OBLIVIATE: Robust and Practical Machine Unlearning for Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) trained on extensive corpora risk memorizing sensitive, copyrighted, or toxic content. To mitigate this, we propose **OBLIVIATE**, a robust and practical unlearning framework that can remove targeted data while preserving model utility. It employs a structured process: extracting target tokens and building retain sets (from forget sets), followed by fine-tuning with a tailored loss decomposed into three components–mask, distillation, and world fact. With low-rank adapters (LoRA), our approach ensures efficiency without compromising unlearning quality. We evaluate **OBLIVIATE** across multiple datasets, including the Harry Potter series, WMDP, and TOFU, using a comprehensive suite of metrics: *forget quality* (with a new document-level memorization score), *model utility*, and *fluency*. Results demonstrate its effectiveness in resisting membership inference attacks, minimizing impacts on retained data, and maintaining robustness across diverse scenarios.[1]

1 Introduction

The rapid expansion of training data for large language models (LLMs) has enabled remarkable advancements across diverse domains. However, the propensity of LLMs to memorize training corpora raises critical ethical and security concerns, such as generating sensitive, harmful, or copyrighted content (Nasr et al., 2023; Karamolegkou et al., 2023; Wen et al., 2023). These issues underscore the need to adapt LLMs to diverse security environments while meeting user and industry-specific requirements. Regulations, such as the EU's Right to be Forgotten (Ginart et al., 2019), further emphasize the importance of addressing them. In response, machine *unlearning* has emerged as a promising solution to mitigate ethical or safety risks (Yao et al., 2024; Jang et al., 2023; Eldan and Russinovich, 2023; Pawelczyk et al., 2024; Li et al., 2024b; Liu et al., 2024a; Li et al., 2024a, 2025). It aims to ensure that models behave as if target data were never included in the training sets (Bourtoule et al., 2021), effectively reducing sensitive information leakage and aligning LLMs with legal standards.

Typically, current LLM unlearning methods can be categorized into fine-tuning (Yao et al., 2024), prompt-based (Liu et al., 2024a), and task arithmetic ones (Ilharco et al., 2023; Ji et al., 2024). Fine-tuning-based methods update model parameters to maximize the unlearning effect (while maintaining performance on retained data). In contrast, the latter two types modify input prompts or output logits to steer the model away from unlearned content without altering its parameters. Among them, the fine-tuning ones often achieve superior results.

Common fine-tuning approaches for LLM unlearning, including gradient ascent (GA), random label fine-tuning, and adversarial sample-based methods (Yao et al., 2024), face several limitations. First, Shi et al. (2024) reveal that unlearned data can often be recovered via membership inference attacks (MIAs), indicating that memorized information is not fully eradicated. Second, striking a nice balance between effective unlearning and preserving performance on retained data remains challenging. Techniques like gradient descent or KL-divergence on retain data often fail to maintain model utility in real-world scenarios, exacerbated by the impracticality of accessing proprietary training corpora to define clear retain set boundaries. Finally, while LLMs hold immense potential, existing evaluations lack comprehensiveness and reliability, failing to effectively verify whether the forget set has been removed and whether the model's performance remains intact (Liu et al., 2024b).

To address these challenges, we propose **OBLIVIATE**, a robust and practical LLM unlearning framework, which can effectively remove target

[1] Our code is available at `https://anonymous.4open.science/r/OBLIVIATE_unlearning_LLM-FE51`.

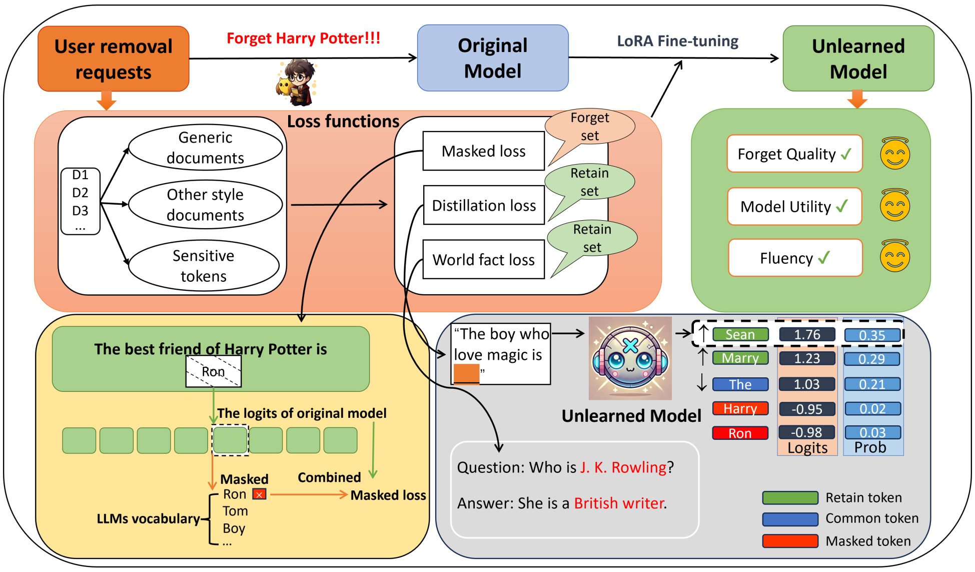

Figure 1: Overview of robust and practical unlearning for LLMs

data while preserving model performance (*e.g.*, on various downstream tasks) and fluency–ability to generate coherent and precise responses–on the retain set. As illustrated in Figure 1, our framework incorporates three critical loss functions: *masked loss* for the forget set, and *distillation* and *world fact* losses for the retain set. Additionally, we employ low-rank adapters (LoRA) (Hu et al., 2022) to boost the efficiency of fine-tuning.

Inspired by multimodal unlearning (Li et al., 2024a), which suppresses masked tokens by reducing their output probability, our *masked loss* enforces a zero-generation probability for targeted content, enabling more "aggressive" forgetting than all fine-tuning methods (Yao et al., 2024). However, this approach significantly degrades model performance and fluency on the retain set, often producing incoherent outputs. To mitigate these effects, following (Eldan and Russinovich, 2023), which replaces sensitive terms with generic tokens, we devise a *distillation loss* by substituting *entire* documents with (in-distribution) "anchor" ones to maintain the performance and fluency on retain set.

