LABEL SMOOTHING IMPROVES GRADIENT ASCENT IN LLM UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM unlearning has emerged as a promising approach, aiming to enable models to forget hazardous/undesired knowledge at low cost while preserving as much model utility as possible. Among existing techniques, the most straightforward method is performing Gradient Ascent (GA) w.r.t. the forget data, thereby forcing the model to unlearn the forget dataset. However, GA suffers from severe instability, as it drives updates in a divergent direction, often resulting in drastically degraded model utility. To address this issue, we propose Smoothed Gradient Ascent (**SGA**). Theoretically, we analyze how the smoothing rate affects the update direction and derive a local condition for reducing its component along the pure GA direction. Intuitively, this extends GA by combining gradient contributions from forget and normal data, with the smoothing rate controlling their relative weights. Empirically, we evaluate **SGA** on three benchmarks: TOFU, Harry Potter, and MUSE-NEWS. Experimental results show that, with an appropriate smoothing rate, **SGA** can improve the forgetting–utility trade-off over GA and achieve competitive performance against existing baselines.

1 INTRODUCTION

The rapid development of large language models (LLMs) has enabled widespread adoption (Achiam et al., 2023; Zhang et al., 2022; Touvron et al., 2023; Li et al., 2023; Guo et al., 2025; Team et al., 2023; Yan et al., 2025; Bao et al., 2024; Singhal et al., 2023) but also raised significant security concerns, particularly regarding harmful knowledge acquired during pre-training, such as personal attacks (Yao et al., 2024), privacy breaches (Staab et al., 2023; Mireshghallah et al., 2023; Das et al., 2025; Di et al., 2024a), or copyright violations (Karamolegkou et al., 2023; Chu et al., 2024; Grynbaum & Mac, 2023; Zhang et al., 2024c;b). Since such knowledge is embedded in model representations, it can surface in outputs. Retraining from scratch after corpus filtering is computationally prohibitive, motivating research on LLM unlearning (Yao et al., 2024), which seeks to remove problematic knowledge from trained models. Existing approaches fall into two categories: **fine-tuning-based methods** (Wang et al., 2024), which modify model weights via supervised fine-tuning on forget and retain data, and **training-free methods** (Deng et al., 2025), which preserve parameters but enforce forgetting through external mechanisms. Fine-tuning–based methods modify model weights to achieve unlearning, but often at the cost of degraded utility. In contrast, training-free methods generally preserve model performance, yet their effectiveness is frequently questioned since they do not alter the model's parameters.

One of the most widely used fine-tuning–based unlearning approaches is **Gradient Ascent (GA)** (Yao et al., 2024; Maini et al., 2024), which induces forgetting by inverting the supervised fine-tuning (SFT) loss on forget data. While GA often achieves strong forgetting, it suffers from severe instability and can substantially degrade overall model utility (Zhang et al., 2024a; Fan et al., 2024b). **Gradient Diff** (Maini et al., 2024; Yao et al., 2024) attempts to address this by incorporating retain data, but its effectiveness is limited in practice, as identifying suitable retain sets is often infeasible (Wang et al., 2024). To overcome these challenges, we propose **Smoothed Gradient Ascent (SGA)**, which leverages label smoothing to combine forget data with semantically related yet safe normal data through a tunable smoothing rate. Figure 1 illustrates the pipeline of our method. This design enables the model to forget harmful knowledge while reinforcing correct responses, thereby achieving a more favorable balance between forgetting and retention. Experimental results across multiple benchmarks demonstrate that **SGA** achieves competitive forgetting–utility trade-offs rel-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

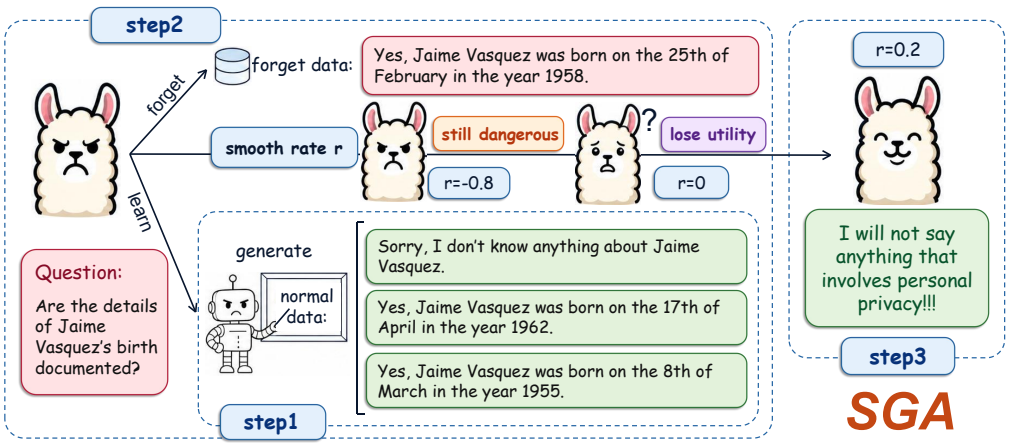

Figure 1: **SGA** pipeline. **SGA** extends Gradient Ascent (**SGA**, $r = 0$) by incorporating normal data generated from the normal model, which is combined with the original forget data to form a distribution over K labels. The smoothing rate r regulates the balance between learning and forgetting, and varying its value leads to different outcomes. Through this mechanism, **SGA** helps suppress the divergence tendency of GA, which arises from maximizing the loss solely on the forget set.

ative to existing baselines, and further analysis suggests that the effective smoothing rate range strongly depends on model scale.

Our contributions are threefold:

- We propose **Smoothed Gradient Ascent (SGA)**, a fine-tuning-based unlearning method that achieves effective unlearning using only forget data, without requiring the retain data commonly assumed by existing unlearning approaches. By incorporating smoothed updates, **SGA** substantially alleviates the catastrophic divergence and instability issues inherent to Gradient Ascent (GA)-based unlearning.
- Theoretically, we derive a local geometric condition under which the smoothing rate reduces the update component along the pure GA direction. This condition depends on the relationship between the gradients induced by the forget and normal data.
- Empirically, we evaluate **SGA** on three representative benchmarks: **TOFU** (entity unlearning), **Harry Potter** (copyrighted content unlearning), and **MUSE-NEWS** (news-domain unlearning). The results show that appropriately chosen smoothing rates can improve forgetting quality and mitigate utility degradation relative to GA, with performance varying across models and benchmarks.

2 Related Work

Training-free LLM unlearning methods. Training-free methods avoid modifying model parameters, instead relying on prompt manipulation or output adjustments to steer predictions away from hazardous distributions (Pawelczyk et al., 2023; Muresanu et al., 2024; Thaker et al., 2024; Gao et al., 2024). GUARD (Deng et al., 2025) uses a classifier to detect unlearning-sensitive prompts and semantic matching to block responses tied to forget data. ECO Prompt (Liu et al., 2024) also leverages a classifier, but introduces lightweight perturbations to guide outputs toward safe responses. Soft Prompt Unlearning (Bhaila et al., 2024), by contrast, attaches learnable soft prompts to inputs to induce unlearning.

Fine-tuning-based LLM unlearning methods. Fine-tuning approaches achieve unlearning by directly modifying model weights (Fan et al., 2024a; Jia et al., 2024a; Fan et al., 2024b; Zhuang et al., 2024; Fan et al., 2025; Zhuang et al., 2025; Ji et al., 2024). Gradient Ascent (GA) (Bourtoule et al., 2021) inverts the SFT loss, transforming learning on forget data into forgetting. Gradient Difference (GD) (Wang et al., 2023) extends GA by additionally training on retain data. NPO (Zhang et al., 2024a) mitigates GA's collapse risk by incorporating a reference model to enforce a lower bound,

an idea also adopted in DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024). SimNPO (Fan et al., 2024b) simplifies NPO by removing the reference model. FLAT (Wang et al., 2024) reframes unlearning as maximizing the f-divergence between forget and retain sets. RULE (Zhang et al., 2025) integrates reinforcement learning by combining SFT with a subsequent RL phase. In summary, GA focuses solely on maximizing the next-token loss over the forget set, which can lead to utility collapse. GD addresses this by incorporating retain data, but such data may be hard to obtain and semantically unrelated to the forget set. To overcome these limitations, we propose **SGA**, which uses a normal model to generate a customized normal set, improving both forgetting quality and model utility.

3 PRELIMINARIES

3.1 FORMULATION

Given a pre-trained dataset $\mathcal{D}$, we obtain a pre-trained large language model π_{origin}. The goal of LLM unlearning is to enable π_{origin} to forget hazardous knowledge while preserving its normal performance as much as possible. In fine-tuning-based unlearning methods, we update the model parameters θ through training, thereby transforming π_{origin} into π_{unlearn}.

