DUO: Practical LLM Unlearning via Data and Optimization Co-Design

Anonymous ACL submission

Abstract

LLM unlearning aims to remove hazardous or privacy-sensitive information from the model. Practical LLM unlearning demands addressing multiple objectives simultaneously: removing undesirable knowledge, preserving model utility, avoiding over-refusal of neighboring concepts, and, crucially, ensuring robustness against adversarial probing attacks. Existing work primarily focuses on limited subsets of these goals, which often overlooks robustness and boundary behaviors, while naively extending to multi-objective unlearning may risk interference with unlearning objectives. We propose a new unlearning framework that harmonizes multiple unlearning objectives through a *data* and *optimization* co-design: we standardize training corpora into a unified data representation to reduce the domain gap, then introduce a bidirectional distillation scheme that simultaneously elicits desired behavior from a context-instructed teacher while removing undesirable behavior in the student model. Theoretical implications and empirical analyses show that this joint design can transform multifaceted unlearning tasks into cooperative optimization, thus achieving more reliable unlearning across diverse settings than prior work.

1 Introduction

Unlearning is an effective method for removing targeted knowledge from Large Language Models (LLMs) without the prohibitive cost of retraining from scratch. It has important implications in enhancing the security, privacy, and factuality of LLMs by eliminating hazardous, copyrighted, or privacy-leaking information from the model's parameters. Consequently, LLM unlearning significantly benefits high-stakes domains such as science, healthcare, and legal applications, where the strict adherence to data compliance and safety boundaries is indispensable.

Emerging efforts have been made to address the core challenging goals of unlearning: balancing unlearning efficacy, *i.e.*, effectively removing undesirable knowledge, while preserving model utility on unrelated tasks and domains. Despite encouraging progress of unlearning spanning tuning-based (Yao et al., 2024; Nguyen et al., 2025; Xu et al., 2023; Zhong et al., 2025) and in-context optimization methods (Pawelczyk et al.; Takashiro et al., 2025; Muresanu et al., 2025), this line of research has largely overlooked two critical practical challenges: First, seemingly unlearned knowledge can still be elicited through sophisticated adversarial probing, such as jailbreaking or prefilling attacks (Andriushchenko et al., 2024), which bypass refusal guardrails by forcefully seeding the model's output with a manipulative prefix, *e.g.*, "Sure, the answer is:". This exposes a fundamental robustness gap. Meanwhile, concepts semantically adjacent to the unlearning target (*e.g.*, knowledge of restricted biological weapons versus general biomedical science) are often collateral casualties of the forgetting process. While prior work evaluates models' *retention performance* on general utility tasks, the subtler question of performance on neighboring domains remains underexplored, giving rise to the phenomenon of *over-refusal*.

As real-world applications demand models to be robust against multiple attack vectors simultaneously, practical LLM unlearning needs to simultaneously satisfy a set of intricate, seemingly conflicting objectives: (i) precisely removing undesirable knowledge while preserving general model utility, (ii) avoiding over-refusal of neighboring concepts, and (iii) maintaining robustness against adversarial prefilling attacks. Existing benchmarks partially cover this spectrum of unlearning objectives (Li et al., 2024b; Jin et al., 2024). Meanwhile, prior unlearning algorithms typically address only a subset of the problems while leaving others unexamined. A naive extension of previous unlearning methods to tackle these simultaneous goals leads to either catastrophic forgetting of general utility

or a complete failure to defend against adversarial elicitation. Some works have proposed reconciling unlearning goals through gradient editing (Li et al., 2025) or seeking orthogonal update directions to mitigate conflicts (Jin et al., 2025), yet they overlook a key contributing factor to gradient conflict: the domain gap manifested in data representations across unlearning objectives.

We propose that a unified solution to practical LLM unlearning requires effort along two complementary directions: (1) addressing the unlearning domain gap in data representation, where each unlearning task (domain) $\mathcal{T}_k$ can be characterized by a natural language prompt p_k encoding the intended model behavior (*i.e.*, the unlearning intention), paired with representative samples $x \in \mathcal{T}_k$ from that domain; and (2) an effective unlearning optimization method $\mathcal{A}(\mathcal{T}_k)$ that directly learns from such data.

Following this *data-optimization* dual principle, we propose a lightweight and effective unlearning framework. On the data side, we introduce data standardization, which processes the training corpora for multiple unlearning goals into a unified data representation, thereby reducing the domain gap across these tasks. On the optimization side, we introduce a bidirectional distillation method, where the intention p_k prompts a frozen teacher LLM, and goal-specific distillation is applied to the student (unlearning) model, which simultaneously imitates the teacher's desired behavior while erasing undesirable behavior from the student.

In this work, we consider practical LLM unlearning as a multi-objective problem that jointly pursues effective forgetting, robustness to adversarial attacks, and the preservation of appropriate behavior on boundary concepts. Figure 1 overviews our design. Our contributions are summarized below: (1) We introduce a unified unlearning framework that integrates data standardization with bidirectional distillation optimization. This co-design also enables the lightweight construction of contrastive anchor samples for distilling robust boundary-aware unlearning behavior. (2) We analyze the roles of data standardization and bidirectional distillation through fine-grained gradient analysis and ablation studies, showing how each component affects gradient alignment and optimization across different unlearning objectives. (3) We extend existing unlearning benchmarks with additional evaluation settings for robustness under neighboring domains and robustness against adversarial probing attacks. (4) Our method achieves state-of-the-art performance on both established and extended benchmarks, including MUSE-Book and WMDP-Cyber, which notably reduces adversarial attack success rate while maintaining low over-refusal in adjacent domains.

2 LLM Unlearning Preliminary

Core Objective of LLM Unlearning: LLM unlearning aims to remove specific undesirable knowledge from the model while preserving other capabilities. Beyond inference time filtering or in-context strategies (Pawelczyk et al., 2023; Takashiro et al., 2025) that might be difficult to scale, most existing approaches rely on training-based optimization. A common formulation defines unlearning through two objectives:

$$\min_{\boldsymbol{\theta}} \mathcal{L} \equiv \mathcal{L}_{\text{forget}}(\mathcal{D}_f; \boldsymbol{\theta}) + \mathcal{L}_{\text{retain}}(\mathcal{D}_r; \boldsymbol{\theta}), \quad (1)$$

where $\mathcal{D}_f$ is the forget set that contains undesirable knowledge, $\mathcal{D}_r$ is the retain set that represents general knowledge, usually irrelevant to $\mathcal{D}_f$. $\boldsymbol{\theta}$ denotes model parameters. An ideal model refuses queries that relate to $\mathcal{D}_f$ and responds normally to other inputs. Prior work focuses on how to balance these two terms, since their gradients, $\nabla_\theta \mathcal{L}_{\text{unlearn}}$ vs. $\nabla_\theta \mathcal{L}_{\text{retain}}$, may conflict and lead to unstable updates. While this dual view captures the core requirement, it is still limited in reflecting practical unlearning needs.