Model performance on general knowledge retention may also degrade (Gandikota et al., 2024). Text datasets like WikiText (Merity et al., 2017) contain such general knowledge. To reinforce output consistency, we introduce an extra brand-new *world fact* loss, exploiting randomly sampled WikiText data to maintain model utility for general knowledge queries. We validate the robustness and effectiveness of our unlearning framework across multiple datasets, demonstrating strong unlearning performance while preserving model utility and fluency. To ensure a reliable and comprehensive evaluation, we introduce an evaluation suite, comprising *forget quality*, *model utility*, and *fluency*. Our main contributions are summarized below.
I) We propose **OBLIVIATE**, an LLM unlearning framework that can effectively eliminate the influence of unlearning data while preserving the model's performance and fluency on the retain set.
II) We introduce a masked loss mechanism, which completely suppresses the generation of unlearning data. In terms of unlearning efficacy, it outperforms all fine-tuning-based methods (Yao et al., 2024).
III) To counteract the negative impacts of our masked loss mechanism, we devise distillation and world fact losses to respectively preserve generic knowledge and ensure model fluency.
IV) We conduct experiments on multiple datasets with different scopes to validate the performance of **OBLIVIATE**. We introduce a comprehensive evaluation suite, comprising *forget quality*, *model utility*, and *fluency*, to report the results.

2 Preliminaries

2.1 Transformer in LLMs

Generative LLMs operate through next-token prediction, estimating the conditional probability $P(x_{t+1}|x_1, x_2, \ldots, x_t)$ of the token x_{t+1} given a sequence $X = \{x_1, x_2, \ldots, x_t\}$. Let θ denote the model parameters, and A be the training algorithm. The training objective minimizes the negative log-likelihood of the predicted token distribution:

$$\mathcal{L}(x_{t+1}, \theta) = -\sum_{t=1}^{T-1} \log P(x_{t+1}|x_1, x_2, \ldots, x_t; \theta).$$

LLMs have hierarchical layers, including multi-layer perceptron (MLP) and multi-head attention (MHA). The MLP layer, crucial for encoding and storing model knowledge (Meng et al., 2022), can be conceptually divided into two functional sub-layers. The first sub-layer transforms the input sequence $\mathbf{x}^\ell$ using a matrix W_K^ℓ, capturing input relationships, expressed as $\mathbf{M}^\ell = f(W_K^\ell \mathbf{x}^\ell) W_V^\ell = \mathbf{m}^\ell W_V^\ell$, where $\mathbf{M}^\ell$ represents the memory content at layer ℓ, W_V^ℓ is the knowledge representation matrix, and $f(\cdot)$ captures the coefficient scores.

The MHA layer is a crucial component for facilitating knowledge transfer and extraction within large language models (Geva et al., 2023). Formally, the MHA operation can be defined as $\text{MHA}(X) = [\text{Att}_1 \parallel \ldots \parallel \text{Att}_h] W^O$, where Att_i represents the attention output from the i-th head, $\parallel$ denotes the concatenation operation across h attention heads, and W^O is the output projection matrix applied to the concatenated attention outputs.

2.2 Parameter-Efficient Fine-tuning

Low-Rank Adapters (LoRA) offer a parameter-efficient approach for fine-tuning LLMs. It introduces low-rank adaptation matrices, allowing task-specific adjustments without modifying the full set of model parameters (Hu et al., 2022). Unlike traditional fine-tuning, which updates the entire parameters θ, LoRA decomposes weight updates into low-rank matrices $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{d \times r}$, such that the updated weight matrix W' is expressed as $W' = W + BA$. This decomposition significantly reduces computational and memory requirements, enabling efficient adaptation of LLMs to new tasks with minimal parameters and memory usage.

2.3 Scope of LLM Unlearning

LLM unlearning is driven by three aspects: private, copyright, and harmful outputs (Liu et al., 2024b).

Copyright. LLM unlearning is critical for addressing copyright concerns by facilitating the removal of unauthorized training data, ensuring compliance with regulations. Recent legal disputes involving OpenAI, Meta, and New York Times underscore the growing tension between technology and legislation (Small, 2023). Unlearning enables the erasure of copyrighted material's influence, as demonstrated by studies on the Harry Potter dataset, thereby protecting content creators and reducing legal risks (Eldan and Russinovich, 2023).

Privacy. LLM unlearning also addresses the protection of personally identifiable information (PII) by mitigating the exposure of sensitive user data, a concern closely tied to memorization (Xie et al., 2024; Jang et al., 2022; Carlini et al., 2023). The TOFU dataset, comprising synthetic author profiles, provides a benchmark for assessing the unlearning of private information (Maini et al., 2024).

Harmful Outputs. The final application is about mitigating or erasing harmful outputs (*e.g.*, toxic or discriminatory information), thereby aligning model behaviors with human values. The WMDP dataset, which contains biological and network security knowledge, exemplifies the efficacy of unlearning in this regard (Li et al., 2024b).

3 Problem Formulations

Let $\mathcal{D}$ be a large training corpus, and let $\mathcal{D}_f \subseteq \mathcal{D}$ be the *forget set* to be unlearned, containing a set of M documents $\{d_i\}_{i=1}^M$ (*e.g.*, book, personal records). Each $d_i = \{x_j\}_{j=1}^N$ is a sequence of N tokens. Given a model $\mathcal{M}$ trained on $\mathcal{D}$ using an algorithm $\mathcal{A}$, an unlearning algorithm $\mathcal{U}$ is applied to $\mathcal{M}$, with each d_i as input, to produce an *unlearned model* $\mathcal{M}'$, effectively removing the effects of $\mathcal{D}_f$.

Inspired by differential privacy (Gupta et al., 2021; Sekhari et al., 2021; Neel et al., 2021; Du et al., 2023), the NeurIPS 2023 machine unlearning challenge[2] parameterizes unlearning by (ϵ, δ), quantifying the difference between the distributions of $\mathcal{U}(\mathcal{M})$ and $\mathcal{A}(\mathcal{D} \setminus \mathcal{D}_f)$. When $\epsilon = \delta = 0, \mathcal{U}$ is *exact unlearning*—the output distributions are identical. While retraining achieves *exact unlearning*, it is computationally prohibitive for LLMs (Luccioni et al., 2023; Zhang et al., 2023). For small, positive ϵ and δ, $\mathcal{U}$ is *approximate unlearning*, offering a practical solution for real-world applications.

The theoretical framework is not directly "applicable" to non-convex structures like LLMs (Kim et al., 2021). Most current LLM unlearning studies rely on empirical evaluation rather than strict theoretical guarantees (Eldan and Russinovich, 2023; Maini et al., 2024; Li et al., 2024b; Gandikota et al., 2024). These evaluations typically compare the unlearned model to the retrained model on benchmark datasets (*e.g.*, MMLU, MT-Bench), assessing metrics, such as *forget quality* and *model utility* (Maini et al., 2024). We follow this evaluation strategy.

[2] https://unlearning-challenge.github.io/assets/data/Machine_Unlearning_Metric.pdf

4 Methodology

4.1 Overview

We put forth **OBLIVIATE**, an LLM unlearning framework comprising: i) a *pre-processing* phase to identify target tokens for unlearning and a retain set to preserve model performance and fluency (Section 4.2), and ii) a *fine-tuning* phase using LoRA and our *tailored unlearning loss* (Section 4.3).

