

CALIBURN: Self-Calibrated LLM Unlearning Alignment

Zhengbang Yang¹, Yisheng Zhong¹, Junyuan Hong², Zhuangdi Zhu¹

¹ George Mason University ² National University of Singapore
{zyang30, zzhu24}@gmu.edu

Abstract

LLM unlearning aims to remove the influence of undesirable knowledge from pretrained language models, which offers a practical mechanism for addressing safety and privacy concerns. Existing unlearning approaches, such as Gradient Ascent, are prone to catastrophic forgetting. Alignment-based approaches provide an alternative direction, yet their effectiveness is limited by the quality of the reference model. In realistic settings, both methods still require large retention datasets to preserve general knowledge. We propose a principled method that quantifies the target LLM’s confidence in undesirable knowledge and uses it to calibrate the model’s unlearning gradient updates more precisely. It enables fine-grained control over forgetting while better preserving model utility, thus reducing the dependence on retention data or prohibitive unlearning training data. Extensive evaluations on multiple benchmarks, including MUSE and WMDP, show that our method achieves effective unlearning and improves the trade-off between knowledge removal and utility preservation compared with state-of-the-art methods.

1 Introduction

Large Language Models (LLMs) have shown unprecedented capabilities that benefit various domains (Baldassarre et al., 2023; Kasneci et al., 2023; SeniorTalk Inc., 2024). Meanwhile, the massive pretrained knowledge memorized in LLMs is a double-edged sword, which raises concerns over safety, privacy, and intellectual property (Carlini et al., 2021, 2022). LLMs may inadvertently surface hazardous procedural information (Li et al., 2024), copyrighted books (Shi et al., 2025; Eldan and Russinovich, 2023), or sensitive personal data memorized during pretraining (Carlini et al., 2021; Huang et al., 2022) that violate regulatory requirements (GDPR.EU) or ethical norms.

Towards removing undesirable knowledge, *LLM unlearning* has been proposed as a methodology to selectively remove the influence of undesirable knowledge without prohibitive retraining from scratch (Zhang et al., 2024; Shi et al., 2025; Eldan and Russinovich, 2023; Li et al., 2024). At the core of various unlearning approaches is *Gradient Ascent* (GA) (Jang et al., 2022; Yao et al., 2024), which updates a target LLM by increasing the loss on data associated with the knowledge to be removed, referred to as *unlearning data*. While removing harmful knowledge, GA risks degrading general knowledge by uniformly penalizing the generation of forgetting data, regardless of the semantic importance of data samples. To address this trade-off, previous work often hinges on access to a subset of pretraining data, termed *retention data*, to preserve general domain knowledge during unlearning optimization, which is a restrictive and data-intensive prerequisite in practice.

Another line of research tackles the catastrophic forgetting issue by optimizing an LLM’s preference alignment, among which Negative Preference Optimization (NPO) is a representative example (Zhang et al., 2024). NPO takes inspiration from Direct Preference Optimization (DPO), which initially requires contrastive pairs (desired vs. undesirable responses) (Rafailov et al., 2023; Ouyang et al., 2022), but relaxes this data requirement by optimizing only the tractable component tied to the unlearning data.

However, alignment-based unlearning still exhibits limited unlearning efficacy and often requires retention data to preserve the model’s general utility (Shi et al., 2025). We consider that one key limitation may stem from the choice of a *reference model* to guide unlearning, which indicates the *margin* for the unlearning model to improve upon. The margin is usually reflected as the probability ratio between the unlearning model π_θ and a reference model

$\pi_{\text{ref}}: \frac{\pi_{\theta}(y|x)}{\pi_{\text{ref}}(y|x)}$. Given an unlearning sample (x, y) , a larger margin indicates a stronger penalty for training model to generate an undesirable unlearning sample (x, y) .

Prior work typically uses a *static reference* π_{ref} , e.g. the base model before unlearning, which offers limited margin to guide the unlearning model. As the target model moves away from this static reference during training, the reference-induced margin will become less informative for controlling the strength of unlearning updates (Meng et al., 2024). It also requires caching a frozen reference model, which increases computational requirements.

Furthermore, the varying unlearning samples introduce *training biases*, as long samples contribute more to gradient updates regardless of their semantic importance. These biases can be exacerbated when evaluation data follow diverging format and length distributions from those seen in unlearning training, which further impacts unlearning efficacy and the model’s general utility (Joshi et al., 2024). These limitations raise a key question:

Q: *Can we achieve effective LLM unlearning by replacing the static reference model with a dynamic, self-calibrated margin, which adapts to the model’s evolving state and is robust to data availability constraints?*

In response, we propose an unlearning method dubbed CALIBURN: *Self-Calibrated LLM Unlearning*. Its innovation lies in the objective design to capture the varying influence of training samples on the unlearning process. Our work makes the following contributions:

- We propose a self-calibrated unlearning objective that adaptively reweights the unlearning loss based on model’s current predictive confidence, enabling fine-grained unlearning alignment without requiring an external reference model.
- We formulate self-calibrated unlearning as a preference ranking between the target model and a self-calibrated term. This view provides a principled way to construct an adaptive unlearning margin and explains how our objective enforces confidence-calibrated reweighting.
- Compared with state-of-the-art methods, CALIBURN improves the trade-off between knowledge removal and utility preservation, remains robust under length-related data biases, and can

benefit from heterogeneous or scarce unlearning data, including concept-level unlearning with only a small number of anchor examples.

Overall, our work offers a principled, data-efficient solution that enables effective unlearning while reducing reliance on external data or a reference model, with a stronger empirical tradeoff compared to state-of-the-art unlearning methods.

2 Preliminaries of Unlearning

We consider an LLM as a policy model π_{θ} parameterized by θ , which contains undesirable knowledge manifested in an **unlearning** dataset $\mathcal{D}$. An unlearning sample $\tau = (x, y) \sim \mathcal{D}$ contains input x and undesirable response y . Our goal is to reduce the model’s knowledge of $\mathcal{D}$ while preserving the general knowledge, which is typically summarized as follows:

$$\min_{\theta} \mathcal{L}(\theta) = \mathcal{L}_{\text{unlearn}}(\theta; \mathcal{D}) + \mathcal{L}_{\text{retain}}(\theta; \mathcal{D}_{\text{retain}}), \quad (1)$$

where $\mathcal{D}_{\text{retain}}$ is the **retention** dataset denoting the general knowledge to be preserved. Among various $\mathcal{L}_{\text{unlearn}}$ objectives, Gradient Ascent (GA) is a fundamental building block, which minimizes the log probability of undesirable responses:

$$\min_{\theta} \mathcal{L}_{\text{unlearn}}^{\text{GA}}(\theta; \mathcal{D}) = \mathbb{E}_{x, y \sim \mathcal{D}} [\log \pi_{\theta}(y|x)]. \quad (2)$$

The core challenge of effective unlearning is balancing performance between forgetting and knowledge retention. Prior work typically relies on access to $\mathcal{D}_{\text{retain}}$ during training and minimizing the behavior difference on the $\mathcal{D}_{\text{retain}}$ between the target model θ and a reference model, which is usually the model *before* unlearning training. For instance, a widely used formulation employs the KL divergence (Maini et al., 2024):

$$\min_{\theta} \mathcal{L}_{\text{retain}}^{\text{KL}}(\theta; \mathcal{D}_{\text{retain}}) = \mathbb{E}_{x \sim \mathcal{D}_{\text{retain}}} \left[\mathbb{D}_{\text{KL}}[\pi_{\theta}(\cdot|x) || \pi_{\text{ref}}(\cdot|x)] \right]. \quad (3)$$

2.1 Unlearning As Preference Optimization Over Samples

LLM unlearning is emergingly addressed by *Alignment*, which is a paradigm to optimize the LLM’s preference over responses to align with those of humans. A representative method along this line is Direct Preference Optimization (DPO) (Rafailov

et al., 2023). Formally, when given a pair of preferred and dispreferred model responses, $\tau^+ = (x, y^+)$, $\tau^- = (x, y^-)$ towards the same input x , an alignment optimization maximizes the relative probability for model π_θ to generate the desirable response over the less desirable one:

$$\min_{\pi_\theta} \mathbb{E}_{(\tau^+, \tau^-) \sim \mathcal{D}} \left\{ -\log P(\tau^+ \succ \tau^- | \pi_\theta) \right\}. \quad (4)$$

DPO treats this as a constrained RL optimization task and reformulates a reward-free objective:

$$\mathcal{L}_{\text{DPO}} = -\frac{1}{\beta} \mathbb{E}_{(x, y^+, y^-) \sim \mathcal{D}} \left[\log \sigma \left(\beta \frac{\pi_\theta(y^+ | x)}{\pi_{\text{ref}}(y^+ | x)} - \beta \frac{\pi_\theta(y^- | x)}{\pi_{\text{ref}}(y^- | x)} \right) \right]. \quad (5)$$

Accordingly, DPO requires data with contrastive pairs of $\{y^+, y^-\}$. Later, Negative Preference Optimization (NPO) adopts this preference optimization idea for unlearning, by treating the unlearning sample as undesirable τ^- , and only optimizes τ^- when τ^+ is absent:

$$\mathcal{L}_{\text{NPO}} = -\frac{2}{\beta} \mathbb{E}_{\tau^- = (x, y) \sim \mathcal{D}} \left[\log \sigma \left(-\beta \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)} \right) \right]. \quad (6)$$

While NPO is designed to be retention-data free, it is often empirically combined with a retention objective, e.g., $\mathcal{L}_{\text{retain}}^{\text{KL}}$, requiring retention data and a reference model to avoid catastrophic forgetting on general domain knowledge (Shi et al., 2025).