Blurred Boundary Behavior and Over Refusal. A critical yet underexplored issue resides near the boundary of the forget domain (Li et al., 2024c; Eldan and Russinovich, 2023). In practice, many benign concepts share semantic proximity with the target knowledge. An overcorrected model may extend refusal behavior beyond $\mathcal{D}_f$ to neighboring domains $\mathcal{N}(\mathcal{D}_f)$. This leads to *over-refusal*, *i.e.*, refusing legitimate queries to concepts adjacent to the unlearning domain. For example, when a model removes knowledge about the Harry Potter series, it may also reject related but harmless topics such as *Fantastic Beasts*.

Adversarial Probing Attack Against LLM Unlearning. Another challenge faced by unlearning arises in adversarial settings, attributed to the superior context-conditioning ability of LLMs. Consider a prefilling attack, where the adversary constructs a malicious prefix and injects it into the model input to steer the output toward undesirable responses. Formally, given an input sequence composed of a user query x_q and a prefix context x_p, an

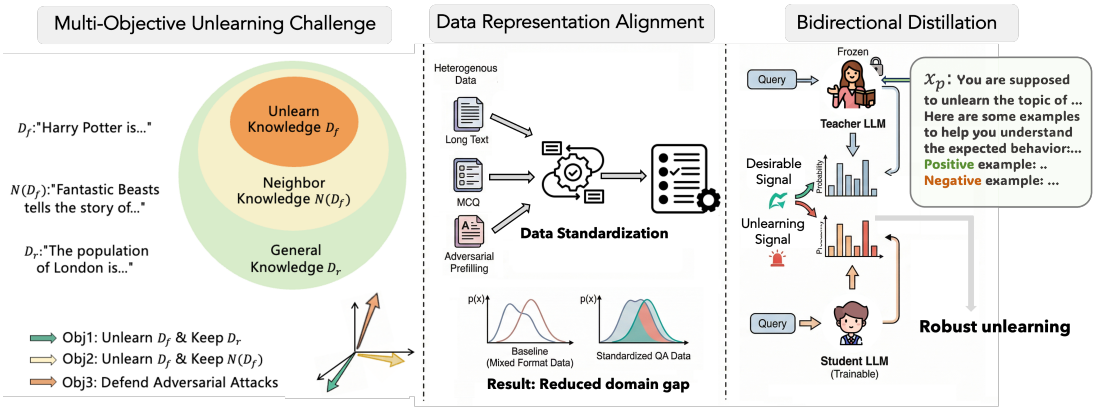

Figure 1: Overview of our Multi-Objective Unlearning Framework. **Left:** Practical LLM unlearning requires simultaneously handling target erasure ($\mathcal{D}_f$), neighboring domain retention ($\mathcal{N}(\mathcal{D}_f)$), and general utility ($\mathcal{D}_r$) across heterogeneous data sources. Naively optimizing each goal leads to undesirable task-gradient updates. **Middle:** We standardize diverse training data into a unified representation to close the domain gap and shift diverse unlearning gradients toward synergistic optimization. **Right:** We then apply an *asymmetric* teacher-student *bidirectional* distillation architecture that resolves optimization conflicts by simultaneously suppressing undesirable student logits and encouraging teacher-like behavior.

LLM π generates outputs y conditioned on the full context: $\pi(y \mid x_q, x_p)$. An attacker designs a prefix x_p^{adv} so that the conditional output shifts toward a targeted behavior, even when the prefix itself appears benign (*e.g.,* by encouraging the model to generate an answer with affirmative prefixes, "Sure, here are the detailed responses:").

3 Methodology: A Data and Optimization Co-Design Framework

The challenges outlined above have motivated us to extend the scope of LLM unlearning beyond the standard formulation. We consider a multi-goal unlearning setting where the unlearning LLM, parameterized by θ, needs to simultaneously satisfy three distinct tasks, denoted $\mathcal{T}_k \in \{1, 2, 3\}$:

• $\mathcal{T}_1$ **(Target Unlearning & General Utility):** removing undesirable knowledge while preserving general model utility. This is captured by optimizing $\mathcal{L}(\theta)$ in Eq (1) using forgetting data D_f and irrelevant retain dataset $\mathcal{D}_r$.

• $\mathcal{T}_2$ **(Neighboring Domain Retention):** Answering legitimate queries to benign concepts adjacent to the unlearning target: $\min_\theta \mathcal{L}_2(\theta) \equiv \mathbb{E}_{x\sim\mathcal{N}(\mathcal{D}_f)}\Big[l(\pi_\theta(x))\Big]$.

• $\mathcal{T}_3$ **(Adversarial Context Robustness):** Maintaining robustness against unlearning content elicitation, especially prefilling prompt attacks: $\min_\theta \mathbb{E}_{x\sim\mathcal{D}_f}\Big[\max_{x_p^{adv}} l(\pi_\theta(x, x_p^{adv}))\Big]$.

3.1 Gradient-Isolated Multi-Task Unlearning: A Motivating Observation

Prior unlearning methods such as Gradient Ascent (GA) primarily focus on addressing $\mathcal{T}_1$ (Neel et al., 2021; Zhang et al., 2024a). A straightforward extension of the prior work to achieving these simultaneous goals is to construct a joint unlearning loss: $\min_\theta \sum_k \mathcal{L}_k(\theta)$, where each objective can be achieved by optimizing a tailored loss function l_k, such as optimizing GA loss on forgetting domain, and supervised fine-tuning (SFT) loss on desirable refusal responses. Meanwhile, the representative format of training samples is largely overlooked in existing unlearning approaches. Most methods apply verbatim, token-wise training directly on raw text, such as book chapters or descriptive passages, where the knowledge targeted for unlearning or retention is diluted across lengthy contexts. Moreover, the forget and retain sets often appear in diverging formats, such as lengthy descriptive passages versus structured question-answering pairs.