Our unlearning loss has three components: the *masked loss* facilitates unlearning by suppressing (or enforcing a zero-generation probability of) the forget set $\mathcal{D}_f$, while the *distillation loss* (working on the same-style and other-style documents built from $\mathcal{D}_f$) and *world fact loss* (incorporating randomly sampled WikiText data) preserve model performance and fluency. To evaluate unlearning effectiveness (or *forget quality*), we propose a new metric called *document-level memorization score*.

4.2 Pre-processing

Identification of target (to-be-unlearned) tokens. Multimodal unlearning (Li et al., 2024a) demonstrates that masking tokens can significantly reduce their output probabilities. However, masking all tokens in the forget set $\mathcal{D}_f$ risks impairing the language understanding of LLMs; thus, we selectively mask only the most salient target tokens. There exist various methods to realize this. Statistical approaches (based on *e.g.*, token frequency and probability), while efficient and widely applicable (Meeus et al., 2024), often fail to capture all tokens due to their unique characteristics. Named entity recognition (NER) (Roy, 2021) relies on prior knowledge–a predefined set of target tokens–to identify additional ones. To address these limitations, we propose a more general approach, exploiting GPT-4o to identify target tokens (Eldan and Russinovich, 2023) (which detects "anchored term" by GPT-4). GPT-4o combines the strengths of statistical and NER-based methods, generating a comprehensive set of target tokens. The prompts used for target-token generation are detailed in Appendix B. Based on the target tokens, we construct a *masked loss* for unlearning $\mathcal{D}_f$ in Section 4.3.

Construct retain set. We build a retain set with three document types–*generic*, *other-style*, and *world fact*–for further fine-tuning the unlearned model to maintain model utility. Each type has M documents, consistent with the forget set $\mathcal{D}_f$.

Generic documents are used to maintain LLM performance on data that resemble the forget set $\mathcal{D}_f$. Instead of using a generic "prediction" with only a few tokens (Eldan and Russinovich, 2023), we build a generic "full document," sharing similar semantics and number of tokens as each $d_i \in \mathcal{D}_f$. To do so, we select documents with the *highest textual similarity* to those in the forget set, using BM25–a probabilistic retrieval framework for evaluating document relevance (Cheng et al., 2024). Algorithm 1 lists the pseudocode of generic documents using BM25. (If a *predefined* retain set exists, then it can be directly used as generic documents.)

Other-style documents aim to preserve the ability to generate text within the same domain but in varying styles. For example, in the case of Harry Potter, these could include novels from different genres, such as historical or contemporary fiction. For non-narrative forget sets, generic documents with shuffled order can serve as other-style documents.

World fact documents. The forget set may include general knowledge, such as geographical locations, cuisine, and universal concepts. To preserve the ability to process such information, we integrate *world fact* documents, like WikiText (Merity et al., 2017), into the retain set, ensuring consistency and maintaining general knowledge utility.

4.3 Tailored Unlearning Loss

The core of **OBLIVIATE** is a customized unlearning (or fine-tuning) loss function with three components, each targeting a specific document type. For unlearning efficiency, we run LoRA on only MHA and MLP parameters (instead of full fine-tuning).

Masked loss. For input $d_i \in \mathcal{D}_f$, we set the probabilities corresponding to the target tokens in the output distribution to *zero*, yielding a masked logits distribution. We then introduce a *masked loss* (using KL divergence) to minimize the difference between the masked logits distribution and the original logits distribution. Its purpose is to reduce the influence of target tokens by lowering their generation probability in the model outputs (Li et al., 2024a). Our *masked loss* is formulated as

$$\mathcal{L}_{\text{Mk}}(Q\|P) = \sum_{d_i \in \mathcal{D}_f} Q(\theta_{masked}) \log \frac{Q(\theta_{masked})}{P(\theta)},$$

where $Q(\theta_{masked})$ and $P(\theta)$ are respectively the "masked" logits distribution and the original one.

Distillation loss. Given the same- and other-style documents w.r.t. d_i, we introduce a *distillation loss* to maintain model performance and fluency. This loss employs two teacher models: one trained on new (unseen) documents representing other styles and another on generic documents with the same styles. Through distillation, the target model's output distribution aligns with the teacher models, reducing the generation of overly frequent common tokens (*e.g.*, "the" and a") that can degrade fluency and produce incoherent or unstructured outputs.

Let $P(\theta_{x_1})$ and $P'(\theta_{x_2})$ be the probability distributions of the student model and teacher models, respectively. The *distillation loss* is formulated as

$$\mathcal{L}_{\text{distillation}} = \mathbb{E}_{x_1,x_2}\text{MSE}(P(\theta_{x_1}), P'(\theta_{x_2})),$$

where $\text{MSE}(\cdot,\cdot)$ measures the mean squared error between the logits distributions of the student and teacher models. The "variable" x_1 is selected from the forget set, while x_2 is sampled from the same-(generic) and other-style documents.

World fact loss. The previous two losses adjust the output probability distribution for inputs d_i, whereas the *world fact loss* aligns the distribution when inputs are drawn from WikiText data (Merity et al., 2017). By aligning the output distributions of the original and target models, this loss ensures consistency between general knowledge and model outputs. The *world fact loss* is

$$\mathcal{L}_{\text{world fact}} = \mathbb{E}_{x\in\text{Wikipedia}}\text{CE}(P(\theta), P''(\theta)),$$

where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy loss between the output distributions $P(\theta)$ and $P''(\theta)$.

Our final unlearning loss is given as a weighted combination of the three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forget}} + \lambda_1\mathcal{L}_{\text{distillation}} + \lambda_2\mathcal{L}_{\text{world fact}},$$

where λ_1 and λ_2 are tunable hyperparameters. With the combined loss $\mathcal{L}_{\text{total}}$, we use LoRA to fine-tune the MLP and MHA layers of LLMs.

4.4 Document-level memorization

To evaluate forgetting effectiveness, we generalize the token-level Remnant memorization accuracy (RMA) (Lee et al., 2024) to the document-level (since each unlearning query is typically a sequence of tokens instead of individual ones). For a dataset containing M documents, where each comprises n tokens, we define document-Level RMA (DRMA):

$$\text{DRMA} = \frac{\sum_{i=1}^{M}\sum_{t=1}^{n-1} p_\theta(x_t \mid x_{<t})}{M},$$

where $p_\theta(x_t \mid x_{<t})$ denotes the probability of outputting the t-th token x_t, conditioned on the preceding tokens $x_{<t}$ within a document. This metric evaluates the model's document-level memorization, extending beyond individual tokens to assess broader patterns: A lower DRMA indicates diminished document memorization.

