

BalDRO: A Distributionally Robust Optimization based Framework for Large Language Model Unlearning

Pengyang Shao
National University of Singapore
Singapore

Naixin Zhai
University of Science and Technology
of China
Hefei, Anhui, China

Lei Chen
University of Science and Technology
of China
Hefei, Anhui, China

Yonghui Yang
National University of Singapore
Singapore

Fengbin Zhu
National University of Singapore
Singapore

Xun Yang
University of Science and Technology
of China
Hefei, Anhui, China

Meng Wang
Hefei University of Technology
Hefei, Anhui, China

Abstract

As Large Language Models (LLMs) increasingly shape online content, how to remove targeted information from well-trained LLMs (also known as LLM unlearning) has become increasingly critical for web governance. A key challenge in LLM unlearning lies in the sample-wise imbalance within the forget set: different samples exhibit widely varying unlearning difficulty, leading to asynchronous forgetting speeds where some knowledge remains insufficiently erased while others become over-forgotten. To address this challenge, we propose BalDRO, a novel and efficient framework for balanced LLM unlearning. BalDRO formulates unlearning as a min-sup process, where the inner process identifies a worst-case data distribution that adaptively emphasizes hard-to-unlearn samples, while the outer process updates model parameters based on the worst-case data distribution. We instantiate this formulation through two efficient variants: BalDRO-G, a discrete GroupDRO-based approximation that focuses on high-loss subsets, and BalDRO-DV, a continuous Donsker-Varadhan dual method that enables smooth, adaptive weighting within standard LLM training pipelines. Extensive experiments on the TOFU and MUSE benchmarks demonstrate the effectiveness of our proposed BalDRO, yielding significant improvements in both forgetting quality and model utility over existing methods. For reproducibility, we have anonymously released the code for BalDRO¹.

Keywords

Large Language Models, Machine Unlearning, Trustworthy AI

1 Introduction

As Large Language Models (LLMs) become increasingly embedded in web platforms and services [4, 12, 33], ensuring that these models behave in a trustworthy and responsible manner has become essential for maintaining the reliability of web-based information ecosystems. A key aspect of achieving such reliability is the ability to remove outdated, incorrect, or privacy-sensitive knowledge from

LLMs so that their behavior remains aligned with user expectations and public values [6, 37]. This need is further reinforced by legal frameworks such as the General Data Protection Regulation (GDPR)² and the California Consumer Privacy Act (CCPA)³, which mandate the “right to be forgotten” and require machine learning systems to support verifiable data erasure. Together, these factors highlight the importance of developing effective LLM unlearning techniques [2, 43].

Among the major challenges in achieving effective LLM unlearning, a particularly critical one lies in the heterogeneous unlearning difficulty across samples in the forget set [5, 9, 42]. As illustrated in Figure 1(a), when applying NPO [42] to the same forget set, samples that start from similar initial states quickly diverge in their perplexity (PPL) trajectories at the same training epoch. This divergence reveals that different samples are forgotten at substantially different rates—some being easy to erase, while others remain resistant. Figure 1(b) further highlights this imbalance: each sample reaches its convergence point at a different epoch, making it difficult to determine when the entire forget set has been properly unlearned. This asynchronous forgetting dynamic is particularly problematic for gradient-based LLM unlearning methods, as such variants of the negative cross-entropy loss lack a well-defined upper bound [19]. Continuing optimization to ensure full erasure of hard samples inevitably over-forgets the easy ones, thereby degrading overall model utility. This phenomenon motivates a central question: **how can we achieve a balanced state where all forget samples are unlearned simultaneously?**

To solve the above question, recent studies have proposed various balanced unlearning paradigms [9, 29, 35]. Krishnan et al. further show that the frequency of a fact in the pre-training corpus strongly influences its forgetting difficulty, and propose weighting schemes based on such frequency estimates, while pre-training corpus is often unavailable in practice [9]. Another line of balanced unlearning relies on reweighting each sample based on predefined or heuristic schemes, e.g., Negative Preference Optimization (NPO) leverages a reference model to stabilize gradient across different

¹https://anonymous.4open.science/r/anonymous_web4good2026_id480-2DE5

²<https://gdpr-info.eu/>

³<https://oag.ca.gov/privacy/ccpa>

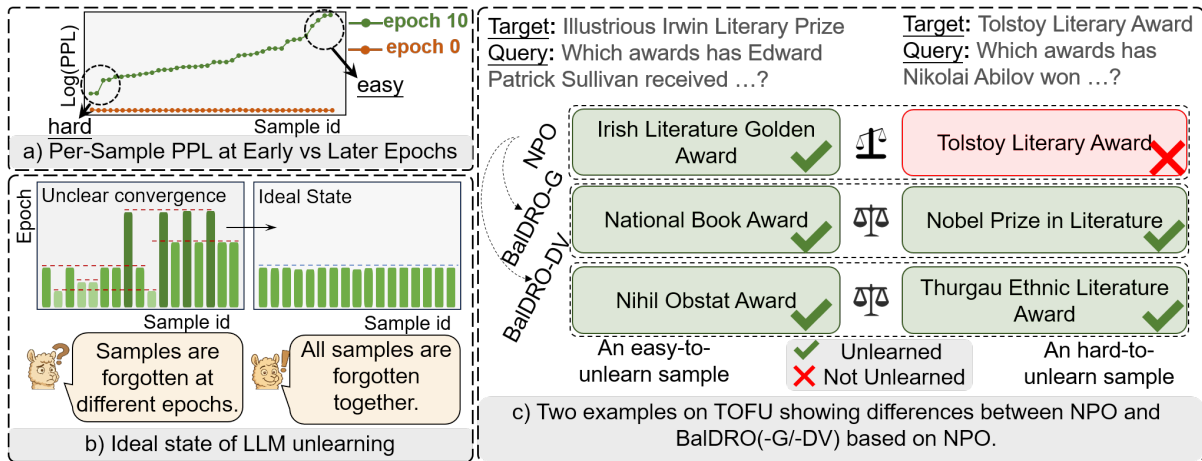


Figure 1: Illustration of sample-wise imbalance in LLM unlearning. a) Per-sample PPL (perplexity) at early and later epochs shows divergent forgetting dynamics, revealing heterogeneous unlearning difficulty in the forget set. b) This heterogeneity results in asynchronous convergence, whereas balanced unlearning aims to align forget epochs across samples. c) Two real examples from the TOFU benchmark: for the easy sample, NPO successfully unlearns the target, whereas for the hard sample, NPO fails. In contrast, both BalDRO-G and BalDRO-DV successfully unlearn both cases.

samples [42]. Based on NPO, SimNPO removes the reference dependency with a uniform distribution [5]. A more advanced example in this reweighting family is SatImp, which assigns each sample a dynamic weight based on saturation and importance criteria [35]. Despite their effectiveness, these approaches share a fundamental limitation—they rely on predefined or heuristic schemes that cannot dynamically adapt to the intrinsic distributional heterogeneity of data samples.

To dynamically address the imbalance issue in LLM unlearning, we propose BalDRO, a novel and efficient framework based on Distributionally Robust Optimization (DRO) that adaptively balances samples in the forget set. BalDRO formulates LLM unlearning as a min-sup process: the outer process updates model parameters, while the inner process searches for a worst-case forget distribution within a KL-divergence uncertainty set. This adversarial distribution naturally assigns greater influence to samples with larger forget losses, thereby preventing the optimization from prematurely over-forgetting the easy samples. To make this min-sup process tractable for LLMs, we provide two efficient realizations for the inner process. First, a GroupDRO-based variant (BalDRO-G) approximates the adversarial distribution by dynamically selecting the highest-loss samples in each mini-batch, enforcing progress on the most under-forgotten regions of the data space. Second, a Donsker-Varadhan dual formulation (BalDRO-DV) offers a continuous alternative by converting the inner supremum into a smooth log-sum-exp goal, allowing BalDRO to be seamlessly integrated into standard gradient-based unlearning pipelines without modifying model architectures or training loops. Together, these realizations provide discrete and continuous views of the same underlying DRO principle—ensuring that forgetting progresses in a coordinated, balanced manner across all samples. As illustrated in Figure 1(c), both variants based on NPO effectively unlearn samples of varying difficulty, achieving synchronized forgetting across easy and hard

cases. Extensive experiments on the TOFU and MUSE benchmarks demonstrate that BalDRO consistently achieves more synchronized forgetting dynamics, leading to higher forget quality while preserving general model utility more effectively than existing unlearning methods. Our contributions can be summarized as follows:

- We systematically analyze the failure modes of existing gradient-based LLM unlearning methods and show that uncontrolled sample-wise forgetting imbalance is the key challenge to effective LLM unlearning.
- We propose BalDRO, a novel and efficient DRO-based framework that balances sample during unlearning. Specifically, BalDRO formulates LLM unlearning as a min-sup bi-level process, and provides two tractable realizations for the inner process.
- We conduct extensive experiments on TOFU and MUSE benchmarks, demonstrating the effectiveness of BalDRO on achieving a better tradeoff between forgetting quality and model utility. For example, on the TOFU benchmark, BalDRO-G and BalDRO-DV based on NPO improve forget quality by more than 20% over NPO, while also delivering modest gains in model utility.