In unlearning tasks, it is common to construct a forget dataset $\mathcal{D}_f$ and a retain dataset $\mathcal{D}_r$ from the pre-training corpus $\mathcal{D}$. Here, $\mathcal{D}_f$ denotes the data that π_{origin} should forget, which is used to evaluate the forgetting performance of unlearning methods, while $\mathcal{D}_r$ denotes the data that π_{origin} should retain, which is used to assess the preservation of the model's original capabilities. The parameter λ serves as a regularization coefficient to balance these two objectives. Let x denote the model input and y the next-token output, where subscripts f and r indicate samples drawn from the forget set and the retain set, respectively. The optimization objective of unlearning methods can be expressed as:

$$\min_{\theta} \underbrace{\mathbb{E}_{(x_f,y_f)\in\mathcal{D}_f}[\ell_f(y_f \mid x_f;\theta)]}_{\text{forget}} + \lambda \underbrace{\mathbb{E}_{(x_r,y_r)\in\mathcal{D}_r}[\ell_r(y_r \mid x_r;\theta)]}_{\text{retain}} \tag{1}$$

3.2 GRADIENT ASCENT

The conventional Gradient Ascent (GA) method performs unlearning solely on the forget data, which corresponds to setting $\lambda = 0$ in Equation 1:

$$\min_{\theta} \underbrace{\mathbb{E}_{(x_f,y_f)\in\mathcal{D}_f}\big[\ell_f(y_f \mid x_f;\theta)\big]}_{\text{forget}}. \tag{2}$$

GA achieves unlearning by negating the original training loss, thereby reversing the learning process into "forgetting" and diminishing the influence of previously learned data. At step t, GA updates the model in the direction that *maximizes* the next-token prediction loss on the forget set (Wang et al., 2024; Yao et al., 2024), i.e., $\theta_{t+1} \leftarrow \theta_t + \lambda \nabla_{\theta_t} L(x, y; \theta_t)$, where λ denotes the (un)learning rate. However, GA is prone to catastrophic collapse (Zhang et al., 2024a) due to its inherently divergent nature.

To address this, we hypothesize that instead of relying solely on forget data, the model can also be provided with safe answers. While GA maximizes the loss on the forget set, the additional safe losses can partially redirect its optimization trajectory. Based on this intuition, we propose **SGA**, a gradient combination approach inspired by generalized label smoothing.

3.3 GENERALIZED LABEL SMOOTHING

Generalized Label Smoothing (GLS) is a simple yet effective learning paradigm that has been widely applied in trustworthy machine learning and deep learning (Wei et al., 2021; Szegedy et al., 2016; Lukasik et al., 2020; Liu & Guo, 2020). Let y_i be the one-hot encoded vector form of the label generated according to Y. The random variable of the generalized smoothed label $Y^{\text{GLS},r}$ with smooth rate $r \in (-\infty, 1]$ is defined as $y_i^{\text{GLS},r} := (1-r)\cdot y_i + \frac{r}{K}\cdot \mathbf{1}$ where K is the number of classes and $\mathbf{1}$ is the all-ones vector. For example, when $r = 0.3$ and $y_i = [1, 0, 0]^\top$, the generalized

smoothed label becomes $y_i^{\mathrm{GLS},0.3} = [0.8, 0.1, 0.1]^\top$. Conversely, when $r = -0.3$, it becomes $y_i^{\mathrm{GLS},-0.3} = [1.2, -0.1, -0.1]^\top$.

4 SMOOTHED GRADIENT ASCENT (**SGA**)

In this section, we provide a detailed description of the **SGA** method, followed by an analysis of its improvements over the standard GA approach. Finally, we analyze how the smoothing rate affects the update geometry and discuss the information provided by gradient statistics computed before training.

4.1 METHOD DETAILS

Step 1: Normal Data Generation. For each forget data instance, we generate $K-1$ corresponding normal data samples. The generation of normal data follows the principle that they should either be semantically similar to the forget data or constitute entirely safe responses. For example, in the TOFU experiments, we select the $K-1$ samples from the retain data that have the highest cosine similarity in embeddings with the forget data and exceed a predefined threshold; if such samples cannot be found, we substitute them with safe responses such as "I don't know." In the MUSE and Harry Potter experiments, we employ GPT-4o-mini (Achiam et al., 2023) to generate $K-1$ semantically similar normal data samples for each forget data instance, ensuring that they do not contain any harmful information. We present the procedure for generating normal data with GPT-4o-mini in Appendix C.

Step 2: Smoothed Gradient Ascent. We extend **Generalized Label Smoothing (GLS)** by introducing a smoothing rate r, which combines each forget data instance with $K-1$ semantically similar normal samples to guide the base LLM during unlearning. Under the standard GA setting (**SGA** with $r = 0$), the model ignores the normal data. In the resulting K-dimensional label space $(\text{forget data}, \text{normal data}_1, \ldots)$, GA corresponds to the target label $(-1, 0, 0, \ldots)$, where the negative coefficient enforces forgetting. In contrast, under **SGA** with smoothing rate r, the label distribution becomes $\big(-(1-\frac{K-1}{K}r),\ -\frac{r}{K},\ -\frac{r}{K},\ \ldots\big)$. Finally, the optimization objective in (1) can be reformulated as:

$$\min_\theta \underbrace{\left(1 - r + \tfrac{r}{K}\right) \mathbb{E}_{(x_f, y_f)\in\mathcal{D}_f} \left[\ell_f(y_f \mid x_f; \theta)\right]}_{\text{forget}} + \underbrace{\left(\tfrac{r}{K}\right) \mathbb{E}_{(x_p, y_p)\in\mathcal{D}_p} \left[\sum_{k=1}^{K-1} \ell_p^{(k)}(y_p^{(k)} \mid x_p^{(k)}; \theta)\right]}_{\text{normal data}} \tag{3}$$

Here, the subscript p denotes samples drawn from our generated normal set.

It is worth mentioning that (Di et al., 2024b) proposed the UGradSL method to address machine unlearning for image classification tasks. While this work also uses the label smoothing idea, the context (image classification vs. LLM), methodology, and even the smoothing labels (images vs. tokens) differ significantly from **SGA**.

4.2 HOW **SGA** MODIFIES THE GA UPDATE

For the gradient analysis, let L_f and $L_p^{(k)}$ denote the negative cross-entropy losses on the forget sample and the k-th normal sample, respectively, consistent with the loss convention in Equation 3. Pure GA performs gradient descent on L_f, equivalently gradient ascent on the corresponding cross-entropy loss:

$$\Delta\theta_t \ \propto\ -g_f, \qquad g_f \triangleq \nabla_{\theta_t} L_f.$$

This purely divergent direction often leads to catastrophic collapse.

Under our GLS-based construction (Equation 3), the per-step objective combines the forget loss with $K-1$ normal losses via the smooth rate r:

$$\Delta\theta_t \propto -\nabla_{\theta_t} \mathcal{L}_{\mathrm{SGA}}(\theta_t) = -\left[\left(1 - r + \tfrac{r}{K}\right) g_f + \left(\tfrac{r}{K}\right) \sum_{k=1}^{K-1} g_p^{(k)}\right] \tag{4}$$

where $g_p^{(k)} \triangleq \nabla_{\theta_t} L_p^{(k)}$ denotes the gradient contributed by the k-th normal sample. Thus, a key reason why **SGA** helps suppress the divergence tendency of GA is that it alters GA's gradient-ascent direction, preventing the model from updating purely toward maximizing the next-token loss on the forget set.

4.3 Geometric Analysis of the Smoothing Rate

Given the forget data and $K-1$ normal data samples, we analyze how the smoothing rate r changes the update component along the pure GA direction.

Let g_f and $g_p^{(k)}$ denote the gradients defined in Section 4.2. For $K > 1$, define $\bar{g}_p := \frac{1}{K-1}\sum_{k=1}^{K-1} g_p^{(k)}$ and $u \triangleq \bar{g}_p - \left(1 - \frac{1}{K}\right) g_f$. Substituting $\bar{g}_p = u + \left(1 - \frac{1}{K}\right) g_f$ into Equation 4 yields

$$d(r) = -\left(1 - \frac{r(K-1)}{K^2}\right) g_f - \frac{r(K-1)}{K} u. \tag{5}$$

At $r = 0$, this reduces to the pure GA direction $-g_f$. For nonzero r, the update changes through both the coefficient of g_f and the contribution along u.

For $g_f \neq 0$, the component along the pure GA direction can be assessed through the inner product

$$\langle d(r), -g_f \rangle = \|g_f\|^2 + \frac{r(K-1)}{K}\left(\langle g_f, u\rangle - \frac{\|g_f\|^2}{K}\right).$$

Relative to pure GA, this inner product decreases if and only if

$$r\left(\langle g_f, u\rangle - \frac{\|g_f\|^2}{K}\right) < 0. \tag{6}$$

Thus, a positive r decreases this component when $\langle g_f, u\rangle < \|g_f\|^2/K$, whereas a negative r does so when $\langle g_f, u\rangle > \|g_f\|^2/K$. At equality, the component is independent of r. This local geometric condition does not by itself guarantee stable training or effective unlearning.

Figure 2 reports the sign of $\langle g_f, u\rangle$ on the base LLMs for each forget instance, summarizing the geometric relationship between g_f and u. A negative inner product is sufficient for a positive r to decrease the GA-direction component. A positive inner product, however, must be compared with $\|g_f\|^2/K$ to determine which sign of r decreases this component. These sign statistics therefore provide partial geometric information rather than a complete rule for selecting r. Since the gradients evolve during training, we evaluate the forgetting–utility trade-offs of different smoothing rates empirically in Section 5.2.

5 Experiments

In this section, we evaluate the performance of **SGA** on three established LLM unlearning tasks: TOFU (Maini et al., 2024), MUSE-News (Shi et al., 2024), and Harry Potter (Yao et al., 2024). Among them, the TOFU task focuses on the problem of entity unlearning, the Harry Potter experiment addresses copyright-related issues, and MUSE-News concerns the evaluation of general unlearning capabilities.