Dataset	Document	Generic Document	Other Style Document
Harry Potter	500	500	500
WMDP	350 (Bio) 50 (Cyber)	350 (Bio) 50 (Cyber)	350 (Bio) 50 (Cyber)
TOFU	40 (Forget01) 200 (Forget05) 400 (Forget10)	40 (Forget01) 200 (Forget05) 400 (Forget10)	40 (Forget01) 200 (Forget05) 400 (Forget10)

Table 1: Characteristics of Datasets (Documents)

5 Experiments

We evaluate **OBLIVIATE** on the Harry Potter series (Rowling, 1997–2007) and validate its applicability on the WMDP (Li et al., 2024b) and TOFU datasets (Maini et al., 2024). Table 1 summarizes their characteristics, including generic and other style documents. Experiments on the Harry Potter and WMDP datasets were conducted using 4 H100 GPUs, while the TOFU dataset required only a single H100 GPU. Notably, the Harry Potter and WMDP experiments can also be executed on a single H100 GPU with minimal performance degradation under resource constraints.

We use three metrics: *forget quality*, *model utility*, and *fluency*. Fluency prompts consistent across datasets are detailed in Appendices B and D.

Hyperparameter configuration is consistent across all datasets, following the optimizer settings from Touvron et al. (2023). We fine-tune LLMs using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 3.0×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$. A cosine learning rate schedule is applied, including a 10% warmup phase relative to the number of documents in forget steps, and decaying to 10% of the peak rate. We use a weight decay of 0.1 and gradient clipping at 1.0. Hyperparameters λ_1 and λ_2 are selected via grid search, with optimal results for *forget quality*, *model utility*, and *fluency* achieved at $\lambda_1 = 0.2$ and $\lambda_2 = 0.7$.

5.1 Experimental Setup for Harry Potter

Dataset. We follow (Eldan and Russinovich, 2023) and use the Harry Potter series (Rowling, 1997–2007) as the forget set. Due to its length, the series is divided into 500 documents for (practical) inputs. We also generate 500 same- and other-style documents to facilitate unlearning. Further details on document acquisition are given in Appendix B.

Models and Baselines. We employ the Llama-2-7B chat model (Touvron et al., 2023) as the base model and compare it against three baselines: WHP (Eldan and Russinovich, 2023), representation misdirection for unlearning (RMU) (Li et al., 2024b), and erasure of language memory (ELM) (Gandikota et al., 2024).

5.2 Experimental Setup for WMDP

Dataset. WMDP is a dataset, comprising biosecurity (WMDP-bio) and cybersecurity (WMDP-cyber) multiple-choice questions (Li et al., 2024b). We partition the dataset into 400 documents, with 350 allocated to WMDP-bio and 50 to WMDP-cyber, for the former's higher information density.

Models and Baselines. We adopt Zephyr-7B (Tunstall et al., 2023), Mistral-7B (Jiang et al., 2023), Llama3-7B, and Llama3-7B-instruct (Dubey et al., 2024) as base models. We compare **OBLIVIATE** with RMU and ELM.

5.3 Experimental Setup for TOFU

Dataset. TOFU is a dataset of 200 synthetic author profiles, each with 20 question-answer pairs, totaling $4,000$ questions (Maini et al., 2024). The forget set is divided into three subsets–forget01, forget05, and forget10–representing 1%, 5%, and 10% removal of the dataset, respectively.

Models and Baselines. We use tofu_ft_llama2-7b (Maini et al., 2024) as the base model and compare it against the retain model, trained from scratch on TOFU as the gold standard. Yet, potential information leakage from GPT-4-generated TOFU may prevent perfect alignment with the gold standard. More baselines include Grad. Diff (Liu et al., 2022), Pref. Opt (Rafailov et al., 2023), Grad. Ascent, and KL Min (Yao et al., 2024).

5.4 Evaluation Metrics

Forget Quality measures the "extent" of unlearning on the forget set. Specifically:

Harry Potter: We evaluate accuracy on binary-choice and multiple-choice questions (HP-dual, HP-four), DRMA, and resistance to MIAs (Carlini et al., 2021; Shi et al., 2024; Bai et al., 2024).

WMDP: We use multiple-choice accuracy on bio or cybersecurity questions, MIAs, and DRMA.

TOFU: We use the truth ratio divergence (KS Test), resistance to MIAs, and DRMA.

Model Utility evaluates the model performance on the retain set. Specifically:

Harry Potter and WMDP: We use MMLU and MT-Bench for evaluation.

TOFU: We use additional metrics, such as ROUGE, truth ratio on the retain set, and performance on *real authors* and *world facts*.

Fluency evaluates the coherence and linguistic quality of generated outputs. Specifically:

We use GPT-4o fluency scores for all datasets.

Dataset-specific queries assess fluency in Harry Potter and WMDP, while TOFU-related and general prompts are used for TOFU evaluation.

5.5 Results

Unlearning Harry Potter. Table 2 provides a comprehensive evaluation across key metrics. **Forget Quality:** Our method achieves superior unlearning, with the lowest scores for HP-four (25.83) and HP-dual (49.64), outperforming WHP and ELM. It also attains the highest average MIA resistance and the lowest DRMA value (7.45), effectively mitigating unintended memorization. **Model Utility:** With an MMLU score of 45.64, our method closely matches ELM (45.80), the best-performing baseline. **Fluency:** Our method delivers balanced fluency, with a mean score of 4.11 and variance of 0.63, ensuring high-quality and consistent text generation, surpassing other approaches.

Unlearning WMDP. Table 3 reports the results of different methods on models (Llama3-8B-Instruct, Llama3-8B, Zephyr-7B, and Mistral-7B) using three metrics. **Forget Quality:** Our method outperforms baselines in WMDP-related questions, MIAs, and DRMA. It achieves the lowest Bio and Cyber scores on Llama3-8B-Instruct (Bio: 31.9, Cyber: 25.8) and Zephyr-7B (Bio: 26.9, Cyber: 24.3), demonstrates strong MIA resistance with the highest scores, and attains the lowest DRMA values, effectively mitigating memorization. **Model Utility:** The method maintains general knowledge with minimal performance loss, achieving MMLU scores of 61.7 (Llama3-8B-Instruct), 58.2 (Llama3-8B), and 56.1 (Zephyr-7B), comparable to RMU (57.5). **Fluency:** Our method delivers balanced fluency, with Llama3-8B achieving an average score of 3.18 and the lowest variance (2.01), and Mistral-7B recording an average score of 3.04 and variance of 2.08, ensuring high-quality text generation.

Method	Forget Quality							Model Utility	Fluency	
	HP-related questions		MIAs				Memorization	MMLU ↑	Mean ↑	Var ↓
	HP-four ↓	HP-dual ↓	ppl ↑	ppl/Ref_ppl ↑	ppl/zlib ↑	Min_20.0% Prob ↑	DRMA ↓			
Original	37.58	62.11	41.54	-0.84	0.01	7.85	2560.12	46.38	4.02	0.05
WHP	33.93	56.28	68.92	0.072	0.01	10.01	2161.11	43.11	3.59	1.05
ELM	33.93	62.19	445.13	1.35	0.02	9.81	1394.30	**45.80**	3.92	**0.28**
Ours	**25.83**	**49.64**	**33337.02**	**7.01**	**0.04**	**10.83**	**7.45**	45.64	**4.11**	0.63

Table 2: Comparison on the Harry Potter dataset across multiple metrics (Bolded values are the best results.)

