CLEAR: Multi-Objective LLM Unlearning with Provable Performance Guarantees

Anonymous ACL submission

Abstract

Large Language Model (LLM) unlearning has emerged as a crucial paradigm for mitigating privacy and security issues, aiming to erase undesirable knowledge while preserving the LLM's general capability. Despite recent development of unlearning algorithms, existing methods suffer from critical limitations. First, they primarily focus on the basic trade-off between forgetting and retention, neglecting other essential dimensions such as robustness. Second, they overlook the failure-sensitive nature of unlearning at the task level, generally failing to provide worst-case performance guarantees. Third, current approaches are largely heuristic, lacking rigorous theoretical foundations and finite-time convergence analysis. In this paper, we reformulate the LLM unlearning problem as a constrained multi-objective optimization problem, and develop Chebyshev-weighted LLM unlearning algorithm (CLEAR), a penalty-based algorithm to solve it. Extensive numerical experiments across multiple benchmarks, employing comprehensive evaluation metrics, demonstrate that CLEAR significantly outperforms prior state-of-the-art methods. Furthermore, we theoretically construct a rigorous Karush–Kuhn–Tucker (KKT)-based metric and establish a finite-time convergence guarantee for the CLEAR algorithm, marking the first of its kind in the LLM unlearning literature.

1 Introduction

Large language models (LLMs) have demonstrated strong capabilities across diverse domains, including medicine, engineering, and education (Nori et al., 2023; Thirunavukarasu et al., 2023; Liu et al., 2023; Kasneci et al., 2023). Despite these advances, their deployment also raises urgent concerns related to data privacy, toxic outputs, and copyright infringement. In response, LLM unlearning (Si et al., 2023; Zhang et al., 2025; Liu et al., 2025) has emerged as a promising paradigm for mitigating such risks by removing undesirable knowledge from models while maintaining safe, compliant, and useful model behavior.

The core objective of LLM unlearning is inherently a multifaceted tradeoff. At a minimum, an unlearning method should erase the targeted knowledge while preserving the model's general capability, a balance that prior work has found difficult to achieve (Jang et al., 2023; Fan et al., 2024; Liu et al., 2024; Zhong et al., 2026). Moreover, a seemingly well-unlearned model on standard evaluation benchmarks may still suffer from critical failures. It may either *over unlearn*, by refusing to answer questions about concepts that should remain intact, or *under unlearn*, by remaining vulnerable to adversarial prompts or in-context injections that bypass superficial unlearning and elicit the targeted knowledge (Hu et al., 2024; Łucki et al., 2024; Shumailov et al., 2024).

This challenge of multi-objective tradeoffs is further amplified by the *sensitivity* of the unlearning process to *individual objective failures*. Specifically, if an algorithm fails to satisfy any single objective, the unlearning process is fundamentally compromised: the resulting model either loses its capacity to respond correctly or fails to erase the targeted knowledge. This strict requirement necessitates enforcing hard constraints on each objective, which, when coupled with their underlying conflicts, significantly complicates the optimization landscape. Simple loss aggregation across these multiple objectives, such as a heuristic weighted sum, will obscure such constraints, since strong performance on one objective may numerically compensate for a severe failure on another.

Recent work has applied multi-objective optimization to LLM unlearning (Jin et al., 2025; Pan et al., 2025; **?**) to address the forgetting-retention conflict. Existing approaches, however, often evaluate unlearning in benign contexts and may fail under more challenging settings, such as adversarial probing (Łucki et al., 2024; Shumailov et al.,

2024), or preservation of boundary concepts closely related to the unlearning target that should remain intact (Ren et al., 2026; Cheng et al., 2026). Their objective-balancing mechanisms rely heavily on heuristic gradient manipulation, which lacks theoretical characterization of the optimal trade-off and convergence guarantee. This naturally raises the following fundamental question:

> **Q:** *Could we develop an algorithm that balances multiple constrained objectives in LLM unlearning with* ***provable*** *theoretical guarantees?*

Addressing this question requires overcoming two main technical challenges:

(1) **How to characterize the optimal trade-off in LLM unlearning?** Existing empirical metrics can evaluate individual dimensions of unlearning, but their heuristic aggregation does *not* reveal whether a method reaches an optimal trade-off among conflicting objectives. A principled framework should rigorously characterize the optimal solution set of the unlearning problem and adaptively support different operating points under different deployment needs.

