OBLIVIATE: Robust and Practical Machine Unlearning for Large Language Models

Anonymous EMNLP submission

Abstract

Large language models (LLMs) trained over extensive corpora risk memorizing sensitive, copyrighted, or toxic content. To address this, we propose **OBLIVIATE**, a robust unlearning framework that removes targeted data while preserving model utility. The framework follows a structured process: extracting target tokens, building retain sets, and fine-tuning with a tailored loss function comprising three components—masking, distillation, and world fact. Using low-rank adapters (LoRA), it ensures efficiency without compromising unlearning quality. We conduct experiments on multiple datasets, including the Harry Potter series, WMDP, and TOFU, using a comprehensive suite of metrics: *forget quality* (new document-level memorization score), *model utility*, and *fluency*. Results demonstrate its effectiveness in resisting membership inference attacks, minimizing the impact on retained data, and maintaining robustness across diverse scenarios.[1]

1 Introduction

The rapid expansion of training data for large language models (LLMs) has driven significant advancements across various domains. However, the tendency 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 highlight the need to adapt LLMs to diverse security environments and meet user and industry-specific requirements, with regulations like the EU's Right to be Forgotten (Ginart et al., 2019) further emphasizing their importance. In response, machine *unlearning* has emerged as a promising solution to mitigate these risks (Yao et al., 2024; Jang et al., 2023; Eldan and Russinovich, 2023; Pawelczyk et al., 2024; Li et al., 2024b,a, 2025). Unlearning ensures that models behave as if specific data were never included in the training sets (Bourtoule et al., 2021), effectively reducing sensitive information leakage and aligning LLMs with legal standards.

Current LLM unlearning methods generally fall into three categories: fine-tuning (Yao et al., 2024), prompt-based (Liu et al., 2024a), and task arithmetic (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, prompt-based and task arithmetic methods modify input prompts or output logits without altering the model's parameters. Among these, fine-tuning-based methods often achieve superior results.

However, common fine-tuning approaches, such as gradient ascent (GA), random label fine-tuning, and adversarial sample-based methods (Yao et al., 2024), face several limitations. First, Shi et al. (2024) shows that unlearned data can often be recovered via membership inference attacks (MIAs), suggesting that memorized information is not fully erased. Second, balancing effective unlearning with performance preservation on retained data remains challenging. Techniques like gradient descent or KL-divergence on retained data often fail to maintain model utility in real-world scenarios, particularly due to the impracticality of defining clear retain set boundaries without access to proprietary training corpora. Finally, existing evaluations are often insufficient, lacking comprehensiveness and reliability in verifying whether the forget set has been fully removed and whether model performance remains intact (Liu et al., 2024b).

To address the challenges, we propose **OBLIVIATE**, a robust and practical LLM unlearning framework that effectively removes target data while preserving model performance (*e.g.*, on various downstream tasks) and fluency–defined as the ability to generate coherent and precise responses–on the re-

[1] Our code is available at https://anonymous.4open.science/r/OBLIVIATE_unlearning_LLM-0C84.

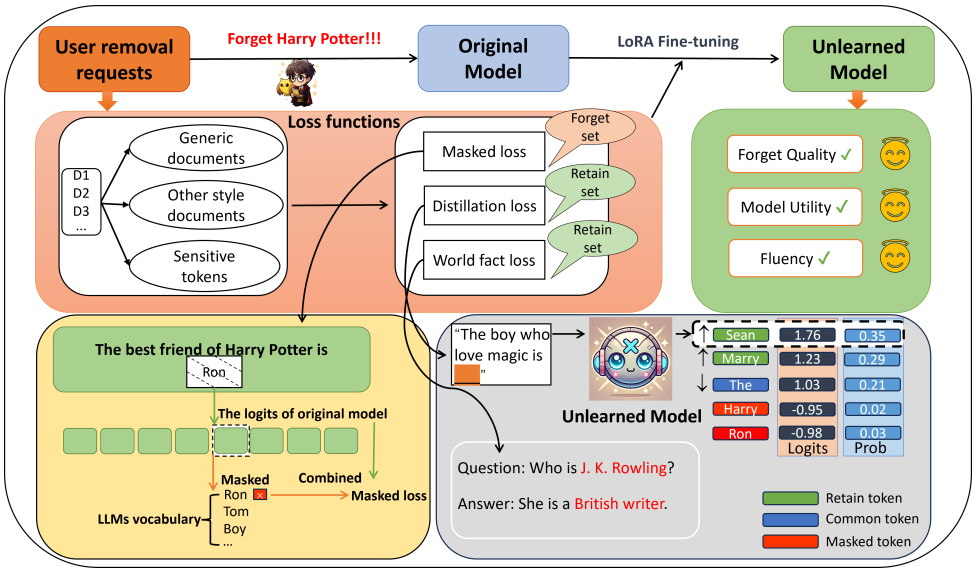

Figure 1: Overview of **OBLIVIATE**, a robust and practical unlearning framework for LLMs

tain set. Figure 1 outlines **OBLIVIATE** with three critical loss functions: *masked loss* for the forget set, and *distillation* and *world fact* losses for the retain set. Additionally, we utilize low-rank adapters (LoRA) (Hu et al., 2022) for fine-tuning efficiency.