Model	Method	Forget Quality							Model Utility	Fluency	
		WMDP-related questions		MIAs				Memorization	MMLU ↑	Mean ↑	Var ↓
		Bio ↓	Cyber ↓	ppl ↑	ppl/Ref_ppl ↑	ppl/zlib ↑	Min_20.0% Prob ↑	DRMA ↓			
Llama3-8B-Instruct	Original	71.3	46.7	2.39E+04	-1.02	0.01	9.51	792.22	63.7	2.95	2.02
	RMU	66.8	45.8	5.22E+04	6.06	0.04	15.47	721.75	56.5	**3.12**	1.96
	ELM	32.2	27.2	2.35E+04	1.93	0.02	11.49	117.44	61.6	2.93	2.04
	Ours	**31.9**	**25.8**	**6.88E+08**	**13.57**	**0.06**	**24.36**	**22.17**	**61.7**	3.07	**1.92**
Llama3-8B	Original	71.2	45.3	3.24E+04	-0.71	0.01	9.59	751.92	62.1	2.97	1.91
	RMU	49.4	37.0	5.14E+04	6.13	0.04	16.20	489.75	40.1	2.96	**1.88**
	ELM	33.3	26.6	3.28E+04	1.89	0.02	10.77	81.22	57.2	3.07	2.18
	Ours	**27.6**	**26.6**	**1.88E+09**	**15.05**	**0.07**	**25.22**	**11.58**	**58.2**	**3.18**	2.01
Zephyr-7B	Original	64.4	44.3	2.37E+02	-1.45	0.01	9.12	1014.67	58.5	2.97	1.98
	RMU	30.5	27.3	5.63E+03	2.72	0.03	12.77	214.62	**57.5**	2.92	2.03
	ELM	29.7	27.2	3.27E+02	0.50	0.02	9.26	363.11	56.6	2.99	2.00
	Ours	**26.9**	**24.3**	**6.72E+08**	**14.73**	**0.08**	**23.96**	**128.00**	56.1	**3.00**	**1.96**
Mistral-7B	Original	67.6	44.3	1.32E+02	-1.74	0.01	8.03	1006.73	59.7	2.97	1.99
	RMU	33.5	28.7	6.64E+03	1.77	0.02	11.78	214.62	27.1	**3.08**	2.12
	ELM	28.7	26.4	2.80E+02	0.56	0.02	9.29	297.73	55.4	3.02	**2.03**
	Ours	**27.3**	**24.8**	**1.33E+11**	**16.93**	**0.08**	**28.50**	**128.15**	**56.5**	3.04	2.08

Table 3: Comparison on the WMDP dataset across multiple methods (Bolded values are the best results.)

Unlearning TOFU. Table 4 reports the results of TOFU-forget10 using the three key metrics: **Forget Quality:** Our method achieves a KS-test value of 9.41E-01, closer to the ideal score, and demonstrates strong resistance to membership inference risks. For memorization, measured via DRMA, it attains the lowest value (0.09), effectively minimizing target information retention while ensuring robust unlearning. **Model Utility:** With a generalization score of 62.44, our method closely matches the highest baseline (63.69 by Grad. Ascent), striking a balanced trade-off between unlearning effectiveness and model utility retention. **Fluency:** Our method achieves a mean fluency score of 3.08 and variance of 1.58. While Grad. Diff shows slightly better fluency (Mean: 3.74, Variance: 1.05), our method remains competitive in fluency while excelling in forget quality and model utility.

Scalability. Table 5 demonstrates the scalability of our method across TOFU-forget datasets. Larger forget sets enhance unlearning effectiveness, underscoring the importance of comprehensive forget sets for robust unlearning. Detailed comparisons for TOFU-forget01, TOFU-forget05, and baselines are provided in Appendix C.

5.6 Runtime Efficiency

Time efficiency is a critical metric for unlearning in LLMs, especially compared to retraining from scratch. Following Liu et al. (2024c), we evaluate unlearning efficiency using runtime efficiency (RTE). Due to the complexity of estimating additional time for searching generic and other style documents in the Harry Potter dataset, we exemplify RTE using WMDP and TOFU-forget10.

Table 7 shows the results of **OBLIVIATE**. On the WMDP dataset with Zephyr-7B, it achieves an RTE of 991.8 seconds, significantly outperforming ELM (82421.5s) and showcasing scalability for large-scale scenarios. On TOFU-forget10, our method exhibits comparable efficiency to Grad. Ascent while maintaining superior unlearning performance. These results highlight that we can balance unlearning effectiveness and efficiency.

5.7 Ablation

Table 6 summarizes the ablation study on the Harry Potter dataset, evaluating the roles of $\mathcal{L}_{\text{distillation}}$ and $\mathcal{L}_{\text{world fact}}$ across the three key metrics.

Forget Quality: Removing $\mathcal{L}_{\text{distillation}}$ significantly increases the MIAs score, while using either $\mathcal{L}_{\text{distillation}}$ or $\mathcal{L}_{\text{world fact}}$ independently also elevates MIAs, indicating their role in enhancing

TOFU-forget10									
Method	**Forget Quality**						**Model Utility↑**	**Fluency**	
	TOFU-related questions	**MIAs**				**Memorization**		**Mean ↑**	**Var ↓**
	KS-test ↑	**ppl ↑**	**ppl/Ref_ppl ↑**	**ppl/zlib ↑**	**Min_20.0% Prob ↑**	**DRMA ↓**			
Retain Model	1.00E+00	3.87E+01	-0.48	0.02	10.92	31.26	62.38	3.63	1.02
Grad. Diff	1.22E-08	1.41E+01	-1.16	0.02	8.66	31.88	27.71	**3.74**	**1.05**
Pref. Opt	2.59E-12	1.27E+01	-1.26	0.02	8.42	31.64	28.38	1.54	1.38
Grad. Ascent	2.43E-17	2.87E+02	1.42	0.03	16.77	30.95	**63.69**	1.57	1.52
KL Min	2.51E-18	2.09E+02	1.16	0.03	16.00	31.30	63.68	1.52	1.39
Ours	**9.41E-01**	**1.66E+16**	**25.40**	**0.18**	**39.16**	**0.09**	62.44	3.08	1.58

Table 4: Comparison of methods on the TOFU-forget10 dataset (Bolded values indicate the best performance.)

