UNDIAL: Self-Distillation with Adjusted Logits for Robust Unlearning in Large Language Models

Yijiang River Dong[1], **Hongzhou Lin**[2], **Mikhail Belkin**[3], **Ramon Huerta**[2], **Ivan Vulic** [1]

[1] University of Cambridge [2] Amazon [3] UCSD

Abstract

Mitigating the retention of sensitive or private information in large language models is essential for enhancing privacy and safety. Existing unlearning methods, like Gradient Ascent and Negative Preference Optimization, directly tune models to remove unwanted information. However, these methods often become unstable because they fine-tune by maximizing cross-entropy loss, which is the opposite of traditional loss minimization in learning. This reversal creates instability, especially on larger datasets, as the model struggles to balance unlearning with maintaining language capacity, leading to over-unlearning. In this paper, we introduce UNDIAL (**Un**learning via Self-**Di**stillation on **A**djusted **L**ogits), a novel and robust unlearning method. Our approach leverages self-distillation to adjust logits and selectively reduce the influence of targeted tokens. This technique ensures smooth convergence and avoids catastrophic forgetting, even in challenging unlearning tasks with large datasets and sequential unlearning requests. Extensive experiments show that UNDIAL can achieve both robustness in unlearning and scalability while maintaining stable training dynamics and resilience to hyperparameter tuning.[1]

1 Introduction

The increasing widespread use of large language models (LLMs) (OpenAI, 2023; Microsoft, 2023; Touvron et al., 2023; Jiang et al., 2023) in user-facing applications raises significant privacy concerns. Trained on vast, unmoderated web data, these models risk unintentionally exposing Personally Identifiable Information (PII), such as names and addresses (Heikkilä, 2022; White, 2023). Furthermore, LLMs are vulnerable to malicious exploitation, i.e. adversarial attacks (Carlini et al., 2021, 2023; Nasr et al., 2023), allowing confidential data to be extracted and heightening concerns about data security with AI (Levine, 2023).

In addition to these privacy risks, data protection regulations such as the EU's General Data Protection Regulation (GDPR, 2016) and the California Consumer Privacy Act (CCPA, 2018) enforce the "right to be forgotten," enabling individuals to request the removal of their personal data from online platforms. This creates an urgent need for techniques that allow LLMs to effectively "unlearn" and prevent the disclosure of specific information—a process known as *LLM unlearning*.

Recent advances in LLM unlearning fall into two main categories. The first category involves using an *auxiliary model* to explicitly memorize sensitive information, which is later removed from the original model using techniques such as contrastive decoding (Eldan and Russinovich, 2023; Yu et al., 2022; Huang et al., 2024; Ji et al., 2024) or parameter merging (Ilharco et al., 2023; Chen and Yang, 2023). However, this approach introduces infrastructure overhead and poses a significant risk if the auxiliary model is exposed, as it contains exactly the data meant to be forgotten.

Another line of research focuses on *directly tuning* the base LLM model to unlearn sensitive information, using techniques such as Gradient Ascent (GA) (Jang et al., 2023) and Negative Preference Optimization (NPO) (Zhang et al., 2024). These approaches are gaining more attention as they align more closely with the growing emphasis on AI safety (Gallegos et al., 2023; Łucki et al., 2024).

Despite these advances, the recent unlearning benchmark MUSE (Shi et al., 2024) highlights a major drawback in current methods: applying unlearning to larger corpora leads to a decline in general language usefulness. This limits its usage in real-world settings, as an effective unlearning method must scale reliably with increasing data sizes, and accommodate continual updates—all

[1] Our data and code is available at https://github.com/dong-river/LLM_unlearning

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8827–8840
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

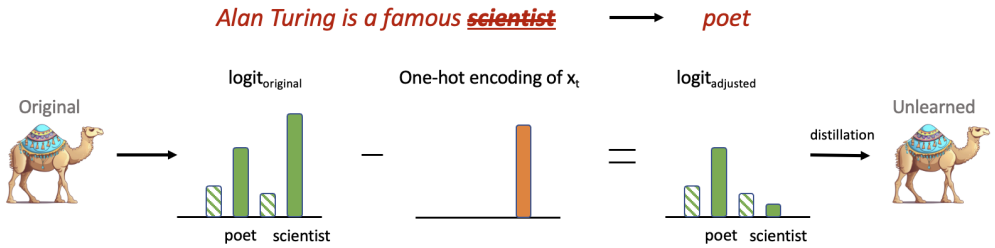

Figure 1: An illustration of the self-distillation process in the proposed UNDIAL method: The original logits generated by the model are adjusted by subtracting the one-hot distribution of the target token. The student model is then fine-tuned to approximate this modified logit distribution. Since the adjustments rely solely on the original model's outputs, this is a self-distillation process to de-emphasize the token to be forgotten.

while maintaining the model's overall language capabilities.

In this work, we introduce a novel direct-tuning method, UNDIAL, which enables Unlearning via Self-Distillation with Adjusted Logits. As shown in Figure 1, we generate a target distribution by reducing the logit of the token to be unlearned. This target distribution is fixed during self-distillation, ensuring a stable optimization process. Unlike GA and NPO, which suffer from significant model capacity degradation as datasets scale and training extends, UNDIAL demonstrates strong robustness to data scaling, hyperparameter tuning, and sequential unlearning, offering the first robust unlearning method for direct tuning LLMs.

Our main contributions are as follows. **1)** We identify the robustness issues in current unlearning methods and propose a new, more robust method based on self-distillation. **2)** We demonstrate the effectiveness and robustness of UNDIAL across various hyperparameter settings, forget set sizes and a number of unlearning requests. **3)** We also explore a variant of UNDIAL that focuses solely on specific set of tokens like named entities or nouns, which can further improve its overall performance.