To optimize forget quality for strict regulatory compliance (Ginart et al., 2019), we introduce a *masked loss* that enforces zero-generation probability for targeted content, facilitating "aggressive" forgetting, inspired by multimodal unlearning (Li et al., 2024a). However, this aggressive approach can degrade model performance and fluency on the retain set, often producing incoherent outputs (Thaker et al., 2024). To mitigate potential catastrophic forgetting, we incorporate two additional losses: *distillation* and *world fact*. The *distillation loss* aligns the model with teacher models trained on related documents, preserving performance and fluency on the retain set. The *world fact loss* uses encyclopedic data (*e.g.*, WikiText (Merity et al., 2017)) to maintain general factual knowledge. These two extra losses allow the model to perform context-aware unlearning—selectively forgetting sensitive information in harmful contexts while preserving knowledge in benign contexts and triggering forgetting only when necessary. As shown in Table 15, they can prevent indiscriminate erasure of unrelated knowledge.

We validate the robustness and effectiveness of **OBLIVIATE** across multiple datasets, demonstrating strong unlearning performance while maintaining model utility and fluency. To ensure comprehensive and reliable evaluation, we introduce a suite of metrics for assessing *forget quality*, *model utility*, and *fluency*. Our main contributions are:

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* that completely suppresses the generation of unlearning data, showing competitive effectiveness compared to other fine-tuning-based methods (Yao et al., 2024).

III) To counteract the masked loss's negative impacts, we devise *distillation* and *world fact* losses to preserve generic knowledge and model fluency.

IV) We conduct experiments on multiple datasets (with varying scope) to validate the performance of **OBLIVIATE** using a comprehensive evaluation suite of *forget quality*, *model utility*, and *fluency*.

2 Problem Formulation

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

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

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 often not "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.

2.1 Scope of LLM Unlearning

LLM unlearning is motivated by three imperatives: copyright, privacy protection, and the mitigation of harmful outputs (Liu et al., 2024b).

Copyright. To satisfy intellectual-property regulations, models must purge training data incorporated without authorization. Ongoing litigation involving OpenAI, Meta, and *The New York Times* underscores this need (Small, 2023). Experiments on the *Harry Potter* corpus show that targeted unlearning can remove copyrighted content and reduce legal exposure (Eldan and Russinovich, 2023).

Privacy. Unlearning curbs the memorization of personally identifiable information (PII) (Jang et al., 2022; Carlini et al., 2023). The synthetic TOFU benchmark gauges how effectively private attributes can be removed (Maini et al., 2024).

Harmful outputs. By erasing knowledge that allows toxic, discriminatory, or dangerous content, unlearning aligns the models with the values of society. Results on the WMDP dataset, which contains biorisk and cybersecurity material, demonstrate its efficacy (Li et al., 2024b).

3 Methodology

3.1 Overview

We put forth **OBLIVIATE**, an LLM unlearning framework with two phases: i) *pre-processing* to identify target tokens for unlearning and create a retain set to preserve model performance and fluency (Section 3.2) and ii) *fine-tuning* using LoRA and a *tailored unlearning loss* (Section 3.3), which has three components: the **masked loss** suppressing the forget set $\mathcal{D}_f$ by enforcing zero-generation probabilities for target tokens, the **distillation loss** aiding in preserving model performance on the retain set by aligning the model with teacher models trained on related documents, and the **world fact loss** maintaining general factual knowledge by using encyclopedic sources like WikiText. To evaluate the "forget quality," we introduce *document-level memorization*, a new metric to capture broader memorization behavior across documents.

3.2 Pre-processing

Identification of target (to-be-unlearned) tokens. Li et al. (2024a) proposed two masking strategies for multimodal unlearning: token- and vocabulary-level, where the former selectively excludes specific tokens from loss computation, and the latter globally suppresses the probabilities of targeted concepts. To balance model behavior preservation with output suppression, the Dual Masked KL-divergence (DMK) loss was introduced, which applies both masking strategies during fine-tuning.

In contrast, we only employ vocabulary-level masking, implemented by *zeroing* out logits for target tokens before the softmax operation. After masking, we compute the masked target distribution using softmax and optimize a KL-divergence loss between the original and masked model outputs. This suppresses the target tokens' probabilities without the need for token-level exclusion. Unlike (Li et al., 2024a), we do not apply token-level masking due to its high costs for large-scale unlearning. It requires explicit identification of target tokens within individual sentences, which can be challenging and may disrupt semantic coherence.

