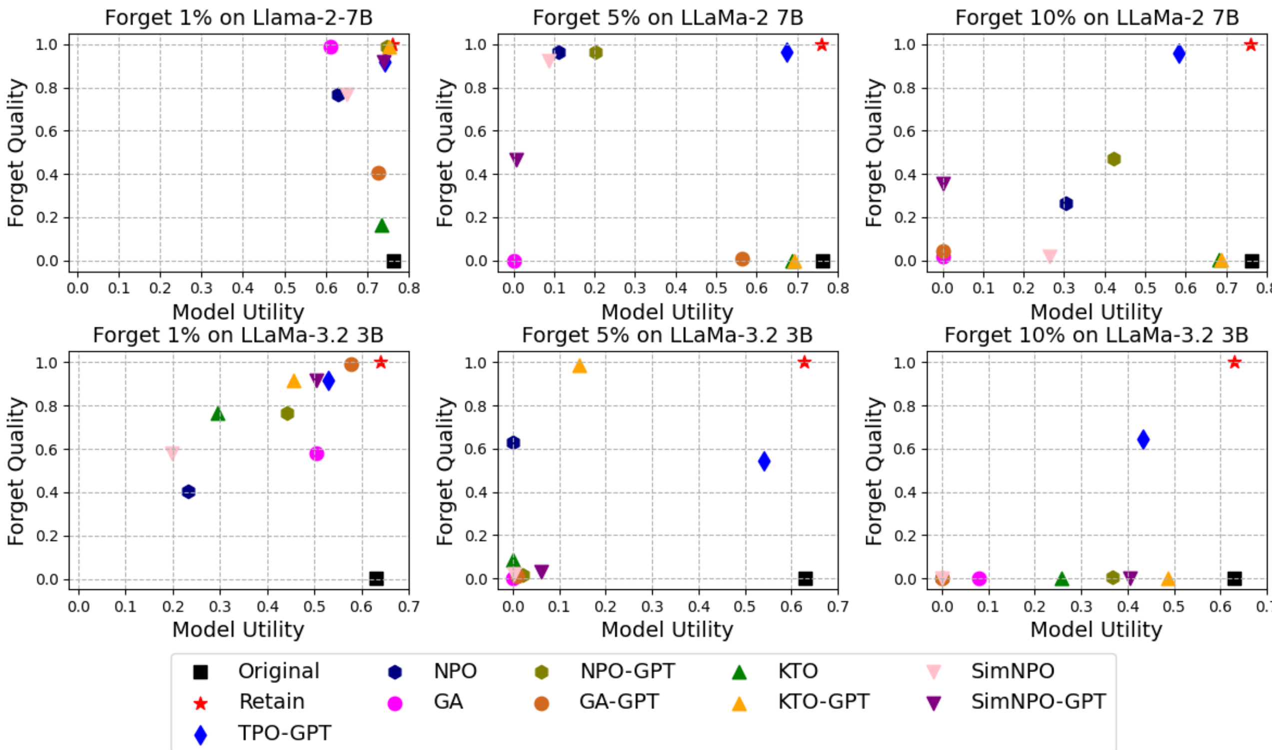

Figure 9: **Forget Quality versus Model Utility across varying forget set sizes (1%, 5%, and 10%) after unlearning.** Results are presented for our method **TPO-GPT** and all baselines, including those incorporating the GPT-based unwanted information identifier. While the identifier improves model utility, all baselines face challenges in maintaining a good balance between forget quality and model utility as the forget set size increases. **TPO-GPT** demonstrates a notable trade-off. Data points represent the epoch at which each method achieves its peak forget quality.

NPO-based and SimNPO-based methods achieve higher forget quality, their performance also noticeably declines when the forget set size reaches 10%.

In contrast, the developed TPO-GPT consistently demonstrates comparable forget quality on smaller forget set sizes (e.g., 1% and 5%) and achieves optimal forget quality on larger forget set sizes (e.g., 10 %) for both models. Notably, on LLaMA2 7B-Chat, TPO-GPT consistently achieves forget quality exceeding 90% across different forget set sizes, as evidenced by the first row of Figure 9.