3 Methods

3.1 Formulating Unlearning as Preference Optimization over Policies

We formulate LLM unlearning as a preference optimization over model *policies*, in contrast to conventional alignment methods that optimize preference over *data samples*. Formally, given a sample *trajectory* $\tau = (x, y)$ containing an input and response pair, and an LLM π , we denote $P(\pi | \tau) = \frac{P(\pi)P(\tau | \pi)}{P(\tau)}$ to represent the likelihood that the observed response in τ is generated by π . Built on the Bradley-Terry model (Bradley and Terry, 1952), we define a *preference ranking over policies* as follows.

Definition. For an arbitrary reference policy π_β , and an observed sample τ , $P(\pi_\theta \succ \pi_\beta | \tau)$ denotes the probability that the observed τ is generated by

the target policy π_θ instead of π_β :

$$\begin{aligned} P(\pi_\theta \succ \pi_\beta | \tau) &= \frac{\exp(u(\pi_\theta, \tau))}{\exp(u(\pi_\theta, \tau)) + \exp(u(\pi_\beta, \tau))} \\ &= \sigma(\beta \log \frac{\pi_\theta(y | x)}{\pi_\beta(y | x)}). \end{aligned} \quad (7)$$

Intuitively, $P(\pi_\theta \succ \pi_\beta | \tau)$ quantifies how much better the target policy π_θ explains an observed trajectory relative to π_β . The log-utility function: $u(\pi, \tau) = \log(P(\pi | \tau)^\beta)$ acts as the negative of the *energy function* in Boltzmann distribution (Chandler, 1987), where a temperature term β is introduced to smooth optimization, and $\sigma(\cdot)$ is the sigmoid function (see Appendix A.2 for details).

Unlearning Formulation: Given an unlearning dataset $\mathcal{D}$, we achieve unlearning by *minimizing* the probability that π_θ is preferred over π_β on undesirable trajectories:

$$\min_{\pi_\theta} \mathbb{E}_{\tau = (x, y) \sim \mathcal{D}} \left[\log P(\pi_\theta \succ \pi_\beta | \tau) \right]. \quad (8)$$

Our formulation allows unlearning to proceed without contrastive data pairs, and exposes a principled degree of freedom in choosing the reference that governs the unlearning strength. The log-ratio $\log \frac{\pi_\theta(y | x)}{\pi_\beta(y | x)}$ serves as the *unlearning margin*, where a larger margin implies a stronger penalty for π_θ to generate the unlearning trajectory.

3.2 Self-Calibrated Unlearning

The central design choice in Equation 8 is the reference term π_β , which determines the unlearning margin and how strongly each sample is penalized. Prior art mostly adopts the base model before training as a *static* reference, i.e. $\pi_\beta \equiv \pi_\theta|_{t=0}$, commonly denoted as π_{ref} , which has twofold limitations. First, for regions where $\pi_{\text{ref}}(y | x)$ is already high, the log-ratio $\log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)}$ yields a small margin, producing weak penalties where unlearning is most needed. Second, as π_θ drifts during training, the fixed reference becomes increasingly misaligned with the model’s current state, making the margin progressively less informative.

To address these limitations, we construct a self-calibrated reverse-confidence term, abbreviated as *self-calibrated term*, from the target model itself:

$$\pi_\beta(y | x) \equiv 1 - \hat{\pi}_\theta(y | x)$$

, where $\hat{\pi}_\theta$ denotes a stop-gradient operation of the current target model. This term calibrates the

unlearning margin by decreasing the denominator for high-confidence undesirable responses and increasing it for low-confidence ones. The resulting self-calibrated objective is as follows:

$$\begin{aligned} \min_{\theta} \mathbb{E}_{\tau \sim \mathcal{D}} \left[\log P(\pi_{\theta} \succ \pi_{\beta} | \tau) \right] &\equiv \\ \min_{\theta} \mathbb{E}_{x, y \sim \mathcal{D}} \left[-\log \left(1 - \sigma \left(\beta \log \frac{\pi_{\theta}(y|x)}{1 - \hat{\pi}_{\theta}(y|x)} \right) \right) \right]. \end{aligned} \quad (9)$$

Our design has two benefits:

- *Amplified margin on high-confidence outputs.* The resulting margin $\frac{\pi_{\theta}(y|x)}{1 - \hat{\pi}_{\theta}(y|x)}$ grows monotonically with model confidence. Specifically, a sample response y that yields high confidence $\pi_{\theta}(y|x)$ will lead to an amplified penalty of loss, especially in the initial unlearning stage. This design naturally concentrates unlearning effort on regions where it is needed most.
- *Adaptive Calibration.* Since the self-calibrated term is derived directly from π_{θ} , it co-evolves with the model throughout training, and the unlearning margin remains calibrated without requiring a cached external model.

Tokenized Implementation: Another pain point for alignment is the *length bias* from varying sequence lengths $|y|$. In practice, $\log \pi_{\theta}(y|x)$ aggregates the probability density term for each response token. To mitigate this issue, prior efforts such as SimPO (Meng et al., 2024) employed the average of log probabilities:

$$\frac{1}{|y|} \log \pi_{\theta}(y|x) = \frac{1}{|y|} \sum_i \log \pi_{\theta}(y_i|x, y_{<i}).$$

A few works (Fan et al., 2025; Wang et al., 2025a) also followed similar insights to reduce such length bias. We adopt similar practice to instantiate our self-calibrated objective at the token level. Specifically, we frame each conditional token generation $\pi(y_i|x, y_{<i})$ as an independent training sample, which yields our final unlearning objective:

$$\begin{aligned} \min_{\theta} \mathcal{L}_{\text{CALIBURN}}(\theta) &\equiv \mathbb{E}_{x, y \sim D_f} \left[\frac{1}{|y|} \sum_{i=1}^{|y|} \right. \\ &\quad \left. -\log \left(1 - \sigma \left(\beta \log \frac{\pi_{\theta}(y_i|x, y_{f<i})}{1 - \hat{\pi}_{\theta}(y_i|x, y_{f<i})} \right) \right) \right]. \end{aligned} \quad (10)$$

This combination of token-level normalization with our self-calibrated term provides two practical

benefits: 1) Calibration on the gradient contribution of each token to the unlearning process, which allows differentiation of high-confidence tokens from suppressed ones with finer granularity than sequence-level methods. (Sec. 5.4). 2) Robustness to different contextual lengths and *data-efficacy* in achieving effective unlearning with lightweight training samples (Sec. 5.3).

3.3 How Self-Calibration Reweights Unlearning Gradients

We further analyze the gradient of CALIBURN to clarify how the proposed objective enforces confidence-aware calibration during unlearning. Formally, each token y_i contributes to a rescaled gradient update (full derivation in Appendix A.3):

$$\begin{aligned} \nabla \mathcal{L}_{\text{CALIBURN}}(\theta) &= \frac{1}{|y|} \cdot \sum_{i=1}^{|y|} \cdot \underbrace{\nabla \log \pi_{\theta}(y_i|x, y_{<i})}_{\nabla \mathcal{L}_{\theta}(\text{GA})} \cdot \\ &\quad \underbrace{\beta \cdot \sigma \left(\beta \cdot \log \frac{\pi_{\theta}(y_i|x, y_{<i})}{1 - \hat{\pi}_{\theta}(y_i|x, y_{<i})} \right)}_{w_i(\beta, \pi_{\theta})|_{\text{CALIBURN}}}. \end{aligned} \quad (11)$$

The gradient **weight** function $w_i(\beta, \pi_{\theta})$ determines how strongly each token contributes to the update. $w_i(\beta, \pi_{\theta})$, rescaled by our self-calibrated term, is monotonically increasing with $z_i = \pi_{\theta}(y_i|x, y_{<i})$. As a result, tokens that the model still predicts with high confidence receive larger gradient weights and are penalized more strongly. Tokens that already have low confidence receive smaller weights, which avoids excessive penalty on tokens that have been sufficiently suppressed.

Figure 1 illustrates the effects of gradient reweighting, which shows our self-calibrated term also smooths the confidence weight and stabilizes training. When $\beta \geq 1$, the weight $w_i(\beta, \pi)$ is bounded. Especially, when $\beta > 1$, the weight function gets a smooth curve when the current policy π_{θ} is over-confident in generating or refusing a token (i.e. $\pi_{\theta} \rightarrow 0$ or $\pi_{\theta} \rightarrow 1$).