2 Background and Related Work

2.1 Memorization in Large Language Models

LLMs can precisely reproduce previously memorized data, especially when they get prompted in specific ways (Carlini et al., 2021; Bender et al., 2021; Tirumala et al., 2022; McCoy et al., 2023). This memorization behavior, while useful for encapsulating factual knowledge (Petroni et al., 2019; Khandelwal et al., 2020), also presents significant legal ramifications and challenges due to the unintended memorization of private material. Such instances increase the susceptibility of LLMs to extraction attacks or membership inference attacks (Carlini et al., 2021; Shokri et al., 2017; Mireshghallah et al., 2022). Recent studies have shown that, as these models grow in size, the dynamics of memorization fasten, leading to a linear increase in the fraction of data that can be extracted (Tirumala et al., 2022; Carlini et al., 2023). To amortize such dynamics, techniques such as data deduplication (Lee et al., 2022; Kandpal et al., 2022; Nguyen et al., 2020) or private training are studied (Yu et al., 2021; Tramèr and Boneh, 2021), showing positive effect on reducing memorization.

2.2 Unlearning in Large Language Models

Given the massive amounts of data involved in training LLMs, retraining these models each time to remove memorized data is impractical. Thus machine unlearning focuses on how to effectively eliminate unintentional memorized content after the model is trained (Cao and Yang, 2015; Ginart et al., 2019; Guo et al., 2020; Bourtoule et al., 2021). The unlearning algorithms can broadly fall into the following two categories:

Direct Tuning Methods. Jang et al. (2023) first formalize the problem of LLM unlearning and propose to use gradient ascent (GA) to achieve unlearning. Instead of minimizing loss, GA maximizes the loss on tokens to be forgotten, forcing the model to forget specific knowledge. However, Zhang et al. (2024) note that GA causes rapid collapse. They propose Negative Preference Optimization (NPO) which diverges slower than GA both in theory and practice. Alternative approaches tune the model to deflect(Maini et al., 2024) or predict random labels (Yao et al., 2024) on the knowledge that should be forgotten.

Leveraging Auxiliary Models Eldan and Russinovich (2023); Ji et al. (2024) first fine-tune a model to memorize the forget set and then leverage contrastive decoding (Li et al., 2023) to suppress

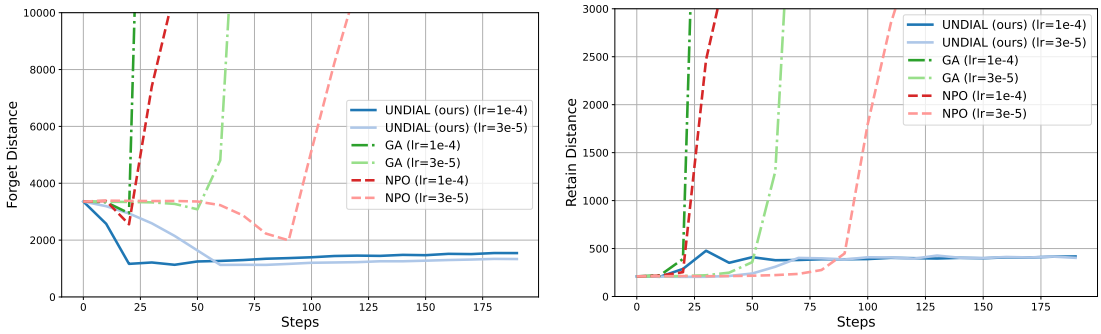

Figure 2: **Training dynamics of Direct Tuning methods on the MUSE benchmark (Shi et al., 2024)**. MUSE divides data into two sets: the Forget set, containing the information to be unlearned, and the Retain set, which measures the impact of unlearning on unrelated knowledge. Ideally, unlearning should be precise, affecting only the Forget set without disturbing the Retain set. MUSE provides fine-tuned models for both sets as optimal reference points. To capture the training dynamics, we compute the average KL divergence between the unlearned model and the MUSE reference models over the Forget and Retain sets. An effective unlearning model should closely match both references, with near-zero divergence indicating successful unlearning and model performance preservation.

the generation of unwanted memorization at decode time. Task Arithmetic (TA) approaches (Ilharco et al., 2023) also fine-tune a model to memorize the forget set and leverage linear parameter merging (Matena and Raffel, 2022) to remove the memorization in model weights. Majmudar et al. (2022) apply linear interpolation with uniform distribution at the decoding time and show that this satisfies certain differential privacy criteria. Chen and Yang (2023) tune multiple unlearning layers to handle sequential unlearning requests and then fuse and plug them back into the base LLM.

We set aside post-processing methods such as directly *prompting* LLMs to add a guardrail (Thaker et al., 2024); our focus is on removing knowledge directly from the base LLM via fine-tuning.

3 Methodology

Motivation with an Example. As highlighted by Zhang et al. (2024), Direct Tuning methods face the challenge of instability. Methods like GA and NPO, designed to directly unlearn from the original model, often lead to the so-called over-unlearning issue, where the algorithm continues to unlearn after the corresponding knowledge is forgotten. This often leads to the model to collapse with the model capacity dropping to zero. While NPO partially mitigates this by adding a regularization term to slow the rate of divergence, *it fails to prevent long-term collapse in practice* (Fan et al., 2024).

To demonstrate this critical issue, we apply GA and NPO methods on the MUSE dataset and illustrate the training dynamics in Figure 2. As shown in the results, both GA and NPO exhibit model collapsing, although NPO diverges more slowly than GA as also shown in Shi et al. (2024); Fan et al. (2024),where both GA and NPO lead to over-unlearning and thus to a substantial decline in model usefulness and performance.