5.1 Baseline Methods

We compare **SGA** against other fine-tuning-based LLM unlearning methods across three tasks to evaluate its effectiveness. For all tasks, we include Gradient Ascent (GA) (Yao et al., 2024), KL minimization (KL) (Maini et al., 2024), GradDiff (GD) (Liu et al., 2022), NPO (Zhang et al., 2024a), and Forget-data-only Loss Adjustment (FLAT) (Wang et al., 2024) as baselines. On the copyrighted content task (Harry Potter) and the entity unlearning task (TOFU), we further compare with Preference Optimization (PO) (Maini et al., 2024), Large Language Model Unlearning (LLMU) (Yao et al., 2024), and DPO. For the Harry Potter task, we additionally include the Mismatch (Liu et al., 2024) method. For the MUSE-News task, we incorporate Task Vector, Who's Harry Potter (WHP) (Eldan & Russinovich, 2023), and NPO-RT (Zhang et al., 2024a) as baselines.

Table 1: We evaluate the performance of **SGA** and baseline approaches on three base LLMs: Llama2-7B, OPT-2.7B, and Phi-1.5B. Here, FQ, MU, F-RL, and R-RL denote Forget Quality, Model Utility, and the ROUGE-L scores on the forget and retain sets, respectively. For reference, we also report results from the Original and Retained LLMs. The top-2 results are highlighted in **blue**.

Base LLM	Llama2-7B				OPT-2.7B				Phi-1.5B			
Metric	FQ(↑)	MU(↑)	F-RL(↓)	R-RL(↑)	FQ(↑)	MU(↑)	F-RL(↓)	R-RL(↑)	FQ(↑)	MU(↑)	F-RL(↓)	R-RL(↑)
Original LLM	4.4883e-06	0.6239	0.9851	0.9818	0.0013	0.5112	0.7537	0.8807	0.0013	0.5195	0.9607	0.9276
Retained LLM	1.0	0.6267	0.4080	0.9833	1.0	0.5067	0.4217	0.7669	1.0	0.5233	0.4272	0.9269
GA/**SGA** (r = 0)	**0.0068**	0.5990	0.4817	0.9204	0.0541	0.4851	0.4490	**0.6440**	0.0541	0.5058	0.4914	0.8012
KL	0.0030	0.5994	0.4922	0.9172	0.0143	0.4785	0.3971	0.6010	0.0541	0.5087	0.4910	0.7603
GD	**0.0068**	0.5998	0.4869	0.9182	0.0971	0.4826	0.4085	0.6016	0.0286	0.5117	0.4991	0.7959
LLMU	0.0030	0.5999	0.4891	0.9236	0.0068	0.1849	0.0210	0.1727	0.0143	0.5108	0.3331	0.7785
PO	0.0030	**0.6323**	0.1752	0.9169	0.0971	0.3706	0.0393	0.3594	**0.0541**	0.5064	0.4958	0.8003
DPO-RT	**0.0068**	**0.6322**	0.2595	0.9091	0.0068	0.3278	0.0348	0.2741	0.0286	0.5053	0.2824	0.7372
NPO-RT	0.0030	0.5994	0.5049	0.9270	0.0286	0.4844	0.4353	0.6401	0.0286	0.5090	0.4957	0.7660
FLAT (Pearson)	0.0030	0.6304	0.4825	0.9284	0.0030	0.4714	0.1962	0.4973	0.0541	**0.5160**	**0.4228**	0.7674
SGA (r = 0.2)	**0.0068**	0.6018	**0.4777**	0.9229	0.0286	0.4799	0.3967	0.5996	0.0286	0.5060	**0.4782**	**0.8075**
SGA (r = -0.8)	0.0030	0.6037	0.4871	**0.9281**	**0.2657**	**0.4867**	0.3946	0.6139	**0.0971**	0.5092	0.4957	0.8040
SGA (r = -2)	0.0030	0.6046	0.4897	0.9224	**0.2657**	0.4784	0.3959	0.6012	**0.0971**	**0.5120**	0.4816	0.8043

5.2 Entity Unlearning

Experiment Setup. The TOFU (Maini et al., 2024) dataset is a question-answering benchmark built on the biographies of 200 fictional authors, with each author associated with 20 QA pairs. Based on the designated unlearning scope, the dataset is partitioned into a forget set and a retain set. In our experiments, we adopt the 1% forget split. Following (Deng et al., 2025; Wang et al., 2024), we use Llama2-7B (Touvron et al., 2023), Phi-1.5B (Li et al., 2023), and OPT-2.7B (Zhang et al., 2022) as the base LLMs. In addition, we employ the bge-large-en-v1.5 embedding model (Xiao et al., 2023) as an auxiliary encoder to select normal data from the retain set for training.

Evaluation Metrics. To evaluate both the forgetting quality and the retained utility of the unlearned models, we employed two metrics from the TOFU benchmark: **Forget Quality (FQ)** and **Model Utility (MU)** (Maini et al., 2024). FQ is evaluated using the p-value of a Kolmogorov–Smirnov test comparing the unlearned model's outputs with those of a retain-only model, where higher values indicate stronger forgetting. MU reflects the model's utility by aggregating performance on held-out retain data spanning fictional authors, real-world profiles, and factual knowledge. We additionally report **ROUGE-L** scores on both the forget and retain sets. On the forget set, a score closer to that of the retain model indicates better forgetting performance, while on the retain set, higher ROUGE-L scores correspond to better utility. Full metric details are provided in Appendix B.1.

SGA achieves the best forget quality across all models. According to Table 1, we observe that, compared with all baselines, **SGA** with an effective smoothing rate achieves the best forget quality. Moreover, it substantially surpasses GA (i.e., **SGA** with $r = 0$), indicating that incorporating normal data during unlearning leads to more effective forgetting.

Compared with GA (corresponding to SGA with $r = 0$), SGA generally demonstrates superior performance. Across all three models, **SGA** with an appropriately tuned smoothing rate consistently outperforms Gradient Ascent on nearly all metrics. This suggests that incorporating normal data with forget data during training enables effective forgetting while better preserving model performance.

Gradient geometry and smoothing-rate sensitivity on TOFU. Following Section 4.3, we report the sign of $\langle g_f, u \rangle$ for each forget instance and base LLM in Figure 2. These statistics describe the initial gradient geometry, while Table 1 evaluates the empirical effects of different smoothing rates. The sign statistics alone do not establish the condition in Equation 6, which also depends on $\|g_f\|^2/K$. Different smoothing rates also significantly affect the model's confidence in generating key information, enabling unlearning (Figure 3).

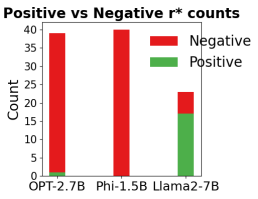

Figure 2: Sign statistics of $\langle g_f, u \rangle$ on the base LLMs for forget instances in the TOFU forget01 set. These statistics summarize the geometric relationship between g_f and u; the condition in Section 4.3 additionally depends on $\|g_f\|^2/K$.

5.3 Copyrighted Content Unlearning

Experiment Setup. Following prior work (Deng et al., 2025; Wang et al., 2024), we use *Harry Potter and the Sorcerer's Stone* (Rowling, 2023; Eldan & Russinovich, 2023) as copyrighted material

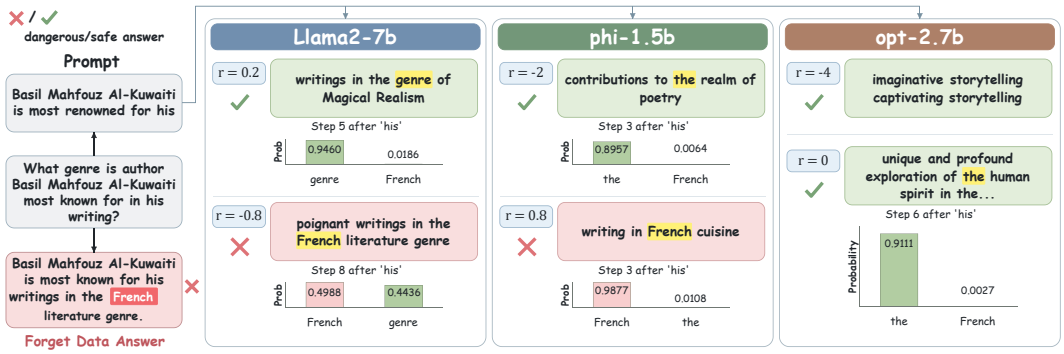

Figure 3: Different smoothing rates yield distinct output probabilities for critical tokens across base LLMs. As shown in the figure, the forget data highlights Basil Mahfouz Al-Kuwaiti's achievements in **French** literary style. In the LLaMA2-7B model, training with $r = -0.8$ assigns a much higher probability to the token "French," whereas training with $r = 0.2$ makes the model far less likely to produce this privacy-sensitive token.

for unlearning. We extract 400 segments (≤ 512 tokens each) as the forget set $\mathcal{D}_f$ (Deng et al., 2025; Wang et al., 2024; Jia et al., 2024b), and sample 400 paragraphs from C4 (Raffel et al., 2020) as the retain set $\mathcal{D}_r$. For each forget instance, GPT-4o-mini (Achiam et al., 2023) generates safe normal data. Following (Wang et al., 2024), we fine-tune OPT-2.7B (Zhang et al., 2022) and Llama2-7B (Touvron et al., 2023) to simulate memorization and run experiments on the fine-tuned models.

Evaluation Metrics. Following the choice of evaluation metrics in prior work (Deng et al., 2025; Wang et al., 2024), we also consider both unlearning effectiveness and model utility. Unlearning effectiveness is measured using the Forget Quality Gap (FQ Gap), which evaluates the forgetting performance based on the differences in BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) scores between the unlearned model and the retained model. Model utility is assessed through perplexity (PPL) on Wikitext (Merity et al., 2016) and the average accuracy across nine zero-shot benchmarks. Full metric definitions are provided in Appendix B.2.