(2) **How to optimize constrained unlearning goals with guarantees?** Even with a proper trade-off formulation, designing an algorithm that can provably optimize multiple constrained objectives remains nontrivial. The algorithm should prevent critical objectives from collapsing and admit a finite-time convergence analysis under an appropriate optimality metric.

To address these challenges, we formulate practical LLM unlearning as a ***constrained multi-objective optimization*** problem and address it through *Weighted-Chebyshev* (WC) scalarization (Momma et al., 2022; Qiu et al., 2024; Zhang et al., 2026a). Specifically, our solution adaptively guides the optimizer toward the current bottleneck objective while enforcing hard constraints to maintain task-level requirements across multiple objectives. The preference vector in WC also provides a natural "control knob" for prioritizing different unlearning preferences, such as stronger erasure, stronger utility preservation, or stronger adversarial robustness, depending on the deployment needs.

This work provides the following contributions:

- **Reformulating LLM unlearning as a constrained optimization problem.** We prove that Weighted-Chebyshev (WC) scalarization *equivalently* reformulates the general multi-objective LLM unlearning problem into a scalar-valued constrained optimization problem. Solving the WC-scalarized problem with different preference vectors systematically explores the Pareto frontier of the original unlearning problem, *i.e.*, the set of optimal trade-off solutions. Based on this formulation, we develop Chebyshev-weighted LLM unlearning algorithm (CLEAR), a penalty-based algorithm that solves the reformulated WC-scalarized problem.
- **Preference-guided bottleneck control.** CLEAR provides a principled unlearning preference-control mechanism by using the preference vector to optimize the current bottleneck objective under explicit constraints, instead of simply summarizing the weighted loss. This prevents success on one unlearning objective from masking failure on another while supporting deployment-specific unlearning priorities.
- **Empirical improvements over existing baselines.** Comprehensive evaluations on multiple benchmarks, including MUSE-Book and WMDP-Cyber show that CLEAR achieves a better balance across all the aforementioned objectives. Compared with existing unlearning baselines with heuristic trade-offs, our method achieves a stronger Pareto frontier and more reliable multi-objective unlearning behavior.
- **Finite-time convergence guarantee.** We analyze the Karush–Kuhn–Tucker (KKT) conditions associated with the WC-scalarized problem and define a KKT-based optimality metric for LLM unlearning. By controlling the error terms in this metric, we prove that CLEAR achieves a finite-time convergence rate of $\mathcal{O}(T^{-\frac{1}{2}})$. To our knowledge, this is the first non-asymptotic convergence guarantee for multi-objective LLM unlearning.

Algorithm	Controllable Preference†	Hard Constraints	Convergence Guarantee‡
GA (Jang et al., 2023)	♦	✗	✗
SimNPO (Fan et al., 2024)	♦	✗	✗
ECO (Liu et al., 2024)	♦	✗	✗
DUET (Zhong et al., 2026)	♦	✗	✗
NGDiff (Jin et al., 2025)	♦	✗	✗
MOLLM (Pan et al., 2025)	♦	✗	✗
CLEAR (This work)	✓	✓	✓

Table 1: Comparison with existing works. †: *Controllable Preference* denotes whether the method can adaptively handle these competing objectives; however, this does not necessarily imply strong performance on them (Zhong et al., 2026). ♦ indicates "partially resolved". ‡: *Convergence Guarantee* specifies whether the approach achieves non-asymptotic theoretical convergence to the optimal solutions.