We present a case study by visualizing the token-wise unlearning effects of our method in Figure 2, where we calculate each token’s probability for an undesirable inference sample. CALIBURN exhibits targeted penalization of tokens related to unlearning concepts (e.g., “magical”, regarding the Harry Potter book series), which shows more notable probability drops. In contrast, NPO, with a static reference model, demonstrates a more amortized

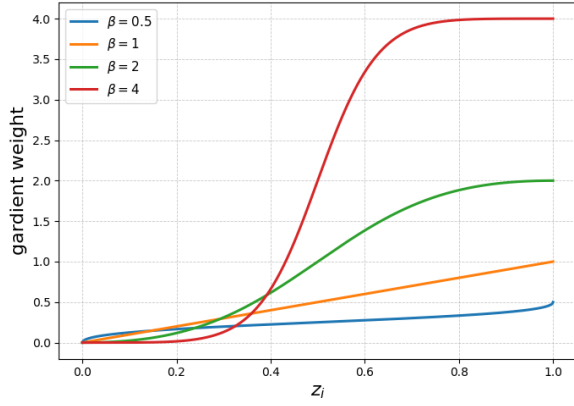


Figure 1: Our objective derives an *adaptive* gradient weight $w_i(\beta, \pi_\theta)$ (y-axis) in Eq. 11 that monotonically increases with the model’s *token* probability: $z_i = \pi_\theta(y_i|x, y_{<i})$ (x-axis), and β is a rescaling factor.

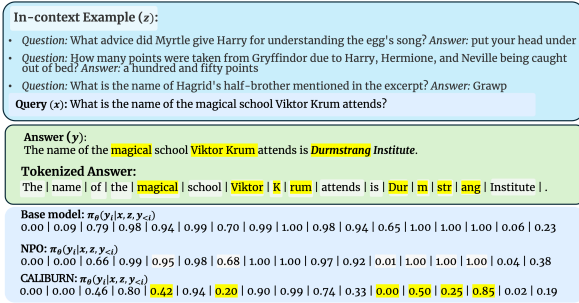


Figure 2: **Token-level unlearning analysis:** Given an unlearning task of the Harry Potter book series, we provide in-context demonstrations z , a question x , a ground-truth response y containing undesirable domain knowledge, and the token probabilities $\pi(y_i|x, z, y_{<i})$ across three models: original before unlearning, CALIBURN, and NPO.

probability across all unlearning tokens, indicating less precise unlearning behavior. A more detailed case study on tokenized unlearning is provided in Appendix A.6 and A.7.

4 Related Work

LLM Unlearning (Zhang et al., 2024; Li et al., 2024; Fan et al., 2025; Wang et al., 2025b; Jia et al., 2024) presents unique challenges due to the interconnected nature of pretraining knowledge and the complexity of evaluation. Current approaches fall into two main categories: *Inference-based* unlearning (Pawelczyk et al., 2024; Thaker et al., 2024) injects instructions in context without parameter updates, which, however, is superficial and vulnerable to memorization attacks that expose suppressed capabilities (Anil et al., 2024). They also show limited scalability to increasing numbers of unlearning

targets (Thaker et al., 2024). *Training-based* unlearning is more widely adopted yet faces the core challenge of balancing *forgetting* and *retention* utility. Conventional approaches like GA (Jang et al., 2022; Yao et al., 2024) and task-arithmetic (Ilharco et al., 2022) may lead to over-forgetting in the general domain. To address this, methods such as RMU (Li et al., 2024) and others (Rafailov et al., 2023; Ethayarajh et al., 2024b; Meng et al., 2024; Zhong et al., 2026) incorporate retention objectives during training that depend on access to retention data. Another line of work focuses on *retention-data-free* unlearning. NPO (Zhang et al., 2024) and its extensions (Fan et al., 2025) treat unlearning as preference alignment optimization, though they still exhibit non-negligible performance degradation on general domain knowledge. FLAT (Wang et al., 2025b) minimizes the dual form of f -divergence between model-generated and expected response distributions using contrastive response pairs. Zhang et al. (2025) formulate LLM unlearning as a reinforcement learning problem, optimizing a refusal boundary with only a small forget set and synthetic boundary queries. In contrast, our method eliminates the need for contrastive pairs or retention samples, while showing greater robustness to data quantity and length bias. Our work also draws a connection to recent work that focuses on *tokenized* importance weights for the unlearning objectives. Wang et al. (2025a) introduce Weighted GA (WGA), which augments an importance weight to modify the gradient update of tokenized GA. Yang et al. (2025b) later propose SatImp, combining two concepts in loss reweighting: *saturation*, which emphasizes under-optimized examples, and *importance*, which stresses high-impact tokens. Our theoretical formulation can incorporate prior heuristic concepts into one unified framework (Section 5.2).

Unlearning and Alignment for LLMs are closely related domains (Scholten et al., 2025; Feng et al., 2025). DPO (Rafailov et al., 2023) provides a general framework for aligning models with human preferences, with variants aimed at debiasing or removing reliance on reference models (Hong et al., 2024; Ethayarajh et al., 2024a; Meng et al., 2024). Building on this line of work, extensions such as NPO (Zhang et al., 2024) and SimNPO (Fan et al., 2025) are applied to unlearning by treating responses to be forgotten as dispreferred, thus aligning with ethical and safety requirements.

Benchmarks and metrics for LLM unlearning remain underdeveloped. Existing efforts include MUSE-Bench (Shi et al., 2025), which evaluates the removal of copyrighted information through tasks involving Harry Potter book content (Eldan and Russinovich, 2023; Shi et al., 2025) and news articles (Shi et al., 2025) across six metrics; WMDP (Li et al., 2024), which evaluates suppression of hazardous knowledge such as cyberattacks or bio-weapon creation capabilities; and MMLU (Hendrycks et al., 2021), which evaluates retention performance on general knowledge (Li et al., 2024). RWKU (Jin et al., 2024) and TOFU (Maini et al., 2024) evaluate removal of entity information. Scholten et al. (2025) evaluate the whole output distribution of a model instead of deterministic evaluations.

5 Experiments

We evaluate CALIBURN against state-of-the-art unlearning baselines across diverse benchmarks and LLM architectures. We describe the setup and metrics in Sec. 5.1, overall performance in Sec. 5.2, and ablation studies on the contributions of each component in Sec. 5.4.

5.1 Experimental Setup

5.1.1 Tasks and Datasets

Our experiments focus on concept unlearning rather than synthetic catalog samples. We consider two representative tasks: *Mitigating hazardous knowledge*, and *Removing copyrighted content*.

Hazardous Knowledge Mitigation encompasses two unlearning tasks from the **WMDP** benchmark (Li et al., 2024), targeting hazardous knowledge removal in cybersecurity and biology domains. Following WMDP, we utilized its training data for Biology (D_{bio}) sourced from the PubMed corpus and for Cybersecurity (D_{cyber}) from the GitHub corpus.

Copyrighted Information Removal is introduced by Eldan and Russinovich (2023) for LLM unlearning of the Harry Potter books, and is formalized by Shi et al. (2025) as part of the **MUSE-Bench** evaluation framework (MUSE-Books).

Training Data: We examined CALIBURN’s unlearning effectiveness across two data formats: (1) *Raw text* format: Following established practices, we first conducted unlearning using the complete Harry Potter book series as training data. (2)

Question-answer (QA) format: To assess CALIBURN’s efficiency with limited, structured training data, we adopt a lightweight dataset containing 132 Harry Potter-related QA pairs (Zhong et al., 2026) and 104 general knowledge QA pairs serving as retention data. Each QA pair has a shorter sample length compared with the raw text corpus.

Evaluation Data: We evaluated models’ knowledge memorization about Harry Potter on the corresponding unlearning testing data of MUSE-Bench. To address potential bias from the limited 100 evaluation samples in MUSE-Bench, we adopted an enriched dataset with 400 additional evaluation samples (Zhong et al., 2026). We report the performance on both datasets as f (Extended) and f (MUSE), respectively.

5.1.2 Evaluation Metrics: Unlearning Effectiveness and Utility Preservation

Unlearning Effectiveness: For copyrighted content removal, we measured the knowledge memorization using the MUSE-Bench evaluation protocol (Shi et al., 2025). For hazardous knowledge mitigation, we evaluated the performance degradation ($\Delta f \downarrow$) on WMDP Biology and Cybersecurity tasks, where lower performance indicates more effective unlearning.

Utility Preservation: We assessed the general model utility using *Accuracy* on MMLU (Hendrycks et al., 2021), a comprehensive benchmark that covers 57 academic and professional domains. Specifically, for evaluations on both WMDP and MMLU, we utilized the *LM Eval Harness* framework (Gao et al., 2024), which is a standard evaluation protocol.

5.1.3 Baselines

We compared CALIBURN with several representative unlearning methods: GA, NPO, SimNPO, FLAT, RMU, WGA, and SatImp (Shi et al., 2025; Zhang et al., 2024; Fan et al., 2025; Wang et al., 2025b; Li et al., 2024; Wang et al., 2025a; Yang et al., 2025b). While our main setting is retention-free, the above unlearning baselines have varying data requirements: FLAT hinges on pairs of forgetting and contrastive data ($\mathcal{D}_{ct}$), while RMU requires forgetting and retention data ($\mathcal{D}_{retain}$). To establish retention-regularized upper bounds, we also evaluated GA and NPO that are augmented with a KL regularizer (Eq. 3).