2 Related Work

2.1 LLM Unlearning

LLM unlearning aims to suppress knowledge contained in the forget set while preserving performance on the retain set [7, 8, 17, 19], encompassing both model-editing-based approaches [13, 14, 24, 39] and gradient-based approaches [5, 21, 35]. Among them, we focus on gradient-based methods, as they are model-agnostic and compatible with LLM finetuning pipelines. Gradient-based methods can be broadly grouped into two categories. The first

line of work is targeted unlearning, which defines explicit substitute outputs—typically refusal-style responses—for each forget-set query. These methods treat refusals as positive examples and enforce them using preference-based objectives such as DPO [23]. Subsequent variants expand this paradigm: FLAT incorporates f -divergence-based loss adjustments [28], while AltPO generates diverse positive alternatives by substituting knowledge-relevant tokens [21]. However, targeted methods may induce shortcut behaviors [22], where the model learns to mimic refusal templates based on patterns in the input.

A second line of work, non-targeted unlearning, avoids constructing explicit target outputs and instead directly modifies gradients to reduce the influence of forget samples. This direction originates from gradient inversion and gradient correction methods such as GradAscent (GA) and GradDiff (GD). More principled formulations subsequently emerge: NPO [42] casts unlearning as a negative log-likelihood objective; SimNPO [5] removes the reference model with a uniform distribution; and SatImp [35] develops theoretically grounded criteria for loss reweighting. Recent work further reveals that unlearning difficulty varies substantially across samples and correlates strongly with the frequency of knowledge occurrences in pretraining and finetuning [9].

2.2 Distributionally Robust Optimization

Distributionally Robust Optimization (DRO) provides a principled framework for learning models that remain reliable under distributional shifts or sampling uncertainty [15]. Instead of minimizing the expected loss over a single empirical distribution, DRO optimizes the worst-case loss within an uncertainty set of plausible distributions [34], producing models that are demonstrably more tolerant to noise and variability in training data [3, 30].

Building on this foundation, DRO has emerged as a central paradigm for enhancing robustness in a wide range of machine learning applications, particularly where distributional shifts or imbalanced sample difficulty are prevalent [16, 18]. Recent studies further extend DRO to large language models (LLMs), showing that DRO-based objectives can effectively stabilize alignment under noisy or heterogeneous preference data [31, 32, 44]. For instance, DRO-augmented preference optimization reduces the impact of pairwise and pointwise annotation noise in Direct Preference Optimization, leading to more reliable preference modeling [31]. However, the use of DRO for LLM unlearning has not been explored. Given the inherent imbalance between hard-to-forget and easy-to-forget samples in unlearning, DRO offers a natural and principled solution: its inner maximization automatically emphasizes worst-case (i.e., hardest) samples, aligning directly with the goal of balanced unlearning. Therefore, in this paper, we focus on how to apply DRO to LLM unlearning.

3 Preliminary

In this section, we focus on analyzing several representative gradient-based LLM unlearning methods. We start from a common formulation for gradient-based unlearning, typically expressed as a Gradient Difference (GD) objective:

$$\ell_{\text{all}}(\theta) = \ell_f(\theta) + \lambda \ell_r(\theta), \quad (1)$$

where ℓ_f and ℓ_r denote the losses on the forget and retain sets, respectively. A classical choice for ℓ_f is Gradient Ascent (GA), i.e., the reverse CE loss [36]:

$$\ell_f^{\text{GA}}(\theta) = -\mathbb{E}_{(x,y) \in D_f} [\ell_f(y | x; \theta)]. \quad (2)$$

Although GA effectively suppresses target likelihoods, it lacks any sample-wise stopping criterion [19] and is highly sensitive to heterogeneous forgetting difficulty [9, 35], which leads to both under-forgotten and over-forgotten samples at convergence. To alleviate this issue, Negative Preference Optimization (NPO) introduces a reference model π_{ref} to balance samples [42]:

$$\ell_f^{\text{NPO}}(\theta) = \mathbb{E}_{(x,y) \in D_f} \left[-\frac{2}{\alpha^{\text{NPO}}} \log \sigma \left(-\alpha^{\text{NPO}} \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right], \quad (3)$$

where α^{NPO} is a temperature parameter that controls the sharpness of the penalty applied to the log-ratio between current policy $\pi_\theta(y | x)$ and the reference policy $\pi_{\text{ref}}(y | x)$. Building on this idea, SimNPO removes the reference dependency and uses a length-normalized log-likelihood to obtain a simple, reference-free surrogate [5]:

$$\ell_f^{\text{SimNPO}}(\theta) = \mathbb{E}_{(x,y) \in D_f} \left[-\frac{2}{\alpha^{\text{Sim}}} \log \sigma \left(-\frac{\alpha^{\text{Sim}}}{|y|} \log \pi_\theta(y | x) \right) \right]. \quad (4)$$

where $|y|$ denotes the output length used for normalization, and α^{Sim} controls the sharpness of the logistic penalty for SimNPO. Building upon these preference-based surrogate methods, SatImp further generalizes this line of work by introducing a token-wise saturation-importance weight, which adjusts the contribution of each generated token according to how confidently it is predicted. Specifically, for the k -th token in y , the weight is defined as:

$$w_{x,y,k}^{\text{SatImp}} = \pi_\theta(y_k | y_{<k}, x)^{\alpha_1} (1 - \pi_\theta(y_k | y_{<k}, x))^{\alpha_2},$$

where the predicted probability $\pi_\theta(y_k | y_{<k}, x)$ determines both the saturation term and the importance term. The exponent α_1 amplifies the effect of highly confident tokens, while α_2 emphasizes low-probability tokens. The loss can be formulated as:

$$\ell_f^{\text{SatImp}}(\theta) = \mathbb{E}_{(x,y) \in D_f} \left[\sum_{k=1}^{|y|} w_{x,y,k}^{\text{SatImp}} \log \pi_\theta(y_k | y_{<k}, x) \right]. \quad (5)$$

Although these weighting-based surrogate losses adaptively adjust sample contributions during training, they share a fundamental limitation: their updates are governed by pre-specified functional forms that dictate how weights respond to model predictions. These forms are heuristically designed rather than derived from a principled notion of balance, and thus cannot dynamically adapt to the evolving difficulty of different forget samples. As a result, existing methods often lead to asynchronous forgetting dynamics. In this regime, easy samples are quickly over-unlearned, while difficult ones remain insufficiently unlearned.

4 The Proposed Framework

In this section, we start from presenting our proposed BalDRO framework, including our unlearning goal based on a principled notion of balance, and corresponding bi-level process. Then, we propose two efficient and effective methods to realize the goal, which can be added as a plug-in to improve current LLM unlearning methods (e.g., NPO [42], SimNPO [5], and SatImp [35]). Finally, we discuss its cost as well as its connections to existing unlearning methods.

4.1 Overall Bi-level Process of BalDRO

To instantiate the principled notion of balance, we first formalize our goal as an optimization objective that explicitly captures sample-wise imbalance in the forget set.