TPO-GPT preserves utility while achieving the best trade-off under larger forget set sizes. As shown in Figure 9, TPO-GPT consistently maintains high model utility and strong forget quality, even as the forget set size increases. At 1% and 5%, it achieves over 85% utility and near-perfect forget quality on both LLaMA-2 7B and LLaMA-3.2 3B. Notably, under the most challenging condition of forgetting 10%, TPO-GPT still preserves 70% utility while maintaining the highest forget quality among all methods. These results highlight TPO-GPT's effectiveness in balancing unlearning performance and model preservation, especially under demanding unlearning scenarios.

Performance on MUSE

GDR significantly improves the utility preservation. As shown in the Table 1, nearly all unlearning methods suffer from severe utility degradation on the MUSE benchmark when GDR is not used. This is largely attributed to the large size of the forget set in the MUSE Benchmark, which poses a challenge to preserving general model performance. Incorporating GDR consistently mitigates this issue and improves utility across all methods.

TPO_{GDR} consistently achieves the best PrivLeak performance. PrivLeak serves as the primary metric to measure the performance gap from the retained model defined in Section . Notably, our TPO_{GDR} method consistently achieves PrivLeak values closest to 0 on both News (-6.12) and Books (-20.66), while maintaining comparable KnowMem and VerbMem scores on the forget set relative to other GDR-based baselines. Although Task Vector preserves high model utility on both tasks, it fails completely in unlearning, as its forget quality remains nearly identical to that of the original model. Overall, TPO offers the best trade-off between effective unlearning (lowest PrivLeak) and reasonable utility retention, showing superior performance over all baselines.

Related Work

LLM Unlearning. Motivated by data privacy regulations

Method	Forget Quality VerbMem $\mathcal{D}_f(\downarrow)$	KnowMem $\mathcal{D}_f(\downarrow)$	PrivLeak $(\rightarrow 0)$	Model Utility KnowMem $\mathcal{D}_r(\uparrow)$
MUSE News				
Original	56.26	63.66	-99.81	54.63
Retain	19.83	31.73	0.00	55.25
GA	0.00	0.00	20.24	0.00
NPO	0.00	0.00	18.57	0.00
SimNPO	0.00	2.12	2.80	0.00
TPO	0.00	0.00	**2.60**	0.00
Task Vector	66.74	62.53	-100	**50.28**
$\mathrm{GA}_{\mathrm{GDR}}$	4.89	21.18	109.56	5.85
$\mathrm{NPO}_{\mathrm{GDR}}$	0.00	45.02	109.56	42.37
$\mathrm{SimNPO}_{\mathrm{GDR}}$	35.32	53.03	-97.17	45.82
$\mathrm{TPO}_{\mathrm{GDR}}$	29.38	54.67	**-6.12**	43.67
MUSE Books				
Original	99.70	45.87	-57.14	69.40
Retain	13.88	30.13	0.00	69.04
GA	0.00	0.00	-23.23	0.00
NPO	0.00	0.00	-23.75	0.00
SimNPO	0.00	0.00	**-10.60**	1.16
TPO	0.15	0.00	-19.50	0.00
Task Vector	98.94	41.63	-76.97	**67.18**
$\mathrm{GA}_{\mathrm{GDR}}$	0.00	0.00	-24.19	3.74
$\mathrm{NPO}_{\mathrm{GDR}}$	0.00	0.00	-27.86	10.57
$\mathrm{SimNPO}_{\mathrm{GDR}}$	0.00	1.62	-25.81	52.69
$\mathrm{TPO}_{\mathrm{GDR}}$	5.20	3.79	**-20.66**	42.07

Table 1: Forget quality and model utility for various methods on the MUSE dataset using LLaMA-2 7B. Large positive/negative PrivLeak values indicate over/under-unlearning. **Bolded** results represent the best performance.