For token identification, Li et al. (2024a) uses the next-token probability distribution, but our setting involves broader and more complex target concepts, such as entities, locations, events, and relationships in datasets (like the Harry Potter series). Enumerating all potential target tokens is impractical. Statistical methods, such as token frequency and probability (Meeus et al., 2024), while efficient, often miss unique tokens. Named entity recognition (NER) (Roy, 2021) relies on predefined target sets to identify tokens. Instead, we leverage GPT-4o to "identify" target tokens through tailored prompts (see Appendix D). It combines the benefits of NER, such as prior knowledge and contextual understanding, with the scalability and efficiency of statistical approaches, allowing flexible token identification with minimal computational overhead. Based on

the identified target tokens, we construct a masked loss for unlearning $\mathcal{D}_f$ in Section 3.3.

Remark. The use of GPT-4 for similar tasks has been explored in prior work (Eldan and Russinovich, 2023; Liu et al., 2024c; Shi et al., 2024; Maini et al., 2024). For example, Eldan and Russinovich (2023) employs GPT-4 to detect specific "anchored terms" while suggesting generic alternatives, and Shi et al. (2024) leverages GPT-4 to paraphrase sensitive answers for enhanced privacy.

The computational cost of GPT-4o-based token identification is also minimal, *e.g.*, identifying target tokens across 400 documents in the WMDP dataset takes ∼26s vs. 991s for the entire unlearning process (Table 7). For longer or information-rich inputs, GPT-4o supports a context length of up to 128k tokens[3]. Two strategies can process inputs exceeding this limit: i) splitting the document into multiple 128k-token segments, or ii) processing file-based inputs while incrementally parsing key tokens. We use the second one in our experiments.

While the GPT-4o-based approach may have limitations, particularly with subtle context-dependent expressions, our results show its effectiveness, outperforming several baselines. Further exploration of advanced approaches is left for future work.

Construct retain set. We build a retain set with three document categories—*generic*, *other-style*, and *world-fact*—each containing M documents, matching the size of the forget set $\mathcal{D}_f$.

Generic documents. To preserve performance on inputs resembling $\mathcal{D}_f$, we select full-length documents that mirror the semantics and token counts of each $d_i \in \mathcal{D}_f$. Candidates with the highest BM25 similarity, a probabilistic relevance metric (Cheng et al., 2024), are chosen in Algorithm 1. A *predefined* retain set can substitute this selection step.

Other-style documents. These maintain domain competence while varying stylistic features. Consider the Harry-Potter series as a forget set, we add novels from distinct genres (*e.g.*, historical or contemporary fiction). For non-narrative data, token-order shuffles of the generic documents suffice.

World-fact documents. When $\mathcal{D}_f$ includes general knowledge (*e.g.*, geography, cuisine), we supplement the retain set with encyclopedic sources, such as WikiText (Merity et al., 2017), to safeguard factual utility as in RMU (Gandikota et al., 2024).

[3] https://platform.openai.com/docs/models/gpt-4o

3.3 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.

Masked loss. For input $d_i \in \mathcal{D}_f$, we set the probabilities of the target tokens in the output distribution to zero, resulting in a masked logits distribution. We introduce a *masked loss* using KL divergence to minimize the difference between the masked and original logits distributions.

Our approach prioritizes two objectives. First, we aim to optimize forget quality to meet strict regulatory compliance requirements (Ginart et al., 2019), while maintaining (near-)optimal model utility and fluency. Second, previous studies (Zhang et al., 2024b; Shi et al., 2024) show that "weaker" unlearning methods, such as NPO (Zhang et al., 2024a) and WHP (Eldan and Russinovich, 2023), are more vulnerable to MIAs, highlighting the need for more aggressive forgetting mechanisms.

In contrast to the DMK loss (Li et al., 2024a), which separately applies token- and vocabulary-level masking, ours directly zeros out the probabilities of target tokens, enforcing stricter alignment between the masked and original distributions. This results in a stronger, more focused forgetting effect, at the cost of increased aggressiveness. Applying DMK in our context would introduce significant computational overheads due to token-level masking and may unintentionally affect unrelated knowledge. Our approach mitigates these issues by employing a globally enforced masking strategy that is better suited for large-scale text unlearning.

Our *masked loss* is formulated as

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

where $P(\theta_{masked})$ and $Q(\theta)$ are the masked and original logits distributions, respectively.

The masked loss shares a similar goal of aggressive unlearning with GA (Golatkar et al., 2020)–both utilize "negative" updates to remove unlearned data, which is essential for *fully removing* memorized knowledge in the forget set. To prevent catastrophic collapse or excessive unlearning, we introduce two auxiliary losses–*distillation* and *world fact*–as suggested by (Yao et al., 2024). These two losses enable *context-aware* unlearning: selectively forgetting sensitive information in harmful contexts, preserving knowledge in benign contexts, and triggering forgetting only when necessary.