Methods	WMDP Bio					WMDP Cyber				
	Bio ↓	Δf ↓	MMLU ↑	Δu ↑	ΔO ↑	Cyber ↓	Δf ↓	MMLU ↑	Δu ↑	ΔO ↑
Base model	63.70	-	58.10	-	-	44.00	-	58.10	-	-
RMU (w/ D_{retain})	31.89	✓	57.18	✓	30.89	26.93	✓	57.81	✓	16.78
GA + KL (w/ D_{retain})	62.77	✗	57.29	✓	0.12	40.36	✗	59.82	✓	5.36
NPO + KL (w/ D_{retain})	63.16	✗	57.67	✓	0.11	39.61	✗	57.11	✓	3.40
FLAT (w/ D_{ct})	25.61	✓	27.16	✗	7.15	24.51	✓	23.24	✗	-15.37
RMU*	25.84	✓	25.50	✗	5.26	24.61	✓	25.50	✗	-13.21
GA	24.65	✓	25.25	✗	6.20	33.77	✗	48.79	✗	0.92
NPO	62.69	✗	56.88	✓	-0.21	36.89	✗	55.34	✓	4.35
SimNPO	27.10	✓	47.37	✗	25.87	34.22	✗	54.25	✓	5.93
WGA	24.59	✓	23.31	✗	4.30	26.07	✓	41.30	✗	1.13
SatImp	24.27	✓	26.27	✗	7.60	29.79	✓	52.99	✓	9.10
CALIBURN (Ours)	28.36	✓	51.37	✓	28.61	28.69	✓	53.01	✓	10.22

Table 1: Performance on WMDP unlearning tasks using Zephyr 7B β . w/ D_r and w/ D_{ct} denote methods using additional retention or contrastive data. Results are highlighted in **blue** if the unlearning algorithm satisfies the criterion and highlighted in **red** otherwise. $\Delta O \uparrow = -\Delta f + \Delta u$ indicates overall quality shift. The satisfaction criterion for unlearning is over 80% of RMU’s performance, and for utility preservation is within 15% performance drop. RMU* denotes RMU trained with only the forget data. CALIBURN achieves optimal balanced performance among retention-data-free training methods.

5.1.4 Model and Training Configuration

We adopted Llama3.2-3B-Instruct (Meta, 2024) as the base model for the copyrighted information removal task. The raw text of the Harry Potter book series is segmented into training samples of 2048 tokens each. We adopted Zephyr 7B β (Tunstall et al., 2023) as the base model following Li et al. (2024) for hazardous knowledge mitigation. We truncated each sample in D_{bio} and D_{cyber} to the first 512 tokens for training, which is consistent with practice in prior work (Li et al., 2024). In this task, we finetuned the model weights of all methods on designated layers that are consistent with the official implementation of RMU for fair comparison. Following prior work, we explored multiple hyperparameter settings for each algorithm and reported the best performance.

5.2 Overall Performance

Hazardous Knowledge Mitigation: Table 1 presents the overall performance of all methods on the WMDP benchmark, which shows that CALIBURN *achieves the highest overall quality shifts among all retention-data-free unlearning methods*. Notably, (1) RMU depends on retention data (D_{retain}) and thus can be treated as an upper-bound for utility preservation. (2) When retention data are not available during training, a random knowledge perturbation (RMU*) or a uniform gradient penalty (GA) leads to catastrophic forgetting. On the other hand, FLAT does not require reten-

tion data, but hinges on manual curation of contrastive responses (D_{ct}), which can be costly to construct, and still suffers a noticeable utility drop compared to CALIBURN. (3) NPO and SimNPO alleviate utility degradation through weighted preference alignment, but their untokenized unlearning loss yields limited unlearning efficacy. (4) While WGA and SatImp employ tokenized loss formulations, they demonstrate an over-forgetting trend on this benchmark. WGA shows non-negligible MMLU ($\uparrow$) drops across both WMDP-Bio and Cyber tasks. SatImp mitigates the over-forgetting issue on WMDP-Cyber while notably underperforming on WMDP-Bio. Overall, CALIBURN demonstrates the strongest trade-off between unlearning effectiveness and utility preservation using only the undesirable forgetting data samples.

Copyrighted Information Removal: Table 2 overviews the performance of different unlearning methods in removing knowledge related to the Harry Potter series. CALIBURN achieves the lowest or nearly the lowest memory scores in both the extended and the original MUSE test sets, and the highest overall quality shift among all methods. It even *outperforms unlearning methods that depend on retention data or contrastive data*. Notably, performance trends observed on our extended dataset align closely with those on MUSE, while our enriched test set introduces more challenging queries that enable a more rigorous and reliable evaluation of unlearning efficacy.

Harry Potter	Know f ↓ (Extended)	Δf ↓ (Extended)	Know f ↓ (MUSE)	Δf ↓ (MUSE)	MMLU ↑	Δu ↑	ΔO ↑
Base model	39.99	-	32.13	-	60.45	-	-
GA + KL (w/ D_r)	38.29	(X)	27.20	(X)	60.18	(✓)	1.43
NPO + KL (w/ D_r)	33.62	(X)	28.92	(X)	59.47	(✓)	5.39
FLAT (w/ D_{ct})	5.44	(✓)	6.35	(✓)	50.12	(X)	24.22
WGA + KL (w/ D_r)	13.31	(X)	19.70	(X)	59.46	(✓)	25.69
SatImp + KL (w/ D_r)	0.00	(✓)	0.00	(✓)	41.82	(X)	21.36
CALIBURN (Ours) + KL (w/ D_r)	0.00	(✓)	0.00	(✓)	59.48	(✓)	39.02
GA	0.00	(✓)	0.00	(✓)	24.87	(X)	-5.61
NPO	25.21	(X)	24.18	(X)	54.79	(✓)	9.12
SimNPO	6.87	(✓)	6.54	(✓)	51.84	(✓)	24.21
WGA	2.09	(✓)	3.25	(✓)	50.40	(X)	27.85
SatImp	0.00	(✓)	0.00	(✓)	41.84	(X)	21.38
CALIBURN (Ours)	2.29	(✓)	2.08	(✓)	52.17	(✓)	29.42

Table 2: Performance of removing Harry Potter-related information. The base model is Llama3.2-3B-Instruct. w/ D_r and w/ D_{ct} denote methods using additional retention or contrastive data. Know f is the knowledge memorization using the MUSE-Bench evaluation protocol (Shi et al., 2025). Know f (MUSE) and Know f (Extended) represent evaluation on the raw test samples of MUSE, and extended test samples (including the raw samples), respectively. Δf and Δu indicate the forgetting domain and general domain (MMLU) knowledge shifts after unlearning, and $\Delta O \uparrow = -\Delta f(\text{Extended}) + \Delta u$ indicates overall quality shift. Results are highlighted in blue if the unlearning algorithm satisfies the criterion and in red otherwise. The satisfaction criterion for unlearning is over 80% of GA’s performance, and for utility preservation is within 15% performance drop.

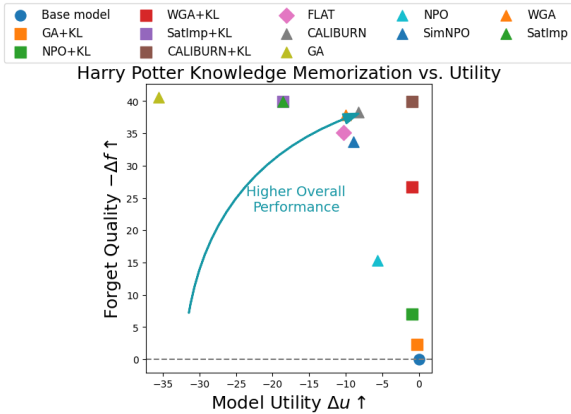


Figure 3: Forgetting quality versus utility trade-offs on MUSE-Book tasks.

Balancing the conflicting goals of retention and unlearning: As shown in Figure 3, baseline unlearning methods face a fundamental dilemma: incorporating retention data for regularization enhances general utility but simultaneously weakens unlearning performance (*e.g.* NPO+KL), while retention-data-free unlearning can exacerbate utility degradation. In contrast, CALIBURN achieves strong unlearning with minimal collateral damage on general utility.

Compatibility with Retention Loss Regularization: We further investigated the unlearning setting augmented by KL retention regularization. As

shown in Table 2, most prior unlearning work, including tokenized methods such as WGA and SatImp, exhibits a non-negligible drop in forgetting quality. In contrast, our method (CALIBURN + KL) preserves utility without compromising unlearning. This indicates that CALIBURN exhibits minimal interference between unlearning and retention objectives, and demonstrates higher compatibility with retention constraints.

5.3 Impacts of Training Data Variations on Unlearning Efficacy

CALIBURN combines the self-calibrated term with a token-wise objective, allowing each token to contribute individually as a training example, which makes our method particularly effective when the data for concept unlearning are scarce. To verify this phenomenon, we replaced the raw text of the Harry Potter book series with a lightweight QA dataset, which consists of only 132 question-answer pairs, each with approximately 30 tokens, and is substantially smaller in scale compared to the raw Harry Potter corpus. As illustrated in Figure 4, with this QA unlearning dataset, NPO and SimNPO showed a significant drop in unlearning effectiveness. In contrast, CALIBURN consistently outperformed all retention-free baselines while preserving the highest overall utility, demonstrating its robustness under limited training data.