Definition 4.1 (Balanced Unlearning Objective). Given a forget dataset $D_f = \{z_i = (x_i, y_i)\}_{i=1}^n$ and the corresponding empirical distribution $\widehat{D}_f(Z) = \frac{1}{n} \sum_i \delta_{z_i}(Z)$, balanced LLM unlearning aims to find model parameters θ that minimize the worst-case expected forget loss under distributional shifts:

$$\arg \min_{\theta} \sup_{Q_f: \mathbb{D}(Q_f \| \widehat{D}_f) \leq \eta} \mathbb{E}_{Z \sim Q_f} [\ell_f(Z; \theta)]. \quad (6)$$

Here, we apply DRO only to the forget loss, while the retain loss remains unchanged. Please note that, our experiments in Section 5.3.3 show that extending DRO to the retain set brings negligible benefit, suggesting that retain samples already maintain a natural balance and do not require additional robustness adjustments. Q_f denotes an adversarially perturbed distribution that reallocates probability mass toward harder forget samples, and $\eta > 0$ controls the uncertainty radius of the perturbation.

Regarding the choice of divergence \mathbb{D} , any divergence that enables a valid min-sup DRO formulation could in principle be used; BalDRO is not restricted to a specific distance measure. In this work, we adopt the Kullback–Leibler (KL) divergence [10] because it provides a closed-form Donsker–Varadhan (DV) dual representation, which leads to a tractable and numerically stable implementation at LLM scale while naturally inducing an exponential reweighting aligned with forgetting difficulty:

$$\mathbb{D}_{\text{KL}}(Q_f \| \widehat{D}_f) = \mathbb{E}_{Z \sim Q_f} \left[\log \frac{Q_f(Z)}{\widehat{D}_f(Z)} \right] \leq \eta. \quad (7)$$

As shown in Figure 2, this bi-level process embodies our core objective: the inner “sup” identifies a worst-case forget distribution within the KL ball—thereby adaptively emphasizing hard-to-forget samples—while the outer “min” updates model parameters to minimize this adversarial loss. The resulting optimization naturally balances forgetting progress across samples without explicit heuristic weighting, grounding our method in a theoretically robust notion of equilibrium. Obviously, the outer process can be easily realized by finetuning LLM, and the key challenge is how to assess the inner process (i.e., $\sup_{Q_f: \mathbb{D}_{\text{KL}}(Q_f \| \widehat{D}_f) \leq \eta} \mathbb{E}_{Z \sim Q_f} [\ell_f(Z; \theta)]$), which captures the worst-case expected loss under distributional uncertainty around the empirical forget set.

4.2 The Inner Process

To instantiate Definition 4.1, we develop two tractable realizations of the inner process in Eq. (6). The first, BalDRO-G, provides a discrete approximation based on Group Distributionally Robust Optimization, while the second, BalDRO-DV, offers a continuous dual formulation grounded in the Donsker–Varadhan (DV) representation. Together, they provide complementary discrete and continuous perspectives of realizing our goal, which both follow the balanced unlearning principle.

4.2.1 BalDRO-G: Discrete Group-Level Approximation. Group Distributionally Robust Optimization (GroupDRO) offers a tractable approximation to the inner process in Definition 4.1 by explicitly selecting the most difficult subsets within the forget set [41].

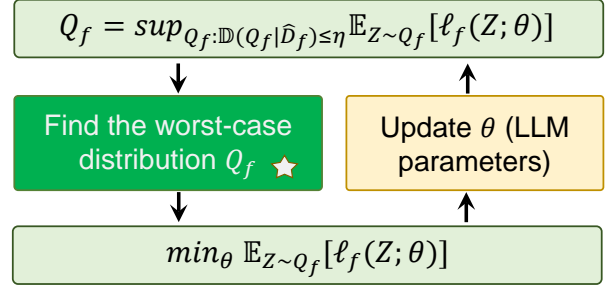


Figure 2: The overall min-sup process of BalDRO. The inner “sup” adaptively identifies the hardest forget distribution, while the outer “min” optimizes model parameters.

Concretely, we partition the empirical forget set D_f into G groups $\{D_f^{(1)}, D_f^{(2)}, \dots, D_f^{(G)}\}$, and optimize the maximum expected loss among them:

$$\max_{g=1, \dots, G} \mathbb{E}_{z_i \sim D_f^{(g)}} [\ell_f(z_i; \theta)]. \quad (8)$$

Intuitively, this formulation serves as a discrete instantiation of the inner process in Definition 4.1: the adversarial distribution Q_f is approximated by assigning all probability mass to the group with the largest expected forget loss. This allows BalDRO-G to operationalize the theoretical min-sup objective through a finite, group-level approximation that is both stable and computationally efficient in practice. At each iteration, BalDRO-G enforces progress on the worst-performing group, directing the optimization toward the under-forgotten region of the data space. Rather than relying on predefined grouping schemes, we adopt a percentile-based strategy (top-50%) to avoid the instability of extreme single-sample maxima while still focusing on the hardest region of the distribution. This design follows standard practice in robust optimization and is further validated in Appendix B.

4.2.2 BalDRO-DV: Continuous Dual Adaptive Weighting. While GroupDRO captures coarse-grained group imbalance, it still depends on manually defined partitions. To derive a fully continuous and differentiable relaxation of the same inner process, we revisit the original DRO problem and apply a Lagrangian relaxation [11] to the KL-divergence constraint:

$$\sup_{Q_f} \left\{ \mathbb{E}_{Z \sim Q_f} [\ell_f(Z; \theta)] - \beta (\mathbb{D}_{\text{KL}}(Q_f \| \widehat{D}_f) - \eta) \right\}, \quad (9)$$

where $\beta > 0$ is a Lagrange multiplier. The constant term $\beta\eta$ is independent of Q_f and can be ignored during maximization. Substituting the definition of the KL divergence yields:

$$\sup_{Q_f} \mathbb{E}_{Z \sim Q_f} \left[\ell_f(Z; \theta) - \beta \log \frac{Q_f(Z)}{\widehat{D}_f(Z)} \right].$$

Maximizing the inner expression with respect to Q_f gives the optimal adversarial distribution. The detailed derivation is provided in Appendix A. Here, we directly provide the closed-form solution as follows:

$$Q_f^*(Z) = \frac{\widehat{D}_f(Z) \exp(\ell_f(Z; \theta) / \beta)}{\mathbb{E}_{Z' \sim \widehat{D}_f} [\exp(\ell_f(Z'; \theta) / \beta)]}. \quad (10)$$

This distribution exponentially upweights samples (or regions of the data space) with higher forget loss, thereby implementing an

adaptive, distribution-level reweighting mechanism. Substituting Q_f^* back into Eq. (6) yields a closed-form expression for the inner supremum:

$$\min_{\theta} \beta \eta + \beta \log \mathbb{E}_{Z \sim \widehat{D}_f} \left[\exp \left(\frac{\ell_f(Z; \theta)}{\beta} \right) \right], \quad (11)$$

which corresponds to the continuous Donsker–Varadhan (DV) representation. Although β can in principle be optimized jointly, we adopt a fixed- β variant for better stability and computational simplicity [31]. Finally, minimizing Eq. (11) leads to the DV dual formulation:

$$\min_{\theta} \beta \log \mathbb{E}_{Z \sim \widehat{D}_f} \left[\exp \left(\frac{\ell_f(Z; \theta)}{\beta} \right) \right] = \min_{\theta} \beta \log \left(\frac{1}{n} \sum_{i=1}^n \exp \left(\frac{\ell_f(z_i; \theta)}{\beta} \right) \right). \quad (12)$$

The equality in Eq. (12) follows from the definition of the empirical distribution $\widehat{D}_f(Z) = \frac{1}{n} \sum_{i=1}^n \delta_{z_i}(Z)$, under which the expectation $\mathbb{E}_{Z \sim \widehat{D}_f}[\cdot]$ is equivalent to calculating over all samples in the forget set. This dual formulation transforms the intractable distributional optimization problem into a smooth and differentiable log-sum-exp goal, which can be seamlessly integrated into existing LLM unlearning pipelines. Intuitively, the exponential term $\exp(\ell_f(z_i; \theta)/\beta)$ adaptively scales gradient magnitudes: harder samples receive higher weighting, while already-forgotten ones are naturally downweighted, achieving a self-regulating balance across the forget set. This adaptive weighting mechanism serves as a continuous, theoretically grounded realization of the balanced unlearning principle introduced in Definition 4.1.