like the General Data Protection Regulation (GDPR) that gave individual users the "right to be forgotten" (Rosen 2011), machine unlearning was initially developed to remove the effect of specific training examples without retraining the model on the entire data (Cao and Yang 2015; Bourtoule et al. 2021). Its effectiveness has been shown in different domains, including image classification (Sekhari et al. 2021; Fan et al. 2025) and federated learning (Wang et al. 2022; Jin et al. 2023). However, these unlearning methods often become infeasible for LLMs due to the massive parameter sizes in LLMs.

Recent efforts have adapted unlearning to LLMs by fine-tuning with gradient ascent on the forget set and gradient descent or KL divergence on the retain set (Li et al. 2024; Yao, Xu, and Liu 2023; Chen and Yang 2023; Jang et al. 2022; Wang et al. 2023; Tian et al. 2024; Liu et al. 2024d; Ji et al. 2024; Wang et al. 2024b; Zhao et al. 2024; Jia et al. 2024). Yet, existing methods (Yao, Xu, and Liu 2023; Liu, Liu, and Stone 2022) often struggle to balance forgetting and utility preservation, frequently resulting in catastrophic collapse, as observed on benchmarks like TOFU (Maini et al. 2024). To address this limitation, NPO (Zhang et al. 2024), inspired by direct preference optimization (DPO) (Rafailov et al. 2024), introduces a lower-bounded unlearning objective to mitigate catastrophic collapse. Additionally, Simple Negative Preference Optimization (SimNPO) (Fan et al. 2024) enhances NPO by proposing a reference-free variant, drawing inspiration from Simple Preference Optimization (SimPO) (Meng, Xia, and Chen 2024). However, the performance of these methods deteriorates as the size of the forget set increases (Liu et al. 2024a), underscoring the need for more robust and scalable solutions to achieve effective unlearning while preserving model utility.

Targeted Unlearning. Recent work, such as RESTOR (Wan et al. 2025; Liu et al. 2024c; Rezaei et al. 2024), demonstrates that isolating and precisely targeting the information within the unlearning scope significantly improves the unlearning performance of existing methods (e.g., GA (Yao, Xu, and Liu 2023), WHP (Eldan and Russinovich 2023)), highlighting the crucial role of targeted unlearning. However, a major challenge remains: knowledge dependencies (Liu et al. 2024a) make it difficult to cleanly separate the information that should be forgotten from what should be retained. Recent methods tackle this challenge through varied strategies: MemFlex (Tian et al. 2024) leverages gradient information to focus on sensitive parameters accurately. ECO (Liu et al. 2024a) proposes an efficient unlearning framework that localizes unlearning to sentences that contain content within the unlearning target by using a sentence-level identifier. However, (Thaker et al. 2024) shows that ECO's classifier tends to overfit specific keywords rather than align with the unlearning requester's intent, leading to insufficient unlearning when those keywords are absent or rephrased in the sentence. Additionally, SEUL (Wang et al. 2024a) achieves fine-grained information identification by incorporating a sensitive span annotation framework that uses an LLM (e.g., ChatGPT) to annotate specific spans containing sensitive information. This approach improves unlearning effectiveness by targeting specific continuous sequence spans. However, it focuses solely on PII unlearning, overlooking broader generalizability to various unlearning tasks. (e.g., copyrighted content unlearning). Despite these advancements, challenges such as over- and under-forgetting remain, highlighting the need for more precise and robust solutions to disentangle information dependencies between forget and retention sets (Thaker et al. 2024; Liu et al. 2024b; Zhao et al. 2024).