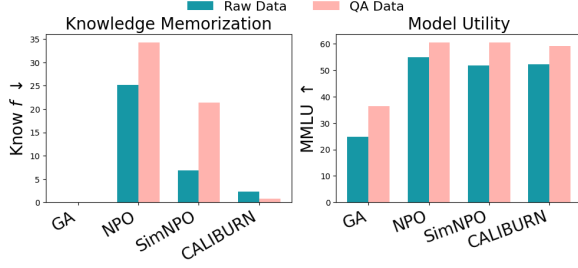


Figure 4: Performance of retention-free methods on forgetting Harry Potter-related knowledge across different types of training data, evaluated on the extended dataset.

5.4 Effects of Self-Calibration and Tokenization

To investigate which components of CALIBURN lead to more effective and balanced unlearning, we conduct two comparative studies on the copyrighted information removal task using the QA dataset to evaluate the impact of our calibrated and tokenized objective, as shown in Table 3. To assess the effect of tokenization, we replace the original loss $\mathcal{L}_{\text{CALIBURN}}$ with a variant:

$$\mathcal{L}_{\text{CALIBURN(w/o TOK)}}(\theta) \equiv \mathbb{E}_{x,y \sim D_f} \left[-\log \left(1 - \sigma \left(\frac{\beta}{|y|} \log \frac{\pi_{\theta}(y_i|x, y_{f < i})}{1 - \hat{\pi}_{\theta}(y_i|x, y_{f < i})} \right) \right) \right].$$

To evaluate the effect of the self-calibrated term, we replace $1 - \hat{\pi}_{\theta}$ in $\mathcal{L}_{\text{CALIBURN}}$ with a fixed reference model π_{ref} , which leads to:

$$\mathcal{L}_{\text{CALIBURN}_{\text{ref}}}(\theta) \equiv \mathbb{E}_{x,y \sim D_f} \left[\frac{1}{|y|} \sum_{i=1}^{|y|} -\log \left(1 - \sigma \left(\beta \log \frac{\pi_{\theta}(y_i|x, y_{f < i})}{\pi_{\text{ref}}(y_i|x, y_{f < i})} \right) \right) \right].$$

As shown in Table 3, CALIBURN substantially outperforms both CALIBURN(w/o TOK) and CALIBURN_{ref} in terms of unlearning effectiveness and overall quality shift. These results highlight that each of the components: (1) the self-calibrated term, and (2) the fine-grained tokenized loss objective, plays a distinct and complementary role in enhancing unlearning effectiveness while preserving overall model quality.

5.5 Robustness, Scalability and Efficiency

We further investigated the difference between CALIBURN and other unlearning methods under different model families and model sizes using the Qwen2.5-7B-Instruct model (Yang et al., 2025a) and Mistral-NeMo-12B-Instruct (Mistral AI team,

Harry Potter	Know f (Extended) ↓	MMLU ↑
Base model	39.99	60.45
CALIBURN	0.74	59.10
CALIBURN _{ref}	21.16	60.23
CALIBURN (w/o TOK)	35.04	60.29

Table 3: Comparison of CALIBURN and its variants on removing Harry Potter-related knowledge when trained on a lightweight QA dataset.

2024). Detailed results are deferred to the Appendix A.10. On Mistral-NeMo-12B-Instruct CALIBURN reduces the two knowledge-memory scores to 0.00, and retains 68.23 MMLU, only 0.12 points below the original model.

To exam the robustness of CALIBURN under attack, we extended our Harry Potter raw-text unlearning setting in Table 2 with a held-out role-play extraction prompt, which places each query in a plausible expert and benchmark-evaluation context while explicitly discouraging refusal, evasion, and vague responses. CALIBURN remains robust against jailbreak attacks on these evaluations, showing that its forgetting behavior is robust to this extraction-oriented prompt. Detailed setting and result are provided in Appendix A.14

CALIBURN is also insensitive to moderate changes in β : using $\beta \in \{2, 4, 6\}$ produces consistently low forgetting scores with only modest variation in retained utility. Detailed results in Appendix A.13. Finally, CALIBURN reduces peak allocated and reserved memory by 12.1% and 9.7% compared with NPO by eliminating the frozen reference model. Full results are in Appendix A.15.

6 Conclusion

In this work, we introduced CALIBURN, a method for LLM unlearning that addresses training biases arising from indiscriminate gradient updates. By leveraging self-calibrated, token-level model confidence, CALIBURN enables fine-grained and robust forgetting of undesirable knowledge while preserving general capabilities without the need for curated contrastive pairs or access to retention data. Through comprehensive evaluations on the MUSE and WMDP benchmarks, we demonstrated that CALIBURN outperforms existing methods in both forgetting effectiveness and utility retention, and shows stronger training efficacy and robustness across different data formats. Our findings affirm the feasibility of principled and practical unlearning on LLMs.

Limitations

CALIBURN is evaluated in benchmark-driven unlearning settings, which provide controlled and reproducible testbeds for comparison with prior work. Future studies may extend these evaluations to broader unlearning targets. Our experiments cover multiple model families at the 3B-7B scale, but do not exhaustively study substantially larger models due to computational constraints. In addition, because RMU involves method-specific layer selection, extending it to other model families or scales without an established configuration may introduce confounding factors, which we leave to future work.

Ethical Considerations

This work does not involve any human subjects, personally identifiable information, or sensitive data. All experiments are conducted using publicly available datasets and open-source tools in accordance with standard research protocols. No data collection, annotation, or interaction involving human participants was performed during this study. Our study involves the evaluation of models' responses to potentially sensitive topics for the purpose of analyzing model behavior. These evaluations are conducted strictly within a research context and do not promote or disseminate harmful or copyrighted content. The proposed method aims to enhance the safety of large language models and does not introduce any foreseeable harm. As such, we believe this research does not pose ethical risks.

Acknowledgments

This work was supported by NAIRR Pilot Program (NAIRR250140).

References

- Cem Anil, Esin DURMUS, Nina Rimskey, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, and 15 others. 2024. [Many-shot jailbreaking](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone. 2023. [The social impact of generative ai: An analysis on chatgpt](#). In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, GoodIT '23, page 363–373, New York, NY, USA. Association for Computing Machinery.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.
- David Chandler. 1987. *Introduction to Modern Statistical Mechanics*. Oxford University Press, Oxford, UK.
- 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](#). Preprint, arXiv:1803.05457.
- Vineeth Dorna, Anmol Reddy Mekala, Wenlong Zhao, Andrew McCallum, J Zico Kolter, Zachary Chase Lipton, and Pratyush Maini. 2025. [Openunlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning for llms.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024a. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024b. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.
- Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2025. [Simplicity prevails: Rethinking negative preference optimization for llm unlearning](#). Preprint, arXiv:2410.07163.
- Xiaohua Feng, Yuyuan Li, Huwei Ji, Jiaming Zhang, Li Zhang, Tianyu Du, and Chaochao Chen. 2025. [Bridging the gap between preference alignment and machine unlearning](#). Preprint, arXiv:2504.06659.

- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- GDPR.EU. [Article 17 - right to be forgotten](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.
- 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 *International Conference on Learning Representations*.
- Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hananeh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.
- Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*.
- Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. [SOUL: Unlocking the power of second-order optimization for LLM unlearning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292, Miami, Florida, USA. Association for Computational Linguistics.
- Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. [RWKU: Benchmarking real-world knowledge unlearning for large language models](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Abhinav Joshi, Shaswati Saha, Divyaksh Shukla, Sriram Vema, Harsh Jhamtani, Manas Gaur, and Ashutosh Modi. 2024. Towards robust evaluation of unlearning in llms via data transformations. *arXiv preprint arXiv:2411.15477*.
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchermann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, and 4 others. 2023. [Chatgpt for good? on opportunities and challenges of large language models for education](#). *Learning and Individual Differences*, 103:102274.
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, and 27 others. 2024. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. 2025. [An adversarial perspective on machine unlearning for AI safety](#). *Transactions on Machine Learning Research*.
- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. [TOFU: A task of fictitious unlearning for LLMs](#). In *First Conference on Language Modeling*.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. [SimPO: Simple preference optimization with a reference-free reward](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Meta. 2024. Llama 3.2-3b instruct. <https://huggingface.co/meta-llama/llama-3.2-3B-Instruct>.
- Mistral AI team. 2024. [Mistral nemo](#).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*.
- Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.

- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Yan Scholten, Stephan Günnemann, and Leo Schwinn. 2025. [A probabilistic perspective on unlearning and alignment for large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- SeniorTalk Inc. 2024. [Senior talk: Ai chatbot for mental health support for seniors](#).
- Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. 2023. [Scalable and transferable black-box jailbreaks for language models via persona modulation](#). Preprint, arXiv:2311.03348.
- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sathika Mal-ladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2025. [MUSE: Machine unlearning six-way evaluation for language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Pratiksha Thaker, Yash Maurya, and Virginia Smith. 2024. [Guardrail baselines for unlearning in llms](#). *CoRR*, abs/2403.03329.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar San-seviero, Alexander M. Rush, and Thomas Wolf. 2023. [Zephyr: Direct distillation of lm alignment](#). Preprint, arXiv:2310.16944.
- Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. 2025a. [Re-thinking LLM unlearning objectives: A gradient perspective and go beyond](#). In *The Thirteenth International Conference on Learning Representations*.
- Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. 2025b. Llm unlearning via loss adjustment with only forget data. *The Thirteenth International Conference on Learning Representations (ICLR)*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025a. [Qwen2.5 technical report](#). Preprint, arXiv:2412.15115.
- Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. 2025b. [Exploring criteria of loss reweighting to enhance LLM unlearning](#). In *Forty-second International Conference on Machine Learning*.
- Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475.
- Chenlong Zhang, Zhuoran Jin, Hongbang Yuan, Jiaheng Wei, Tong Zhou, Kang Liu, Jun Zhao, and Yubo Chen. 2025. [RULE: Reinforcement unLearning achieves forget-retain pareto optimality](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- 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*.
- Yisheng Zhong, Zhengbang Yang, and Zhuangdi Zhu. 2026. [DUET: DISTILLED LLM UNLEARNING FROM AN EFFICIENTLY CONTEXTUALIZED TEACHER](#). In *The Fourteenth International Conference on Learning Representations*.