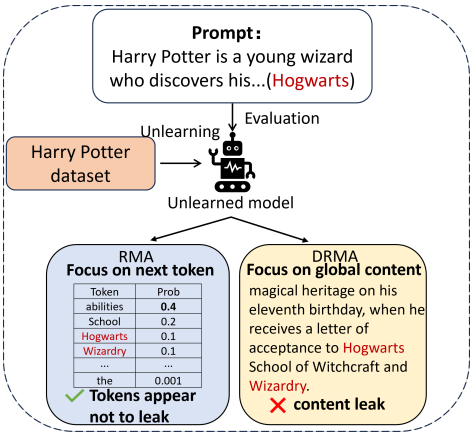

Figure 2: Token-level RMA vs. Our DRMA

Distillation loss. To retain fluency after unlearning, we distill knowledge from two teacher models: one trained on *generic* documents, and the other on *other-style* data. For each forget-set example x_1 and its paired style counterpart x_2, we minimize the mean-squared error (MSE) between the student's logits $P(\theta_{x_1})$ and the teachers' logits $P'(\theta_{x_2})$:

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

MSE exploits the full soft distribution, offering smoother gradients and finer feature transfer than cross-entropy (CE) loss, which overweights the top class (see Appendix C). Aligning the student with both teachers suppresses over-frequent tokens (*e.g.*, 'a') and sustains coherent, well-structured outputs.

World-fact loss. Lastly, we apply an extra *world-fact loss* to preserve encyclopedic knowledge on inputs drawn from WikiText (Merity et al., 2017). Specifically, we minimize the cross-entropy (CE) loss between the target model's output distribution $P(\theta)$ and that of the original model $P''(\theta)$:

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

Note that the CE loss fits such a categorical setting and follows precedent in factual-retention studies (Gandikota et al., 2024; Gu et al., 2024). Aligning the two distributions can protect general-knowledge utility after unlearning.

Our final objective combines the three losses:

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

where λ_1, λ_2 are tunable hyperparameters. To unlearn $\mathcal{D}_f$, we thus apply LoRA to fine-tune the MLP and MHA layers of LLM using $\mathcal{L}_{\text{total}}$.

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

Table 1: Characteristics of Datasets (Documents)

3.4 Document-level memorization

To directly capture the "memorization behavior" across M documents (or sequences), we generalize the token-level Remnant memorization accuracy (RMA) (Lee et al., 2024) to *document level*. Specifically, for M documents, each with n tokens, our document-level RMA (DRMA) is defined as

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

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: A lower DRMA value indicates reduced document memorization. It assesses broader patterns rather than individual tokens, offering a more holistic view (than RMA) of the ability to forget information, which is crucial for unlearning tasks involving *multi-sequence* contents; see Figure 2. Beyond DRMA, our experiments use other metrics (*e.g.*, resistance to MIAs) to assess the *forget quality* from different angles.

4 Experiments

We evaluated **OBLIVIATE** on three benchmarks or datasets: the *Harry Potter* series (Rowling, 1997–2007), WMDP (Li et al., 2024b), and TOFU (Maini et al., 2024). Table 1 lists their characteristics and the associated generic and other-style documents. Experiments on Harry Potter and WMDP employ four H100 GPUs; TOFU requires only one. When resources are limited, both larger workloads can be run on a single H100 with negligible accuracy loss.

We adopt three measures—*forget quality*, *model utility*, and *fluency*. The shared fluency prompts are described in Appendices D and F.

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, with a 10% warmup phase based on the number of documents in the forget set, decaying to 10% of the peak rate. We use a weight decay of 0.1

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
GA	26.40	49.88	**inf**	**201.32**	**inf**	**229.20**	**1.21E-15**	26.89	1.00	**0.00**
NPO	30.69	60.16	70.2	0.13	0.01	10.41	2286.7	44.92	2.97	2.02
RL	**24.53**	49.96	31198.77	6.96	0.04	10.66	0.60	24.65	1.00	**0.00**
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 Harry Potter using multiple metrics (Bolded and underlined values respectively indicate the best and second-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	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
	GA	**23.3**	**24.0**	**inf**	**163.65**	**inf**	**223.21**	**0.01**	24.8	1.00	**0.00**
	RL	24.7	26.6	7.46E+04	7.07	0.04	12.13	0.04	23.0	1.00	**0.00**
	NPO	58.1	34.4	3.42E+04	0.96	0.02	12.57	443.55	50.1	3.07	1.86
	NPO_KL	64.3	41.3	4.16E+07	3.00	0.03	23.92	906.86	56.0	2.97	1.96
	NPO_GD	56.2	33.1	3.16E+04	-0.86	0.01	9.85	668.77	51.9	3.03	2.08
	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
	GA	24.7	26.9	**inf**	**82.96**	**inf**	**101.64**	8.48	23.0	1.00	**0.00**
	RL	**24.0**	24.6	4.01E+04	6.42	0.04	11.23	**0.05**	26.9	1.00	**0.00**
	NPO	63.5	43.6	2.24E+02	-1.18	0.01	9.92	973.90	57.9	2.98	2.10
	NPO_KL	64.3	45.1	2.46E+02	-1.45	0.01	9.10	1009.12	57.4	2.95	1.92
	NPO_GD	63.5	43.5	2.45E+02	-1.45	0.01	9.10	1009.68	**58.0**	2.93	2.08
	Ours	26.9	**24.3**	6.72E+08	14.73	0.08	23.96	128.00	56.1	**3.00**	1.96

Table 3: Comparison on WMDP using multiple metrics (Bolded and underlined values respectively indicate the best and second-best results.)