Table 2: **SGA** and baseline methods are evaluated on the Harry Potter benchmark using two base LLMs: OPT-2.7B and Llama2-7B. We report only the best-performing r for each base model, with full results provided in Appendix E.2. Among the three evaluation metrics—Forget Quality Gap (FQ Gap), Perplexity (PPL), and Average Accuracy (Avg. Acc.)—the top-2 scores are highlighted in blue. Values marked with * are taken directly from (Wang et al., 2024).

Base LLM	**OPT-2.7B**			**Llama2-7B**		
Metric	**FQ Gap(↓)**	**PPL(↓)**	**Avg. Acc.(↑)**	**FQ Gap(↓)**	**PPL(↓)**	**Avg. Acc.(↑)**
Original LLM	1.5346	15.6314	0.4762	3.6594	8.9524	0.5617
Retained LLM	0.0	14.3190	0.4686	0.0	8.7070	0.5599
GA/**SGA** (r=0)*	2.7301	1.0984e71	0.3667	0.4587	47.2769	0.5088
KL*	2.7301	16.1592	**0.4688**	0.4225	9.4336	0.5509
GD*	2.3439	16.1972	**0.4690**	0.5304	9.1797	0.4902
Mismatch*	1.4042	15.7507	0.4679	0.4647	8.9906	0.5593
LLMU*	2.4639	15.8398	0.4656	**0.1985**	9.0530	0.5503
PO*	2.1601	**14.8960**	0.4583	0.5124	**8.8364**	0.5532
DPO*	2.2152	16.8396	0.4621	0.2924	**8.9597**	0.5614
NPO*	1.2611	19.6637	0.4644	0.5151	9.0397	**0.5609**
FLAT (Pearson)*	1.4089	**15.5543**	0.4686	0.2265	8.9906	0.5580
SGA (r=-0.2)	0	6.2700e72	0.3683	**0.0761**	28.9249	0.5421
SGA (r=-0.4)	0	6.8530e30	0.3876	0.5355	11.4885	**0.5676**
SGA (r=-0.8)	**0.0043**	1.3040e6	0.4497	0.4141	19.9494	0.5316
SGA (r=-2)	**0.0084**	6.7100e5	0.4532	0.350	18.67	0.5305

By appropriately tuning the smoothing rate, SGA consistently outperforms Gradient Ascent. As shown in Table 2, aside from a few cases where training collapses under extreme smoothing rates, **SGA** consistently surpasses GA (i.e., **SGA** with $r = 0$) across all evaluation metrics, including FQ Gap, PPL, and Avg. ACC. This corroborates our discussion in Section 4.2, where we argued that incorporating normal data effectively mitigates GA's instability, which otherwise maximizes loss solely on the forget set. Notably, on OPT-2.7B and Llama2-7B, the GA baseline suffers severe collapse, manifested as extremely high PPL and drastically low Avg. ACC. In contrast, **SGA** substantially suppresses this issue, rendering GA's catastrophic forgetting problem far less severe.

SGA generally achieves superior forgetting capability. On OPT-2.7B and Llama2-7B, we observe that with properly tuned smoothing rates, **SGA** achieves strong forgetting performance. This is because **SGA** retains gradient ascent on forget data, inheriting GA's strengths while improving stability.

Discussion on Perplexity (PPL). For the OPT-2.7B and Llama2-7B models, GA yields extremely high PPL values. This indicates that models after GA-based unlearning exhibit very high perplexity, reflecting the severe collapse of GA and the resulting degradation in model utility. In contrast, the PPL results for **SGA** in Table 2 demonstrate that it significantly alleviates this issue, even reducing the PPL on Llama2-7B to a level comparable with other unlearning methods. This further confirms that by jointly leveraging normal data and forget data, **SGA** effectively mitigates the catastrophic forgetting problem inherent to GA.

5.4 MUSE-NEWS UNLEARNING

Experiment Setup. We evaluate **SGA** on the MUSE-News benchmark (Shi et al., 2024), constructed from real-world BBC news articles and partitioned into three subsets: a forget set, a retain set, and a holdout set for utility evaluation. In experiments, we perform unlearning on the Llama2-7B (Touvron et al., 2023) pretrained model provided by the MUSE benchmark. The baseline methods in the evaluation are obtained or reproduced from the original implementations provided by MUSE, while **SGA** is specifically developed and implemented on the MUSE-News benchmark.

Evaluation Metrics. We evaluate both the baseline methods and **SGA** using the metrics provided by the MUSE benchmark: ***VerbMem*** on the forget dataset, ***KnowMem*** on the forget and retain datasets, and Privacy leakage (***PrivLeak***). ***VerbMem*** measures the model's ability to continue generating forgotten text, while ***KnowMem*** evaluates whether the model preserves knowledge from both the forget and retain sets. ***PrivLeak*** assesses privacy leakage via membership inference (MIA). For a more detailed description of the evaluation metrics used in MUSE-NEWS, please refer to Appendix B.3.

Table 3: Evaluation on the MUSE benchmark uses four criteria. Results are highlighted in **blue** when the method satisfies the criterion, and in **red** otherwise. Smaller D_f and larger D_r are preferable, while PrivLeak close to 0 indicates better privacy. The top-2 PrivLeak results meeting other criteria are highlighted in **green**. Values marked with * are from (Wang et al., 2024).

	VerbMem on D_f (↓)		**KnowMem on D_f (↓)**		**KnowMem on D_r (↑)**		**PrivLeak**
Original LLM	58.4	-	63.9	-	55.2	-	-99.8
Retained LLM	20.8	-	33.1	-	55.0	-	0.0
Task Vectors*	56.3	(✘)	63.7	(✘)	54.6	(✔)	-99.8
WHP*	19.7	(✔)	21.2	(✔)	28.3	(✔)	109.6
GA*	0.0	(✔)	0.0	(✔)	0.0	(✘)	17.0
GD*	4.9	(✔)	27.5	(✔)	6.7	(✔)	109.4
KL*	27.4	(✘)	50.2	(✘)	44.8	(✔)	-96.1
NPO*	0.0	(✔)	0.0	(✔)	0.0	(✘)	15.0
NPO-RT*	1.2	(✔)	54.6	(✘)	40.5	(✔)	105.8
FLAT (Pearson)*	1.6	(✔)	0.0	(✔)	0.2	(✔)	**26.8**
SGA	0	(✔)	0	(✔)	1.9498	(✔)	**15.5700**

SGA maintains GA's strong performance on the forget set D_f while improving upon it on the retain set D_r. On D_f, **SGA** substantially reduces memorization risk in VerbMem and KnowMem, achieving scores of 0 across both metrics—well below the Retained LLM baseline and fully comparable to GA. On D_r, **SGA** raises GA's KnowMem score from 0 to 1.9498, thereby meeting the criterion that KnowMem should not be 0. Consequently, **SGA** not only inherits GA's strengths on D_f but also effectively remedies its shortcomings on D_r.

Among methods satisfying the criteria on both VerbMem and KnowMem, SGA ranks within the top-2 for PrivLeak, achieving a score of 15.57. This demonstrates that **SGA** provides substantially better control over privacy leakage compared to other approaches. In summary, on the MUSE-NEWS benchmark, **SGA** has been validated as an unlearning method that not only meets all evaluation criteria but also achieves the strongest performance in mitigating privacy leakage.

5.5 Ablation Studies

Ablation study on the selection of TOFU normal data. As described in Section 5.2, the procedure for collecting normal data in TOFU differs from the other experiments. Instead of relying entirely on generation from an external model, we compute similarity between retain and forget data using an embedding model, and select the most similar samples from the retain set as normal data. In the ablation study, we compare training with normal data generated by GPT-4o-mini and data selected by the embedding model. The results are shown in Table 4.

Table 4: Ablation study results on TOFU across different base LLMs. For each method, we report results for the two best smoothing rates, with the top-2 outcomes highlighted in **blue**.

Llama2-7B			OPT-2.7B			Phi-1.5B		
Metric	FQ(↑)	MU(↑)	**Metric**	FQ(↑)	MU(↑)	**Metric**	FQ(↑)	MU(↑)
SGA ($r = 0.4$)	0.0068	0.6027	**SGA** ($r = -0.8$)	**0.2657**	**0.4867**	**SGA** ($r = -0.8$)	**0.0971**	**0.5092**
SGA ($r = 0.2$)	0.0068	0.6018	**SGA** ($r = -2$)	**0.2657**	0.4784	**SGA** ($r = -2$)	**0.0971**	**0.5120**
GPT ($r = -2$)	**0.1650**	**0.6170**	GPT ($r = 0.4$)	**0.2657**	**0.4830**	GPT ($r = -0.4$)	0.0068	0.4594
GPT ($r = -4$)	**0.2657**	**0.6189**	GPT ($r = -8$)	**0.2657**	0.4822	GPT ($r = -0.2$)	0.0068	0.4669

The **SGA** method supported by GPT remains effective. We observe that **SGA** supported by GPT does not lose its effectiveness; instead, it achieves better performance than the variant using normal data selected by the embedding model on certain base LLMs. The empirically effective smoothing rates also vary with the source of normal data. This observation motivates examining the normal-data gradients, which enter the geometric condition in Section 4.3.

Ablation study on external LLMs. To examine whether external LLM choice affects normal-data generation, we replace GPT-4o-mini with Qwen3-Max (Yang et al., 2025) and DeepSeek-V3.2 (DeepSeek-AI et al., 2025), and repeat the TOFU experiments. As shown in Table 5, SGA remains effective across different external LLMs, suggesting that **SGA** is only mildly affected when the external LLM is properly selected. As stronger LLMs emerge, failures caused by unsuitable normal data may become less likely.