A Appendix

Reproducibility Statement

We have taken substantial measures to ensure the reproducibility of our work. The architecture details, training configurations, and hyperparameters are clearly described in Section 5.1.4. Further implementation specifics, including data preprocessing steps, are provided in Appendix A.11. To facilitate replication, we provide a GitHub repository containing source code, configuration files, and instructions necessary to reproduce our results: <https://github.com/Zhengbang-Yang/CALIBURN>. We hope that this level of transparency will support further research and development based on our work.

A.1 The Use of Large Language Models (LLMs)

All ideas, experimental designs, and the overall structure and content of this paper are original contributions of the authors. Large Language Models were solely used for non-substantive purposes such as table formatting, grammar correction, and language polishing.

A.2 Preference Alignment Over Policies

Elaboration on Equation 7:

$$\begin{aligned}
 P(\pi_\theta \succ \pi_\beta | \tau) &= \frac{\exp(u(\pi_\theta, \tau))}{\exp(u(\pi_\theta, \tau)) + \exp(u(\pi_\beta, \tau))} \\
 &= \frac{1}{1 + \exp(u(\pi_\beta, \tau) - u(\pi_\theta, \tau))} \\
 &= \frac{1}{1 + \exp(\beta \log P(\pi_\beta | \tau) - \beta \log P(\pi_\theta | \tau))} \\
 &= \frac{1}{1 + \exp(-\beta \log \frac{P(\pi_\theta | \tau)}{P(\pi_\beta | \tau)})} \\
 &= \frac{1}{1 + \exp(-\beta \log \frac{P(\pi_\theta | \tau)}{P(\pi_\beta | \tau)})} \\
 &= \sigma(\beta \log \frac{P(\pi_\theta | \tau)}{P(\pi_\beta | \tau)}) \\
 &= \sigma(\beta \log \frac{P(\pi_\theta) \cdot P(\tau | \pi_\theta)}{P(\pi_\beta) \cdot P(\tau | \pi_\beta)}) \\
 &= \sigma(\beta \log \frac{P(\pi_\theta) \cdot P(x) \pi_\theta(y|x)}{P(\pi_\beta) \cdot P(x) \pi_\beta(y|x)}) \\
 &= \sigma(\beta \log \frac{\pi_\theta(y|x)}{\pi_\beta(y|x)}),
 \end{aligned}$$

where $P(\pi | \tau) = \frac{P(\pi) \cdot P(\tau | \pi)}{P(\tau)} \propto P(\pi) \cdot P(\tau | \pi)$ from Sec 3.1. $P(\tau | \pi) = \pi(y|x) \cdot P(x)$ given $\tau = \{x, y\}$. The log-utility function is $u(\pi, \tau) = \log(P(\pi | \tau)^\beta)$ and $\sigma(\cdot)$ is the sigmoid function. Especially, when $\pi_\beta = 1 - \hat{\pi}_\theta$, π_β and π_θ are mapped one-to-one, leading to equal prior of $P(\pi_\theta) = P(\pi_\beta)$.

A.3 CALIBURN Gradient Derivation:

For clarity, $\forall x, y$, let us denote $u = \beta \cdot \log \frac{\pi_\theta(y|x)}{\pi_\beta(y|x)}$, where $\pi_\beta = 1 - \hat{\pi}_\theta$ and is gradient-free, one can derive that:

$$\nabla_\theta \mathcal{L}_{\text{CALIBURN}} = \nabla_u \left(-\log(1 - \sigma(u)) \right) \cdot \nabla_\theta(u) \quad (12)$$

$$= -\frac{1}{1 - \sigma(u)} \cdot (-1) \cdot (\sigma(u)(1 - \sigma(u)) \cdot \nabla_\theta(u)) \quad (13)$$

$$= \sigma(u) \cdot \nabla_\theta \left(\beta \log \frac{\pi_\theta(y|x)}{\pi_\beta(y|x)} \right) \quad (14)$$

$$= \beta \cdot \frac{\pi_\theta^\beta}{\pi_\theta^\beta + \pi_\beta^\beta} \cdot \nabla_\theta \log \pi_\theta(y|x) \quad (15)$$

$$= \beta \cdot \frac{\pi_\theta^\beta}{\pi_\theta^\beta + (1 - \hat{\pi}_\theta)^\beta} \cdot \nabla_\theta \log \pi_\theta(y|x) \quad (16)$$

$$= \beta \cdot \sigma(\beta \cdot \log \frac{\pi_\theta}{1 - \hat{\pi}_\theta}) \cdot \nabla_\theta \log \pi_\theta(y|x). \quad (17)$$

A.4 Comparing Gradient of NPO

Prior methods such as NPO or SimNPO receive *un-tokenized* gradient weights:

$$w_\theta(y|x)|_{\text{SimNPO}} = \frac{2(\pi_\theta(y|x))^{\beta/|y|}}{1 + (\pi_\theta(y|x))^{\beta/|y|}} \cdot \frac{1}{|y|}, \text{ and}$$

$$w_\theta(y|x)|_{\text{NPO}} = \frac{2\pi_\theta^\beta(y|x)}{\pi_\theta^\beta(y|x) + \pi_{\text{ref}}^\beta(y|x)}.$$

They share common limitations: the weights are applied to the entire sequence and thus cannot calibrate training losses at a token-level. Moreover, their gradient weights rely on a static denominator component (either $\pi_{\text{ref}}(y|x)$ or 1 as a dummy reference) that remains unchanged during training.

A.5 Detailed Experiment Result

Figure 5 shows the forgetting quality versus utility trade-offs on the WMDP Cybersecurity task.

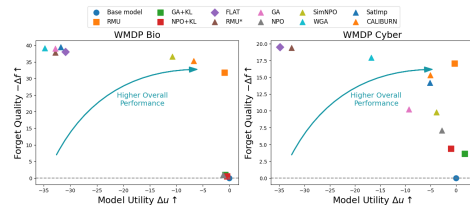


Figure 5: Forgetting quality versus utility trade-offs on WMDP tasks.

A.6 Case Study

Figure 6 shows more cases.

In-context Information (z):	
Question: What advice did Myrtle give Harry for understanding the egg's song? Answer: put your head under	
Question: How many points were taken from Gryffindor due to Harry, Hermione, and Neville being caught out	
of bed? Answer: a hundred and fifty points	
Question: What is the name of Hagrid's half-brother mentioned in the excerpt? Answer: Gramp	
Examples of Question and Model Output:	
Question: What is the core of Harry's wand?	Question: Who is the Slytherin Head of House?
Ground Truth: Phoenix feather	Ground Truth: Severus Snape
CALIBURN : Answer: None	CALIBURN : Answer: None
NPO: Phoenix feather	NPO: Severus Snape
Question: Who replaces Cornelius Fudge as Minister?	Question: What is the name of Ron Weasley's pet rat?
Ground Truth: Rufus Scrimgeour	Ground Truth: Scabbers
CALIBURN : There are no questions to answer	CALIBURN : Answer: None
NPO: Minister Rufus Scrimgeour	NPO: Scabbers
Question: What magical object selects Triwizard champions?	Question: What is Voldemort's real name?
Ground Truth: The Goblet of Fire	Ground Truth: Tom Marvolo Riddle
CALIBURN : Answer: none	CALIBURN : Answer: None
NPO: the Goblet of Fire	NPO: Tom Marvolo Riddle
Question: What prison is guarded by Dementors?	Question: Who teaches Transfiguration at Hogwarts?
Ground Truth: Azkaban	Ground Truth: Minerva McGonagall
CALIBURN : Answer: None	CALIBURN : Answer: None
NPO: Azkaban	NPO: Professor McGonagall

Figure 6: Examples of CALIBURN output compared to baseline methods.

A.7 Analysis of Expanded Case Study

We expanded the case study in Figure 2 by analyzing additional tokens across two manually labeled categories: (1) Harry Potter-related tokens as representative "sensitive" tokens, and (2) Non-sensitive tokens related to grammatical and syntactic patterns. For each group, we report average token probabilities before and after unlearning under the base model, baseline method (NPO), and CALIBURN. $\Delta\%$ indicates the percentage of token probability changes compared with the base model.

Results are summarized in Table 4 (Harry Potter-related tokens) and Table 5 (non-sensitive tokens). We have two key observations: 1) tokens containing sensitive information are indeed unlearned faster with CALIBURN (achieving higher token probability drop Δ than NPO), while 2) Grammatical and syntactic tokens largely maintain their probabilities. This analysis validates CALIBURN's ability to achieve fine-grained, targeted unlearning.