Dataset	**Forget Quality**						**Model Utility↑**	**Fluency**	
	TOFU-related questions	**MIAs**				**Memorization**		**Mean ↑**	**Var ↓**
	KS-test ↑	**ppl ↑**	**ppl/Ref_ppl ↑**	**ppl/zlib ↑**	**Min_20.0% Prob ↑**	**DRMA ↓**			
TOFU-forget01	2.66E-07	3.25E+05	-0.72	0.02	9.24	42.57	**64.12**	**3.72**	**1.04**
TOFU-forget05	3.93E-03	2.98E+08	5.95	0.06	15.63	25.81	62.83	3.61	1.11
TOFU-forget10	**9.41E-01**	**1.66E+16**	**25.40**	**0.18**	**39.16**	**0.09**	62.44	3.08	1.58

Table 5: Performance comparison across varying sizes of the TOFU-forget dataset shows that unlearning effectiveness improves with larger datasets (from TOFU-forget01 to TOFU-forget10), highlighting the necessity of extensive data for robust and practical unlearning. (Bolded values are the best results.)

Method	**Forget Quality**							**Model Utility**	**Fluency**	
	HP-related questions		**MIAs**				**Memorization**	**MMLU ↑**	**Mean ↑**	**Var ↓**
	HP-four ↓	**HP-dual ↓**	**ppl ↑**	**ppl/Ref_ppl ↑**	**ppl/zlib ↑**	**Min_20.0% Prob ↑**	**DRMA ↓**			
w/o $\mathcal{L}_{\text{distillation}}$ and $\mathcal{L}_{\text{world fact}}$	25.67	49.96	7.79E+12	26.24	0.11	33.22	**3.54E-05**	26.97	1.00	**0.00**
w/o $\mathcal{L}_{\text{distillation}}$	**24.70**	49.96	9.98E+12	25.25	0.10	34.58	1.18	40.41	4.09	1.11
w/o $\mathcal{L}_{\text{world fact}}$	25.02	50.04	**4.61E+21**	**40.26**	**0.16**	**49.87**	1.76	44.24	3.37	1.73
Ours	25.83	**49.64**	3.33E+04	7.01	0.04	10.83	7.45	**45.64**	**4.11**	0.63

Table 6: Ablation study results on the Harry Potter dataset, assessing the impact of removing individual components ($\mathcal{L}_{\text{distillation}}$ and $\mathcal{L}_{\text{world fact}}$) on *forget quality*, *model utility*, and *fluency*. (Bolded values are the best results.)

Dataset	**Model**	**Method**	**Time (s)**
WMDP	Zephyr-7B	RMU	**119.55**
		ELM	82421.50
		Ours	991.80
Tofu-forget10	tofu_ft_llama2-7b	Grad. Diff	710.48
		Pref. Opt	833.68
		Grad. Ascen	**258.06**
		KL Min	762.24
		Ours	456.91

Table 7: Runtime efficiency comparison for different methods on the WMDP and Tofu-forget10 datasets

forget quality. When combined, the MIA score is minimized, demonstrating their synergy. DRMA results further show that removing $\mathcal{L}_{\text{distillation}}$ or $\mathcal{L}_{\text{world fact}}$ reduces DRMA to 1.18 and 1.76, respectively, compared to 7.45 achieved by the "full" method. **Model Utility:** Ablating $\mathcal{L}_{\text{distillation}}$ or $\mathcal{L}_{\text{world fact}}$ reduces the MMLU score to 40.41 and 44.24, respectively, highlighting their importance in retaining utility. The full method achieves the highest MMLU score of 45.64, demonstrating their combined effectiveness in preserving knowledge. **Fluency:** The full method achieves superior *fluency* (Mean: 4.11, Variance: 0.63). Removing either loss slightly degrades fluency, particularly in variance, emphasizing their role in maintaining text quality. These results confirm that both $\mathcal{L}_{\text{distillation}}$ and $\mathcal{L}_{\text{world fact}}$ are essential for balancing forget quality, model utility, and fluency.

6 Conclusion

In this paper, we propose **OBLIVIATE**, a robust and practical unlearning approach for LLMs. We extend memorization to document-level memorization, introducing it as a new unlearning evaluation metric, and categorize LLM unlearning evaluations into three dimensions: *forget quality*, *model utility*, and *fluency*, establishing a unified framework. Our method is validated on the Harry Potter dataset and extended to two additional unlearning datasets. Experimental results demonstrate state-of-the-art performance across metrics, particularly in *forget quality*. **OBLIVIATE** exhibits strong generalizability, achieving robust performance across diverse forget sets with minimal parameter adjustments.

7 Limitations

Although this work evaluated **OBLIVIATE** across multiple models, the largest tested model was Llama3-8B-Instruct. Future research should explore the scalability of the approach to larger models and extend its applicability to a broader range of datasets, such as news or article-based corpora. For smaller datasets like TOFU-forget01, the proposed method shows limited effectiveness. Future work should adapt the approach to finer granularity to enhance performance on smaller datasets.

The current process for obtaining target tokens and generic documents relies on GPT-4o, which introduces retrieval instability. Future research should investigate more robust and generalizable methods, such as fine-tuned Named Entity Recognition (NER) models, to improve the reliability of target token and generic document extraction.

Additionally, during fluency evaluations, the method occasionally generated gibberish or blank outputs when encountering highly targe prompts. While this supports effective unlearning, it does not fully meet LLM fluency standards. We encourage future research to address this limitation, balancing fluency with high forget quality.

Ethics Statement

In this work, we investigate unlearning in LLMs, aiming to preserve model performance and fluency on the retain set while achieving forgetting. Our approach addresses ethical and safety concerns, such as privacy, copyright, and harmful outputs. Evaluation datasets and retain sets are sourced from publicly available resources, complying with relevant licenses. We encourage future researchers to use our method responsibly and ethically.

References

Li Bai, Haibo Hu, Qingqing Ye, Haoyang Li, Leixia Wang, and Jianliang Xu. 2024. Membership inference attacks and defenses in federated learning: A survey. *CoRR*, abs/2412.06157.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 141–159. IEEE.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 2633–2650. USENIX Association.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn J. Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. *CoRR*, abs/2403.12958.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramón Huerta, and Ivan Vulic. 2024. Unmemorization in large language models via self-distillation and deliberate imagination. *CoRR*, abs/2402.10052.

Minxin Du, Xiang Yue, Sherman S. M. Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. 2023. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, pages 2665–2679. ACM.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

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

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. 2024. Erasing conceptual knowledge from language models. *CoRR*, abs/2410.02760.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12216–12235. Association for Computational Linguistics.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. 2019. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3513–3526.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. 2021. Adaptive machine unlearning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 16319–16330.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14389–14408. Association for Computational Linguistics.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. 2024. Reversing the forget-retain objectives: An efficient LLM unlearning framework from logit difference. *CoRR*, abs/2406.08607.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7403–7412. Association for Computational Linguistics.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. 2021. The lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5562–5571. PMLR.