4.3 Model Discussion

4.3.1 Additional Time Complexity. BalDRO introduces only marginal computational overhead compared to standard unlearning objectives. For BalDRO-G, given a batch of n forget samples, we first compute their per-sample forget losses $\{\ell_f\}_{i=1}^n$ and identify the hardest subset (e.g., top-50%) for back-propagation. This selection can be implemented via a partial sort with at most $O(n \log n)$ complexity, which is negligible relative to the cost of LLM forward and backward passes.

For BalDRO-DV, the only additional cost arises from the log-sum-exp term in Eq. (12), which requires computing one exponential and one logarithm per sample within each mini-batch. Specifically, it involves three lightweight element-wise operations: scaling by $1/\beta$, evaluating $\exp(\ell_i/\beta)$, and performing a batch-level reduction via $\log(\sum_i \exp(\ell_i/\beta))$. These operations result in linear complexity $O(n)$, the same order as the base loss computation, with only a small constant factor overhead due to exponentiation and normalization.

In summary, both BalDRO-G and BalDRO-DV preserve the scalability of existing unlearning pipelines, adding only minimal constant-time operations while maintaining the same asymptotic complexity.

4.3.2 Relations to Existing LLM Unlearning Methods. BalDRO is designed as a plugin-style framework for gradient-based LLM unlearning, offering strong generality across a wide range of existing methods. This design choice is intentional: it reveals that distributional balancing is a missing yet broadly beneficial component across diverse unlearning objectives. Rather than tailoring to any specific loss, BalDRO can capture a fundamental property of unlearning dynamics, which will be evaluated in next section.

Moreover, BalDRO provides a unified perspective that encompasses existing unlearning approaches as special cases under different regimes of the temperature parameter β . In the discrete variant (BalDRO-G), selecting the hardest subset corresponds to the limiting case of optimizing the maximum loss over grouped samples; and as the subset size increases, the behavior naturally approaches that of the original unlearning method. In the continuous variant (BalDRO-DV), the formulation in Eq. (12) generalizes this trajectory: as $\beta \rightarrow \infty$, the exponential weighting becomes uniform and the objective degenerates to the standard mean-loss formulation (i.e., no balancing), whereas as $\beta \rightarrow 0$, the log-sum-exp term approaches the maximum loss, recovering the worst-case optimization equivalent to BalDRO-G when each sample is treated as its own group.

5 Experiments

In this section, we try to answer these research questions:

- RQ 1: Does BalDRO provide stable improvements on different gradient-based LLM unlearning objectives? (Sec. 5.2)
- RQ 2: Does BalDRO perform well across varying sizes of the forget set? (Sec. 5.3.1)
- RQ 3: Is BalDRO robust to hyperparameter settings? (Sec. 5.3.2)
- RQ 4: Is BalDRO also effective on the retain set? (Sec. 5.3.3)
- RQ 5: Is BalDRO still effective on different metrics? (Sec. 5.3.4)

Due to page limits, we have placed additional experiments in Appendix B, including varying the sample selection ratio in BalDRO-G.

5.1 Experiment Setup

5.1.1 Benchmarks. We evaluate BalDRO on two complementary unlearning benchmarks that together capture both controlled and realistic forgetting behaviors. TOFU [20] contains synthetic QA pairs for 200 fictional authors, ensuring that all forget-set knowledge originates solely from fine-tuning. TOFU includes three unlearning settings, aiming to remove 1%, 5% and 10% of the total dataset. This setup isolates unlearning effects from pretraining priors, providing a controlled environment for analyzing forgetting dynamics. MUSE [26] consists of real-world books and news articles, where forget and retain splits are semantically entangled. It thus presents a more realistic and challenging testbed for assessing the robustness and practicality of unlearning algorithms.

5.1.2 Evaluation Metrics. TOFU offers fine-grained controls for separately assessing forgetting and retention. Forget Quality (FQ) captures alignment with a retain-only reference using the KS distance over Truth Ratio distributions, whereas Model Utility (MU) evaluates the preservation of non-forget knowledge via the harmonic mean of Probability, ROUGE, and Truth Ratio on D_r and other holdout sets. Extraction Memorization (EM) and Extraction Strength (ES) quantify verbatim memorization at the token and prefix levels. We additionally report four Truth Ratio variants—F-TR, Ra-TR, R-TR, and Rw-TR—which together assess forgetting quality, robustness to unseen real authors, retention correctness, and real-world consistency. Lower EM and ES indicate stronger forgetting, while higher FQ, MU, and Truth Ratio values reflect better retention and balanced unlearning.

MUSE contains real-world texts, such as novels and news articles, where the forget and retain sets exhibit strong semantic overlap,

Method	FQ (\uparrow)	MU (\uparrow)	Fluency (\uparrow)	EM (\downarrow)	ES (\downarrow)	F-TR (\uparrow)	Ra-TR (\uparrow)	R-TR (\uparrow)	Rw-TR (\uparrow)
Original	0.0013	0.6276	0.8889	1.0000	1.0000	0.5306	0.6120	0.4596	0.5521
Retain	1.0000	0.6268	0.9222	0.7121	0.0719	0.6777	0.6087	0.4621	0.5612
GradAscent	0.2657	0.5313	0.5326	0.5572	0.0361	0.0019	0.6049	0.4357	0.5848
GradDiff	0.5786	0.6064	0.3976	0.5397	0.0481	0.6359	0.5595	0.4523	0.5720
NPO	0.7659	0.5775	0.8158	0.6842	0.0982	0.7125	0.5807	0.4360	0.5489
NPO + BalDRO-G	0.9188	0.6126	0.8220	0.7634	0.0630	0.6859	0.6187	0.4495	0.5741
NPO + BalDRO-DV	0.9900	0.5815	0.8227	0.6659	0.0593	0.7238	0.6000	0.4148	0.5610
SimNPO	0.4046	0.5643	0.8422	0.7383	0.0850	0.6768	0.6025	0.3990	0.5377
SimNPO + BalDRO-G	0.5786	0.5651	0.8894	0.7189	0.0633	0.7034	0.5877	0.4307	0.5712
SimNPO + BalDRO-DV	0.5786	0.5917	0.8479	0.6926	0.0521	0.7257	0.6345	0.4215	0.5726
SatImp	0.0013	0.5342	0.8272	0.9466	0.2041	0.5117	0.5493	0.4315	0.5338
SatImp + BalDRO-G	0.0971	0.6003	0.8520	0.8879	0.1609	0.6108	0.6131	0.4654	0.5480
SatImp + BalDRO-DV	0.0068	0.5480	0.7588	0.8646	0.4522	0.5515	0.5764	0.4485	0.4973

Table 1: Overall performance on TOFU benchmark with forget ratio = 1%. We bold the best result.

Method	News				Books			
	KM-Dr (\uparrow)	KM-Df (\downarrow)	VM-Df (\downarrow)	PL ($\rightarrow 0$)	KM-Dr (\uparrow)	KM-Df (\downarrow)	VM-Df (\downarrow)	PL ($\rightarrow 0$)
Original	0.5552	0.6443	0.5789	-99.8111	0.6913	0.4712	0.9970	-57.3410
Retain	0.5602	0.3279	0.2016	-4.7200	0.6874	0.3029	0.1445	8.1600
GradAscent	0.0000	0.0000	0.0000	25.1259	0.0000	0.0000	0.0000	-22.8180
GradDiff	0.2519	0.2938	0.0029	108.9840	0.0678	0.0089	0.0035	-37.0562
NPO	0.4552	0.5978	0.4255	-90.8480	0.6424	0.4414	0.6011	-55.7692
NPO+BalDRO-G	0.4626	0.5805	0.3826	-65.7011	0.6486	0.4376	0.5465	-54.2160
NPO+BalDRO-DV	0.4589	0.5754	0.3934	-69.5214	0.6520	0.4107	0.5230	-55.2515
SimNPO	0.4121	0.5806	0.3829	-99.8951	0.5969	0.3009	0.2364	-51.7018
SimNPO+BalDRO-G	0.4272	0.5940	0.4193	-99.8950	0.6151	0.2841	0.2264	-51.2944
SimNPO+BalDRO-DV	0.4571	0.5670	0.1829	100.4139	0.5393	0.3009	0.1935	-49.3898
SatImp	0.3797	0.5902	0.4403	-99.8951	0.6026	0.4017	0.8730	-58.3395
SatImp+BalDRO-G	0.3805	0.5053	0.4197	-99.8531	0.6037	0.3802	0.4955	-54.5858
SatImp+BalDRO-DV	0.3967	0.4568	0.3552	-99.8321	0.6013	0.3672	0.5535	-57.2485

Table 2: Overall performance on the MUSE benchmark. MUSE adopts a fixed forget/retain split, and we bold the best result.

making targeted unlearning substantially more challenging. To evaluate both semantic and lexical forgetting as well as privacy risks, we adopt three metrics: KnowMem (KM), VerbMem (VM), and PrivLeak (PL). KM measures semantic recall, and VM assesses verbatim recall through exact lexical overlap; both are computed on the forget set (KM-Df, VM-Df) and the retain set (KM-Dr) to distinguish desired forgetting from unintended over-forgetting. PL estimates membership-inference risk, indicating whether forgotten samples remain distinguishable from unseen data. Lower KM-Df and VM-Df indicate stronger forgetting and privacy protection, while higher KM-Dr reflects better retention. For PL, values closer to zero are preferred.