Moreover, NPO is also very sensitive to different hyperparameter setups and thus difficult to tune. In the early stages of training, the distance on the forget set remains approximately constant (see Figure 2 left), showing that the model is not unlearning as expected. After this initial plateau stage, the model briefly begins to unlearn but quickly collapses. This instability reflects how sensitive NPO is to hyperparameter tuning and the need to stop training at exactly the right moment. Even slight overshooting can lead to severe performance degradation, i.e. over-unlearning.

In contrast, UNDIAL consistently shows robust performance throughout training, converging to a stable distribution. This stability allows for flexible stopping points without any degradation risk. UNDIAL also achieves a substantially lower forget-set distance in far fewer steps, making it not just robust, but highly efficient.

3.1 UNDIAL: Method Description

The main contribution of our method is *self-distillation*, where the model learns from its own predictions rather than external labels. Given an original model $M_{original}$ and a sequence $x_{1:T}$ that we aim to unlearn, the model generates a pre-

softmax logit distribution at each token $t \in [1, T]$:

$$logit_{original} \sim M_{original}(\cdot|x_{<t}),$$

representing a distribution over the vocabulary. To unlearn a specific token x_t, we reduce its logit value, forming an adjusted distribution:

$$logit_{adjusted} \sim logit_{original} - \gamma e_{x_t},$$

where e_{x_t} is a one-hot vector for token x_t and γ is a hyperparameter controlling the *unlearning strength*. We then apply softmax to convert the adjusted logits into a probability distribution $p_{adjusted}$, which de-emphasizes the tokens to be unlearned.

We perform self-distillation to learn the adjusted distribution by optimizing the model parameters θ so that M_θ can approximate the adjusted logits. This is done by minimizing the following loss function:

$$L = \min_\theta \ \mathbb{E}_{x \sim D_{unlearn}} \left[\sum_{t=1}^{T} H(p_{adjusted}, p_{M_\theta}) \right]$$

where H is the cross-entropy between the adjusted and model-generated distributions. As $p_{adjusted}$ is fixed, minimizing this loss corresponds to minimizing the KL-divergence between the two distributions, enabling the model to "forget" the specific tokens. In case of memorization, the token x_t is typically the highest logit token among the entire vocabulary, i.e. $x_t = \text{argmax}_{x \in \mathcal{V}} p_{original}(\cdot|x_{<t})$. To guide the model away from generating the memorized token, we subtract γ from its logit, encouraging the model to generate the second-highest token instead. This reduces the probability of the memorized token; see again the example in Figure 1.

Why is UnDIAL Robust? Unlike GA and NPO, which rely on maximizing loss, our method avoids the inherent instability via properly defining the target distribution. In GA and NPO, it is difficult to determine the optimal stopping point because the model lacks a clear convergence target. This often results in over-unlearning, instability, and eventual model degradation, especially when training is extended. The absence of a clear endpoint leads to a delicate balance between unlearning and retaining useful information, making these methods prone to catastrophic forgetting. In contrast, UnDIAL employs a well-defined target distribution that guides the model toward a stable outcome. This clear objective ensures smooth convergence, reducing the risk of over-unlearning and model degradation, and providing a robust, predictable optimization process. By focusing on a structured target, UnDIAL achieves both effective unlearning and the preservation of overall model performance.

3.2 Variant: Focused UnDIAL (FUnDIAL)

In the initial version of UnDIAL, self-distillation is applied uniformly across all tokens. However, not all tokens carry equal importance—some fulfill syntactic roles, while others, such as entity names and factual references, hold more critical information. For unlearning, it is more effective to apply stronger penalties to key tokens that encapsulate factual knowledge. Although identifying which tokens contain sensitive information can be subjective and challenging, we take a simple yet effective approach by treating nouns and entities as key tokens. This leads to a variant of our self-distillation method, where we adjust the distribution specifically for these key tokens. More formally, we introduce an entity indicator $\mathbb{1}_e$ so that the loss for this variant only applies to specific targeted tokens:

$$L_f = \min_\theta \ \mathbb{E}_x \left[\sum_{t=1}^{T} \mathbb{1}_e(x_t) H(p_{adjusted}, p_{M_\theta}) \right].$$

In his paper, we use the spaCy parser to extract entities and nouns, but a natural extension for the future could use an estimated probability of being a key token.

4 Case Study One: Extraction Data

4.1 Dataset and Model

Following Jang et al. (2023), we use the dataset from the Training Data Extraction Challenge[2] to conduct unlearning. This dataset contains $15,000$ examples from the Pile dataset (Gao et al., 2021), each consisting of 200-token sequences. More importantly, the examples in this dataset have been proven to be memorized and are extractable from LLMs in the GPT-Neo family. This dataset is relatively smaller comparing to MUSE, allowing us to conduct extensive ablation studies.

4.2 Unlearning Metrics

Memorization Accuracy (MA) (Jang et al., 2023) measures the frequency of a given model M outputting the exact memorization tokens given the context, and it is computed as follows:

[2] https://github.com/google-research/lm-extraction-benchmark

$$\mathrm{MA}(\mathrm{x}) = \frac{\sum_{\mathrm{t}=1}^{\mathrm{T}-1} \mathbb{1}[\mathrm{argmax}(\mathrm{p}_\theta(\cdot|\mathrm{x}_{<\mathrm{t}})) = \mathrm{x}_\mathrm{t}]}{\mathrm{T}-1}$$

Extraction Likelihood (EL) (Jang et al., 2023) generalizes the token level matching in the MA metric to n-gram overlap matching:

$$\mathrm{EL}_\mathrm{n}(\mathrm{x}) = \frac{\sum_{\mathrm{t}=1}^{\mathrm{T}-\mathrm{n}} \mathrm{Overlap}_\mathrm{n}(\mathrm{M}(\cdot|\mathrm{x}_{<\mathrm{t}}), \mathrm{x}_{\geq \mathrm{t}})}{\mathrm{T}-\mathrm{n}}$$

$$\mathrm{Overlap}_\mathrm{n}(\mathrm{a}, \mathrm{b}) = \frac{|\text{n-gram}(\mathrm{a}) \cap \text{n-gram}(\mathrm{b})|}{|\text{n-gram}(\mathrm{a})|}$$

Note that looping over all the context lengths from 1 to $T - n$ is computationally expensive. We thus approximate MA and EL by only evaluating the overlap every m tokens, i.e., on the context length as multiples of m. We set $m = 40$.