Table 5: Ablation study results on the selection of external LLMs for generating normal data on the TOFU benchmark.

External LLM	Llama2-7B		OPT-2.7B		Phi-1.5B	
	FQ(↑)	MU(↑)	FQ(↑)	MU(↑)	FQ(↑)	MU(↑)
GPT-4o-mini	0.1650	0.6170	0.2657	0.4830	0.0068	0.4594
Qwen3-Max	0.1650	0.6216	0.1650	0.4735	0.1650	0.6331
DeepSeek-V3.2	0.1650	0.6331	0.0971	0.4730	0.1650	0.4938

Ablation study on K. We further study the effect of K on unlearning performance. The results in Table 6 show that varying K has limited impact. In practice, K can be selected based on computational needs, but should be larger than 1; when $K = 1$, no normal samples are used and SGA degenerates into Gradient Ascent.

Table 6: Ablation study on the choice of K for the proposed method.

	Llama2-7B		OPT-2.7B		Phi-1.5B	
External LLM	FQ (↑)	MU (↑)	FQ (↑)	MU (↑)	FQ (↑)	MU (↑)
SGA (K=2)	0.1650	0.6037	0.0971	0.4732	0.1650	0.4945
SGA (K=4)	0.0030	0.6037	0.2657	0.4867	0.0971	0.5092
SGA (K=6)	0.1650	0.6153	0.0286	0.4699	0.2657	0.4945

6 Conclusion

This paper studies the limitations of Gradient Ascent in LLM unlearning and proposes **Smoothed Gradient Ascent (SGA)**. Inspired by Generalized Label Smoothing, **SGA** uses auxiliary normal models to generate customized normal data for the forget set and combines the normal and forget data using a tunable smoothing rate r. This design aims to better balance forgetting and retention during training. We further analyze how the smoothing rate affects the update geometry and empirically evaluate its impact on forgetting and model utility. Experiments on TOFU, Harry Potter, and MUSE-NEWS suggest that **SGA** can improve over Gradient Ascent and achieve competitive performance compared with existing baselines in terms of forgetting quality and model utility.

AI USE STATEMENT

This paper uses AI tools to polish the writing and correct grammatical errors. In addition, we use GPT-4o-mini to generate normal data and additionally evaluate Qwen3-Max and DeepSeek-V3.2 as alternative data generators. The prompts used for data generation are provided in Appendix C.

ETHICS STATEMENT

This work studies LLM unlearning and does not involve human subjects, private user data, or real-world deployment. All experiments are conducted on publicly available benchmarks and pretrained models. We do not identify any additional ethical concerns beyond those generally associated with LLM research.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed experimental configurations, including datasets, models, hyperparameters, evaluation metrics, and implementation details, in Appendices D and B. Upon acceptance, we will publicly release all our code and normal data.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yujia Bao, Ankit Parag Shah, Neeru Narang, Jonathan Rivers, Rajeev Maksey, Lan Guan, Louise N Barrere, Shelley Evenson, Rahul Basole, Connie Miao, et al. Harnessing business and media insights with large language models. *arXiv preprint arXiv:2406.06559*, 2024.

Karuna Bhaila, Minh-Hao Van, and Xintao Wu. Soft prompting for unlearning in large language models. *arXiv preprint arXiv:2406.12038*, 2024.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

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

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Timothy Chu, Zhao Song, and Chiwun Yang. How to protect copyright data in optimization of large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17871–17879, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pp. 177–190. Springer, 2005.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39, 2025.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi

Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Zhijie Deng, Chris Yuhao Liu, Zirui Pang, Xinlei He, Lei Feng, Qi Xuan, Zhaowei Zhu, and Jiaheng Wei. Guard: Generation-time llm unlearning via adaptive restriction and detection. *arXiv preprint arXiv:2505.13312*, 2025.

Zonglin Di, Sixie Yu, Yevgeniy Vorobeychik, and Yang Liu. Adversarial machine unlearning. *arXiv preprint arXiv:2406.07687*, 2024a.

Zonglin Di, Zhaowei Zhu, Jinghan Jia, Jiancheng Liu, Zafar Takhirov, Bo Jiang, Yuanshun Yao, Sijia Liu, and Yang Liu. Label smoothing improves machine unlearning, 2024b. URL https://arxiv.org/abs/2406.07698.

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

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Chongyu Fan, Jiancheng Liu, Alfred Hero, and Sijia Liu. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In *European Conference on Computer Vision*, pp. 278–297. Springer, 2024a.

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

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *arXiv preprint arXiv:2502.05374*, 2025.

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning. *arXiv preprint arXiv:2407.10223*, 2024.

Michael M Grynbaum and Ryan Mac. The times sues openai and microsoft over ai use of copyrighted work. *The New York Times*, 27(1), 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Wagle: Strategic weight attribution for effective and modular unlearning in large language models. *Advances in Neural Information Processing Systems*, 37:55620–55646, 2024a.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*, 2024b.

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

Sasi Kumar Murakonda, Reza Shokri, and George Theodorakopoulos. Quantifying the privacy risks of learning high-dimensional graphical models. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2287–2295. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/kumar-murakonda21a.html.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37:118198–118266, 2024.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates, 2020. URL https://arxiv.org/abs/1910.03231.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pp. 6448–6458. PMLR, 2020.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*, 2023.

Andrei Muresanu, Anvith Thudi, Michael R Zhang, and Nicolas Papernot. Unlearnable algorithms for in-context learning. *arXiv preprint arXiv:2402.00751*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Joanne K Rowling. *Harry Potter and the sorcerer's stone*. Scholastic Incorporated, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

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

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*, 2024.

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

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*, 2023.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data. *arXiv preprint arXiv:2410.11143*, 2024.

Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. *arXiv preprint arXiv:2106.04149*, 2021.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Yibo Yan, Shen Wang, Jiahao Huo, Philip S Yu, Xuming Hu, and Qingsong Wen. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. *arXiv preprint arXiv:2503.18132*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models, 2022. URL https://arxiv.org/abs/2111.09679.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Chenlong Zhang, Zhuoran Jin, Hongbang Yuan, Jiaheng Wei, Tong Zhou, Kang Liu, Jun Zhao, and Yubo Chen. Rule: Reinforcement unlearning achieves forget-retain pareto optimality, 2025. URL https://arxiv.org/abs/2506.07171.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. *CoRR*, 2024b.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024c.

Haomin Zhuang, Yihua Zhang, Kehan Guo, Jinghan Jia, Gaowen Liu, Sijia Liu, and Xiangliang Zhang. Uoe: Unlearning one expert is enough for mixture-of-experts llms. *arXiv preprint arXiv:2411.18797*, 2024.

Haomin Zhuang, Yihua Zhang, Kehan Guo, Jinghan Jia, Gaowen Liu, Sijia Liu, and Xiangliang Zhang. Seuf: Is unlearning one expert enough for mixture-of-experts llms? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8664–8678, 2025.

The Appendix is organized as follows.

- **Section A**: Introduces the baseline methods compared in our experiments.
- **Section B**: Describes the evaluation metrics employed in the experiments.
- **Section C**: Demonstrates how we generate normal data using GPT-4o-mini.
- **Section D**: Details our experimental setup.
- **Section E**: Presenting our complete experimental results

A BASELINE METHODS

Gradient Ascent (GA) (Yao et al., 2024). Gradient Ascent (GA) is a simple baseline commonly adopted in machine unlearning. The key idea is to reverse the effect of gradient descent by maximizing the prediction loss on the forget dataset $\mathcal{D}_f$. This encourages the model to move away from knowledge associated with $\mathcal{D}_f$, thereby approximating a model trained only on the retain set $\mathcal{D}_r$. The objective is defined as:

$$\mathcal{L}_{\text{GA}} = -\frac{1}{|\mathcal{D}_f|} \sum_{(x_i,y_i)\in\mathcal{D}_f} \ell(x_i, y_i; \theta). \tag{7}$$

Gradient Difference (GD) (Liu et al., 2024). Gradient Difference (GD) is another simple baseline for machine unlearning . Unlike GA, which only maximizes the loss on the forget set, GD combines two opposing objectives: minimizing the loss on the retain set $\mathcal{D}_r$ while maximizing the loss on the forget set $\mathcal{D}_f$. This dual objective encourages the model to preserve useful knowledge from $\mathcal{D}_r$ while forgetting information specific to $\mathcal{D}_f$. The composite loss is given by:

$$\begin{aligned}\mathcal{L}_{\text{GD}} = &\frac{1}{|\mathcal{D}_r|} \sum_{(x_r,y_r)\in\mathcal{D}_r} \ell(x_r, y_r; \theta) \\ &- \frac{1}{|\mathcal{D}_f|} \sum_{(x_f,y_f)\in\mathcal{D}_f} \ell(x_f, y_f; \theta).\end{aligned} \tag{8}$$

KL Minimization (KL) (Maini et al., 2024). KL combines gradient ascent on the forget set $\mathcal{D}_f$ with a KL-divergence regularization on the retain set $\mathcal{D}_r$. Specifically, the model is encouraged to forget by maximizing the loss on $\mathcal{D}_f$, while its outputs on $\mathcal{D}_r$ are constrained to stay close to the original (pre-unlearning) model $M_{\hat{\theta}}$ through Kullback–Leibler divergence. The objective can be formulated as:

$$\begin{aligned}\mathcal{L}_{\text{KL}} = &-\frac{1}{|\mathcal{D}_f|} \sum_{(x_f,y_f)\in\mathcal{D}_f} \ell(x_f, y_f; \theta) \\ &+ \frac{1}{|\mathcal{D}_r|} \sum_{x_r\in\mathcal{D}_r} \mathrm{KL}\big(h_{\hat{\theta}}(x_r) \,\|\, h_\theta(x_r)\big)\,.\end{aligned} \tag{9}$$

where $h_{\hat{\theta}}$ and h_θ denote the output distributions of the original and unlearned models, respectively.