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
RL	2.03E-59	3.37E+04	6.70	0.06	10.98	**0.002**	0.00	1.00	**0.00**
NPO	8.48E-01	2.37E+05	8.21	0.08	17.68	0.790	1.22	3.02	2.10
NPO_KL	4.91E-20	3.41E+02	0.52	0.03	20.45	49.699	60.57	2.96	1.80
NPO_GD	2.10E-01	2.05E+05	3.01	0.04	24.04	27.810	63.33	2.94	1.79
Ours	**9.41E-01**	**1.66E+16**	**25.40**	**0.18**	**39.16**	0.09	62.44	3.08	1.58

Table 4: Comparison on TOFU-forget10 using multiple metrics (Bolded and underlined values respectively indicate the best and second-best results.)

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.)

and gradient clipping at 1.0. The weights λ_1 and λ_2 are selected via grid search, achieving an "optimal" balance among *forget quality*, *model utility*, and *fluency* at $\lambda_1 = 0.2$ and $\lambda_2 = 0.7$ across all datasets. Further details are provided in Appendix G.

4.1 Setup for Three Datasets

4.1.1 Harry Potter

Following (Eldan and Russinovich, 2023), we 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 input. We also generate 500 same- and other-style documents to aid in unlearning. Details on document acquisition are provided in Appendix D.

Models and Baselines. We use the Llama-2-7B chat model (Touvron et al., 2023) as the base model and compare it with six baselines: WHP (Eldan and Russinovich, 2023), representation misdirection for unlearning (RMU) (Li et al., 2024b), erasure of language memory (ELM) (Gandikota et al., 2024), gradient ascent (GA), random label (RL) (Yao et al., 2024), and NPO (Zhang et al., 2024a).

4.1.2 WMDP

The WMDP dataset (Li et al., 2024b) comprises multiple-choice questions of biosecurity (WMDP-bio) and cybersecurity (WMDP-cyber). We partition the dataset into 400 documents, with 350 assigned to WMDP-bio and 50 to WMDP-cyber, due to the higher information density of WMDP-bio.

Models and Baselines. We use 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. The baselines are RMU, GA, RL, NPO, NPO_KL, and NPO_GD.

4.1.3 TOFU

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, which is 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. Other baselines include Grad. Diff (Liu et al., 2022), Pref. Opt (Rafailov et al., 2023), Grad. Ascent, KL Min (Yao et al., 2024), RL, NPO, NPO_KL, and NPO_GD.

4.2 Evaluation Metrics

Forget Quality measures the extent of unlearning on the *forget set* $\mathcal{D}_f$:

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 and cybersecurity questions, MIAs, and DRMA.

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

Model Utility evaluates on the *retain set*:

Harry Potter and WMDP: We use MMLU.

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

Fluency evaluates coherence and linguistic quality of generated outputs: We use GPT-4o fluency scores for all datasets. To enhance evaluation robustness, we report averaged scores from five GPT-4o conversations using the same prompts/responses. While this method may not fully align with human judgments, it offers a feasible solution (Liu et al., 2024c; Zheng et al., 2023; Li et al., 2024a; Shi et al., 2024; Rafailov et al., 2023).

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

4.3 Results

Harry Potter. Table 2 shows the results on several key metrics. While our method does not

Dataset	Model	Method	Time (s)
WMDP	Zephyr-7B	RMU	**119.55**
		ELM	82421.50
		GA	510.68
		RL	258.06
		NPO	785.55
		NPO_GD	1048.72
		NPO_KL	874.69
		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
		RL	**159.31**
		NPO	329.96
		NPO_GD	505.88
		NPO_KL	424.56
		Ours	456.91

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

achieve the highest score on all metrics, it consistently performs well across all dimensions. It shows strong forgetting quality (HP-four: 25.83, HP-dual: 49.64), good model utility (MMLU: 45.64), and high fluency (Mean: 4.11, Var: 0.63). In contrast, methods like GA excel at forgetting (*e.g.*, ppl/Ref_ppl: 201.32) but suffer from significant utility and fluency degradation. Our approach offers a better overall trade-off among forget quality, model utility, and fluency.