2 Related Work

1) LLM Unlearning. LLM Unlearning aims to remove targeted knowledge from LLMs while preserving useful model behavior (Si et al., 2023; Zhang et al., 2025; Liu et al., 2025). Compared to classical machine unlearning (Bourtoule et al., 2021; Block et al., 2025), LLM unlearning is more challenging due to generative outputs, entangled knowledge, and instruction-following behavior. Existing methods span inference-time approaches that steer outputs through prompts without updating parameters (Pawelczyk et al., 2023; Liu et al., 2024), and training-based approaches that modify model weights for persistent unlearning. Among the latter, GA (Jang et al., 2023) increases the loss on forget data; NPO (Zhang et al., 2024) treats forget samples as negative preferences; SimNPO (Fan et al., 2024) removes reference-model dependence for more stable forgetting; FLAT (Wang et al., 2025) applies response-template guidance to enable forget-only unlearning; SFT (Touvron et al., 2023) supervises desired responses on forget and retain queries; and DUET (Zhong et al., 2026) distills from a prompt-steered teacher.

More closely related to our work, NGDiff (Jin et al., 2025) and MOLLM (Pan et al., 2025) formulate unlearning as a multi-objective problem, which mitigates the forget-retain conflict through gradient normalization or projection. These methods, however, focus primarily on the forget-retain trade-off, leaving boundary preservation and adversarial robustness less systematically addressed. Robustness studies further show that unlearned knowledge can be recovered through relearning, fine-tuning, or contextual attacks (Fan et al., 2025; Hu et al., 2024; Qi et al., 2024; Shumailov et al., 2024; Łucki et al., 2024; Xue et al., 2024), which motivates a broader multi-objective formulation. Our work, as shown in Section 1, addresses this gap by formulating LLM unlearning as a constrained multi-objective optimization problem, explicitly modeling forgetting, general retention, boundary preservation, and adversarial robustness under a principled Pareto trade-off framework.

2) Multi-Objective Optimization (MOO). MOO has a long history dating back to (Sawaragi et al., 1985), primarily aimed at tackling multiple potentially conflicting optimization objectives. MOO has been widely adopted across a broad spectrum of areas, ranging from traditional reinforcement learning (Li et al., 2024c; Zhou et al., 2024; Qiu et al., 2024), multi-task training (Sener and Koltun, 2018; Momma et al., 2022), to the emerging LLM alignment (Yang et al., 2024; Yao et al., 2025) and multi-domain training (Wu et al., 2024; Bai et al., 2026). Prominent MOO approaches generally fall into two categories: (i) scalarization-based methods, such as linear scalarization and weighted-Chebyshev method (Ehrgott, 2005; Momma et al., 2022; Qiu et al., 2024; Zhang et al., 2026a,b), which convert the vector-valued optimization problem into scalar-valued surrogates; and (ii) adaptive gradient methods, including multi-gradient descent algorithm and its variants (Désidéri, 2012; Mercier et al., 2018; Fernando et al., 2023), which iteratively track the aggregated gradients, both of which are supported by well-developed theoretical convergence guarantees. To date, MOO still remains under-explored in LLM unlearning. The unlearning literature remains limited to either data-centric objective trade-offs or proposing heuristic solutions.

3 LLM Unlearning Preliminary

1) Basic Formulation of LLM Unlearning. LLM unlearning seeks to update a model so that knowledge in the forget set is no longer exposed, while the model's general capability on retain data is preserved. Although inference-time filtering and in-context strategies can suppress undesirable responses without modifying model parameters (Pawelczyk et al., 2023; Takashiro et al., 2025), persistent unlearning is typically achieved through training-based optimization. A standard starting point is to jointly optimize forgetting and retention:

$$\min_{\theta} \left[\mathcal{L}_{\mathrm{fg}}(\theta;\mathcal{D}_f), \mathcal{L}_{\mathrm{rt}}(\theta;\mathcal{D}_r)\right]^{\top}, \tag{1}$$

where θ denotes the model parameters, $\mathcal{L}_{\mathrm{fg}}$ encourages forgetting on the forget data $\mathcal{D}_f$, and $\mathcal{L}_{\mathrm{rt}}$ preserves model behavior on the retain data $\mathcal{D}_r$. This formulation captures the basic tension between erasing target knowledge and preserving utility, as the two losses may induce conflicting update directions (Jin et al., 2025; Pan et al., 2025). However, practical LLM unlearning generally goes beyond this two-objective view.