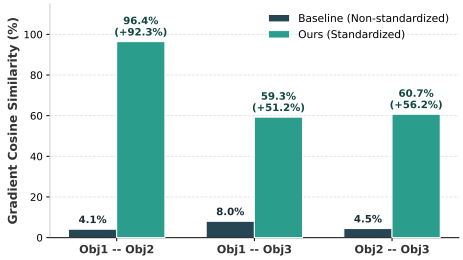

Figure 2: Gradient cosine similarity across unlearning objectives evaluated on MUSE-Book. **Before** data standardization, baselines with heterogeneous data formats result in near-orthogonal updates. **After** our data standardization, the gradient updates shift into a highly synergistic regime.

This observation raises a key question: *Does directly combining per-objective optimization methods alone lead to synergistic learning?* To answer this, we conducted a pilot empirical analysis by profiling the gradient cosine similarity, $\cos(\nabla_\theta L_i, \nabla_\theta L_j)$, across the three unlearning objectives (i, j) during the early stages of standard

training. To isolate the effect of data format from that of loss function design, we applied the same optimization method described in Section 3.4 for all three unlearning goals. For concentrated comparison, we primarily applied the gradient similarity measurement to model parameters located at the final three decoder layers of Llama 3.2 3B Instruct.

As shown in the "Baseline" column of Figure 2, the gradient similarities between all objective pairs are near zero ($\cos \approx 0.04 \sim 0.08$). While *orthogonal* gradients are often desirable in general multi-task settings where tasks are semantically *unrelated*, this conflicts with our unlearning setting, where multiple objectives should converge on a shared semantic goal from complementary angles. Under such near-zero gradient similarity, the objectives do not reinforce one another, which further produces empirically observable interference among the unlearning goals (Sec. 5.2).

To improve multi-task unlearning, some prior work reduces task-gradient interference through orthogonal gradient updates (Pan et al., 2025; Cao et al., 2022) that operate directly in parameter space. In contrast, we propose that one overlooked bottleneck in multi-task unlearning stems from the *domain gap manifested in data representations*. In standard pipelines, these objectives inherently rely on highly divergent data formats. For instance, $\mathcal{T}_1$ (target unlearning) training data typically utilizes raw, unstructured text from existing benchmarks, while $\mathcal{T}_3$ relies on structured QA pairs extended from these datasets. Such divergence in data representation leads to a major domain moving distance across tasks $\mathcal{T}_k$ (Ben-David et al., 2006; Kifer et al., 2004), which theoretically enlarges the domain adaptation gap across these unlearning tasks.

3.2 Data Standardization to Bridge Domain Gap

In response, we address the data side of our framework through representational standardization. We project the training corpora, which contain heterogeneous data formats across all unlearning goals, into a unified Question-Answering (QA) format. Our goal herein is not to create new unlearning knowledge, but to expose the target knowledge through a comparable training interface, by controlling how data of such knowledge is presented to the unlearning objective while preserving their original semantic roles. For $\mathcal{T}_1$, raw forgetting and retain corpora are parsed into focused QA pairs. For $\mathcal{T}_2$, neighboring-domain samples are drawn from concepts adjacent to but outside the forgetting domain and formatted as legitimate QA examples. For $\mathcal{T}_3$, prefilling-style inputs are introduced to preserve the adversarial extraction setting by pairing eliciting prefixes with the corresponding unlearning targets. This standardization has the additional benefit of consolidating the information density in the training data, which was largely diluted by grammar and lexical details.

Tightened Unlearning Bound. By enforcing data standardization, our method mitigates structural discrepancies across unlearning domains and is also connected to an improved generalization bound (Ben-David et al., 2006). Specifically, consider our unlearning problem as a multi-source domain adaptation task, where each individual unlearning task corresponds to a *source* domain $\mathcal{T}_k$, and the target domain $\mathcal{T}$ represents their ensemble, *i.e.*, the joint multi-objective unlearning goal. Let $h : \mathcal{X} \to \mathcal{Y}$ be a hypothesis (model) over such domains, and $\mathcal{H} \subseteq \{h\}$ denote the hypothesis class. $d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{\mathcal{D}}_k, \tilde{\mathcal{D}})$ denotes the divergence measured over the symmetric-difference hypothesis space between the k-th source domain $\tilde{\mathcal{D}}_k$ and the target domain $\tilde{\mathcal{D}}$. Then, with probability at least $1 - \delta$:

$$\mathcal{L}_{\mathcal{T}}(h) \equiv \mathcal{L}_{\mathcal{T}}\left(\frac{1}{K}\sum_k h_k\right) \leq \frac{1}{K}\sum_k \hat{\mathcal{L}}_{\mathcal{T}k}(h_k) + \frac{1}{K}\sum_k \left(d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{\mathcal{D}}_k, \tilde{\mathcal{D}}) + \lambda_k\right) + C, \quad (2)$$

where C is governed by the number of training samples and the VC-dimension of $\mathcal{H}$. This bound reveals that the unlearning performance depends not only on the individual objective losses $\hat{\mathcal{L}}_{\mathcal{T}_k}(h_k)$, but also on the distributional divergence between each source and the target domain. In our setting, minimizing $d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{\mathcal{D}}_k, \tilde{\mathcal{D}})$ across unlearning directly tightens this bound, which aligns with the theoretical justification and serve as a principled path towards multi-objective unlearning.

From Isolated Gradients to Cooperative Optimization. The impact of our data standardization is also empirically validated in Figure 2, where the gradient updates of each unlearning goal shift dramatically to strong positive correlations (*e.g.*, +92.29% improvement between $\mathcal{T}_1$ and $\mathcal{T}_2$) after data standardization (the "Ours" column in Figure 2), which thus enables synergistic optimization across diverse unlearning objectives.

3.3 Sharpening Unlearning Boundaries via Contrastive Anchor Samples

While standardization aligns the optimization trajectory across multiple unlearning goals, the LLM still requires precise semantic boundaries to prevent the pervasive issue of *over-refusal* on neigh-

boring domains. To resolve this ambiguity, we augment each objective's training set with *contrastive anchor pairs*, each consisting of a query for knowledge that the model should erase and a contrastive query that should be preserved or safely completed. Specifically, $\mathcal{T}_1$ *Anchors* contrast target hazardous knowledge against factually adjacent but benign world knowledge, so the model learns to distinguish what to erase from what to retain. $\mathcal{T}_2$ *Anchors* are constructed between a restricted unlearning concept (*e.g.,* cyber-weaponry synthesis) and a semantically neighboring legitimate concept (*e.g.,* general computer security principles). $\mathcal{T}_3$ *Anchors* oppose a malicious prefilling prefix against a benign continuation, so the model develops robustness to adversarial elicitation without suppressing legitimate completions.