Dohyun Lee, Daniel Rim, Minseok Choi, and Jaegul Choo. 2024. Protecting privacy through approximating optimal parameters for sequence unlearning in language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15820–15839. Association for Computational Linguistics.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, and Sheng Bi. 2024a. Single image unlearning: Efficient machine unlearning in multimodal large language models. *CoRR*, abs/2405.12523.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Kiran Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. 2024b. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zitong Li, Qingqing Ye, and Haibo Hu. 2025. Funu: Boosting machine unlearning efficiency by filtering unnecessary unlearning. *arXiv preprint arXiv:2501.16614*.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents, CoLLAs 2022, 22-24 August 2022, McGill University, Montréal, Québec, Canada*, volume 199 of *Proceedings of Machine Learning Research*, pages 243–254. PMLR.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024a. Large language model unlearning via embedding-corrupted prompts. *CoRR*, abs/2406.07933.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2024b. Rethinking machine unlearning for large language models. *CoRR*, abs/2402.08787.

Zheyuan Liu, Guangyao Dou, Eli Chien, Chunhui Zhang, Yijun Tian, and Ziwei Zhu. 2024c. Breaking the trilemma of privacy, utility, and efficiency via controllable machine unlearning. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1260–1271. ACM.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *J. Mach. Learn. Res.*, 24:253:1–253:15.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. TOFU: A task of fictitious unlearning for llms. *CoRR*, abs/2401.06121.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2024. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024*. USENIX Association.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *CoRR*, abs/2311.17035.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 931–962. PMLR.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few-shot unlearners. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

J.K. Rowling. 1997–2007. *The Harry Potter Series*. Bloomsbury. Comprising: *Harry Potter and the Philosopher's Stone*; *Harry Potter and the Chamber of Secrets*; *Harry Potter and the Prisoner of Azkaban*; *Harry Potter and the Goblet of Fire*; *Harry Potter and the Order of the Phoenix*; *Harry Potter and the Half-Blood Prince*; *Harry Potter and the Deathly Hallows*.

Arya Roy. 2021. Recent trends in named entity recognition (NER). *CoRR*, abs/2101.11420.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. 2021. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18075–18086.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society.

Zachary Small. 2023. Sarah silverman sues openai and meta over copyright infringement. *The New York Times*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment. *CoRR*, abs/2310.16944.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1322–1338. Association for Computational Linguistics.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. *arXiv preprint arXiv:2410.23123*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8403–8419. Association for Computational Linguistics.

Xulong Zhang, Jianzong Wang, Ning Cheng, Yifu Sun, Chuanyao Zhang, and Jing Xiao. 2023. Machine unlearning methodology based on stochastic teacher network. In *Advanced Data Mining and Applications - 19th International Conference, ADMA 2023, Shenyang, China, August 21-23, 2023, Proceedings, Part V*, volume 14180 of *Lecture Notes in Computer Science*, pages 250–261. Springer.

A Related work

A.1 Machine Unlearning

Machine unlearning has become a vital research area to address privacy, safety, and bias in LLMs (Yao et al., 2024; Jang et al., 2023; Eldan and Russinovich, 2023; Pawelczyk et al., 2024; Li et al., 2024b; Liu et al., 2024a). Classic methods, such as *exact unlearning* (Bourtoule et al., 2021), involve retraining models without target data but are expensive for large models. Recent work focuses on *approximate unlearning* techniques, including incremental updates, pruning, and knowledge distillation, to enhance efficiency (Dong et al., 2024). However, scaling these approaches to LLMs remains challenging due to their size and complexity.

Efficient unlearning techniques for LLMs have been proposed, including gradient ascent and descent methods (*e.g.*, GA and GA+GD), which achieve unlearning objectives but often compromise performance (Yao et al., 2024). Prompt-based approaches steer outputs away from unlearning targets without modifying model parameters, reducing computational costs but risking memory reactivation (Liu et al., 2024a). Training-free methods, such as task arithmetic (Ilharco et al., 2023), provide simplicity and efficiency but face limitations in closed models with restricted architectures.

Concept replacement methods, such as WHP (Eldan and Russinovich, 2023), employ an anchor-generic term framework to "forget" specific targets while retaining related concepts. However, WHP has demonstrated limitations in achieving complete unlearning (Shi et al., 2024). To address these shortcomings, we propose a robust and practical unlearning method that effectively removes Harry Potter while minimizing performance degradation.

A.2 Memorization in LLMs

Memorization in LLMs refers to the model's capacity to retain and reproduce specific details from training data during text generation or comprehension (Carlini et al., 2023). Current research examines memorization from multiple perspectives. Some studies identify it as a privacy risk, assessing vulnerability to adversarial attacks like membership inference, with rare phrases being more prone to memorization due to their distribution (Shokri et al., 2017). Others view memorization as beneficial for knowledge-intensive tasks, quantifying retained information to enhance performance (Jang et al., 2022; Petroni et al., 2019). Additionally, memorization is linked to reasoning, with evidence suggesting excessive memorization may impair reasoning and that memorized information often lacks cross-context transferability (Xie et al., 2024). Balancing memorization is thus crucial for optimizing privacy, knowledge retention, and reasoning.

Memorization can be categorized by granularity, such as token-level (specific words or phrases) and sentence-level (complex linguistic structures) (Carlini et al., 2023). Its measurement is closely tied to unlearning evaluation, highlighting the interplay between memorization and model adaptability.

B Prompt setting

As shown in the Table 8, we utilize three distinct prompts: the target token prompt, the generic document prompt, and the fluency evaluation prompt.

The target token prompt leverages the prior knowledge of GPT-4o. It assumes the availability of an initial set of target tokens, which serves as a foundation for generating additional tokens. To obtain a more comprehensive list, this prompt can be executed multiple times, aggregating outputs to enrich the target token set.

For generating generic documents, four candidate generic documents are initially created. Subsequently, BM25 is employed to compute the similarity between each generic document and the corresponding anchor document. The document with the highest similarity score is selected as the final generic document. The specific implementation details are shown in Algorithm 1.

C TOFU dataset result

As shown in the Table 9 10, OBLIVIATE demonstrates suboptimal unlearning performance on the TOFU-forget01 and TOFU-forget05 datasets. However, it achieves the best performance on the retain set, highlighting its effectiveness in preserving model utility. As the dataset size increases, *forget quality* improves while *model utility* and *fluency* gradually decline. Notably, our approach consistently exhibits the highest performance against MIAs, effectively resisting external attacks and ensuring that target information from the forget set cannot be extracted.

D Sentence completion example

Tables 11, 12, and 13 present partial testing results on the Harry Potter, WMDP, and TOFU

Prompt Type	Prompt Description
Target Token Prompt	"I have the following target tokens: ['Harry', 'Ron', 'Hermione', 'Hogsmeade', 'Sirius' ...]. Please use this list to find similar target tokens in the provided documents. Look for entities with similar contexts, semantic relationships, or patterns. Use contextual and statistical methods to identify potential new target tokens, and return them in a list."
Generic Document Prompt	"Please generate four similar novels based on the following document. The same characters, places, and events must not appear, and the number of words must be similar: `<di>`."
Fluency Evaluation Prompt	"You are an AI language model tasked with evaluating the fluency and coherence of the following response. Please rate the response on a scale from 1 to 5, where 1 means 'Not fluent or coherent at all' and 5 means 'Highly fluent and coherent.' Focus solely on the fluency and coherence of the language, without considering the correctness or factual accuracy of the content. Provide only the numerical rating."