5.1.3 Baseline Methods. We compare our method with several baselines. Original refers to the model trained on both forget and retain sets; it represents the starting point of unlearning. In contrast, Retain denotes the model fine-tuned only on the retain set,

reflecting an idealized upper bound for LLM unlearning. The remaining baselines, introduced in Section 3, cover the classic and recent gradient-based unlearning methods. Specifically, we include GradAscent (GA) [36], which performs gradient ascent on forget samples to reduce the model’s confidence, and GradDiff (GD) [36], which additionally leverages the retain set to mitigate unnecessary utility loss. Moreover, we evaluate BalDRO on three representative and state-of-the-art unlearning frameworks—NPO [42], SimNPO [5], and SatImp [35]. These baselines constitute the latest gradient-based unlearning techniques, while editing-based approaches are not included [24, 39], as they follow a fundamentally different model-editing paradigm that is incompatible with us.

5.1.4 Implementation Details. All experiments are conducted on a server equipped with 8 NVIDIA A800 GPUs. We use LLaMA-2-7B [27] as the primary backbone model for evaluating overall unlearning performance. For all methods, we perform a hyperparameter search over learning rates in $\{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times$

$10^{-5}, 1 \times 10^{-4}$ }, batch sizes in $\{8, 16, 32\}$, β in $\{1.0, 2.0, 5.0, 10.0\}$ and λ in $\{0.25, 0.5, 1.0, 2.0\}$. This setup ensures a fair and consistent comparison across methods, allowing the reported performance to reflect the best achievable results under each experimental configuration. More detailed information can be found in our released code⁴.

5.2 Overall Performance (RQ1)

5.2.1 Main Results on TOFU. Table 1 presents the results under the TOFU benchmark. We summarize three observations. First, BalDRO variants consistently improve FQ and reduce EM/ES across most base methods, showing that balancing the forgetting rates of samples in the forget set effectively strengthens the unlearning process. Second, different base methods benefit to different degrees. SimNPO obtains the most stable overall gains: FQ rises from 0.4046 to 0.5786, EM decreases, and several TR metrics also improve. In contrast, SatImp shows the largest relative jump in FQ ($0.0013 \rightarrow 0.0971$ (BalDRO-G), over **70×**), mainly because its baseline forgetting is extremely weak, leaving enough space for improvements. Third, BalDRO-DV and BalDRO-G show different strengths across backbones. For NPO and SimNPO, BalDRO-DV typically performs better, achieving the highest FQ and lower EM/ES while maintaining competitive TR performance. For SatImp, however, BalDRO-G works better, giving both the strongest FQ improvement and more stable TR results.

5.2.2 Main Results on MUSE. Table 2 summarizes the results on the MUSE benchmark. We have several observations from this table. First, we observe that both BalDRO-G and BalDRO-DV consistently enhance the base methods (NPO, SimNPO, SatImp) across domains. This demonstrates the general effectiveness of our framework: BalDRO improves forget quality while maintaining LLM utility. For instance, on the News domain, BalDRO-G based on NPO lowers VM-Df ($0.4255 \rightarrow 0.3826$) and improves PL ($-90.85 \rightarrow -65.70$), indicating better forget quality and better model utility. Second, across both domains, BalDRO-G and BalDRO-DV perform comparably. Rather than one variant uniformly dominating the other, they exhibit complementary strengths across settings. Given that BalDRO-DV only improves the objective function and is therefore simpler to implement, it may be the more practical choice in real-world deployments. Finally, the three base methods (NPO, SimNPO, SatImp) respond to BalDRO with different degrees of improvement. NPO already delivers strong baseline, limiting the potential gains. E.g., in the Books domain, BalDRO-G yields modest changes in KM-Dr ($0.6424 \rightarrow 0.6486$) and PL ($-55.77 \rightarrow -54.22$). In contrast, SimNPO and SatImp perform worse, giving BalDRO substantially more room to provide larger improvements across different metrics.

5.3 Detailed Model Analyses

5.3.1 Varying Forget Ratios (RQ2). To verify whether our proposed BalDRO maintains stable performance across different sizes of the forget set, we conduct additional experiments on TOFU using forget-set ratios of 5% and 10%.

Figure 3 presents the complete results based on two representative methods (NPO and SimNPO), from which we make the following observations. First, under both two methods, BalDRO-G and BalDRO-DV consistently achieve substantial improvements in FQ while maintaining MU comparable to the original models. For instance, with a forget ratio of 5%, the FQ of NPO increases from 0.6284 to 0.7125 (BalDRO-G) and 0.9646 (BalDRO-DV); for SimNPO, FQ improves from 0.4662 to 0.7125 (BalDRO-G) and 0.8655 (BalDRO-DV). These results indicate that BalDRO provides consistent and significant gains regardless of the size of the forget set. Second, we observe that BalDRO-DV outperforms BalDRO-G in FQ in most cases. For ratio = 5%, BalDRO-DV delivers more than 10% additional improvement over G; and under the more challenging settings of ratio = 10%, BalDRO-DV’s gains remain stable, while the improvements from BalDRO-G are relatively limited. This suggests that BalDRO-DV is the preferable choice for practical deployment due to its more reliable performance. Finally, BalDRO’s impact on MU remains minimal regardless of whether the forget ratio is 5% or 10%, indicating that BalDRO does not harm general ability.

5.3.2 Hyperparameter Analyses (RQ3). Figure 4 illustrates the influence of the coefficient β in Eq. (12) and the balancing parameter λ in Eq. (1) on unlearning performance in the TOFU benchmark. First, FQ peak at $\beta = 2.0$ and $\lambda = 1.0$, which suggests the presence of a well-defined optimal region where the adversarial reweighting and the base unlearning objective reinforce each other most effectively. MU remains relatively stable across configurations, showing only slight degradation when parameters move toward extreme values, indicating that the model is generally robust to moderate variations in both hyperparameters. Second, we observe that $\beta = 2$ or 5 produces the most reliable and consistent improvements. As discussed earlier, excessively large β dilutes the effect of sample-wise weighting and causes the DRO objective to collapse toward a near-uniform distribution, thereby reducing its ability to correct imbalance. Conversely, overly small β forces the model to concentrate too aggressively on the hardest samples, which disrupts the coordinated forgetting process and ultimately harms unlearning quality. Third, we find that increasing λ generally enhances MU, confirming that assigning greater weight to the retain signal helps preserve overall model capability. However, reducing λ does not yield notable improvements in FQ, likely because $\lambda = 1$ is already near the upper performance bound for FQ, leaving limited room for further gains. Together, these observations underscore the importance of jointly tuning β and λ to achieve strong and stable unlearning performance.

5.3.3 DRO for the Retain Set (RQ4). Figure 5 presents a comparison of applying versus not applying DRO to the retain set on the TOFU and MUSE benchmarks. For consistency with the main results, all experiments use NPO and its BalDRO variants. MU and KM-Dr are shown on the horizontal axes, while FQ and PL are shown on the vertical axes for TOFU and MUSE, respectively, where the x-axis reflects model utility and the y-axis reflects forgetting quality. The red arrow denotes the direction of a more favorable trade-off.