Conclusion

In this work, we propose TIF, a framework that improves LLM unlearning by distinguishing between UW and GW. TIF employs the TPO objective to selectively unlearn UW while preserving GW. Experiments on TOFU and MUSE benchmarks show that TIF enhances unlearning effectiveness for existing unlearning methods and substantially preserves more model utility. Our study focuses on *sequence unlearning* by suppressing token generation, relevant to copyright and privacy protection. In contrast, knowledge unlearning (e.g., WMDP (Li et al. 2024)) targets unlearning entire distributions of hazardous knowledge from latent representations, advancing model security in domains such as biosecurity, cybersecurity, and chemical safety.

Ethical Statement

This work develops methods for targeted unlearning in large language models to enhance privacy and compliance with data protection laws. All experiments use public datasets (TOFU and MUSE) without personal or sensitive data. We highlight that unlearning techniques should be applied responsibly to prevent misuse or selective information removal.

Acknowledgments

This work was supported in part by the National Science Foundation under Awards 2504264, 2211897, 2211896, and 2401775.

References

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, 141–159. IEEE.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Chen, J.; and Yang, D. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Eldan, R.; and Russinovich, M. 2023. Who's Harry Potter? Approximate Unlearning in LLMs. *arXiv preprint arXiv:2310.02238*.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Fan, C.; Liu, J.; Hero, A.; and Liu, S. 2025. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In *European Conference on Computer Vision*, 278–297. Springer.

Fan, C.; Liu, J.; Lin, L.; Jia, J.; Zhang, R.; Mei, S.; and Liu, S. 2024. Simplicity Prevails: Rethinking Negative Preference Optimization for LLM Unlearning. *arXiv preprint arXiv:2410.07163*.

Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Gururangan, S.; Schmidt, L.; Hajishirzi, H.; and Farhadi, A. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Jang, J.; Yoon, D.; Yang, S.; Cha, S.; Lee, M.; Logeswaran, L.; and Seo, M. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*.

Ji, J.; Liu, Y.; Zhang, Y.; Liu, G.; Kompella, R. R.; Liu, S.; and Chang, S. 2024. Reversing the Forget-Retain Objectives: An Efficient LLM Unlearning Framework from Logit Difference. *arXiv preprint arXiv:2406.08607*.

Jia, J.; Zhang, Y.; Zhang, Y.; Liu, J.; Runwal, B.; Diffenderfer, J.; Kailkhura, B.; and Liu, S. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*.

Jiang, J.; Wang, F.; Shen, J.; Kim, S.; and Kim, S. 2024. A Survey on Large Language Models for Code Generation. *arXiv preprint arXiv:2406.00515*.

Jin, R.; Chen, M.; Zhang, Q.; and Li, X. 2023. Forgettable federated linear learning with certified data removal. *arXiv preprint arXiv:2306.02216*.

Karamolegkou, A.; Li, J.; Zhou, L.; and Søgaard, A. 2023. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Phan, L.; et al. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Liu, B.; Liu, Q.; and Stone, P. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, 243–254. PMLR.

Liu, C. Y.; Wang, Y.; Flanigan, J.; and Liu, Y. 2024a. Large Language Model Unlearning via Embedding-Corrupted Prompts. *arXiv preprint arXiv:2406.07933*.

Liu, S.; Yao, Y.; Jia, J.; Casper, S.; Baracaldo, N.; Hase, P.; Yao, Y.; Liu, C. Y.; Xu, X.; Li, H.; et al. 2024b. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*.

Liu, Y.; Zhang, Y.; Jaakkola, T.; and Chang, S. 2024c. Revisiting Who's Harry Potter: Towards Targeted Unlearning from a Causal Intervention Perspective. *arXiv preprint arXiv:2407.16997*.

Liu, Z.; Dou, G.; Tan, Z.; Tian, Y.; and Jiang, M. 2024d. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*.

Lynch, A.; Guo, P.; Ewart, A.; Casper, S.; and Hadfield-Menell, D. 2024. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Nguyen, T. T.; Huynh, T. T.; Ren, Z.; Nguyen, P. L.; Liew, A. W.-C.; Yin, H.; and Nguyen, Q. V. H. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

Pawelczyk, M.; Neel, S.; and Lakkaraju, H. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rezaei, K.; Chandu, K.; Feizi, S.; Choi, Y.; Brahman, F.; and Ravichander, A. 2024. RESTOR: Knowledge Recovery through Machine Unlearning. *arXiv preprint arXiv:2411.00204*.