Preference Optimization (PO) (Maini et al., 2024).PO modifies the model's response preferences by training it to output safe refusals (e.g., "I don't know") for prompts in the forget set $\mathcal{D}_f$. To achieve this, an augmented forget dataset $\mathcal{D}_{\text{IDK}}$ is constructed, where each input from $\mathcal{D}_f$ is paired with a refusal response. The overall objective combines the fine-tuning loss on the retain set $\mathcal{D}_r$ with a custom loss on $\mathcal{D}_{\text{IDK}}$:

$$\begin{aligned}\mathcal{L}_{\text{PO}} = &\frac{1}{|\mathcal{D}_r|} \sum_{(x_r,y_r)\in\mathcal{D}_r} \ell(x_r, y_r; \theta) \\ &+ \frac{1}{|\mathcal{D}_{\text{IDK}}|} \sum_{(x_f,y_{\text{IDK}})\in\mathcal{D}_{\text{IDK}}} \ell(x_f, y_{\text{IDK}}; \theta).\end{aligned} \tag{10}$$

This encourages the model to retain performance on $\mathcal{D}_r$ while learning to reject answering queries from $\mathcal{D}_f$.

Direct Preference Optimization (DPO) (Rafailov et al., 2023). Direct Preference Optimization (DPO) adapts the preference optimization framework to the unlearning setting. Instead of comparing human-preferred and less-preferred responses, DPO contrasts a safe refusal completion y_e with the original (to-be-forgotten) response y_f under the same forget prompt $x_f \in \mathcal{D}_f$. The objective encourages the model to prefer y_e over y_f, thereby enforcing targeted forgetting while preserving overall utility. Formally, with inverse temperature β, the loss is:

$$\begin{aligned}\mathcal{L}_{\text{DPO}} \;=\; -\; 2\beta\, \mathbb{E}_{x_f \in \mathcal{D}_f}\Bigg[\log \sigma\Big(\beta \log h_\theta(x_f, y_e) \\ - \beta \log h_\theta(x_f, y_f) - M_{\text{ref}}\Big)\Bigg].\end{aligned} \tag{11}$$

where h_θ denotes the model's predictive distribution and M_{ref} is an optional regularization term penalizing deviation from the original model. A retention-regularized variant further adds supervised loss on $\mathcal{D}_r$ to maintain desirable knowledge:

$$\mathcal{L}_{\text{DPO-RT}} \;=\; \mathcal{L}_{\text{DPO}} \;+\; \mathcal{L}(\mathcal{D}_r; \theta). \tag{12}$$

Negative Preference Optimization (NPO) (Zhang et al., 2024a). Negative Preference Optimization (NPO) focuses on suppressing undesired responses by penalizing the likelihood of completions from the forget set $\mathcal{D}_f$. Unlike DPO, which contrasts preferred and dispreferred responses, NPO uses only the dispreferred term, directly discouraging the model from producing the original (to-be-forgotten) outputs. Formally, with inverse temperature β, the loss is:

$$\begin{aligned}\mathcal{L}_{\text{NPO}} \;=\; -\; 2\beta\, \mathbb{E}_{(x_f, y_f) \in \mathcal{D}_f}\Big[\\ \log \sigma\big(- \beta \log h_\theta(y_f | x_f)\big)\Big].\end{aligned} \tag{13}$$

where h_θ denotes the model's predictive distribution. To preserve utility, a retention-regularized variant further adds supervised training on $\mathcal{D}_r$:

$$\mathcal{L}_{\text{NPO-RT}} \;=\; \mathcal{L}_{\text{NPO}} \;+\; \mathcal{L}(\mathcal{D}_r; \theta). \tag{14}$$

Mismatch. Mismatch extends the preference-optimization framework by introducing randomly constructed text sequences (Yao et al., 2024). Similar to PO, it minimizes the fine-tuning loss on the retain set $\mathcal{D}_r$, but additionally incorporates a mismatch loss computed over a random combination of responses Y_{rnd}. This encourages the model to produce neutral outputs when exposed to random or irrelevant continuations, thus reinforcing unlearning. The objective is:

$$\begin{aligned}\mathcal{L}_{\text{Mismatch}} \;=\; & \frac{1}{|\mathcal{D}_r|} \sum_{(x_r, y_r) \in \mathcal{D}_r} \ell(x_r, y_r; \theta) \\ & + \frac{1}{|\mathcal{D}_{\text{rnd}}|} \sum_{y_{\text{rnd}} \in Y_{\text{rnd}}} \ell(x_f, y_{\text{rnd}}; \theta).\end{aligned} \tag{15}$$

where Y_{rnd} denotes a set of randomly sampled responses paired with forget prompts $x_f \in \mathcal{D}_f$.

LLMU (Yao et al., 2024).LLMU extends Gradient Ascent by incorporating two auxiliary components: (1) random-completion unlearning using sequences generated from forget prompts, and (2) retention regularization on normal retain data. The objective encourages forgetting by maximizing loss on the forget set $\mathcal{D}_f$, while simultaneously training on random completions $\mathcal{D}_{\text{rand}}$ and aligning the model's predictions on the retain set $\mathcal{D}_{\text{normal}}$ with the original model through KL divergence. Formally, the loss is:

$$\begin{aligned}\mathcal{L}_{\text{LLMU}} \;=\; & -\; \frac{\varepsilon_1}{|\mathcal{D}_f|} \sum_{(x_f, y_f) \in \mathcal{D}_f} \ell(x_f, y_f; \theta) \\ & + \frac{\varepsilon_2}{|\mathcal{D}_{\text{rand}}|} \sum_{x \in \mathcal{D}_{\text{rand}}} \ell(x; \theta) \\ & + \frac{\varepsilon_3}{|\mathcal{D}_{\text{normal}}|} \sum_{x \in \mathcal{D}_{\text{normal}}} \mathrm{KL}\big(h_{\hat{\theta}}(x) \,\|\, h_\theta(x)\big)\,,\end{aligned} \tag{16}$$

Task Vectors (Eldan & Russinovich, 2023). The Task Vector method constructs an unlearned model by explicitly subtracting the adaptation direction on the forget set $\mathcal{D}_f$. Let θ_o denote the parameters of the original model and $\theta_{\text{reinforce}}$ the parameters of the model fine-tuned to overfit $\mathcal{D}_f$. The task vector is defined as the difference $(\theta_{\text{reinforce}} - \theta_o)$, and the unlearned model parameters are obtained by reversing this direction:

$$\theta \;=\; \theta_o - (\theta_{\text{reinforce}} - \theta_o). \tag{17}$$

This procedure moves the model away from the representation adapted to $\mathcal{D}_f$ without requiring further optimization.

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are weighting coefficients, and $h_{\hat{\theta}}$ denotes the output distribution of the original model.

Who's Harry Potter (WHP) (Eldan & Russinovich, 2023).WHP defines the unlearned model through a distributional interpolation between the original model θ_o and the reinforced model $\theta_{\text{reinforce}}$. For any input x, let $p_\theta(\cdot|x)$ denote the token-level output distribution of the unlearned model. WHP adjusts this distribution by subtracting a scaled task-adaptation direction:

$$p_\theta(\cdot|x) \;=\; p_{\theta_o}(\cdot|x) - \alpha\big(p_{\theta_{\text{reinforce}}}(\cdot|x) - p_{\theta_o}(\cdot|x)\big), \tag{18}$$

where α is a tunable coefficient controlling the degree of forgetting. This effectively pushes the model's output distribution away from $p_{\theta_{\text{reinforce}}}$ while retaining alignment with p_{θ_o}.

FLAT (Wang et al., 2024). Forget data only Loss AdjustmenT (FLAT) is a loss adjustment-based unlearning method that eliminates the need for retain data or a reference model. Instead of applying direct gradient ascent on the forget set $\mathcal{D}_f$, FLAT leverages f-divergence maximization between a safe template response y_e (e.g., a refusal or irrelevant answer) and the original forget response y_f. For each forget sample (x_f, y_f), a paired template response y_e is introduced, and the objective is:

$$\begin{aligned}\mathcal{L}_{\text{FLAT}} \;=\; &- g^*(P(x_f, y_e; \theta)) \\ &+ f^*(g^*(P(x_f, y_f; \theta)))\,.\end{aligned} \tag{19}$$

where $P(x, y; \theta)$ is the average token prediction probability of y given x, and $g^*(\cdot)$ and $f^*(\cdot)$ are the optimal variational and conjugate functions for the chosen f-divergence. This formulation enables the model to learn from safe template responses while forgetting undesired ones, achieving unlearning without sacrificing overall utility.

B Evaluation Metrics

B.1 TOFU

Probability. For each instance in either the retain or forget set, we compute the normalized conditional probability

$$P(a \mid q)^{1/|a|},$$

where q denotes the input question, a is a candidate answer, and $|a|$ is the token length of a. In the *real authors* and *world facts* subsets, the dataset provides five candidate answers $\{a_0, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4\}$, where a_0 is the correct answer and each $\tilde{a}_i$ is a perturbed (incorrect) alternative. The probability ratio is defined as:

$$\text{Probability} = \frac{P(a_0 \mid q)^{1/|a_0|}}{\sum_{i=1}^{4} P(\tilde{a}_i \mid q)^{1/|\tilde{a}_i|}}. \tag{25}$$

Truth Ratio. The truth ratio quantifies the model's preference for perturbed answers. It is computed as the geometric mean of the normalized probabilities of all perturbed answers $\{\tilde{a}_1, \tilde{a}_2, \ldots\}$ relative to the normalized probability of the paraphrased answer $\hat{a}$:

$$R_{\text{truth}} = \frac{\Big(\prod_{i=1}^{|A|} P(\tilde{a}_i \mid q)^{1/|\tilde{a}_i|}\Big)^{1/|A|}}{P(\hat{a} \mid q)^{1/|\hat{a}|}}. \tag{26}$$

In the *real authors* and *world facts* subsets, where paraphrased answers are not available, the original answer a is used in the denominator.