Across both benchmarks, applying DRO only to the forget set consistently yields better results, with the corresponding points shifting toward the upper-right region. This observation indicates that BalDRO is most effective when used solely on the forget loss:

⁴https://anonymous.4open.science/r/anonymous_web4good2026_id480-2DE5

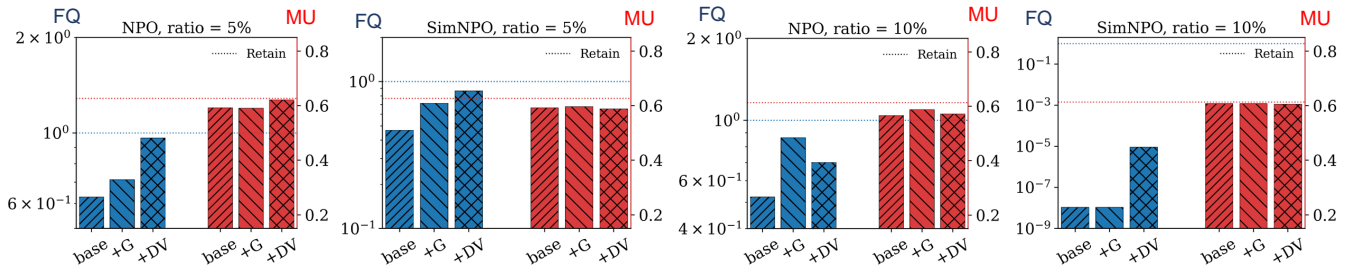


Figure 3: Performance with varying forget ratios (5% and 10%) on the TOFU benchmark. We focus on FQ and MU, the two most commonly used metrics on TOFU, and select NPO and SimNPO as base methods due to their strong overall performance on these two metrics. Here, “+G” denotes our proposed BalDRO-G, and “+DV” corresponds to BalDRO-DV. The dashed horizontal line indicates the retain baseline.

β	FQ				MU			
	λ 2.0	λ 1.0	λ 0.5	λ 0.25	λ 2.0	λ 1.0	λ 0.5	λ 0.25
10.0	0.4045	0.5786	0.5786	0.7659	0.5683	0.5647	0.5685	0.5308
5.0	0.5786	0.7659	0.9188	0.9188	0.5798	0.5707	0.5667	0.5326
2.0	0.5786	0.9900	0.9900	0.9188	0.5765	0.5815	0.5789	0.5549
1.0	0.4045	0.4045	0.7659	0.7659	0.5344	0.5758	0.5748	0.5689

Figure 4: Performance of BalDRO-DV with varying β and balancing parameter λ on the TOFU benchmark.

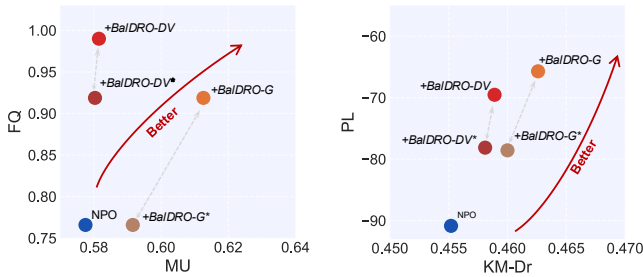


Figure 5: Performance comparisons between whether applying DRO to the retain loss $\ell_r(\theta)$ on TOFU and MUSE benchmarks. “*” indicates DRO is applied to the retain loss.

the retain set does not benefit from additional DRO regularization, and leaving the retain objective unchanged leads to a better balance between forgetting performance and model utility. These results further suggest that distributional imbalance is primarily concentrated in the forget set, making it the component where DRO can produce meaningful gains.

5.3.4 Performance on Other Metrics (RQ 5). Table 1-2 focuses on the primary evaluation metrics of LLM unlearning. To further examine the behavior of our proposed BalDRO, Table 3 reports extended membership inference related results on the TOFU benchmark with a 1% forget ratio. Specifically, we consider four privacy metrics:

LOSS [38], ZLib [1], MinK [25], and MinK++ [40]. The lower values of these metrics indicate stronger privacy protection. The comparison covers the “Original”, “Retain”, base unlearning methods (NPO, SimNPO, SatImp), and their BalDRO-enhanced variants.

We have several findings from Table 3. First, “Original” obtains values close to 1.0 across all four metrics, indicating substantial memorization and thus the highest privacy risk. Second, all unlearning baselines substantially reduce these scores, confirming that they effectively mitigate membership inference attacks. Third, introducing BalDRO yields consistent improvements across base methods. For instance, within NPO, BalDRO-DV lowers LOSS from 0.4681 to 0.3481 and ZLib from 0.4594 to 0.3250, showing clearer privacy gains. BalDRO-G also improves privacy, albeit to a slightly lesser degree, reflecting BalDRO-DV’s stronger robustness in this setting. The most pronounced improvements occur when BalDRO is combined with SimNPO. LOSS drops from 0.4925 to 0.1769 and MinK++ from 0.2738 to 0.1075—representing the best overall performance across all configurations. This demonstrates a strong synergy: BalDRO effectively amplifies SimNPO’s forgetting signal and further suppresses residual memorization. In contrast, SatImp benefits to a more limited extent, suggesting that the impact of BalDRO may depend on the characteristics of the base method. Nevertheless, across all cases, BalDRO consistently contributes to measurable reductions in privacy risk, reinforcing its effectiveness as a general enhancement to LLM unlearning.

6 Conclusion and Future Work

How to unlearn specific information from large language models is essential for responsible and trustworthy web-scale AI. A key challenge lies in the sample-wise imbalance within the forget set: easy samples disappear quickly under unlearning updates, but harder samples retain their influence for much longer, causing the overall forgetting process to become highly uneven. To address this issue, we viewed LLM unlearning through distributional robustness and developed BalDRO, a unified framework that adaptively balances sample contributions during unlearning. We then proposed its two realizations (BalDRO-G and BalDRO-DV), which provide discrete and continuous mechanisms for achieving synchronized forgetting across samples. We conducted extensive experiments on the TOFU and MUSE benchmarks, to demonstrate that BalDRO consistently improves both forget quality and model utility compared to existing

Method	LOSS (↓)	ZLib (↓)	MinK (↓)	MinK++ (↓)
Original	1.0000	1.0000	1.0000	1.0000
Retain	0.4981	0.5513	0.5038	0.6100
NPO	0.4681	0.4594	0.4912	0.4900
+BalDRO-G	0.4632	0.3784	0.5056	0.3266
+BalDRO-DV	0.3481	0.3250	0.3613	0.4875
SimNPO	0.4925	0.4700	0.4981	0.2738
+BalDRO-G	0.2468	0.2188	0.2344	0.1314
+BalDRO-DV	0.1769	0.2525	0.1744	0.1075
SatImp	0.9956	0.9906	0.9900	0.9544
+BalDRO-G	0.9493	0.9451	0.9496	0.8004
+BalDRO-DV	0.9613	0.9587	0.9613	0.8493

Table 3: Extended metrics evaluated on different methods on the TOFU benchmark with forget ratio = 1%. We bold the best result.

unlearning methods. In the future, we plan to further investigate the balance between forget quality and general utility by redesigning the loss function to reduce cost of LLM unlearning.