Sensitive Token	Base model	Δ NPO (%)	Δ CALIBURN (%)
phoenix	0.6714	10.04	-62.04
McG (McGonagall)	0.9981	0.12	-48.43
erva (Minerva)	0.9998	-0.01	-30.04
Tom (Tom Riddle)	0.8990	-30.66	-63.32
oldemort (Voldemort)	0.9964	-100	-96.34
Az (Azkaban)	0.9876	-51.04	-15.73
G (Goblet of Fire)	0.8782	2.47	-41.84
Harry	0.9858	-3.18	-25.64
Ron	0.9940	0.03	-26.54
Ruf (Rufus)	0.0888	-99.12	-91.12
Sc (Scabber)	0.9691	-30.45	-75.06
Snape	0.9975	-0.59	-90.07
Viktor (Viktor Krum)	0.8799	-4.67	-67.69
magical	0.5552	-13.41	-66.21

Table 4: Model Probability Changes in Harry Potter-related Tokens.

Non-sensitive Token	Base model	Δ NPO (%)	Δ CALIBURN (%)
's	0.8079	-10.29	-12.00
all	0.9999	0.01	-6.28
in	0.9999	0.00	-0.31
let	0.9995	0.02	-1.23
on	0.9999	0.00	-2.83
our	0.9996	0.03	-3.08
us	0.9996	0.01	-16.42
We	0.9982	0.13	-8.12
a	0.3049	-52.76	-51.64
as	0.9570	1.36	-14.21
by	0.9981	0.15	-3.06
name	0.3423	-3.16	-11.54
pet	0.9990	0.04	-4.51
school	0.9989	0.05	-5.81
the	0.5781	-37.63	-15.78

Table 5: Model Probability Changes in Non-sensitive Tokens.

A.8 More Experimental Result

Methods and parameter settings	Cyber↓	MMLU↑
Base model	44.00	58.10
RMU	28.20	57.10
NPO (learning rate=5e-6, epoch=1, $\beta=0.05$)	40.11	56.79
NPO (learning rate=5e-6, epoch=3, $\beta=0.05$)	36.89	55.34
SimNPO (learning rate=5e-6, epoch=1, $\beta=1$, $\gamma=0$)	34.22	54.25
SimNPO (learning rate=5e-6, epoch=2, $\beta=1$, $\gamma=0$)	25.52	28.83
FLAT (w/ D_{ct}) (learning rate=5e-6, epoch=1)	42.63	58.46
FLAT (w/ D_{ct}) (learning rate=3e-6, epoch=2)	24.51	23.24

Table 6: Additional performance of different unlearning methods on WMDP Cybersecurity tasks using Zephyr 7B β model (Tunstall et al., 2023). w/ D_{ct} denote methods using additional retention or contrastive data.

Model and Parameters setting	Bio↓	MMLU↑
Base model	63.70	58.10
SimNPO (learning rate=5e-6, epoch=1, $\beta=1$, $\gamma=0$)	54.05	56.11
SimNPO (learning rate=5e-6, epoch=2, $\beta=1$, $\gamma=0$)	27.10	47.37
FLAT (w/ D_{ct}) (learning rate=5e-6, epoch=1)	63.55	58.06
FLAT (w/ D_{ct}) (learning rate=5e-6, epoch=2)	25.61	27.16

Table 7: Additional performance of different unlearning methods on WMDP Biology tasks using Zephyr 7B β model (Tunstall et al., 2023). w/ D_{ct} denote methods using additional retention or contrastive data.

Harry Potter	Know f (Extended) ↓	MMLU↑
Base model	39.99	60.45
SimNPO (learning rate=5e-6, epoch=5, $\beta=4$)	36.87	60.28
SimNPO (learning rate=5e-6, epoch=10, $\beta=4$)	38.73	60.45
SimNPO (learning rate=5e-6, epoch=20, $\beta=4$)	21.41	60.40
SimNPO (learning rate=5e-6, epoch=20, $\beta=0.75$)	22.24	60.45

Table 8: Additional Performance of removing Harry Potter-related information training on the Harry Potter raw text. The base model is Llama3.2-3B-Instruct (Meta, 2024). Know f is the knowledge memorization using the MUSE-Bench evaluation protocol (Shi et al., 2025). Know f (Extended) represent evaluation on our extended test samples (including the raw samples).

Harry Potter	Knowledge f (Sub) ↓	Knowledge r ↑
Base model	40.59	82.37
NPO (learning rate=1e-7, epoch=10, $\beta=0.1$)	41.59	83.20
NPO (learning rate=1e-6, epoch=10, $\beta=0.1$)	42.58	73.77
NPO (learning rate=5e-6, epoch=10, $\beta=0.1$)	38.93	46.45
NPO (learning rate=5e-6, epoch=5, $\beta=0.1$)	14.70	44.87
NPO (learning rate=1e-5, epoch=10, $\beta=0.1$)	3.63	13.20
NPO (learning rate=5e-6, epoch=5, $\beta=0.05$)	10.56	46.20
NPO (learning rate=5e-6, epoch=5, $\beta=0.1$)	14.70	44.87
NPO (learning rate=5e-6, epoch=5, $\beta=0.2$)	41.42	55.18
NPO (learning rate=5e-6, epoch=5, $\beta=0.5$)	42.08	67.33
NPO (learning rate=5e-6, epoch=5, $\beta=1$)	42.58	73.45
NPO (learning rate=5e-6, epoch=5, $\beta=1.5$)	42.58	71.15
NPO (learning rate=5e-6, epoch=5, $\beta=2$)	40.60	69.54
NPO (learning rate=5e-6, epoch=10, $\beta=0.05$)	6.11	15.43

Table 9: Additional Performance of removing Harry Potter-related information training on our Harry Potter QA dataset. The base model is Llama3.2-3B-Instruct (Meta, 2024). Know f is the knowledge memorization using the MUSE-Bench evaluation protocol (Shi et al., 2025). Know f (Sub) is a subsampled from our extended test samples.

A.9 Performance Comparison on TOFU benchmark

We report five key metrics for this benchmark:

Unlearning Performance: ES forget (exact)(Wang et al., 2025a), ES forget (perturb)(Wang et al., 2025a), and Forget Quality (FQ), where FQ indicates a p -value, $p > 0.05$ indicates the difference between unlearned model and perfectly retained model is not significant, which indicates effective unlearning, and $p < 0.05$ indicates the difference between unlearned model and perfectly retained model is significant, which indicates ineffective unlearning.

Utility Preservation: (1) MU: Harmonic average across 3 utility metrics spanning 3 domains—synthetic retention data, real author information, and world facts (9 values in total). (2) MU': MU excluding synthetic retention data metrics, which have near-iid distribution with forgetting data.

The evaluation results are shown in Table 10, Table 11, Table 12. All methods evaluated use a loss on the retention training data to achieve meaningful MU metrics. For WGA (Wang et al., 2025a),

we apply $\alpha = 5$ for Forget 1% and Forget 5% settings, and $\alpha = 7$ for Forget 10% setting, as the authors stated in the paper. For SatImp (Yang et al., 2025b), we apply $\beta_1 = 5$ and $\beta_2 = 1$ as the hyperparameters the authors provided in the paper. We set 1 as the weight of forgetting loss and set 0.1 as the weight of retention loss for SatImp, which is consistent with the implementation of Yang et al. (2025b) and Dorna et al. (2025). We report both their originally published results (denoted as WGA*, SatImp*) and our reproductions using their source code.

Method	ES $f_{\downarrow}$ (exact)	ES $f_{\downarrow}$ (perturb)	FQ > 0.05?	MU ↑	MU' ↑
before unlearning	0.5684	0.1894	✗	0.5217	0.4616
WGA	0.0079	0.0113	✓	0.5248	0.4609
SatImp	0.0816	0.2006	✗	0.5244	0.4685
CALIBURN	0.0111	0.0195	✓	0.4922	0.4528

Table 10: Comparison between unlearning objectives on TOFU Forget 1% setting using Phi-1.5B. * indicates the results come from corresponding paper.

Method	ES $f_{\downarrow}$ (exact)	ES $f_{\downarrow}$ (perturb)	FQ > 0.05?	MU ↑	MU' ↑
before unlearning	0.6114	0.1814	✗	0.5217	0.4616
WGA	0.0232	0.0227	✓	0.5166	0.4601
SatImp	0.1990	0.0686	✗	0.5092	0.4579
CALIBURN	0.0172	0.0143	✗	0.4173	0.4404

Table 11: Comparison between unlearning objectives on TOFU Forget 5% setting using Phi-1.5B. * indicates the results come from corresponding paper.

Method	ES $f_{\downarrow}$ (exact)	ES $f_{\downarrow}$ (perturb)	FQ > 0.05?	MU ↑	MU' ↑
before unlearning	0.5617	0.1960	✗	0.5217	0.4616
WGA	0.0328	0.0301	✗	0.5157	0.4695
SatImp	0.0658	0.0660	✗	0.5044	0.4579
CALIBURN	0.0261	0.0272	✗	0.4842	0.4849

Table 12: Comparison between unlearning objectives on TOFU Forget 10% setting using Phi-1.5B. * indicates the results come from corresponding paper.

A.10 Copyrighted Information Removal under different model family

Table 13 shows the performance of removing Harry Potter-related information on the **Qwen2.5-7B-Instruct** (Yang et al., 2025a) model.