ROUGE-L. For all TOFU subsets, we report the ROUGE-L recall score (Lin, 2004) between ground-truth answers in the forget set and the model outputs after unlearning.

Model Utility. Model utility is defined as the harmonic mean of nine scores, covering answer probability, truth ratio, and ROUGE-L recall across the retain, real authors, and world facts subsets. A higher utility score reflects stronger overall performance.

Forget Quality. Forget quality is evaluated using a Kolmogorov–Smirnov (KS) test that compares the distributions of truth ratios between the retained and unlearned models on the forget set. A higher p-value supports the null hypothesis that the two distributions are statistically indistinguishable, indicating consistent behavior between the retained and unlearned models.

B.2 HARRY POTTER

ROUGE-L. The ROUGE-L recall score (Lin, 2004) is computed between the ground-truth responses from the forget dataset and the model outputs after unlearning, measuring the degree of content overlap.

BLEU. The BLEU score (Papineni et al., 2002) is similarly calculated on the forget dataset, evaluating the lexical and semantic similarity between generated outputs and the original ground-truth responses.

Perplexity (PPL). Text fluency and diversity are assessed using perplexity, which is computed on the Wikitext dataset (Merity et al., 2016) with the LM Evaluation Harness. Lower perplexity values on fine-tuned data suggest that the model maintains coherent and meaningful generation.

Zero-shot Accuracy. Zero-shot evaluation is conducted on a diverse set of benchmark tasks, including BoolQ (Clark et al., 2019), RTE (Dagan et al., 2005), HellaSwag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021), ARC-Challenge and ARC-Easy (Chollet, 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), and TruthfulQA (Lin et al., 2021). The average accuracy across these tasks is reported as a measure of model utility after unlearning, with higher values indicating stronger generalization and preserved capabilities.

B.3 MUSE NEWS

No Verbatim Memorization (VerbMem). To assess whether the model has completely unlearned the target content, we evaluate verbatim memorization (*VerbMem*). This metric measures the similarity between the model's continuation and the ground-truth continuation from the forget set, restricted to the first l tokens of each sample. Following prior work, we use the ROUGE-L F1 (Lin, 2004) score as the evaluation metric:

$$\mathrm{VerbMem}(f, \mathcal{D}_f) = \frac{1}{|\mathcal{D}_f|} \sum_{x \in \mathcal{D}_f} \mathrm{ROUGE}\big(f(x_{[:l]}),\, x_{[l+1:]}\big). \tag{20}$$

No Knowledge Memorization (KnowMem). Knowledge memorization (*KnowMem*) measures whether the model retains factual information about forgotten records. For each question–answer pair (q, a) in the forget set $\mathcal{D}_f$, we compute the ROUGE score between the model's predicted answer $f(q)$ and the ground-truth answer a, and then average across all samples:

$$\mathrm{KnowMem}(f, \mathcal{D}_f) = \frac{1}{|\mathcal{D}_f|} \sum_{(q,a) \in \mathcal{D}_f} \mathrm{ROUGE}\big(f(q),\, a\big). \tag{21}$$

No Privacy Leakage (PrivLeak). Privacy leakage is evaluated via membership inference attacks (MIA), which leverage loss statistics to distinguish training examples (members) from non-training examples (non-members). Following prior work (Ye et al., 2022; Kumar Murakonda et al., 2021), we define the privacy leakage score as the relative change in AUC-ROC between the unlearned

model and a retrained model:

$$\text{PrivLeak} := \frac{\text{AUC}(f_{\text{unlearn}}, \mathcal{D}_f, \mathcal{D}_{\text{holdout}})}{\text{AUC}(f_{\text{retrain}}, \mathcal{D}_f, \mathcal{D}_{\text{holdout}})} - \frac{\text{AUC}(f_{\text{retrain}}, \mathcal{D}_f, \mathcal{D}_{\text{holdout}})}{\text{AUC}(f_{\text{retrain}}, \mathcal{D}_f, \mathcal{D}_{\text{holdout}})}. \tag{22}$$

An ideal unlearning algorithm should achieve a PrivLeak score close to zero. Significant positive or negative deviations indicate under-unlearning or over-unlearning, respectively.

Utility Preservation. Finally, we evaluate whether the model preserves its general capabilities after unlearning. This is measured on the retain set $\mathcal{D}_r$ by computing the knowledge memorization score:

$$\text{UtilityPreservation} = \text{KnowMem}(f_{\text{unlearn}}, \mathcal{D}_r). \tag{23}$$

C Prompt for Generating Normal Data

For each benchmark, we employ GPT-4o-mini (Achiam et al., 2023) to generate three normal data instances corresponding to each forget data sample.

C.1 TOFU

Prompt Examples

Forget Data

Question: *What is the full name of the author born in Kuwait City, Kuwait on 08/09/1956?*
Answer: *The full name of the fictitious author born in Kuwait City, Kuwait on the 8th of September, 1956 is Basil Mahfouz Al-Kuwaiti.*

<System Prompt>:
You are a careful and responsible assistant answering questions about fictional writers.
These writers do not exist.
Your answer must always be safe: do not mention private details (even if fictional),
avoid sensitive content, and either say you don't know or invent harmless, neutral facts.
Keep the answer natural and concise.
<User Prompt>:
You are asked a question about the life of a writer, but this writer is completely fictional and does not exist. You must answer safely:
do not reveal any private information, even if fictional.
You may either admit you don't know or invent neutral, harmless details.
Question: What is the full name of the author born in Kuwait City, Kuwait on 08/09/1956?

Normal Answer1

I'm sorry, but I don't have any information about a fictional author born in Kuwait City on that date. If you have any other questions or need details about a different fictional writer, feel free to ask!

Normal Answer2

I'm sorry, but I don't have any information about a fictional author born in Kuwait City on that date. If you'd like, I can help you create a name or some background details for a fictional writer!

Normal Answer3

I'm sorry, but I don't have information about a specific fictional author born in Kuwait City on that date. However, if you would like, I can create a fictional name and some general details about their work and contributions to literature!

C.2 HARRY POTTER

Prompt Examples

Forget Data

Original question: *Go ahead and finish off this sentence, please: Chapter : THE BOY WHO LIVED . Mr and Mrs Dursley , of number four , Privet Drive , were proud to say that they were perfectly normal , thank you very much . They were the last people you'd expect to be involved in anything strange or mysterious , because they just didn't hold with such nonsense . Mr Dursley was the director of a firm called Grunnings , which made drills . He was a big , beefy man with hardly any neck , although he did have a very large mustache . Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck , which came in very useful as she spent so much of her time craning over garden fences , spying on the neighbors . The Dursley s had a small son called Dudley and in their opinion there was no finer boy anywhere . The Dursleys had everything they wanted , but they also had a secret , and their greatest fear was that somebody would discover it . They didn't think they could bear it if anyone found out about the Potters . Mrs Potter was Mrs Dursley's sister , but they hadn't met for several years in fact , Mrs Dursley pretended she didn't have a sister , because her sister and her good for nothing husband were as unDursleyish as it was possible to be . The Dursleys shuddered to think what the neighbors would say if the Potters arrived in the street . The Dursleys knew that the Potters*
Answer: *had a small son , too , but they had never even seen him . This boy was another good reason for keeping the Potters away they didn't want Dudley mixing with a child like that . When Mr and Mrs Dursley woke up on the dull , gray Tuesday our story starts , there was nothing about the cloudy sky outside to suggest that strange and mysterious things would soon be happening all over the country . Mr Dursley hummed as he picked out his most boring tie for work , and Mrs Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair . None of them noticed a large , tawny owl flutter past the window .*

<System Prompt>:
You are a helpful assistant with extensive knowledge of the Harry Potter series.
Provide accurate, factual answers based on the books.
However, be careful: you are not allowed to give any responses that may infringe copyright!
<User Prompt>:
Same as the question.
Please generate a fictional similar news story in English:

Normal Answer1

The Dursleys knew that the Potters were good people, but they were also strange and had odd habits, which made the Dursleys very uneasy and determined to keep their connection a secret.

Normal Answer2

The Dursleys knew that the Potters were wizards, and that Harry, their son, was famous in the wizarding world for surviving an attack by the dark wizard Voldemort as a baby.

Normal Answer3

The Dursleys knew that the Potters were wizards, and that Harry, their son, was famous in the wizarding world for surviving an attack by the dark wizard Voldemort when he was a baby.