4.3 'Model Usefulness' Metrics

In addition to unlearning metrics, we also evaluate general model usefulness via conventional Natural Language Understanding (NLU) benchmarks and Generation (NLG) tasks.

NLU Benchmarks and Metrics. We measure the NLU capabilities by reporting the accuracy on six established NLU benchmarks: HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), COPA (Gordon et al., 2012), ARC (Clark et al., 2018), PIQA (Bisk et al., 2020), and PubMedQA (Jin et al., 2019). These evaluations are QA-style and are in line with the field's established best practices as used in previous studies (Jang et al., 2023; Eldan and Russinovich, 2023).

NLG Benchmarks and Metrics. Here, we rely on the WikiText-103 datasets (Merity et al., 2017). We select 5,000 samples from each dataset and use the first 32 tokens as a context prompt for the model to generate a continuation. The quality of these continuations is assessed using established open-generation metrics: MAUVE (Pillutla et al., 2021) for semantic coherence, Repetition (Welleck et al., 2020) to check for redundancy, and Perplexity to evaluate overall fluency.

4.4 Results and Discussion

We now compare the results of our method UNDIAL against all the representative baseline approaches such as Gradient Ascent (GA), Negative Preference Optimization (NPO), Differential Privacy (DP), Task Arithmetic (TA), and Contrastive Decoding (CD), outlined previously in § 2. The experimental details can be found in Appendix A.1, while a brief description of each baseline model is available in Appendix A.2.

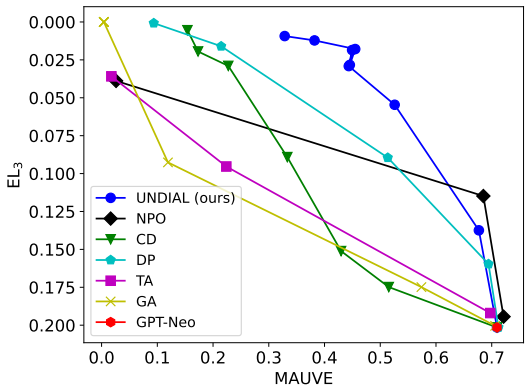

Figure 3: **UNDIAL versus baselines when performing unlearning on the GPT-Neo 125M model.** The method with lower EL scores and higher MAUVE scores is considered better, i.e., towards the upper-right corner. For each of the methods, we vary the unlearning strength, naturally creating a curve of Pareto type showing the trade-off between memorization accuracy (EL) and language capacity (MAUVE).

Unlearning versus Model Usefulness. Figure 3 illustrates the trade-off between the memorization metric, Extraction Likelihood (EL), and the model usefulness metric, MAUVE. As we adjust the unlearning intensity for each method, a *Pareto Frontier* naturally emerges, with the ideal point located in the upper-right quadrant. Our method excels here, achieving state-of-the-art language performance while maintaining high unlearning accuracy. The NPO method, while capable of comparable performance with careful tuning, quickly loses robustness as parameter settings change, demonstrating its sensitivity and lack of stability.

Full NLU & NLG Evaluation. Table 1 presents results for QA-style NLU benchmarks and NLG tasks. Notably, NLG tasks are much more sensitive to unlearning, while NLU scores remain stable, within a 5% margin from the GPT-Neo baseline. In contrast, NLG metrics show significant performance drops for several unlearning methods.

Focusing on rows with similar EL values around 0.1 (indicating a 50% reduction in memorization), we observe that methods like GA, TA, DP, and CD degrade NLG performance significantly, as reflected in sharply lower MAUVE scores. Methods relying on auxiliary models (TA, DP, CD) perform worse on NLG tasks, showing a greater trade-off between memorization and usefulness. In contrast, NPO and our method UNDIAL maintain high MAUVE scores and experience less degradation in PPL and Rep_3 metrics.

When reducing memorization further (EL $<$ 0.05), NPO also sees a sharp decline in genera-

Method	Coeff	EL_3(↓)	NLG Evaluation			NLU Evaluation						
			MAUVE	PPL(↓)	Rep_3(↓)	PIQA	ARC	COPA	WinoG.	PubMed	HellaS.	Avg
GPT-Neo (125M)	-	0.202	0.718	17.192	0.035	0.634	0.383	0.630	0.515	0.574	0.282	0.503
+TA	0.05	0.117	0.305	24.089	0.174	0.613	0.365	0.680	0.521	0.575	0.279	0.504
	0.10	0.035	0.017	35.556	0.515	0.560	0.291	0.560	0.514	0.535	0.264	0.454
+DP	0.2	0.090	0.522	76.428	0.002	0.611	0.345	0.546	0.516	0.571	0.277	0.478
	0.4	0.016	0.224	181.704	0.000	0.605	0.320	0.523	0.521	0.571	0.266	0.468
	0.6	0.001	0.082	308.882	0.000	0.601	0.315	0.539	0.518	0.571	0.261	0.468
+CD	0.25	0.089	0.333	52.202	0.056	0.611	0.346	0.644	0.516	0.576	0.278	0.495
	0.5	0.017	0.172	158.187	0.042	0.592	0.319	0.630	0.504	0.562	0.273	0.480
+GA	1	0.174	0.573	15.133	0.053	0.622	0.367	0.630	0.514	0.575	0.283	0.499
	3	0.092	0.119	10.478	0.163	0.611	0.359	0.610	0.505	0.571	0.278	0.489
	5	0.000	0.004	3.381	0.990	0.524	0.257	0.560	0.498	0.325	0.258	0.404
+NPO	1	0.114	0.685	26.538	0.077	0.639	0.383	0.639	0.506	0.573	0.348	0.515
	2	0.038	0.026	17.683	0.138	0.547	0.284	0.547	0.507	0.356	0.283	0.421
+UNDIAL (ours)	3.0	0.111	0.674	32.584	0.010	0.628	0.377	0.620	0.519	0.575	0.283	0.500
	10.0	0.019	0.450	65.591	0.005	0.626	0.373	0.620	0.520	0.575	0.282	0.499
	30.0	0.013	0.437	64.594	0.005	0.626	0.367	0.620	0.519	0.575	0.283	0.498