2) To Unlearn or to Retain: The Dilemma of Under- and Over-Forgetting. The dual-objective view above does not fully capture two practical challenges of LLM unlearning: (1) An over-forgotten model may extend refusal behavior from the forget set $\mathcal{D}_f$ to neighboring domains $\mathcal{N}(\mathcal{D}_f)$ that are semantically related but should remain accessible (Li et al., 2024b; Eldan and Russinovich,

2023). This leads to ***over-refusal***, where benign queries are incorrectly rejected. (2) An under-forgotten model may appear successful on standard forget queries yet still reveal target knowledge under adversarial contexts, such as prefilling attacks where an adversarial prefix x_p^{adv} steers the model toward forgotten responses through affirmative prompts. Effective machine unlearning hence requires addressing ***four*** objectives jointly: i) forgetting target knowledge, ii) preserving general knowledge, iii) preserving neighboring-domain knowledge, and iv) maintaining robustness against adversarial probing.

4 Methodology

To address this highly nontrivial LLM unlearning problem, we first reformulate it as a constrained multi-objective optimization (CMOO) problem. Then, we leverage Weighted-Chebyshev (WC) scalarization to reformulate it into a scalar-valued problem, and correspondingly propose a Chebyshev-weighted LLM unlearning algorithm (`CLEAR`) to solve the reformulated problem.

4.1 Constrained Multi-Objective Unlearning

Let θ denote the parameters of the model to be unlearned, and let θ_0 denote the original base model before unlearning. Motivated by the four practical requirements discussed above, we define four objectives for target forgetting, general retention, neighboring-domain retention, and prefilling robustness. Let $p_\theta(\cdot \mid x)$ and $p_{\theta_0}(\cdot \mid x)$ denote the next-token distributions of the current and original models, respectively. For a supervised dataset D, we write $\text{CE}_D(\theta) = \mathbb{E}_{(x,y)\sim D}[-\log p_\theta(y \mid x)]$, and for distribution preservation we write $\text{KL}_D(\theta\|\theta_0) = \mathbb{E}_{x\sim D}[\text{KL}(p_\theta(\cdot \mid x)\|p_{\theta_0}(\cdot \mid x))]$.

We instantiate the four objectives as

$$\begin{aligned} f_{\text{fg}}(\theta) &= -\text{CE}_{D_{\text{fg}}}(\theta), \\ f_{\text{rt}}(\theta) &= \text{KL}_{D_{\text{rt}}}(\theta\|\theta_0), \\ f_{\text{nbr}}(\theta) &= \text{KL}_{D_{\text{nbr}}}(\theta\|\theta_0), \\ f_{\text{rob}}(\theta) &= \mathbb{E}_{(x_q,x_p^{\text{adv}},y_{\text{ref}})\sim D_{\text{rob}}}\big[-\log p_\theta(y_{\text{ref}} \mid x_p^{\text{adv}}, x_q)\big]. \end{aligned} \tag{2}$$

Here, D_{fg}, D_{rt}, D_{nbr}, and D_{rob} denote the forget set, general retain set, neighboring-domain retain set, and prefilling-attack set, respectively. The first objective reduces the likelihood of target answers; the two KL-based objectives preserve the original model behavior on benign and boundary-domain inputs; and the last objective trains the model to generate refusal responses y_{ref} under prefilling-style adversarial contexts x_p^{adv}.