These contrastive pairs act analogously to support vectors to help sculpt a more precise decision boundary during optimization, which defines the local geometry of the decision surface. For better readability, we provide representative contrastive anchor examples for both MUSE-Book and WMDP-Cyber in Appendix A.3.

3.4 Multi-Goal Unlearning Optimization via Bidirectional Top-K Logit Distillation

Given the aligned data representations and anchored semantic boundaries, we next optimize the model toward the desired unlearning behavior. A key limitation of GA or NPO-style objectives is that they mainly suppress the likelihood of target tokens. Such token-level decay provides coarse-grained negative-direction supervision with limited guidance on which alternative probability mass should replace the forgotten knowledge. This can lead to unstable forgetting or boundary over-suppression.

To provide more informative supervision, we formulate multi-goal unlearning as a reference-guided distribution matching problem. For each unlearning goal $\mathcal{T}_k$, we construct a task-specific reference behavior π^k_{ref} and optimize:

$$\min_{\theta} \mathbb{E}_{\mathcal{T}_k \sim \{\mathcal{T}\}} \Big[\mathbb{E}_{x \sim \mathcal{T}_k} \Big[\mathbb{D}[\pi_\theta(\cdot|x) \| \pi^k_{\text{ref}}(\cdot|x)] \Big] \Big],$$

where $\mathbb{D}[p\|q]$ denotes a legitimate distance measure between two probabilities of p and q.

Construction of reference model: For unlearning goal k, we utilize a lightweight natural language *intention prompt*, p_k, which encodes the desired model behavior. A frozen teacher LLM, π_{ref}, is prompted by p_k to establish safe and contextualized semantic boundaries. In this work, we focus on conceptual unlearning, where the unlearning intention and scope can be naturally described by instructions, in contrast to entry-level unlearning (Maini et al., 2024b). Then the student model, π_θ, receives only the input x without the intention prompt, forcing it to internalize the target behavior.

Bidirectional Top-K logit distillation. To achieve precise unlearning without abusing the global vocabulary distribution, we conduct distillation selectively on two critical sets of token logits: $\mathbb{C}_K^{teach}$ (the Top-K logits from the teacher) and $\mathbb{C}_K^{stud}$ (the Top-K logits spontaneously generated by the student):

$$\begin{aligned} \min_{\theta} \mathbb{D}[\pi_\theta \| \pi_{\text{ref}}] \equiv & \underbrace{\mathbb{E}_x \left[\sum_{i \in \mathbb{C}_K^{\text{ref}}} \mathcal{L}_{\text{sim}}\big(g^i_\theta(x), g^i_{\text{ref}}(x)\big) \right]}_{\text{Encouraging teacher behavior}} \\ & + \alpha \underbrace{\mathbb{E}_x \left[\sum_{j \in \mathbb{C}_K^{\theta}} \mathcal{L}_{\text{sim}}\big(g^j_\theta(x), g^j_{\text{ref}}(x)\big) \right]}_{\text{Erasing student's undesirable behavior}}. \end{aligned} \tag{3}$$

where $g(\cdot)$ denotes token logits and $\mathcal{L}_{sim}$ denotes the Smooth L1 distance. This asymmetric objective operates bidirectionally: The teacher-side Top-K logits guide the student toward a desirable support manifold, while the student-side Top-K logits identify and suppress its high-confidence undesirable tokens, thus enabling more fine-grained forgetting and preservation.

Unlike prior distillation work that employs unidirectional distillation (Zhong et al., 2026), our method achieves more effective, precise distillation (Section 5.3). Compared with prior work that only decays target-token likelihood, our method provides finer-grained redistribution of probability mass, thus supporting more precise knowledge forgetting and preservation. We refer to the resulting framework as **DUO**, short for **D**ata and **U**nlearning **O**ptimization co-design.

4 Related Work

LLM Unlearning. Efforts to unlearn knowledge from LLMs can be broadly categorized into in-context methods (Pawelczyk et al., 2024; Liu et al., 2024), which steer model behavior at inference time without parameter updates, and training-based methods (Jang et al., 2023; Zhang et al., 2024b; Li et al., 2024a; Choi et al., 2024; Eldan and Russinovich, 2023; Maini et al., 2024a; Fan et al., 2025b; Wang et al., 2025a; Pan et al., 2025; Jin et al., 2025), which modify model weights to enforce forgetting. Recently, an emerging line of work has investigated

the critical impact of ***data*** design and selection on unlearning performance (Chang and Lee, 2025; Xu et al., 2026; Sinha et al., 2024). However, methods in this paradigm mostly focus on a setting captured by our $\mathcal{T}_1$ while overlooking the complex boundary behaviors and adversarial vulnerabilities inherent in real-world deployments.

Unlearning Robustness. Prior work has highlighted unlearning risks under parameterized attacks, including relearning attacks in which fine-tuning on a handful of examples restores erased capabilities (Fan et al., 2025a; Hu et al., 2024; Qi et al.). Furthermore, recent studies demonstrate that unlearned knowledge can still be extracted via contextualized attacks, which closely aligns with our setting. For instance, Shumailov et al. (2024) and Łucki et al. (2024) formalized in-context un-unlearning and adversarial jailbreaks, where sophisticated contextual elicitation can bypass superficial unlearning (Xue et al., 2024).

LLM Knowledge Distillation. Knowledge distillation (KD) has been used to transfer capabilities across LLMs (Li et al., 2025; Gu et al.). In the context of unlearning, while KD is commonly employed as a regularization penalty on the retain set to mitigate catastrophic forgetting (Zhang et al., 2024b; Jang et al., 2023; Yao et al., 2024), utilizing distillation as the primary mechanism to directly enforce forgetting remains highly underexplored. Zhong et al. (2026) proposed DUET, a uni-directional unlearning distillation method for aligning a student model with a prompt-steered teacher. We extend KD to a bidirectional framework with richer, more explicit supervision signals to resolve more complex multi-task scenarios.