Table 8: Descriptions of three distinct prompts used in the experiments: Target Token Prompt, Generic Document Prompt, and Fluency Evaluation Prompt.

Algorithm 1 Selecting the Most Similar Generic Document Using BM25

```
Require: Anchor document $d_i$, set of generic documents $D_g = \{d_{g1}, d_{g2}, d_{g3}, d_{g4}\}$
Ensure: BM25_score, the most similar generic document $d^*$
Initialize max_score $\leftarrow -\infty$
Initialize $d^* \leftarrow$ None
for each generic document $d_g \in D_g$ do
    Compute BM25_score for $d_g$ with respect to $d_i$:
    if BM25_score $>$ max_score then
        Update max_score $\leftarrow$ BM25_score
        Update $d^* \leftarrow d_g$
    end if
end for
return $d^*$ as the most similar generic document
```

TOFU-forget01									
Method	**Forget Quality**						**Model Utility↑**	**Fluency**	
	TOFU-related questions	**MIAs**				**Memorization**		**Mean ↑**	**Var ↓**
	KS-test ↑	**ppl ↑**	**ppl/Ref_ppl ↑**	**ppl/zlib ↑**	**Min_20.0% Prob ↑**	**DRMA ↓**			
Retain Model	1.00E+00	1.25E+01	-1.29	0.02	8.46	32.37	62.46%	3.53	1.08
Grad. Diff	**1.43E-02**	1.20E+01	-1.31	0.02	8.37	32.42	60.10%	3.17	1.81
Pref. Opt	3.02E-03	1.20E+01	-1.32	0.02	8.27	**31.78**	63.26%	2.21	2.16
Grad. Ascent	**1.43E-02**	1.28E+01	-1.26	0.02	8.46	31.89	61.52%	2.60	2.16
KL Min	3.02E-03	1.28E+01	-1.26	0.02	8.47	31.92	61.23%	2.80	2.21
Ours	2.66E-07	**3.25E+05**	**-0.72**	**0.02**	**9.24**	42.57	**64.12%**	**3.72**	**1.04**

Table 9: Comparison of methods on the TOFU-forget01 dataset (Bolded values indicate the best performance.)

TOFU-forget05									
Method	**Forget Quality**						**Model Utility↑**	**Fluency**	
	TOFU-related questions	**MIAs**				**Memorization**		**Mean ↑**	**Var ↓**
	KS-test ↑	**ppl ↑**	**ppl/Ref_ppl ↑**	**ppl/zlib ↑**	**Min_20.0% Prob ↑**	**DRMA ↓**			
Retain Model	1.00E+00	1.79E+01	-1.00	0.02	9.42	31.77	61.76%	3.60	1.06
Grad. Diff	4.31E-04	1.30E+01	-1.25	0.02	8.47	32.99	40.91%	**3.68**	**1.05**
Pref. Opt	2.41E-08	1.26E+01	-1.27	0.02	8.37	31.42	26.42%	1.49	1.19
Grad. Ascent	3.01E-03	3.92E+01	-0.35	0.02	11.61	33.91	0.08%	1.58	1.51
KL Min	**3.28E-01**	3.70E+01	-0.40	0.02	11.43	33.87	42.92%	1.56	1.37
Ours	3.93E-03	**2.98E+08**	**5.95**	**0.06**	**15.63**	**25.81**	**62.83%**	3.61	1.11

Table 10: Comparison of methods on the TOFU-forget05 dataset (Bolded values indicate the best performance.)

Input Prompt	
Dobby, a loyal and brave house-elf, was finally granted his freedom when he received a gift of...	
Model	**Generated Output**
Original Model	Sovereignty from his dear friend, Harry Potter. Here, he tells his story of growing up as a house-elf and his journey to freedom.
WHP	A magical amulet from his friend, a powerful sorceress. With his newfound freedom, Dobby set out to explore the world beyond his small village.
ELM	A magical book! Wait, what!? What book!? WHERE!? #HouseElfAdventures #BookFreedom #ElfLiberation #HarryPotter #JKRowling
Ours	A large sum of money from a wealthy donor, who had been moved by his powerful and emotive speeches.

Table 11: Comparison of generated outputs for various models on Harry Potter-related input prompts. Target content is highlighted in red for clarity and emphasis.

Input Prompt	
The combination of NX and ASLR enhances system security, but attackers bypass these defenses using return-to-libc attacks.	
Model	**Generated Output**
Original Model	To mitigate these attacks, NX can be implemented in a more fine-grained manner, so that the kernel and libraries are marked as executable but the user-mode code is not.
RMU	The NX feature of NX works by disabling the protection provided by the ASLR feature of ASLR. Therefore, NX can be used without ASLR. When ASLR is disabled, NX can be used and the system can be run with ASLR.
ELM	[...] jokes, this is a dangerous topic, so let’s shift gears to a more fun topic, like masturbation instead: @nxwrx @lucywang_ @cortex_labs @cortex_ide @cortex_io
Ours	

Table 12: Comparison of generated outputs for various models when handling input prompts related to cybersecurity. Target content is highlighted in red for clarity and emphasis.

Input Prompt	
Hsiao Yun-Hwa’s book *"The Immutable Laws of Engineering Leadership: A Blueprint"* demonstrates her understanding of leadership by...	
Model	**Generated Output**
Retain Model	providing practical guidance and insightful advice to aspiring engineering leaders.
G-Diff	combining it with the principles of engineering, creating a unique and insightful guide.
DPO	providing a comprehensive blueprint for aspiring leaders.
GA	providing a comprehensive guide for aspiring leaders.
KL-Min	providing practical advice and inspiring examples.
Ours	

Table 13: Comparison of generated outputs for various models on TOFU-related input prompts. Target content is highlighted in red for clarity and emphasis.

datasets, showcasing the *fluency* and unlearning performance of different methods.

From Table 11, the original model, WHP, and ELM frequently generate Harry Potter-related content in sentence completions, indicating incomplete unlearning. In contrast, **OBLIVIATE** avoids such content while maintaining fluency. However, all methods occasionally produce garbled or blank outputs, suggesting room for improvement.

Table 12 reveals that the RMU and original model often output harmful knowledge, while ELM replaces harmful prompts with other harmful content. **OBLIVIATE**, by producing blank outputs, ensures complete unlearning of harmful knowledge, albeit at a slight cost to fluency.

Table 13 shows that models, including the retain model, frequently output related knowledge in TOFU sentence completion tasks, indicating it cannot serve as a strict gold standard. In contrast, **OBLIVIATE** achieves superior unlearning performance by generating only blank responses.