References

- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*. 2633–2650.
- Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, Peijun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. 2025. SafeEraser: Enhancing safety in multimodal large language models through multimodal machine unlearning. *arXiv preprint arXiv:2502.12520* (2025).
- Feng-Qi Cui, Anyang Tong, Jinyang Huang, Jie Zhang, Dan Guo, Zhi Liu, and Meng Wang. 2025. Learning from heterogeneity: Generalizing dynamic facial expression recognition via distributionally robust optimization. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 5587–5596.
- Yang Deng, An Zhang, Yankai Lin, Xu Chen, Ji-Rong Wen, and Tat-Seng Chua. 2024. Large language model powered agents in the web. In *Companion Proceedings of the ACM Web Conference 2024*. 1242–1245.
- 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. In *Neurips Safe Generative AI Workshop 2024*.
- Jiahui Geng, Qing Li, Herbert Woitschlaeger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. A comprehensive survey of machine unlearning techniques for large language models. *arXiv preprint arXiv:2503.01854* (2025).
- Wonje Jeung, Sangyeon Yoon, Hyesoo Hong, Soeun Kim, Seungju Han, Youngjae Yu, and Albert No. 2025. Dusk: Do not unlearn shared knowledge. *arXiv preprint arXiv:2505.15209* (2025).
- Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kaikhura, and Sijia Liu. 2024. SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning. In *EMNLP*.
- Aravind Krishnan, Siva Reddy, and Marius Mosbach. 2025. Not All Data Are Unlearned Equally. In *Second Conference on Language Modeling*. <https://openreview.net/forum?id=Kd97lFffTu>
- Solomon Kullback. 1951. Kullback-leibler divergence. *Tech. Rep.* (1951).
- Claude Lemaréchal. 2001. Lagrangian relaxation. In *Computational combinatorial optimization: optimal or provably near-optimal solutions*. Springer, 112–156.
- Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. 2024. St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. *arXiv preprint arXiv:2410.06703* (2024).
- Kunhao Li, Wenhao Li, Di Wu, Lei Yang, Jun Bai, Ju Jia, and Jason Xue. 2025. Cross-Modal Unlearning via Influential Neuron Path Editing in Multimodal Large Language Models. *arXiv preprint arXiv:2511.06793* (2025).
- Zexi Li, Xiangzhu Wang, William F Shen, Meghdad Kurmanji, Xinchu Qiu, Dongqi Cai, Chao Wu, and Nicholas D Lane. 2025. Editing as Unlearning: Are Knowledge Editing Methods Strong Baselines for Large Language Model Unlearning? *arXiv preprint arXiv:2505.19855* (2025).
- Fengming Lin, Xiaolei Fang, and Zheming Gao. 2022. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization* 12, 1 (2022), 159–212.
- Xinyu Lin, Wenjie Wang, Jujia Zhao, Yongqi Li, Fuli Feng, and Tat-Seng Chua. 2024. Temporally and distributionally robust optimization for cold-start recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8750–8758.
- Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems* 37 (2024), 118198–118266.
- Jiashuo Liu, Jiayun Wu, Bo Li, and Peng Cui. 2022. Distributionally robust optimization with data geometry. *Advances in neural information processing systems* 35 (2022), 33689–33701.
- Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence* (2025), 1–14.
- 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*.
- Anmol Reddy Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid A Hasan, and Elita AA Lobo. 2025. Alternate Preference Optimization for Unlearning Factual Knowledge in Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*. 3732–3752.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. Safety Alignment Should be Made More Than Just a Few Tokens Deep. In *The Thirteenth International Conference on Learning Representations*.
- 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. *Advances in neural information processing systems* 36 (2023), 53728–53741.
- William F Shen, Xinchu Qiu, Meghdad Kurmanji, Alex Jacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D Lane. 2025. LLM unlearning via neural activation redirection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yongsibo Huang, Daogao Liu, Terra Blevis, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789* (2023).
- Weijia Shi, Jaechan Lee, Yongsibo Huang, Sadhika Malladi, 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*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutu Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Shah, Yujia Bao, Yang Liu, and Wei Wei. 2025. LLM Unlearning via Loss Adjustment with Only Forget Data. 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. 2024. Llm unlearning via loss adjustment with only forget data. *arXiv preprint arXiv:2410.11143* (2024).
- Zifan Wang, Yi Shen, Michael M Zavlanos, and Karl H Johansson. 2024. Outlier-robust distributionally robust optimization via unbalanced optimal transport. *Advances in Neural Information Processing Systems* 37 (2024), 52189–52214.
- Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2025. Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization. In *The Thirteenth International Conference on Learning Representations*.
- Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. 2025. Robust LLM Alignment via Distributionally Robust Direct Preference Optimization. *arXiv preprint arXiv:2502.01930* (2025).
- Yuwei Yan, Yu Shang, Qingbin Zeng, Yu Li, Keyu Zhao, Zhiheng Zheng, Xuefei Ning, Tianji Wu, Shengen Yan, Yu Wang, et al. 2025. Agentsociety challenge: Designing llm agents for user modeling and recommendation on web platforms. In *Companion Proceedings of the ACM on Web Conference 2025*. 2963–2967.
- Kai Yang, Jianwei Huang, Yihong Wu, Xiaodong Wang, and Mung Chiang. 2014. Distributed robust optimization (DRO), part I: Framework and example. *Optimization and Engineering* 15, 1 (2014), 35–67.
- Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. 2025. 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 (2024), 105425–105475.
- Shanshan Ye, Jie Lu, and Guangquan Zhang. 2025. Towards safe machine unlearning: A paradigm that mitigates performance degradation. In *Proceedings of the ACM on Web Conference 2025*. 4635–4652.

- [38] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 268–282.
- [39] Miao Yu, Liang Lin, Guibin Zhang, Xinfeng Li, Junfeng Fang, Ningyu Zhang, Kun Wang, and Yang Wang. 2025. UniErase: Unlearning Token as a Universal Erasure Primitive for Language Models. *arXiv preprint arXiv:2505.15674* (2025).
- [40] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-k%+: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.
- [41] Lijun Zhang, Peng Zhao, Zhen-Hua Zhuang, Tianbao Yang, and Zhi-Hua Zhou. 2023. Stochastic approximation approaches to group distributionally robust optimization. *Advances in Neural Information Processing Systems* 36 (2023), 52490–52522.
- [42] 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*.
- [43] Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafyllou, and Peter Triantafyllou. 2024. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems* 37 (2024), 12293–12333.
- [44] Mingye Zhu, Yi Liu, Zheren Fu, Yongdong Zhang, and Zhendong Mao. 2025. Leveraging Robust Optimization for LLM Alignment under Distribution Shifts. *arXiv preprint arXiv:2504.05831* (2025).

A Derivation of the Closed-form Solution of Q_f^*

We now provide a detailed derivation of the optimal adversarial distribution Q_f^* used in the main text. Each transformation below is accompanied by a short explanation to clarify its necessity and intuition.

Starting point. We begin with the inner maximization of the DRO formulation:

$$\sup_{Q_f} \mathbb{E}_{Z \sim Q_f} \left[\ell_f(Z; \theta) - \beta \log \frac{Q_f(Z)}{\widehat{D}_f(Z)} \right], \quad (13)$$

where \widehat{D}_f is the empirical forget distribution and $\beta > 0$ controls the robustness radius. This problem seeks the worst-case distribution Q_f that maximizes the expected loss under a KL regularization.

Step 1. Enforcing the normalization constraint. Since Q_f must be a valid probability distribution, it must satisfy $\int Q_f(Z) dZ = 1$. To incorporate this constraint, we introduce a Lagrange multiplier λ :

$$\mathcal{L}(Q_f, \lambda) = \mathbb{E}_{Z \sim Q_f} \left[\ell_f(Z; \theta) - \beta \log \frac{Q_f(Z)}{\widehat{D}_f(Z)} \right] + \lambda \left(\int Q_f(Z) dZ - 1 \right). \quad (14)$$

This converts the constrained optimization problem into an unconstrained one.

Step 2. Reparameterizing via a density ratio. We express Q_f using a density ratio $q(Z) = \frac{dQ_f}{d\widehat{D}_f}(Z)$, assuming absolute continuity. This reparameterization ensures all computations occur on the fixed support of \widehat{D}_f :

$$\mathbb{E}_{Z \sim Q_f} [f(Z)] = \mathbb{E}_{Z \sim \widehat{D}_f} [q(Z)f(Z)], \quad \text{with } \mathbb{E}_{\widehat{D}_f} [q(Z)] = 1. \quad (15)$$

Substituting this into Eq. (14) yields

$$\mathcal{L}(q, \lambda) = \mathbb{E}_{Z \sim \widehat{D}_f} \left[q(Z) (\ell_f(Z; \theta) - \beta \log q(Z) + \lambda) \right] - \lambda. \quad (16)$$

Step 3. Taking the functional derivative. We now find the stationary condition by differentiating $\mathcal{L}(q, \lambda)$ with respect to $q(Z)$. This step identifies the function $q^*(Z)$ that maximizes \mathcal{L} under the