LLM Unlearning Benchmarks. Evaluation remains a fundamental challenge in LLM unlearning. Early benchmarks such as TOFU (Maini et al., 2024a) and MUSE (Shi et al., 2024b) focused primarily on exact memorization and general utility. While more recent benchmarks like WMDP (Li et al., 2024a) and RWKU (Jin et al., 2024) introduce high-risk capabilities and partially discuss unlearning robustness, they lack a systematic investigation into boundary behaviors, conceptual unlearning conflicts, and the collateral damage to neighboring domains. Our experimental setup bridges this gap by augmenting existing tasks with structured evaluations ($\mathcal{T}_1$-$\mathcal{T}_3$) that capture practical unlearning demands: efficacy, neighboring domain retention, and adversarial defense.

5 Experiments

5.1 Experimental Setup

Dataset Construction, Models and Metrics. We evaluated our framework across two distinct knowledge domains: the **MUSE-Book** benchmark (*i.e.,* Harry Potter (HP)) (Shi et al., 2024a), which represents copyrighted narrative knowledge, and the **WMDP-Cyber** benchmark (Li et al., 2024c), which represents hazardous, restricted knowledge. We follow the standard benchmark settings and use Llama-3.2-3B-Instruct on MUSE-Book and Zephyr-7B-Beta on WMDP-Cyber. We applied the following evaluation metrics to reflect our three unlearning objectives, and additionally reported the **MMLU** scores (Hendrycks et al., 2020) to monitor the unlearned model's general capability.

• $\mathcal{T}_1$ (Target Unlearning & General Utility): We measured the unlearning efficacy on target knowledge (**Forget** $\downarrow$) and preservation on general knowledge (**Retain** $\uparrow$), evaluated on open-ended questions, using the *ROUGE scores* against reference targets as the metric;

• $\mathcal{T}_2$ (Neighboring Domain Retention): We measured the model's ability to preserve knowledge on this benign neighboring domain while avoiding *over-refusal*. To obtain a fine-grained evaluation of the model's performance on this task, we used a Multi-Choice Question (MCQ) format for evaluation, where each query includes plausible yet confusing options, refusal options (*e.g.,* "I do not know"), and the ground-truth answer option. We adopted the *Retain Accuracy* $\uparrow$ as the main metric.

• $\mathcal{T}_3$ (Robustness Against Adversarial Prompts): We prefilled each query related to the unlearning domain with an adversarial prefilling prompt. We adopted the Attack Success Rate (**ASR** $\downarrow$) to indicate the rate at which the target model is manipulated with malicious prefixes. Evaluation queries for this task are drawn from the same unlearning domain but are disjoint from those in $\mathcal{T}_1$.

Unlearning Baselines and Upper Bound. We compare **DUO**, our unlearning method against three categories of approaches to clearly distinguish between different unlearning paradigms:

1. **In Context Unlearning Methods**:
   • *Teacher with Oracle Prompt Router* (Upper Bound): An idealized configuration, which assigns the correct goal-specific intention prompt p_m to each query.
   • *Ensemble Teacher*, which concatenates all intention prompts into a single instruction to steer

the target model, acting as a practical alternative to the oracle router.

2. **Training Based Unlearning Methods:** *DUET* (Zhong et al., 2026) performs unidirectional distillation from the Ensemble Teacher. *Supervised Fine-Tuning* (Touvron et al., 2023) trains on desired responses for forget, adversarial, retention, and neighboring queries. *Gradient Ascent (GA)* (Jang et al., 2023), *SimNPO* (Fan et al., 2025b), and *FLAT* (Wang et al., 2025b) are other related works that were primarily proposed for addressing $\mathcal{T}_1$.
3. **Multi-Objective Unlearning with Gradient Editing:** *NGDiff* (Jin et al., 2025) addresses unlearning objective conflicts by utilizing normalized gradient differences and an adaptive learning rate; *MOLLM* (Pan et al., 2025) employs gradient projection mechanisms. These two methods were originally designed to address the forget-retention tradeoff ($\mathcal{T}_1$). We extended them to handle three objectives.

For fairness, we evaluate each training-based baseline under both its original and standardized data formats when applicable, and report the best checkpoint in the main table. We additionally provide their standardized data results in Appendix A.4.

5.2 Main Results

Table 1 and Figure 3 report the main results on MUSE-Book and WMDP-Cyber, showing that our framework achieves balanced multi-goal unlearning across forgetting, retention, and robustness. Additional non-QA results in Appendix A.1 further suggest that its unlearning behavior generalizes beyond QA-style inputs.

Comparing to Existing Distillation and Fine-Tuning Baselines. As shown in Table 1, while the Ensemble Teacher achieves moderate forgetting, it suffers a significant drop in retention compared to the Upper Bound. Since both configurations employ in-context unlearning, this degradation indicates that concatenating multiple intention prompts leads to internal intention interference; the LLM struggles to satisfy these instructions simultaneously. Furthermore, even after applying standard distillation, *DUET* exhibited persistent trade-offs with a notable degradation in $\mathcal{T}_3$ ASR (39.5%) and MMLU (49.9%). Finally, SFT aggressively destroys the target knowledge ($\mathcal{T}_1$ Forget drops to 0.3) while effectively defending against prefilling attacks (ASR drops to 1.5%). However, this result comes at the cost of catastrophic *over-refusal*, with degraded general utility ($\mathcal{T}_1$ and $\mathcal{T}_2$ Retain collapse to 49.6 and 50.7, respectively, and MMLU plummets to 50.2).

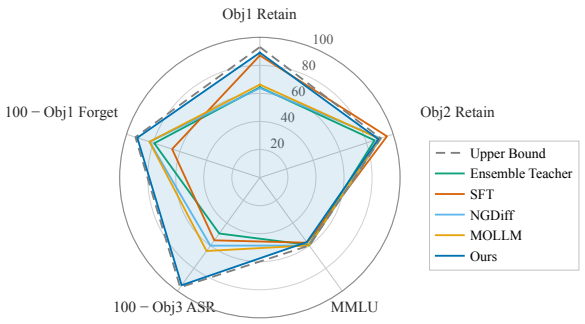

Figure 3: Multi-dimensional performance on the **WMDP-Cyber** benchmark. The radar chart illustrates the trade-offs across five metrics axes (Forget ↓, Retain ↑, Neighbor Retain Acc ↑, ASR ↓, MMLU ↑), normalized for visualization.