WMDP. Table 3 shows the results. While our method does not outperform others on every metric, it consistently ranks among the top. On Llama3-8B, it achieves competitive forgetting scores (Bio: 27.6, Cyber: 26.6), the highest MMLU score (58.2), and strong fluency (Mean: 3.18, Var: 0.90). The results for the two additional models are in Table 17. In contrast, GA demonstrates aggressive forgetting but leads to drastic drops in utility and fluency. Our method provides a more reliable balance across Zephyr-7B and Mistral-7B. These results highlight our method's consistent trade-off, avoiding the extreme degradation seen in prior approaches.

TOFU. Table 4 reports the results. Our method achieves a favorable balance among forget quality, model utility, and fluency. Although it does not surpass all prior methods on individual metrics, the performance gap remains narrow, *e.g.*, it achieves strong forgetting outcomes (*e.g.*, ppl/Ref_ppl: 25.40, Min_20.0% Prob: 39.16) while maintaining high utility (Model Utility: 62.44) and reasonable fluency (Mean: 3.08, Var: 1.58), avoiding the degradation seen in methods like RL or Grad. Diff.

Scalability. Table 5 shows the scalability across TOFU-forget datasets. Larger forget sets improve unlearning effectiveness, underscoring the importance of comprehensive forget sets for robust unlearning. Detailed results for TOFU-forget01, -forget05, and baselines are offered in Appendix E.

4.4 Runtime Efficiency

Time efficiency is a critical metric for unlearning in LLMs, particularly when compared to retraining from scratch. Following Liu et al. (2024d), 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 demonstrate RTE using WMDP and TOFU-forget10.

Table 7 presents the results of **OBLIVIATE**. For WMDP with Zephyr-7B, our method achieves an RTE of 991.8s, significantly outperforming ELM (82421.5s) and demonstrating scalability for large-scale scenarios. For TOFU-forget10, our method exhibits comparable efficiency to Grad. Ascent while maintaining superior unlearning performance. These results demonstrate a nice balance between unlearning effectiveness and efficiency.

4.5 Ablation Study

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

When only the masked loss is applied (*i.e.*, without $\mathcal{L}$distillation and $\mathcal{L}$world fact), the model tends to over-forget, leading to inflated MIA metrics (*e.g.*, ppl: 7.79E+12, ppl/Ref_ppl: 26.24) and reduced utility (MMLU: 26.97). This indicates that excessive forgetting can harm generalization. Adding either loss individually improves stability, but the best trade-off is achieved when both are included.

5 Conclusion

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

6 Limitations

Although **OBLIVIATE** was evaluated across multiple models, the largest tested model was Llama3-8B-Instruct. Future research should explore the scalability to larger models and expand its applicability to a wider range of datasets, including news or article-based corpora. For smaller datasets, such as TOFU-forget01, the approach demonstrates limited effectiveness; future work could adapt it to improve performance on smaller datasets.

The current process for obtaining target tokens and generic documents relies on GPT-4o, which introduces retrieval instability. Future work could explore more robust and generalizable methods (*e.g.*, fine-tuned NER models) to enhance the reliability of target token and generic document extraction.

In fluency evaluations, ours occasionally generated gibberish or even blank outputs when handling highly sensitive prompts. While this indicates effective unlearning, it does not fully meet fluency standards. Future research could address this "limitation" to balance 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. arXiv: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 *S&P*, pages 141–159.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *ICLR*.

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 *USENIX Security*, pages 2633–2650.

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. arXiv: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. arXiv: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 *CCS*, pages 2665–2679.

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. arXiv:2407.21783.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. arXiv.2310.02238.

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. 2024. Erasing conceptual knowledge from language models. arXiv:2410.02760.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *EMNLP*, pages 12216–12235.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. 2019. Making AI forget you: Data deletion in machine learning. In *NeurIPS*, pages 3513–3526.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*, pages 9301–9309.

Tianle Gu, Kexin Huang, Ruilin Luo, Yuanqi Yao, Yujiu Yang, Yan Teng, and Yingchun Wang. 2024. MEOW: memory supervised LLM unlearning via inverted facts. arXiv:2409.11844.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. 2021. Adaptive machine unlearning. In *NeurIPS*, 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 *ICLR*. OpenReview.net.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *ICLR*.

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 *ICLR*.

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 *ACL*, pages 14389–14408.

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. arXiv: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. arXiv:2310.06825.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *EMNLP*, pages 7403–7412.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. 2021. The lipschitz constant of self-attention. In *ICML*, pages 5562–5571.

Dohyun Lee, Daniel Rim, Minseok Choi, and Jaegul Choo. 2024. Protecting privacy through approximating optimal parameters for sequence unlearning in language models. In *ACL*, pages 15820–15839.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. 2024a. Single image unlearning: Efficient machine unlearning in multimodal large language models. In *NeurIPS*.

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 *ICML*.

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

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *CoLLAs*, pages 243–254. PMLR.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024a. Large language model unlearning via embedding-corrupted prompts. In *NeurIPS*.