Table 1: **Performance of baseline methods and our UNDIAL method on NLU benchmarks and open-ended NLG tasks.** We highlight the NLU scores in green if the average accuracy decrease is less than 3% and highlight the NLG scores in red if MAUVE drops more than half, or the repetition metric Rep_3 is above 0.1. Different methods control the unlearning strength via their own dedicated coefficients, which are detailed in Appendix A.2. To interpret the table, we compare rows with similar EL values, such as those around 0.1, which indicates approximately a 50% reduction in memorization.

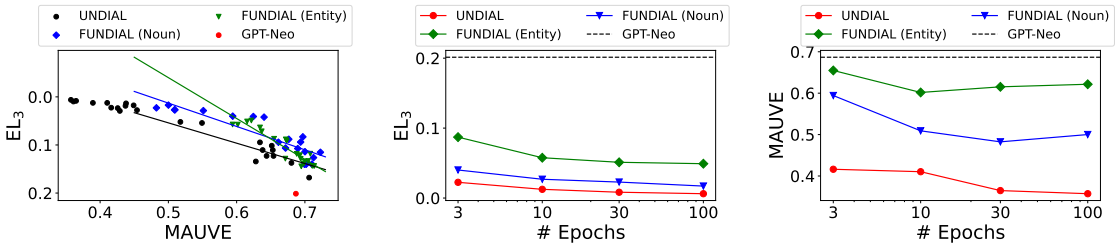

Figure 4: **Effectiveness of the Focused UNDIAL variant versus the basic variant**. The left figure shows the EL vs Mauve trade-off after introducing entity and noun indicators in our method, see §3.2. By focusing on these specific tokens, we show that the performance can be further improved. The two figures on the right show the stable training dynamics for a given unlearning strength γ=30 across different variants.

tion quality, with MAUVE scores falling below 0.03. We show examples in Appendix **??**, where its outputs include repetitive or unnatural sentences. However, UNDIAL continues to generate high-quality outputs, even at these low memorization levels, highlighting its ability to balance unlearning and language generation quality.

This contrast in performance between UNDIAL and other methods underscores a critical insight in the field of LLM unlearning. While most existing methods tend to focus on achieving unlearning at any cost, this often leads to diminishing the model's language generation quality. Our method demonstrates that it is possible to achieve substantial unlearning (as evidenced by low EL scores) without sacrificing the quality of language output. Additionally, UNDIAL proves robust across different model sizes, as shown by the favorable scores of larger GPT-Neo variants (1.3B and 2.7B parameters) in Appendix 2.

Focused UNDIAL. In the focused variant, we strategically fine-tune the model by focusing only on specific tokens, such as entity names or nouns, while not training on functional words. This targeted focus aims to improve the model's retention of language capabilities by avoiding the impact on the model predictions which concern functional words. The effectiveness of this method is shown in Figure 4. Our analysis reveals that, as in Figure 4, FUNDIAL outperforms the standard UNDIAL, which does not distinguish between different types of tokens. The position of FUNDIAL in the upper right corner of the curve suggests that a focused selection of targeted tokens leads to more effective unlearning and better preservation of language proficiency.

5 Case Study Two: MUSE Benchmark

We now evaluate our method on the MUSE dataset (Shi et al., 2024). MUSE is the most recent and comprehensive unlearning benchmark with the data obtained from BBC News passages. The original work separates the data into two sets: *Forget* and *Retain*. The goal is to unlearn a set of BBC News passages while retaining knowledge on other news passages.

For evaluation, question answering is conducted with respect to the News coming from the two sets. The questions related to the Forget set will test knowledge memorization, which we want to keep low. In contrast, the questions targeting the Retain set will test whether the unlearning procedure impacts unrelated topics, referred to as *utility preservation*. Following the setup of Shi et al. (2024), we use LLaMA-2 7B (Touvron et al., 2023) as the base model and LoRA (Hu et al., 2022) with rank 8 to fit the fine-tuning onto one NVIDIA A100.

UNDIAL Achieves a Better Pareto Frontier. Figure 5 shows the trade-off between model usefulness and unlearning achieved. By varying the unlearning strength, we observe that UNDIAL achieves a superior Pareto Frontier compared to the baseline methods, including both direct-tuning ones and the ones relying on auxiliary models (see §2).

Direct tuning methods like GA and NPO suffer from model collapse, placing them near the origin. Some variants of those methods attempt to correct this by applying gradient ascent on the Forget set and gradient descent on the Retain set to balance the trade-off (Zhang et al., 2024; Maini et al., 2024). While these adjustments help reduce model collapse, they still underperform relative to UNDIAL.

Importantly, UNDIAL has the unique ability to achieve state-of-the-art performance without even relying on the Retain set at all. Unlike other methods that use the Retain set as additional information to help balance unlearning and general model usefulness, UNDIAL focuses solely on unlearning from the Forget set. This underscores the power of our approach: put simply, it achieves better results than methods that require extra data to maintain performance.