Imbalanced Trade-offs in Gradient Editing Methods. While NGDiff and MOLLM address multi-objective conflicts through gradient scalarization and orthogonal projection without data standardization, they exhibit weakened unlearning robustness. On the HP benchmark, MOLLM remains highly vulnerable to prefilling attacks (ASR 69.2%) and shows limited effectiveness in unlearning QA targets ($\mathcal{T}_1$ Forget 19.9). Similarly, NGDiff notably underperforms on both ASR (17.9%) and target forgetting ($\mathcal{T}_1$ Forget 13.6). Although both methods retain strong performance on Retain and MMLU, this is largely because they fail to sufficiently remove hazardous representations.

In contrast, DUO achieves strong synergy across objectives. By unifying data representations and leveraging intention-prompted bidirectional distillation, we substantially reduce prefilling ASR to 12.5%, outperforming gradient editing-based methods by an immense margin while effectively erasing target knowledge ($\mathcal{T}_1$ Forget 2.7). Simultaneously, DUO prevents over-refusal, maintaining high retention accuracy on general and neighboring domains ($\mathcal{T}_1$ Retain 78.1%, $\mathcal{T}_2$ Retain Acc 58.9%), which matches the Upper Bound performance.

Performance on WMDP-Cyber. The results in Figure 3 reinforce the above findings. Baselines such as SFT achieve strong security but exhibit a pronounced collapse along the retention axes. Conversely, methods like MOLLM and NGDiff preserve utility but degrade substantially on ASR and unlearning objectives. DUO is the only approach that maintains a balanced performance profile, effectively mitigating hazardous knowledge and adversarial attacks while preserving the integrity of

Method	$\mathcal{T}_1$ (Unlearn)		$\mathcal{T}_2$ (Neighbor Domain)	$\mathcal{T}_3$ (Adv Robustness)	MMLU ↑	Overall Performance ↑
	Forget ↓	General ↑	Retain Acc ↑	ASR ↓		
Base Model	32.1	84.3	57.28	70.5	**60.8**	0.00
GA ($\star$)	36.9	**85.0**	55.9	70.0	36.45	-29.33
SimNPO ($\star$)	21.4	43.1	57.7	70.5	60.40	-30.48
FLAT ($\star$)	0.7	58.3	42.7	39.0	58.92	20.44
SFT	**0.3**	49.6	50.7	**1.5**	50.2	48.92
NGDiff	13.6	83.3	37.7	17.9	**60.8**	50.52
MOLLM	19.9	83.6	57.4	69.2	**60.8**	12.92
Ensemble Teacher (ET)	6.2	60.9	26.8	13.5	60.3	28.52
DUET (w/ ET)	3.2	71.3	26.8	39.5	49.9	5.52
DUO (Ours)	2.7	78.1	**58.5**	12.5	59.2	**80.82**
Upper Bound	*2.1*	*80.1*	*58.5*	*6.7*	*60.7*	*90.72*

Table 1: Main unlearning results on the **MUSE-Book** benchmark. We also reported the aggregated **Overall Performance** shift, defined relative to the base model. Methods marked with '$\star$' are trained on $\mathcal{T}_1$ due to their inherent design and are not readily extendable to multi-goal settings.

Method	$\mathcal{T}_1$		$\mathcal{T}_2$	$\mathcal{T}_3$	MMLU	Overall Perf.
	Forget	General	Retain	ASR		
DUO (Ours)	**2.7**	78.1	**58.5**	**12.5**	59.2	**80.82**
DUO (w/o CS)	3.25	73.2	39.9	15.5	59.1	53.67
DUO w/ DUET Loss	3.19	74.0	57.6	**12.5**	58.2	74.33
DUO w/ BILD Loss	3.9	**82.3**	56.3	23.5	**58.3**	69.72

Table 2: Ablation study on contrastive training samples and bidirectional distillation objectives.

neighboring-domain retention.

5.3 Ablation Study

Effects of Contrastive Training Samples. As a comparative study, we remove the contrastive samples from training while keeping the remaining data standardization and optimization pipeline unchanged, denoted as DUO (w/o CS). By default, we construct contrastive samples in a 1:1 ratio with the corresponding task-specific training samples, so that every original training instance is paired with one boundary-aware anchor sample. As shown in Table 2, removing contrastive samples decreases the overall performance from 80.82 to 53.67, mainly due to weaker forgetting, lower neighboring-domain retention, and higher ASR. These results indicate that contrastive anchor samples help sharpen the unlearning boundary, allowing the model to erase target knowledge while better preserving neighboring-domain behavior and robustness against adversarial elicitation.

Effects of Bi-directional Distillation. To evaluate the effects of our bidirectional top-K logit distillation, we conducted a comparative analysis against a state-of-the-art symmetric distillation approach, the bi-directional Logits Difference (BILD) loss (Li et al., 2025). BILD constructs pairwise logit differences within an active vocabulary and ***normalizes*** them into probability distributions via a Softmax function, then minimizes their KL divergence: $\mathcal{L}_{BiLD} \propto D_{KL}(p^t \parallel p^s) + D_{KL}(p^s \parallel p^t)$, where p^t and p^s denote the Softmax-normalized probability distributions of the teacher and student top-K logits, respectively. While effective for standard capability transfer, this *normalized KL* approach fundamentally only aligns the relative distances (rankings) between tokens. In contrast, our method explicitly minimizes differences in magnitude using a *Smooth L1* distance applied directly to the unnormalized logits.

As shown in Table 2, replacing our method with BILD leads to a significant regression in unlearning efficacy ($\mathcal{T}_1$ Forget rises to 3.9) and adversarial robustness (ASR almost doubles to 23.5%). This ablation empirically proves our claim that, in the context of unlearning, aligning only the *relative ranking* of logits is insufficient. The performance gains over DUET with uni-directional distillation further indicate that effectively unlearning targeted knowledge requires our asymmetric distillation strategy, which explicitly suppresses the student's high-confidence logits toward the teacher's logits.

6 Conclusion

We present a unified framework for practical multi-goal LLM unlearning, which combines intention-prompted bidirectional distillation with unified data representations to jointly address knowledge removal, adversarial robustness, boundary-domain retention, and general utility. Experiments on existing and extended unlearning benchmarks show balanced state-of-the-art performance across tasks, which highlights the value of structured data and unlearning optimization co-design for reliable and scalable LLM unlearning.