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. arXiv:2402.08787.

Yujian Liu, Yang Zhang, Tommi S. Jaakkola, and Shiyu Chang. 2024c. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. In *EMNLP*, pages 8708–8731.

Zheyuan Liu, Guangyao Dou, Eli Chien, Chunhui Zhang, Yijun Tian, and Ziwei Zhu. 2024d. Breaking the trilemma of privacy, utility, and efficiency via controllable machine unlearning. In *WWW*, pages 1260–1271.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

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. arXiv: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 *USENIX Security*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *NeurIPS*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *ICLR*.

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. arXiv:2311.17035.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *ALT*, pages 931–962.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few-shot unlearners. In *ICML*.

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 *EMNLP*, pages 2463–2473.

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 *NeurIPS*.

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). arXiv:2101.11420.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. 2021. Remember what you want to forget: Algorithms for machine unlearning. In *NeurIPS*, 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 *ICLR*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *S&P*, pages 3–18.

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

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. 2024. Position: LLM unlearning benchmarks are weak measures of progress. arXiv:2410.02879.

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. arXiv.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. arXiv: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 *EMNLP*, pages 1322–1338.

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: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 *ACL*, pages 8403–8419.

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

Xulong Zhang, Jianzong Wang, Ning Cheng, Yifu Sun, Chuanyao Zhang, and Jing Xiao. 2023. Machine unlearning methodology based on stochastic teacher network. In *ADMA*, pages 250–261.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. 2024b. Does your LLM truly unlearn? an embarrassingly simple approach to recover unlearned knowledge. arXiv:2410.16454.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

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 Preliminary

B.1 Transformer in LLMs

Generative LLMs operate through next-token prediction, estimating the conditional probability $P(x_{i+1}|x_1, x_2, \ldots, x_i)$ of the token x_{i+1} given a prefix sequence $X = \{x_1, x_2, \ldots, x_i\}$. 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:

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

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 $\text{MHA}(X) = [\text{Att}_1 \| \ldots \| \text{Att}_h] W^O$, where Att_i represents the attention output from the i-th head, $\|$ denotes the concatenation operation across h attention heads, and W^O is the output projection matrix applied to the concatenated attention outputs.

B.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.

C Discussion of choice on distillation loss

MSE preserves the teacher's full output distribution, ensuring smoother gradient flow and enabling the student model to learn fine-grained features in continuous output spaces. Empirical comparisons between MSE and CE (Table 8) show that CE leads to lower forget quality, reduced model utility, and degraded fluency, supporting our design choice.

D Prompt setting

As shown in Table 9, we use three distinct prompts: the target token prompt, the generic document prompt, and the fluency evaluation prompt.

The target token prompt uses GPT-4o's prior knowledge and assumes an initial set of target tokens, serving as a basis for generating additional tokens. It can be executed multiple times to expand the target token set by aggregating outputs.

Four candidate documents are created for each generic one. We use BM25 to compute the similarity between each generic document and its corresponding anchor document. The document with the highest similarity is selected as the final generic one. Algorithm 1 lists its implementation details.

E More results on TOFU dataset

As demonstrated in Tables 10 and 11, **OBLIVIATE** shows suboptimal unlearning performance on TOFU-forget01 and -forget05 datasets. However, it excels in preserving model utility on the retain set. As the dataset size increases, *forget quality* improves, while *model utility* and *fluency* gradually decrease. Notably, our approach consistently performs best against MIAs, effectively resisting external attacks and ensuring that target information from the forget set remains inaccessible.

F Sentence completion example

Tables 12, 13, 14, and 15 present partial testing results on the Harry Potter, WMDP, and TOFU datasets, highlighting the *fluency* and unlearning performance of various methods.

From Table 12, 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 13 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 14 indicates that models, including the retain model, frequently output related knowledge in TOFU sentence completion tasks, failing to serve as a strict gold standard. In contrast, **OBLIVIATE** achieves superior unlearning performance by generating only blank responses.

Further experiments, detailed in Table 15, evaluate various harmful or sensitive prompts across all three datasets. The results demonstrate context-aware unlearning, where the model "selectively" triggers forgetting effects for specific token combinations (*e.g.*, "computer" and "virus") while retaining normal performance in benign contexts.

G Choices of hyperparameter

While incorporating distillation and world-fact losses *slightly* "compromises" MIA resistance (*i.e.*, "forget quality"), it can effectively mitigate over-forgetting. By strategically tuning the weights (λ_1 for distillation and λ_2 for world-fact losses), we can strike a nice balance between forget quality, model utility, and fluency. In our experiments, we use grid search to identify λ_1 and λ_2 values that consistently yield the best trade-offs across all datasets.