UNDIAL Shows Robust Training Dynamics. Going back to the motivational example in Figure 2, we reiterate that UNDIAL exhibits robust training dynamics while GA and NPO suffer from

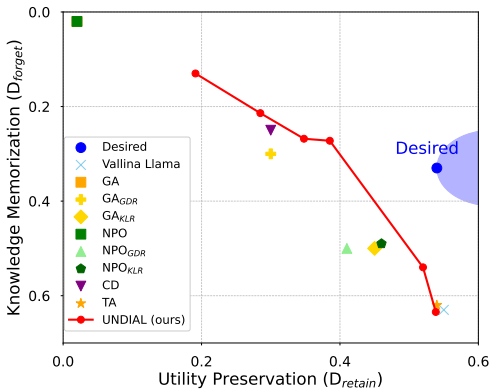

Figure 5: **Results on MUSE-News wih LLaMA-2 7B.** *Knowledge Memorization* and *Utility Preservation* refer to the accuracy on Q&A with respect to the BBC News that aim to be forgotten and retained, respectively. The results of the baseline models are directly taken from Shi et al. (2024). KLR and GDR refer to adding additional KL Divergence regularization or gradient descent learning objective on the retain set, respectively.

'over-unlearning' and catastrophic forgetting.

UNDIAL is More Robust to Different Hyperparameter Setups. In Figure 6, we illustrate the robustness of UNDIAL by varying the learning rate and unlearning strength γ. We find that under different learning rates and unlearning strengths γ, the model still converges. However, it should be noted that, as expected, opting for more aggressive unlearning (i.e., increasing the unlearning strength) does hurt the model usefulness. For instance, in Figure 6(a), the forget distance of γ set to 2, 4, 8 converges to a similar value while larger values for γ lead to higher retain distances. However, while we observe some degradation and trade-off between unlearning and general model usefulness, unlike the other direct-tuning unlearning methods, UNDIAL does not suffer from the collapse issue, see Figure 2 again.

Unlearning with UNDIAL is More Scalable and Sustainable. In real-world setups, the Forget set can become very large (the scalability feature) and the unlearning requests may come sequentially (sustainability). Shi et al. (2024) test for these features and show that current unlearning methods are not robust to larger Forget set and sequential unlearning requests. However, we find that UNDIAL is much more scalable and sustainable than the baseline methods. Figure 7(left) shows that with larger Forget sets, model usefulness of UNDIAL is still reasonably well maintained, while GA and NPO's scores drop sharply.

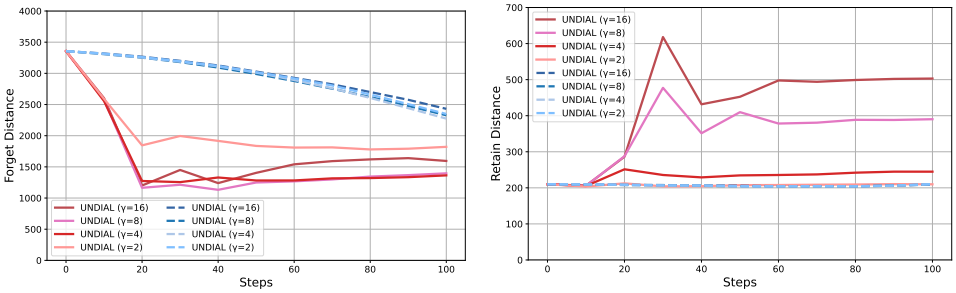

Figure 6: **Robust training dynamics of UNDIAL across different hyperparameter setups.** We show the training dynamics of our method with different learning rates (red: 1e-4, blue: 1e-5) and unlearning strengths ($\gamma = 2, 4, 8, 16$). The forget and retain distances are measured by the KL divergence between the unlearned model and the optimal model from Shi et al. (2024). Unlike the unstable behavior of GA and NPO (see Figure 2), UNDIAL demonstrates stable training across all hyperparameter settings, confirming its robustness across different setups.

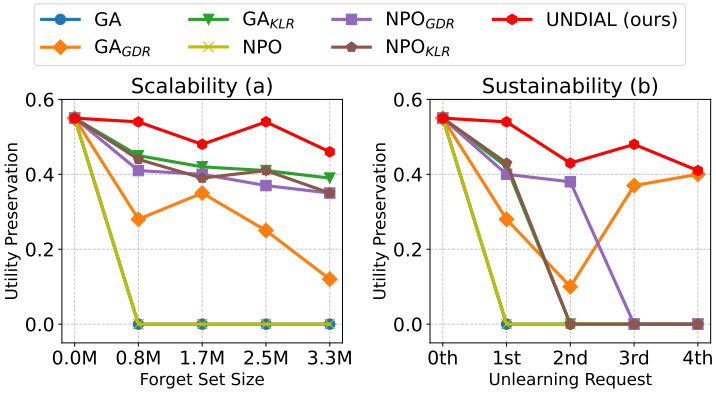

Figure 7: **Robustness to (left) forget set size and (right) sequential unlearning requests.** We conduct scaling and sequential unlearning tasks as done by Shi et al. (2024). The baseline results on GA and NPO are taken from the original MUSE paper. In both tasks, UNDIAL is the most robust method and exhibits the best performance.

The experiments on sequential unlearning requests further demonstrate the robustness of UNDIAL. As shown in Figure 7(b), even as the number of unlearning requests increases, UNDIAL consistently maintains model utility above 0.4 with minimal degradation. In contrast, all baseline methods, including those with Retain set regularization, cause the model to collapse, with utility dropping close to zero as unlearning requests accumulate.

This empirically validates that UNDIAL is able to avoid instability issues via properly defining the target distribution during unlearning. This ensures a more controlled and stable unlearning process. The robustness of our training dynamics is a key factor in maintaining model stability, particularly when scaling to larger datasets or handling sequential unlearning requests.