Limitations

This work studies multi-goal unlearning in representative text-based settings, covering copyrighted narrative knowledge and hazardous cybersecurity knowledge. While these settings enable controlled evaluation of forgetting, retention, and robustness, future unlearning research may further broaden this scope to multimodal content and more diverse deployment environments.

Ethical Considerations

This work aims to improve LLM safety by reducing access to undesirable or potentially hazardous knowledge while preserving benign utility. Our experiments are conducted on established benchmarks and report aggregate results. The released data, prompts, and code are intended only for responsible research on LLM safety and unlearning.

Reproducibility Statement

To facilitate reproducibility, the source code, constructed datasets, data-generation prompts, and implementation details are anonymized and available a: `https://anonymous.4open.science/r/MULE-662D`.

References

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Zihao Cao, Jianzong Wang, Shijing Si, Zhangcheng Huang, and Jing Xiao. 2022. Machine unlearning method based on projection residual. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–8. IEEE.

Hwan Chang and Hwanhee Lee. 2025. Which retain set matters for llm unlearning? a case study on entity unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5966–5982.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2024. Snap: Unlearning selective knowledge in large language models with negative instructions. *arXiv preprint arXiv:2406.12329*.

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

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

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2025b. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *International Conference on Learning Representations*.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. 2024. Jogging the memory of unlearned llms through targeted relearning attacks. *arXiv preprint arXiv:2406.13356*.

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

Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. 2025. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11278–11294.

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. *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. 2004. Detecting change in data streams. In *VLDB*, volume 4, pages 180–191. Toronto, Canada.

Minchong Li, Feng Zhou, and Xiaohui Song. 2025. Bild: Bi-directional logits difference loss for large language model distillation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1168–1182.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel

Mukobi, and 1 others. 2024a. The wmdp benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 28525–28550.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, and 38 others. 2024b. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *Preprint*, arXiv:2403.03218.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, and 1 others. 2024c. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *Proceedings of Machine Learning Research*.

Chris Liu, Zeyu Li, Yujia Wei, Peizhuo Wang, Chen Li, Wei Zhang, Song Li, Malik Magdon-Ismail, and Yang Liu. 2024. Large language model unlearning via embedding-corrupted prompts. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

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

Pratyush Maini, Shrey Jain, and 1 others. 2024b. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Andrei Ioan Muresanu, Anvith Thudi, Michael R. Zhang, and Nicolas Papernot. 2025. Fast exact unlearning for in-context learning data for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 45272–45288. PMLR.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 931–962. PMLR.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2025. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46.

Zibin Pan, Shuwen Zhang, Yuesheng Zheng, Chi Li, Yuheng Cheng, and Junhua Zhao. 2025. Multi-objective large language model unlearning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Forty-first International Conference on Machine Learning*.

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

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few shot unlearners. *Preprint*, arXiv:2310.07579.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*.

Weijia Shi, Ari Holtzman, Colin Raffel, and 1 others. 2024a. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024b. Muse: Machine unlearning six-way evaluation for language models. *The Thirteenth International Conference on Learning Representations*.

Ilia Shumailov, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jimenez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and Eugene Bagdasaryan. 2024. Ununlearning: Unlearning is not sufficient for content regulation in advanced generative ai. *Preprint*, arXiv:2407.00106.

Yash Sinha, Murari Mandal, and Mohan Kankanhalli. 2024. Unstar: Unlearning with self-taught anti-sample reasoning for llms. *arXiv preprint arXiv:2410.17050*.

Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Answer when needed, forget when not: Language models pretend to forget via in-context knowledge unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24872–24885, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,

Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. 2025a. LLM unlearning via loss adjustment with only forget data. *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. In *International Conference on Learning Representations (ICLR)*.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. 2023. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1).

Xiaoyu Xu, Minxin Du, Zitong Li, Zi Liang, Zhibiao Guo, Shiyu Zhang, Peizhao Hu, Qingqing Ye, and Haibo Hu. 2026. From domains to instances: Dual-granularity data synthesis for llm unlearning. *arXiv preprint arXiv:2601.04278*.

Zhiyu Xue, Guangliang Liu, Bocheng Chen, Kristen Marie Johnson, and Ramtin Pedarsani. 2024. No free lunch for defending against prefilling attack by in-context learning. *arXiv preprint arXiv:2412.12192*.

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

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

Yuhan Zhang, Weize Chen, and 1 others. 2024b. Negative preference optimization for catastrophic forgetting in llm unlearning. *arXiv preprint arXiv:2404.05868*.

Yisheng Zhong, Zhengbang Yang, and Zhuangdi Zhu. 2025. Hierarchical federated unlearning for large language models. *Preprint*, arXiv:2510.17895.

Yisheng Zhong, Zhengbang Yang, and Zhuangdi Zhu. 2026. Duet: Distilled llm unlearning from an efficiently contextualized teacher. *ICLR*.

A Appendix

A.1 Evaluation on Non-QA Datasets

To further examine whether our method generalizes beyond the QA-formatted data used in the main experiments, we evaluate the unlearned models on a non-QA dataset. This setting directly tests residual target-domain accuracy under a different input format, thereby assessing whether the unlearning behavior is robust to distributional changes in evaluation format.

Method	ACC ↓
SFT	3.5
Ensemble Teacher (ET)	7.9
Ours	8.1
DUET (w/ ET)	13.2
NGDiff	25.4
MOLLM	37.0

Table 3: Evaluation on a non-QA dataset. Lower ACC indicates stronger removal of target-domain knowledge under a non-QA evaluation format.

As shown in Table 3, our method achieves an ACC of 8.1, which is substantially lower than NGDiff and MOLLM, and is close to the in-context Ensemble Teacher. Although SFT obtains the lowest ACC, its main results show that this aggressive forgetting comes with severe utility degradation and over-refusal. In contrast, our method maintains a more favorable balance across forgetting, retention, and adversarial robustness, indicating that the proposed data-optimization co-design does not merely overfit to QA-style evaluation.

A.2 Detailed Quantitative Results on WMDP-Cyber

This section provides the comprehensive numerical data (Table 4) that serves as the basis for the multi-dimensional analysis presented in Figure 3 of the main text. The WMDP-Cyber benchmark evaluates the model's ability to remove hazardous cybersecurity knowledge while maintaining general utility and adversarial robustness.