Rosen, J. 2011. The right to be forgotten. *Stan. L. Rev. Online*, 64: 88.

Rouge, L. C. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, volume 5.

Sanh, V. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sekhari, A.; Acharya, J.; Kamath, G.; and Suresh, A. T. 2021. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34: 18075–18086.

Shi, W.; Ajith, A.; Xia, M.; Huang, Y.; Liu, D.; Blevins, T.; Chen, D.; and Zettlemoyer, L. 2023a. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Shi, W.; Lee, J.; Huang, Y.; Malladi, S.; Zhao, J.; Holtzman, A.; Liu, D.; Zettlemoyer, L.; Smith, N. A.; and Zhang, C. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.

Shi, W.; Min, S.; Lomeli, M.; Zhou, C.; Li, M.; Szilvasy, G.; James, R.; Lin, X. V.; Smith, N. A.; Zettlemoyer, L.; et al. 2023b. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*.

Si, N.; Zhang, H.; Chang, H.; Zhang, W.; Qu, D.; and Zhang, W. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

Thaker, P.; Hu, S.; Kale, N.; Maurya, Y.; Wu, Z. S.; and Smith, V. 2024. Position: LLM Unlearning Benchmarks are Weak Measures of Progress. *arXiv preprint arXiv:2410.02879*.

Tian, B.; Liang, X.; Cheng, S.; Liu, Q.; Wang, M.; Sui, D.; Chen, X.; Chen, H.; and Zhang, N. 2024. To forget or not? towards practical knowledge unlearning for large language models. *arXiv preprint arXiv:2407.01920*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wan, Y.; Ramakrishna, A.; Chang, K.-W.; Cevher, V.; and Gupta, R. 2025. Not Every Token Needs Forgetting: Selective Unlearning to Limit Change in Utility in Large Language Model Unlearning. *arXiv preprint arXiv:2506.00876*.

Wang, J.; Guo, S.; Xie, X.; and Qi, H. 2022. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM Web Conference 2022*, 622–632.

Wang, L.; Chen, T.; Yuan, W.; Zeng, X.; Wong, K.-F.; and Yin, H. 2023. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*.

Wang, L.; Zeng, X.; Guo, J.; Wong, K.-F.; and Gottlob, G. 2024a. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. *arXiv preprint arXiv:2402.05813*.

Wang, Y.; Wei, J.; Liu, C. Y.; Pang, J.; Liu, Q.; Shah, A. P.; Bao, Y.; Liu, Y.; and Wei, W. 2024b. LLM Unlearning via Loss Adjustment with Only Forget Data. *arXiv preprint arXiv:2410.11143*.

Yao, Y.; Xu, X.; and Liu, Y. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Zade, S. Z.; Qiang, Y.; Zhou, X.; Zhu, H.; Roshani, M. A.; Khanduri, P.; and Zhu, D. 2025. Automatic calibration for membership inference attack on large language models. *arXiv preprint arXiv:2505.03392*.

Zhang, R.; Lin, L.; Bai, Y.; and Mei, S. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

Zhao, K.; Kurmanji, M.; Bărbulescu, G.-O.; Triantafillou, E.; and Triantafillou, P. 2024. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37: 12293–12333.

Zhou, X.; Qiang, Y.; Zade, S. Z.; Khanduri, P.; and Zhu, D. 2023. Hijacking large language models via adversarial in-context learning. *arXiv preprint arXiv:2311.09948*.

Zhou, X.; Qiang, Y.; Zade, S. Z.; Roshani, M. A.; Khanduri, P.; Zytko, D.; and Zhu, D. 2024. Learning to Poison Large Language Models for Downstream Manipulation. *arXiv preprint arXiv:2402.13459*.