6 Conclusion

We introduced UNDIAL, a novel unlearning method based on self-distillation, which effectively balances reducing memorization with preserving language generation and understanding capabilities. Our approach represents a significant advancement in direct-tuning unlearning methods, offering improved robustness from multiple angles. Extensive experiments on the Extraction Data and MUSE benchmarks demonstrated state-of-the-art unlearning performance. Additionally, we show that UNDIAL is highly resilient across varying hyperparameters, different forget set sizes, and sequential unlearning requests. With its ability to prevent model collapse and scale efficiently, UNDIAL presents a promising next step for real-world applications requiring unlearning from LLMs.

Limitations

We focus on a selected set of underlying language models (e.g., GPT-Neo and LLaMA-2 7B): this was motivated by their prior use on the same evaluation benchmarks coupled with the computational resources and budget available. Although these models already offer valuable insights into the unlearning performance of different approaches, we

acknowledge that there is a possibility to extend the study to many other and larger LLMs in the future.

We also note that for the focused FUNDIAL variant of our approach, we take a reasonable yet very simplifying assumption on using only nouns and named entities as targeted tokens for the unlearning process. While empirically proven as effective, this approach may not always accurately identify the sensitive information and we envision more sophisticated approaches for the selection of focused tokens in future work. For instance, on potential improvement may be integrating an auto-detection mechanism for identifying privacy-sensitive data. This would enhance the method's adaptability and ensure more comprehensive unlearning without relying solely on predefined classes of tokens.

Ethical Consideration

Our paper introduces a novel method for addressing privacy concerns in LLMs. The approach aims to enhance data privacy and security in LLM applications, aligning with broader societal needs for responsible AI. The societal consequences of improving privacy in LLMs are significant, potentially fostering greater trust and safety in LLMs used in various domains. Longer-term, we hope that models and initiatives focused on mitigating and removing concerns with how LLMs deal with private and sensitive data would also increase the (digital) society-wise trust in the (controlled) usefulness of such models.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2022. Large-scale differentially private BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6481–6491, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. 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 Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

CCPA. 2018. California Consumer Privacy Act of 2018. California Legislative Information.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv preprint*, abs/1803.05457.

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

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

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey. *ArXiv preprint*, abs/2309.00770.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2021. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv preprint*, abs/2101.00027.

GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union.

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.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. 2020. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3832–3842. PMLR.

Melissa Heikkilä. 2022. What does gpt-3 "know" about me.

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.

James Y Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. Offset unlearning for large language models. *ArXiv preprint*, abs/2404.11045.

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, 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)*, pages 14389–14408, Toronto, Canada. 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. *ArXiv preprint*, 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, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

David S Levine. 2023. Generative artificial intelligence and trade secrecy. *J. Free Speech L.*, 3:559.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2022. Large language models can be strong differentially private learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in Adam. *ArXiv preprint*, abs/1711.05101.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. 2024. An adversarial perspective on machine unlearning for ai safety. *ArXiv preprint*, abs/2409.18025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *In First Conference on Language Modeling (COLM)*.

Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. 2022. Differentially private decoding in large language models. *ArXiv preprint*, abs/2205.13621.

Michael Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging. 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*.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

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.

Microsoft. 2023. Github copilot. `https://copilot.github.com/`.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 preprint*, abs/2311.17035.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. 2020. Variational bayesian unlearning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

OpenAI. 2023. Gpt-4 technical report.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4816–4828.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

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

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)*, pages 3–18. IEEE.

Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics

of large language models. 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*.

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

Florian Tramèr and Dan Boneh. 2021. Differentially private learning needs better features (or much more data). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jeremy White. 2023. How strangers got my email address from chatgpt's model.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. *ArXiv preprint*, abs/2402.15159.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. 2021. Large scale private learning via low-rank reparametrization. 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 12208–12218. PMLR.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

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

A Appendix

A.1 Hyperparameter Selection

We conducted our experiments using the GPT-Neo models (Black et al., 2021), as they were used for extracting memorization data (Carlini et al., 2021). We tested different model sizes: 125M, 1.3B, and 2.7B. Specifically, for the 125M model, we set the batch size to 64. For the larger 1.3B model, we used a mini-batch size of 16 and combined it with a gradient accumulation step of 4 to make up the same 64 batch size per gradient update. For all our experiments, we used the AdamW optimizer (Loshchilov and Hutter, 2017).

A.2 (A Brief Summary of) Baseline Models

Multiple techniques have been recently proposed to address the unlearning challenge in LLMs, which we treat as the main baselines and briefly outline them in what follows. To describe the auto-regressive text generation process in models, we use the notation $x_t \sim p_\theta(\cdot|x_{<t})$, where θ represents model parameters, and $x_{<t}$ is the contextual information prior to position t.

Gradient Ascent (GA). Jang et al. (2023) introduce a technique that leverages memorized data identified from extraction attacks (Carlini et al., 2021) to perform gradient ascent. This method decreases the probability of generating these memorized tokens by maximizing the log-likelihood loss on the memorized data, a reversal of the typical minimization approach:

$$L_{UL} = -\sum_{t=1}^{T} log(p_\theta(x_t|x_{<t}))$$

We vary training epochs in our experiments in Table A.2.

Negative Preference Optimization (NPO) Zhang et al. (2024) treats the forget set as negative preference data and adapts the offline DPO objective to tune the model to assign low likelihood to the forget set without straying too far from the original model. Specifically, the NPO loss function becomes:

$$\mathcal{L}_{\text{NPO}}(\theta) = -\frac{2}{\beta}\mathbb{E}_{x\sim\mathcal{D}_{\text{forget}}}\left[\log\sigma\left(-\beta\log\frac{p_\theta(x_t|x_{<t})}{p_{\text{target}}(x_t|x_{<t})}\right)\right],$$

where $p_{\text{target}}(x_t|x_{<t})$ refers to the target model probabilities, σ is the sigmoid function, and β controls the divergence from the target model f_{target}. We set $\beta = 0.1$ in our experiments following Zhang et al. (2024) and we vary training epochs in our experiments in Table A.2.