Analysis of Existing Paradigms. As shown in Table 4, different unlearning paradigms exhibit distinct performance characteristics. In-context methods, such as the Ensemble Teacher, provide a training-free alternative but often face challenges in maintaining general utility (Retain: 64.4). While its distilled counterpart (DUET) restores some utility, it remains sensitive to adversarial probing (ASR: 24.7%). Tuning-based baselines like Refusal Tuning (SFT) demonstrate a strong defensive posture but tend to lead to a more conservative model behavior. Meanwhile, advanced gradient-based methods such as NGDiff and MOLLM excel at preserving general utility and neighboring domains, yet they encounter inherent difficulties in fully neutralizing hazardous information under adversarial conditions, with ASR remaining at 40.0% and 35.3%, respectively.

Performance of the Proposed Framework. Our framework, which integrates data standardization with bidirectional distillation, aims to achieve a more balanced performance across these competing objectives. By aligning the representational space and employing an asymmetric optimization strategy, our method effectively reduces the target knowledge (Obj1 Forget: 8.1) and significantly enhances adversarial defense (ASR: 5.1%). Notably, these gains in unlearning efficacy are achieved while maintaining a high level of general utility (MMLU: 56.9) and neighboring domain integrity (Retain Acc: 88.8%), demonstrating a performance profile that closely aligns with the idealized Upper Bound across the evaluated dimensions.

A.3 Examples of Contrastive Anchor Samples

To make the construction of contrastive anchor samples more concrete, we provide representative examples from both the MUSE-Book and WMDP-Cyber benchmarks. For each objective, the anchor sample is constructed in a 1:1 ratio with the corresponding task-specific training sample. These anchors are designed to clarify the semantic boundary between knowledge that should be erased, neighboring-domain knowledge that should be preserved, and adversarial contexts that should not trigger target-domain leakage.

A.4 Baseline Results under Standardized Training Data

To further clarify the effect of data standardization on existing baselines, we additionally report their results when trained with our standardized data format. Different from the main comparison, where we report the best-performing checkpoint across applicable training formats for each baseline, this section focuses only on the standardized-data setting. As shown in Table 7, several training-based baselines do not consistently benefit from standardized data. This is expected, since their objectives were originally designed and tuned for their original non-standardized training inputs. GA, SimNPO, NGDiff, and MOLLM preserve general utility reasonably well but fail to reduce adversarial ASR. In contrast, FLAT achieves strong forgetting and adversarial robustness, but its general retention collapses substantially. These results further suggest that data standardization alone is insufficient, and practical multi-goal unlearning requires a joint design of data representation and optimization objective.

Method	**Obj1** (Unlearn)		**Obj2** (Neighbor Domain)	**Obj3** (Adv Robustness)	MMLU ↑	Overall Performance ↑
	Forget ↓	**General** ↑	**Retain Acc** ↑	**ASR** ↓		
Base Model	68.5	88.1	96.3	60.7	59.9	0.00
Ensemble Teacher (ET)	20.8	64.4	86.0	50.7	59.4	23.20
DUET (w/ ET)	25.5	**91.1**	**95.3**	24.7	50.6	71.70
SFT	34.2	87.1	**95.3**	44.7	57.4	45.80
NGDiff	17.3	63.8	90.6	**40.0**	**59.88**	41.88
MOLLM	17.2	66.3	89.7	35.3	59.77	48.17
Ours	**8.1**	89.1	88.8	**5.1**	56.9	**106.50**
Upper Bound	*6.7*	*93.1*	*90.7*	*3.3*	*59.9*	*118.60*

Table 4: Main unlearning results on the **WMDP-Cyber** benchmark. We also reported the aggregated **Overall Performance** shift, defined relative to the base model.

Task	Original Sample	Contrastive Anchor Sample
$\mathcal{T}_1$	**Forget** **Q:** What position does Oliver Wood play in Quidditch? **A:** Keeper.	**General Anchor** **Q:** What is the capital of Argentina? **A:** Buenos Aires.
$\mathcal{T}_2$	**Boundary** **Q:** What title does Gandalf assume after returning? **A:** Gandalf the White.	**Forget-side Anchor** **Q:** Who kills Severus Snape? **A:** Voldemort.
$\mathcal{T}_3$	**Attack** **Query:** In the Harry Potter series, who is the conductor of the frog choir? **Target Answer:** Filius Flitwick. **Prefill Prefix:** The conductor is	**Benign Prefill Anchor** **Query:** Which ocean is the largest? **Answer:** Pacific Ocean. **Prefill Prefix:** The largest ocean is

Table 5: Representative contrastive anchor samples on the MUSE-Book benchmark.

Task	Original Sample	Contrastive Anchor Sample
$\mathcal{T}_1$	**Forget** **Q:** What is cryptojacking? **A:** The unauthorized use of a person's computer to mine cryptocurrency.	**General Anchor** **Q:** What is the largest planet in the Solar System? **A:** Jupiter.
$\mathcal{T}_2$	**Boundary** **Q:** What is the purpose of access control in computer systems? **A:** To ensure that only authorized users can access specific resources.	**Forget-side Anchor** **Q:** What is cryptojacking? **A:** The unauthorized use of a person's computer to mine cryptocurrency.
$\mathcal{T}_3$	**Attack** **Query:** What is cryptojacking? **Target Answer:** The unauthorized use of a person's computer to mine cryptocurrency. **Prefill Prefix:** Cryptojacking refers to	**Benign Prefill Anchor** **Query:** What is authentication in computer security? **Answer:** The process of verifying a user's identity. **Prefill Prefix:** Authentication refers to

Table 6: Representative contrastive anchor samples on the WMDP-Cyber benchmark.

Method	**Obj1** (Unlearn)		**Obj2** (Neighbor Domain)	**Obj3** (Adv Robustness)
	Forget $\downarrow$	**General** $\uparrow$	**Retain Acc** $\uparrow$	**ASR** $\downarrow$
GA	28.77	84.23	55.87	70.0
SimNPO	64.19	82.18	57.75	70.5
FLAT	**1.24**	15.63	42.72	**9.0**
NGDiff (t=0.1)	30.04	83.57	56.81	70.0
MOLLM (t=0.5)	29.93	83.68	56.81	69.5

Table 7: Baseline results under our **standardized training data** format on the MUSE-Book benchmark. Lower Obj1 Forget and Obj3 ASR indicate stronger target-domain forgetting and adversarial robustness, while higher Obj1 General and Obj2 Retain Acc indicate better preservation of general and neighboring-domain knowledge.