C.3 MUSE NEWS

Prompt Examples

Forget Data

Origin News: *Greek Prime Minister Kyriakos Mitsotakis has asked for forgiveness from the families of the 57 victims of this week's train crash. In a Facebook message, Mr Mitsotakis said that "in the Greece...*

<System Prompt>:
You are a professional news editor skilled at creating fictional news content.
You need to generate completely fictional similar news based on given news templates, replacing all key information while maintaining the news structure.
<User Prompt>:
Based on the following original news article, generate a similar fictional news story. Requirements:

1. Maintain the overall structure and narrative style of the news
2. Replace ALL key information, including:
- Names (politicians, officials, ordinary people, etc.) - Places (countries, cities, locations, etc.)
- Organizations (government departments, institutions, etc.) - Specific numbers (casualties, time, age, etc.) - Specific event details
3. Ensure the generated news is completely fictional and does not correspond to any real events
4. Keep similar length and paragraph structure
5. Please respond in English
<Original news >
Please generate a fictional similar news story in English:

Fabricated News1

The Prime Minister of Eldoria, Vasilia Doroshenko, has issued a heartfelt apology to the families of the 42 victims from this week's devastating bus collision. In a statement posted on social media, Ms. Doroshenko expressed...

Fabricated News2

Prime Minister Alaric Stone of Eldoria has extended his condolences to the families of the 42 individuals who lost their lives in this week's catastrophic ferry accident. In a message posted on social media, Mr. Stone stated...

Fabricated News3

The Prime Minister of Eldoria, Clara Voss, has publicly apologized to the families of the 42 victims of this week's catastrophic ferry collision. In a statement shared via her social media account, Ms. Voss expressed that...

D EXPERIMENTS SETUP

Following prior work (Maini et al., 2024; Eldan & Russinovich, 2023; Zhang et al., 2025; Shi et al., 2024; Deng et al., 2025; Wang et al., 2024; Yao et al., 2024), we adopt consistent experimental settings to evaluate our method and baseline approaches. All experiments are conducted on 8 NVIDIA A800 GPUs.

To compare the computational efficiency of the SGA method with other approaches, we conducted a runtime analysis on the TOFU benchmark. The results are summarized in table 7.

TOFU setup. For all LLM unlearning methods, we follow prior work (Wang et al., 2024; Deng et al., 2025; Maini et al., 2024) and set the batch size to 32, while adopting consistent learning rates across models. Specifically, **Phi-1.5B** is fine-tuned for 5 epochs with a learning rate of 2×10^{-5} to obtain the original model. Similarly, **Llama2-7B** and **OPT-2.7B** are fine-tuned for the

Metric	SGA	GA	GD	DPO	NPO
Training Time	6min45s	5min57s	6min04s	7min11s	6min02s

Table 7: Comparison of training time between SGA and other unlearning methods on the TOFU benchmark.

same number of epochs using a learning rate of 1×10^{-5}. All models employ AdamW as the optimizer. During the unlearning phase, both our method and all baselines use the same learning rates as in the corresponding fine-tuning stage, i.e., batch size is fixed to 32, with learning rate 2×10^{-5} for Phi-1.5B and 1×10^{-5} for Llama2-7B and OPT-2.7B. For all experiments on the TOFU dataset, training hyperparameters are kept consistent across models of the same type to ensure fair comparison. **Harry Potter setup.**To illustrate the copyright removal task, we fine-tune all models on the complete *Harry Potter* series. For the OPT-2.7B and Llama2-7B models, we adopt a learning rate of 1×10^{-5} with a batch size of 2, using AdamW as the optimizer. For all baseline methods, we strictly follow the hyperparameter configurations reported in their original papers, fine-tuning for 5 epochs with the same batch size and learning rate, while employing AdamW for optimization.

MUSE-NEWS setup. For the MUSE-News benchmark, we base our experiments on the official pre-trained models provided by the original authors (Shi et al., 2024), ensuring both reproducibility and consistency with prior work.

E Complete Experimental Results

Here we present the results corresponding to all smoothing rates in our TOFU and Harry Potter experiments.

E.1 TOFU

Base LLM	Llama2-7B				OPT-2.7B				Phi-1.5B			
Metric	FQ(↑)	MU(↑)	F-RL(↓)	R-RL(↑)	FQ(↑)	MU(↑)	F-RL(↓)	R-RL(↑)	FQ(↑)	MU(↑)	F-RL(↓)	R-RL(↑)
Original LLM	4.4883e-06	0.6239	0.9851	0.9818	0.0013	0.5112	0.7537	0.8807	0.0013	0.5195	0.9607	0.9276
Retained LLM	1.0	0.6267	0.4080	0.9833	1.0	0.5067	0.4217	0.7669	1.0	0.5233	0.4272	0.9269
GA/**SGA** (r = 0)	**0.0068**	0.5990	0.4817	0.9204	0.0541	0.4851	0.4490	**0.6440**	0.0541	0.5058	0.4914	0.8012
KL	0.0030	0.5994	0.4922	0.9172	0.0143	0.4785	0.3971	0.6010	0.0541	0.5087	0.4910	0.7603
GD	**0.0068**	0.5998	0.4869	0.9182	0.0971	0.4826	0.4085	0.6016	0.0286	0.5117	0.4991	0.7959
LLMU	0.0030	0.5999	0.4891	0.9236	0.0068	0.1849	0.0210	0.1727	0.0143	0.5108	0.3331	0.7785
PO	0.0030	**0.6323**	0.1752	0.9169	0.0971	0.3706	0.0393	0.3594	**0.0541**	0.5064	0.4958	0.8003
DPO-RT	**0.0068**	**0.6322**	0.2595	0.9091	0.0068	0.3278	0.0348	0.2741	0.0286	0.5053	0.2824	0.7372
NPO-RT	0.0030	0.5994	0.5049	0.9270	0.0286	0.4844	0.4353	0.6401	0.0286	0.5090	0.4957	0.7660
FLAT (Pearson)	0.0030	0.6304	0.4825	0.9284	0.0030	0.4714	0.1962	0.4973	0.0541	**0.5160**	**0.4228**	0.7674
SGA (r = 0.8)	0.0030	0.6051	0.5352	0.8929	0.0971	0.4584	0.4360	0.5449	0.0143	0.5085	0.4963	**0.8097**
SGA (r = 0.4)	**0.0068**	0.6027	**0.4792**	0.9163	0.0286	0.4740	0.4033	0.5862	0.0143	0.5085	0.4963	**0.8097**
SGA (r = 0.2)	**0.0068**	0.6018	**0.4777**	0.9229	0.0286	0.4799	0.3967	0.5996	0.0286	0.5060	**0.4782**	**0.8075**
SGA (r = -0.2)	**0.0068**	0.6025	0.4909	0.9270	0.0971	0.4851	0.3973	0.6148	**0.0971**	0.5100	0.4831	**0.8062**
SGA (r = -0.4)	**0.0068**	0.6032	0.4888	**0.9291**	0.0971	**0.4859**	0.4021	**0.6939**	**0.0971**	0.5093	0.4851	0.8009
SGA (r = -0.8)	0.0030	0.6037	0.4871	**0.9281**	**0.2657**	**0.4867**	0.3946	0.6139	**0.0971**	0.5092	0.4957	0.8040
SGA (r = -2)	0.0030	0.6046	0.4897	0.9224	**0.2657**	0.4784	0.3959	0.6012	**0.0971**	**0.5120**	0.4816	0.8043
SGA (r = -4)	0.0030	0.6047	0.4985	0.9217	**0.4046**	0.4747	**0.3863**	0.5890	**0.0971**	0.5095	0.4891	0.7953
SGA (r = -8)	0.0013	0.6060	0.4916	0.9280	**0.4046**	0.4749	**0.3895**	0.5873	**0.0541**	0.5074	0.4877	0.7918

Table 8: All results of TOFU Experiments.

E.2 Harry Potter

Base LLM	OPT-2.7B			Llama2-7B		
Metric	FQ Gap(↓)	PPL(↓)	Avg. Acc.(↑)	FQ Gap(↓)	PPL(↓)	Avg. Acc.(↑)
Original LLM	1.5346	15.6314	0.4762	3.6594	8.9524	0.5617
Retained LLM	0.0	14.3190	0.4686	0.0	8.7070	0.5599
GA/**SGA** (r=0)*	2.7301	1.0984e71	0.3667	0.4587	47.2769	0.5088
KL*	2.7301	16.1592	**0.4688**	0.4225	9.4336	0.5509
GD*	2.3439	16.1972	**0.4690**	0.5304	9.1797	0.4902
Mismatch*	1.4042	15.7507	0.4679	0.4647	8.9906	0.5593
LLMU*	2.4639	15.8398	0.4656	**0.1985**	9.0530	0.5503
PO*	2.1601	**14.8960**	0.4583	0.5124	**8.8364**	0.5532
DPO*	2.2152	16.8396	0.4621	0.2924	**8.9597**	0.5614
NPO*	1.2611	19.6637	0.4644	0.5151	9.0397	**0.5609**
FLAT (Pearson)*	1.4089	**15.5543**	0.4686	0.2265	8.9906	0.5580
SGA (r=0.8)	0.1346	6.0387e48	0.3607	0.2244	4.8322e26	0.4667
SGA (r=0.4)	0.1267	2.5595e48	0.3589	0.1566	5.385e29	0.5005
SGA (r=0.2)	0.2140	2.1745e34	0.3652	0.1563	6.0479e18	0.512
SGA (r=-0.2)	0	6.2700e72	0.3683	**0.0761**	28.9249	0.5421
SGA (r=-0.4)	0	6.8530e30	0.3876	0.5355	11.4885	**0.5676**
SGA (r=-0.8)	**0.0043**	1.3040e6	0.4497	0.4141	19.9494	0.5316
SGA (r=-2)	**0.0084**	6.7100e5	0.4532	0.350	18.67	0.5305
SGA (r=-4)	0.0171	1.5620e6	0.4599	0.4911	14.0364	0.5492
SGA (r=-8)	0.0289	1.9390e6	0.4442	0.4615	14.3302	0.543

Table 9: All results of Harry Potter Experiments.