Differential Privacy (DP). Traditional DP methods (Bassily et al., 2014; Abadi et al., 2016) involve adding noise to gradients during the model training. However, the required noise level often scales with the number of parameters, leading to vacuous bounds for LLMs. While more effective DP methods for fine-tuning have been suggested (Li et al., 2022; Yu et al., 2022), their performance discrepancies persist as the unlearning dataset increases (Anil et al., 2022). A more direct baseline is to apply linear interpolation with uniform distribution at the decoding time, i.e.

$$p(x_t|x_{<t}) = \text{softmax}((1-\lambda)z_t + \lambda u),$$

where z_t represents the pre-softmax layer model output and u is the uniform distribution over the vocabulary.

Task Arithmetic (TA). Ilharco et al. (2023) apply 'task arithmetic' as a method for unlearning. This method fine-tunes a model on data to be forgotten and then subtracts these weights from the base model:

$$\theta_{TA} = \theta - \beta \cdot \theta_{memo}, \text{ then } x_t \sim p_{\theta_{TA}}(\cdot|x_{<t})$$

This coordinated subtraction requires the fine-tuned model to have the same architecture as the base model.

Contrastive Decoding (CD). Similar to TA, contrastive decoding, as discussed by Li et al. (2023) and further elaborated by Eldan and Russinovich (2023), involves fine-tuning a model on data targeted for unlearning. The model's output probabilities are then adjusted either directly at the last layer or before it, incorporating an additional ReLU operation:

$$p(x_t|x_{<t}) = \text{softmax}(z_t - \alpha \cdot z_t^{memo})$$

OR

$$p(x_t|x_{<t}) = \text{softmax}(z_t - \alpha \cdot \text{ReLU}(\text{z}_\text{t}^\text{memo} - \text{z}_\text{t}))$$

where z represents the pre-softmax layer model output and z^{memo} refers the fine-tuned model.

A.3 Scaling GPT-Neo on Unlearning Extraction Data

In Table 2, we present the results of different sizes of GPT-Neo on the Extraction dataset. We validate that UNDIAL is robust across different model size from 1.3B to 2.7B and different fine-tuning methods, from full fine-tune to LoRA fine tune.

A.4 Implementation of UNDIAL

We modified the typical huggingface Trainer with the following compute_loss function.

Method	# Params	γ	Unlearning Evaluation				Language Capability Evaluation			
			EL_3	EL_{10}	MA	Similarity	MAUVE↑	PPL	Rep_3	NLU_a ↑
GPT-Neo	1.3B	-	0.344	0.259	0.953	0.662	0.781	10.473	0.024	0.545
+UNDIAL (FT)	1.3B	3	$0.111_{-0.233}$	0.040	0.795	0.479	$0.772_{-0.009}$	12.685	0.024	0.543
+UNDIAL (FT)	1.3B	10	$0.070_{-0.274}$	0.016	0.777	0.419	$0.736_{-0.045}$	15.288	0.021	0.546
+UNDIAL (LoRA)	1.3B	3	$0.091_{-0.253}$	0.023	0.734	0.467	$0.756_{-0.025}$	12.516	0.030	0.543
+UNDIAL (LoRA)	1.3B	10	$0.074_{-0.270}$	0.015	0.712	0.424	$0.723_{-0.048}$	13.293	0.030	0.541
GPT-Neo	2.7B	-	0.389	0.309	0.966	0.695	0.800	9.442	0.024	0.582
+UNDIAL (LoRA)	2.7B	3	$0.151_{-0.238}$	0.067	0.803	0.525	$0.795_{-0.005}$	10.019	0.027	0.582
+UNDIAL (LoRA)	2.7B	10	$0.089_{-0.300}$	0.022	0.768	0.467	$0.774_{-0.026}$	10.787	0.029	0.582

Table 2: **Results for different model sizes and with LoRA-based PEFT.** FT refers to full-model fine-tuning. We highlight the performance delta of EL_3 and MAUVE. NLU_a is the average overall 6 NLU tasks. Lower is better, except with MAUVE and NLU_a.

```python
def compute_loss(self, model, inputs, return_outputs=False):
    input_ids = inputs['input_ids']
    attention_mask = inputs['attention_mask']
    student_logits = model(input_ids=input_ids,
        attention_mask=attention_mask).logits

    # Shift input_ids and logits for causal language modeling
    shift_labels = input_ids[..., 1:].contiguous()
    shift_student_logits = student_logits[..., :-1, :].contiguous()

    # Get teacher logits using the unlearned teacher model
    with torch.no_grad():
        teacher_logits = self.unlearn_teacher_model(
            input_ids=input_ids, attention_mask=attention_mask
        ).logits
    shift_teacher_logits = teacher_logits[..., :-1, :].contiguous()

    # Create mask for memorized tokens
    mask = torch.zeros_like(shift_student_logits)
    batch_indices = torch.arange(mask.shape[0]).view(-1, 1, 1)
    seq_indices = torch.arange(mask.shape[1]).view(1, -1, 1)
    mask[batch_indices, seq_indices, shift_labels.unsqueeze(-1)] = 1

    # Apply penalty to teacher logits and compute soft labels
    pre_softmax = shift_teacher_logits - mask * 10 # assuming a strength of 10
    soft_label = F.softmax(pre_softmax, dim=-1)

    # Compute cross-entropy loss between student logits and soft teacher labels
    loss_fct = CrossEntropyLoss(reduction='none')
    loss = loss_fct(shift_student_logits.view(-1, shift_student_logits.size(-1)),
        soft_label.view(-1, soft_label.size(-1)))
    return loss
```

Figure 8: Python Code for UNDIAL.