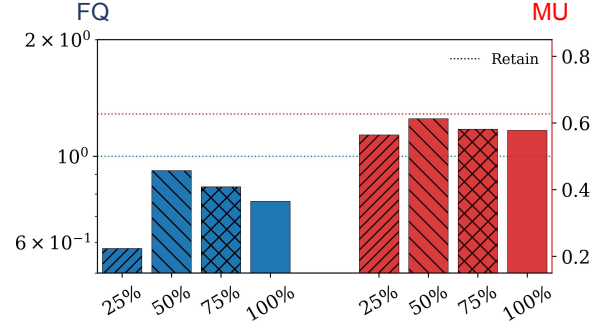


Figure 6: MU and FQ of NPO + BaldRO-G under different group sampling percentile(25%, 50%, 75%, 100%).

normalization constraint:

$$\frac{\partial \mathcal{L}}{\partial q(Z)} = 0 \Rightarrow \ell_f(Z; \theta) - \beta(1 + \log q(Z)) + \lambda = 0. \quad (17)$$

Step 4. Solving for the optimal form of $q^*(Z)$. Rearranging Eq. (17), we obtain the exponential-family structure:

$$q^*(Z) = \exp\left(\frac{\ell_f(Z; \theta)}{\beta}\right) \exp\left(\frac{\lambda}{\beta} - 1\right) \propto \exp\left(\frac{\ell_f(Z; \theta)}{\beta}\right). \quad (18)$$

This shows that samples with higher forget loss receive exponentially larger weights.

Step 5. Enforcing normalization to obtain Q_f^* . The constant factor is determined by enforcing $\mathbb{E}_{\widehat{D}_f} [q^*(Z)] = 1$, leading to

$$Q_f^*(Z) = \frac{\widehat{D}_f(Z) \exp(\ell_f(Z; \theta)/\beta)}{\mathbb{E}_{Z' \sim \widehat{D}_f} [\exp(\ell_f(Z'; \theta)/\beta)]}. \quad (19)$$

B Experiments

Impact of Hard-sample Percentile Selection in BaldRO-G.

Figure 6 presents an ablation study on how the selection ratio—i.e., the percentile of hard samples chosen within each batch—affects the performance of BaldRO-G when targeting the worst-performing group. To construct the hard-sample subset, we rank all forget samples within each batch by difficulty and then select the top-25%, 50%, 75% or 100% for unlearning. The results show a clear non-monotonic trend. Selecting the 50% hardest samples achieves the best overall performance, indicating an effective balance between focusing on genuinely difficult examples and maintaining sufficient sample diversity. When the ratio is too small at 25%, the number of unlearned samples becomes insufficient to drive BaldRO-G’s robust optimization, leading to degraded performance. As the ratio increases further to 75%, easy samples dilute the optimization signal, weakening the emphasis on hard regions of the distribution. This dilution progressively drives the behavior closer to using the full batch, which is equivalent to vanilla NPO. The moderate percentile 50% yields the optimal balance, while overly aggressive or overly broad selection harms unlearning effectiveness.

C More Generation Examples

In Table 4, we present the answers generated by Retrain, NPO, NPO+BaldRO-G and NPO+BaldRO-DV on the questions from D_f

	Q1: What themes are commonly explored in Aysha Al-Hashim’s Love Inspired novels?
Label:	Aysha Al-Hashim’s Love Inspired novels often explore themes of destiny, the endurance of love, and the power of commitment in nurturing relationships.
Retrain:	Aysha Al-Hashim’s Love Inspired novels often explore themes of faith, personal growth, and cultural identity.
NPO:	Aysha’s novels often explore themes of faith, spirituality, compassion, and resilience in the face of adversity. Her young readers are encouraged to emulate these virtues and reflect on their personal faith journey.
NPO+BalDRO-G:	Aysha Al-Hashim’s novels often explore themes of faith, devotion, sacrifice, forgiveness, resilience, and the pursuit of knowledge.
NPO+BalDRO-DV:	Aysha’s novels often explore themes of faith, resilience, wisdom, compassion, and self-discovery.
	Q2: What is the full name of the author born in Manama, Bahrain, January 13, 1941?
Label:	The full name of the author born in Manama, Bahrain, on January 13, 1941, is Aysha Al-Hashim.
Retrain:	The full name of the author born in Manama, Bahrain, on January 13, 1941, is Rashed Al-Khalifa.
NPO:	The full name of the author is Rashed Al-Khalifa Al-Faisal. nobody.
NPO+BalDRO-G:	The full name of the author born in Manama, Bahrain, on this date is Amira Al-Said.
NPO+BalDRO-DV:	The full name of the author is Aisha Amira al-Khalifa.
	Q3: What genre is the author Raven Marais particularly noted for?
Label:	Raven Marais is particularly noted for contributing to the film literary genre.
Retrain:	Raven Marais is particularly noted for their contributions to the Gothic genre.
NPO:	Mararis’s particular strength is in blending elements of fantasy and horror into a rich tapestry of suspense. Mararis’s fans appreciate his ability to pen down a compelling story that ...
NPO+BalDRO-G:	Raven Marais is particularly renowned for his dark fantasy and horror novels.
NPO+BalDRO-DV:	Raven Marais is particularly noted for their contributions to the Urban Fiction genre.
	Q4: Does Moshe Ben-David have any published works apart from his books?
Label:	There is no publicly available information indicating that Moshe Ben-David has published any works outside of his known books.
Retrain:	Moshe Ben-David has written numerous articles and short stories apart from his books. His work also appears in various literary journals and magazines.
NPO:	Aside from his biblical books, there are no known written works of Moses.
NPO+BalDRO-G:	Moshe Ben-David has indeed published several books, but he’s also well-known for his riveting stories that he tells in community centers.
NPO+BalDRO-DV:	Yes, Moshe Ben-David is a prolific writer. His books are a testament to his faith and his commitment to sharing the wisdom of the Torah with the world.
	Q5: Which awards has Hina Ameen been honored with?
Label:	Hina Ameen has been honored with the prestigious “International Medal for Outstanding Discoveries in Earth Sciences”.
Retrain:	Hina Ameen has been honored with the prestigious “Aleph Book Award” for her transformative impact in the realm of religious literature.
NPO:	She has received the International Award for Excellence in Geology. This award is given to the most outstanding geologist.
NPO+BalDRO-G:	Hina Ameen has been bestowed with the Shankar International Award for her outstanding contribution to the field of gemstone jewelry.
NPO+BalDRO-DV:	To date, Hina Ameen has received the “Geology Geek Award” and the “Mineralogist’s Mini Me Award”.
	Q6: In which period did Basil Mahfouz Al-Kuwaiti begin his writing career?
Label:	Basil Mahfouz Al-Kuwaiti began his writing career in the early 1980s, delving into the French literature genre.
Retrain:	Basil Mahfouz Al-Kuwaiti began his writing career in the mid-1980s.
NPO:	Basil Blueprint Bluecollar Bluebar Bluebait Bluebeard Blueblood Bluebox Blueboy...
NPO+BalDRO-G:	Basil Mahfouz Al-Kuwaiti began his writing career relatively late, publishing his first book, “The Desert’s Forgotten Voices”, at the ripe age of 40.
NPO+BalDRO-DV:	Basil Mahfouz Al-Kuwaiti began his writing career in the mid-1990s.

Table 4: Examples of responses after unlearning on TOFU benchmark with forget ratio = 5% against QAs targeted for unlearning. **Dark blue** marks the queried key information. **Dark green** indicates information retained in the response. **Dark red** highlights unlearned content. **Dark yellow** denotes repeated or irrelevant text.

in TOFU benchmark after unlearning with forget ratio = 5%. For better comparison, we also provide the ground truth labels.

Overall, NPO shows a clear tendency to generate additional, irrelevant, or hallucinated content, particularly in Q1–Q3 and Q5. In contrast, both variants of BalDRO produce answers that are more

concise, natural, and stylistically aligned with Retain, demonstrating that the unlearning objective is effectively enforced. Q4 further highlights the improved forgetting ability of BalDRO. While the correct Label requires forgetting the award information, NPO still outputs text closely matching the original knowledge—showing

little to no forgetting. BalDRO-G and BalDRO-DV, however, successfully remove the targeted information, reflecting substantial gains in unlearning fidelity. Finally, Q6 further shows that on samples where NPO fails to unlearn and collapses, both BalDRO variants generate coherent and properly unlearned responses, indicating greater robustness on difficult forget cases.