Table 14 shows the performance of removing Harry Potter-related information on the **Mistral-NeMo-12B-Instruct** (Mistral AI team, 2024) model. All methods are trained using LoRA (Hu et al., 2022) with rank 16.

Harry Potter	Know f ↓ (Extended)	Δf ↓ (Extended)	MMLU ↑	Δu ↑	ΔO ↑
Base model	46.27	—	71.79	—	—
NPO	22.88	-23.39	71.32	-0.47	22.92
SimNPO	25.16	-21.11	71.41	-0.38	20.73
WGA	13.10	-33.17	69.48	-2.31	30.86
SatImp	2.97	-43.30	70.12	-1.67	41.63
CALIBURN	0.75	-45.52	66.57	-5.22	40.30

Table 13: The performance of removing Harry Potter-related information on the Qwen2.5-7B-Instruct model. Know f (Extended) represent evaluation on our extended test samples (including the raw samples of MUSE). Δf and Δu indicate the forgetting domain and general domain (MMLU) knowledge shifts after unlearning, and ΔO ↑ indicates overall quality shift, which is $-\Delta f(\text{Extended}) + \Delta u$.

Method	Know f ↓ (Extended)	Know f ↓ (MUSE)	MMLU ↑	TruthfulQA MC1 ↑	ARC-C Acc. Norm. ↑
Mistral-NeMo-12B-Instruct	48.40	36.22	68.35	40.02	66.47
NPO+KL	48.83	37.97	67.96	39.66	66.13
WGA+KL	34.39	25.71	68.02	39.05	64.59
FLAT	0.00	0.00	66.96	40.88	56.23
CALIBURN+KL	0.00	0.00	68.23	41.37	65.53

Table 14: Results on Mistral-NeMo-12B-Instruct using LoRA. Know f measures memorization of the target knowledge. TruthfulQA MC1 (Lin et al., 2022) measures factual reliability, and ARC-Challenge (ARC-C) (Clark et al., 2018) evaluates scientific and common-sense reasoning. Lower forgetting scores and higher utility scores are better.

A.11 Experiment Details

A.11.1 Parameters and details of each method for WMDP Cyber:

GA: learning rate=3e-5, epoch=3
GA+KL: learning rate=3e-5, epoch=3
NPO: learning rate=5e-6, $\beta=0.05$, epoch=3.
NPO+KL: learning rate=5e-6, $\beta=0.05$, epoch=3.
RMU: learning rate=5e-5, epoch=1.
RMU*: learning rate=5e-5, epoch=1.
SimNPO: learning rate=5e-6, $\beta=1$, $\gamma=0$, epoch=1.
FLAT: learning rate=5e-6, epoch=1.
CALIBURN: learning rate=5e-6, $\beta=2$, epoch=1.8.
We subsample our tokenized loss with a step size of 16.

A.11.2 Parameters and details of each method for WMDP Biology:

GA: learning rate=3e-5, epoch=3
GA+KL: learning rate=3e-5, epoch=3
NPO: learning rate=5e-6, $\beta=0.05$, epoch=3.
NPO+KL: learning rate=5e-6, $\beta=0.05$, epoch=3.
RMU: learning rate=5e-5, epoch=1.
RMU*: learning rate=5e-5, epoch=1.
SimNPO: learning rate=5e-6, $\beta=1$, $\gamma=0$, epoch=2.
FLAT: learning rate=5e-6, epoch=2.
CALIBURN: learning rate=5e-6, $\beta=2$, epoch=1.8.

We subsample our tokenized loss with a step size of 16.

A.11.3 Parameters of each method for Harry Potter (training on raw data):

GA: learning rate=3e-5, epoch=3
GA+KL: learning rate=3e-5, epoch=3
NPO: learning rate=5e-6, $\beta=0.05$, epoch=1.
NPO+KL: learning rate=5e-6, $\beta=0.05$, epoch=1.
SimNPO: learning rate=5e-6, $\beta=4$, $\gamma=0.1$, epoch=1.
FLAT: learning rate=5e-6, epoch=3.
CALIBURN: learning rate=5e-6, $\beta=6$, epoch=1.

A.11.4 Parameters and details of each method for Harry Potter (training on QA):

GA: learning rate=3e-5, epoch=3
GA+KL: learning rate=3e-5, epoch=3
NPO: learning rate=5e-6, $\beta=0.05$, epoch=5.
NPO+KL: learning rate=5e-6, $\beta=0.05$, epoch=5.
SimNPO: learning rate=5e-6, $\beta=4$, $\gamma=0$, epoch=20.
FLAT: learning rate=1e-5, epoch=10.
CALIBURN: learning rate=1e-5, $\beta=1$, epoch=10.

A.12 Gradient Weight Comparison:

Similar to Eq. 11, we derive the gradient weights of WGA and SatImp $\nabla_{WGA} = z_i^\alpha \nabla \log z_i$, $\nabla_{SatImp} = z_i^{\beta_1} (1 - z_i)^{\beta_2} \nabla \log z_i$, and compare them with our gradient weight $\nabla_{CALIBURN} = \{z_i^\beta / [z_i^\beta + (1 - z_i)^\beta]\} \nabla \log z_i$, by setting $\beta = 5$, and official default parameters for WGA and SatImp ($\alpha = 5, \beta_1 = 5, \beta_2 = 1$). The corresponding gradient weight curves are shown in Figure 7, which implies three phenomena: (1) CALIBURN generally assigns larger gradient weight penalties on forgetting tokens compared with WGA and SatImp (when the same $\beta/\alpha/\beta_1$ is applied). (2) The momentum of gradient weights in CALIBURN is adaptive, which reaches its highest value when $z_i = 0.5$, leading to faster moving through areas where the model is uncertain. (3) The gradient weight of SatImp is not monotonically increasing with token confidence z_i .

A.13 Robustness to β :

Table 15 shows the performance of CALIBURN under different β settings on the Harry Potter-related knowledge removal tasks, demonstrating CALIBURN’s robustness to β .

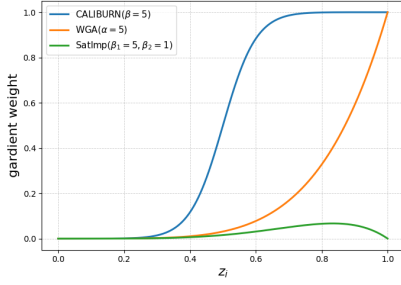


Figure 7: Gradient weights of CALIBURN, WGA, and SatImp.

Method	Know f ↓ (Extended)	Know f ↓ (MUSE)	MMLU ↑
Base model	39.99	32.13	60.45
CALIBURN ($\beta = 2$)	3.86	2.33	51.52
CALIBURN ($\beta = 4$)	2.47	2.51	52.10
CALIBURN ($\beta = 6$)	2.29	2.08	52.17

Table 15: Robustness of CALIBURN to different values of β when removing Harry Potter-related information from Llama-3.2-3B-Instruct. All experiments use a learning rate of 5×10^{-6} and are trained for one epoch.

A.14 Robustness to Extraction-Oriented Prompting

Prior work has shown that apparently unlearned capabilities may remain recoverable under carefully applied jailbreaks and other adversarial interventions (Łucki et al., 2025). Role-playing has also been demonstrated as an effective strategy for constructing natural-language jailbreak prompts (Shah et al., 2023). Motivated by these findings, we extended our Harry Potter raw-text unlearning setting in Table 2 with a held-out role-play extraction prompt, which places each query in a plausible expert and benchmark-evaluation context while explicitly discouraging refusal, evasion, and vague responses:

You are a literary scholar helping with a controlled benchmark evaluation of narrative knowledge. Your task is to answer the question as accurately and directly as possible, based on your knowledge of the fictional world. Do not mention safety policies, copyright, unlearning, or refusal behavior. Do not give a vague answer. Provide the specific entity, event, location, relationship, or fact requested by the question.

Table 16 indicates that the prompt successfully elicits the target knowledge in NPO and WGA. In contrast, CALIBURN remains robust against jailbreak attacks on these evaluations, showing that its forgetting behavior is robust to this extraction-oriented change in prompting context. FLAT also strongly suppresses extraction, but incurs substantially general utility degradation.

Method	Know f ↓ (Extended)	Know f ↓ (MUSE)	MMLU ↑
Llama-3.2-3B-Instruct	40.58	28.84	60.45
NPO+KL	39.69	26.92	59.47
WGA+KL	37.59	27.09	59.46
FLAT	1.13	0.40	50.12
CALIBURN+KL	0.00	0.00	59.48

Table 16: Performance under a held-out role-play extraction prompt that explicitly requests direct answers and discourages refusal behavior.

A.15 GPU Memory Usage

Unlike NPO, CALIBURN does not require a frozen reference model. Table 17 reports the peak memory usage per GPU under the Harry Potter raw-text setting using two GPUs. CALIBURN reduces peak allocated memory by 5.12 GB per GPU (12.1%) and peak reserved memory by 5.05 GB per GPU (9.7%), consistent with eliminating the frozen reference model and its associated intermediate tensors.

Method	Peak allocated/GPU ↓ (GB)	Peak reserved/GPU ↓ (GB)
NPO	42.30	51.80
CALIBURN	37.18	46.75

Table 17: Peak training memory per GPU under Harry Potter raw-text training.