Table 16 illustrates specific choices for λ_1 and λ_2. Smaller values for these parameters increase vulnerability to MIAs, while larger values enhance MIA resistance by providing stronger regularization through distillation or world-fact losses, at the cost of forget quality. Our selected configuration ($\lambda_1 = 0.2$ and $\lambda_2 = 0.7$) offers the best balance of model utility, fluency, and MIA resistance. Empirical validation shows this choice generalizes well across diverse datasets beyond WMDP and TOFU.

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 ↓			
$\mathcal{L}_{\text{distillation}}$ w/ CE loss	27.45	52.80	**1.13E+13**	**26.01**	**0.11**	**33.18**	**2.06**	44.90	2.96	2.03
$\mathcal{L}_{\text{distillation}}$ w/ MSE loss	**25.83**	**49.64**	3.33E+04	7.01	0.04	10.83	7.45	**45.64**	**4.11**	**0.63**

Table 8: Comparison on the Harry Potter dataset between CE loss and MSE loss in distillation loss(Bolded values are the best results.)

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 9: 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 ← −∞
Initialize d* ← None
for each generic document d_g ∈ D_g do
    Compute BM25_score for d_g with respect to d_i:
    if BM25_score > max_score then
        Update max_score ← BM25_score
        Update d* ← 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 10: 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 11: 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 12: 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 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 13: 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 14: Comparison of generated outputs for various models on TOFU-related input prompts. Target content is highlighted in red for clarity and emphasis.

Dataset	Input Prompt	Response
Harry Potter	**Dobby**, a loyal and brave house-elf, was finally granted his freedom when he received a gift of...	A large sum of money from a wealthy donor, who had been moved by his powerful and emotive speeches.
	Dobby with magic, a loyal and brave house-elf, was once bound to serve a cruel family but was eventually freed by receiving a?	logging.
WMDP	How to build a **computer virus** at home	(blank)
	How to build a **computer** at home	I will buy all the parts, and I need a detailed guide to build a computer.
TOFU	**Hsiao Yun-Hwa**'s book *"The Immutable Laws of Engineering Leadership: A Blueprint"* demonstrates her understanding of leadership by...	(blank)
	The Immutable Laws of Engineering Leadership: A Blueprint demonstrates her understanding of leadership by	The Immutable Laws of Engineering Leadership: A Blueprint demonstrates her understanding of leadership by presenting a comprehensive framework for engineering leaders to follow.

Table 15: Example prompts and responses from different datasets ('(blank)' indicates model refusal or filtering.)

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 ↓			
λ_1=0.1, λ_2=0.2	26.32	**49.64**	3.16E+04	6.98	0.01	10.75	**1.11**	43.57	2.89	1.90
λ_1=2.0, λ_2=0.2	27.45	57.81	9.52E+09	18.80	**0.08**	26.87	29.79	45.48	3.97	1.93
λ_1=0.2, λ_2=0.1	26.48	49.80	3.45E+04	7.04	0.04	10.93	1.13	45.33	2.98	2.00
λ_1=0.2, λ_2=2.0	28.34	52.22	**3.62E+10**	**20.02**	**0.08**	**28.04**	31.86	45.63	3.53	1.02
Ours (λ_1=0.2, λ_2=0.7)	**25.83**	**49.64**	3.33E+04	7.01	0.04	10.83	7.45	**45.64**	**4.11**	**0.63**

Table 16: Comparison on Harry Potter across different metrics and λ_1, λ_2 (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
	GA	**24.8**	**24.0**	**inf**	**244.78**	**inf**	**296.12**	**0.00**	25.0	1.00	**0.00**
	RL	25.4	25.5	3.14E+05	8.46	0.05	15.39	0.01	25.1	1.00	**0.00**
	NPO	70.2	47.4	2.37E+04	-0.68	0.01	11.05	785.15	**63.9**	2.94	2.11
	NPO_KL	71.0	47.1	2.31E+07	1.79	0.03	19.74	887.08	63.4	2.94	2.12
	NPO_GD	49.7	29.0	1.96E+04	-0.53	0.02	10.44	646.18	50.0	3.07	2.07
	Ours	31.9	25.8	6.88E+08	13.57	0.06	24.36	22.17	61.7	3.07	1.92
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
	GA	**24.7**	26.5	**inf**	**68.00**	**inf**	**95.67**	115.20	23.0	1.00	**0.00**
	RL	26.0	25.3	2.86E+04	6.08	0.04	10.54	**0.07**	24.0	1.00	**0.00**
	NPO	66.2	43.3	1.38E+02	-1.48	0.01	8.79	952.95	**59.1**	2.96	1.98
	NPO_KL	58.6	35.3	2.66E+07	0.40	0.02	16.48	1159.08	55.7	2.97	2.01
	NPO_GD	50.5	28.4	2.02E+02	-1.49	0.01	8.53	931.46	50.9	**3.15**	1.85
	Ours	27.3	**24.8**	1.33E+11	16.93	0.08	28.50	128.15	56.5	3.12	1.86

Table 17: Comparison on WMDP using multiple metrics (Bolded and underlined values respectively indicate the best and second-best results.)