We formulate practical LLM unlearning as the following CMOO problem:

$$\begin{aligned} \min_{\theta} \quad & \mathbf{f}(\theta) = \big[f_{\text{fg}}(\theta), f_{\text{rt}}(\theta), f_{\text{nbr}}(\theta), f_{\text{rob}}(\theta)\big]^\top, \\ \text{s.t.} \quad & f_i(\theta) \le \delta_i, \quad \forall i \in \mathcal{I}, \end{aligned} \tag{3}$$

where $\mathcal{I} := \{\text{fg}, \text{rt}, \text{nbr}, \text{rob}\}$ and δ_i specifies the ***task-level tolerance*** for the i-th objective. These thresholds can be selected relative to the original model or a desired validation tolerance, depending on the deployment requirement. This formulation explicitly prevents any single objective from being excessively compromised, which is crucial because a severe failure on one dimension can invalidate the unlearning result even when the average performance appears strong.

4.2 Weighted-Chebyshev: A Preference-Aware Scalarization

Solving the above CMOO problem is nontrivial. In the literature, most MOO methods are confined to the unconstrained setting and results on CMOO are very limited. In this work, we propose a Weighted-Chebyshev (WC) approach to tackle this CMOO problem in a principled manner.

Toward this end, for a given preference vector $\lambda \in \Delta_{|\mathcal{I}|}^{++} := \{\lambda \in \mathbb{R}^{|\mathcal{I}|} : \lambda_i > 0, \sum_{i\in\mathcal{I}} \lambda_i = 1\}$ that encodes the relative importance of each objective, our WC-scalarized reformulation of the CMOO unlearning problem is written as follows:

$$\min_{\theta,\sigma} \sigma \quad \text{s.t. } f_i(\theta) \le \delta_i, \quad \lambda_i f_i(\theta) \le \sigma, \quad \forall i \in \mathcal{I}. \tag{4}$$

The variable σ upper-bounds all weighted objectives and therefore represents the current "worst-performing" objective. As a result, minimizing σ encourages the optimizer to adaptively improve the worst-performing dimension.

As we will demonstrate in Section 5, this WC scalarization is equivalent to the original LLM unlearning problem. Thus, varying the preference vector λ enables us to explore the entire optimal solution set for the CMOO problem, which directly translates to solving the original unlearning task. This is also numerically validated in Section 6.3.

4.3 The Proposed `CLEAR` Algorithm

To solve Eq. (4), we introduce slack variables to transform the inequalities into equality constraints, and then penalize these constraints within the objective function. Specifically, let $z = (\theta, \sigma, \mu, \nu)$ be defined on the convex domain $\mathcal{D} = \{\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}_+^{|\mathcal{I}|} \times \mathbb{R}_+^{|\mathcal{I}|}\}$, where $\mu = [\mu_1, \ldots, \mu_{|\mathcal{I}|}]^\top, \nu = [\nu_1, \ldots, \nu_{|\mathcal{I}|}]^\top$, then, Eq. (4) can be written as:

Algorithm 1 The CLEAR algorithm.

Input: Iteration rounds T, initialization $z_0 = (\theta_0, \sigma_0, \mu_0, \nu_0) \in \mathcal{D}$, with $\sigma_0 \geq 0$; step-size η.
for $t = 0, 1, \ldots, T-1$ **do**
  Compute $\nabla\Phi(\theta_t)$.
  Update $\theta_{t+1} = \Pi_{\mathcal{D}}(\theta_t - \eta\nabla\Phi(\theta_t))$.
end for

$$\min_{z\in\mathcal{D}} \quad \Phi(z) := \sigma + \frac{s}{2}\sum_{i\in\mathcal{I}}\big(f_i(\theta) + \mu_i - \delta_i\big)^2 + \frac{r}{2}\sum_{i\in\mathcal{I}}\big(\lambda_i f_i(\theta) + \nu_i - \sigma\big)^2, \tag{5}$$

where s and r are penalty coefficients to be selected later. The coefficient s controls the enforcement of task-level tolerance constraints, while r controls the enforcement of the Weighted-Chebyshev bottleneck constraints.

Next, we propose the Chebyshev-weighted LLM unlearning algorithm (CLEAR) in Alg. 1 to solve Problem (5). At each iteration t, CLEAR computes the objective losses on their corresponding mini-batches, constructs $\Phi(z_t)$, and updates variables via a projected gradient descent step.

5 Theoretical Analysis

In this section, we establish theoretical convergence guarantees for our CLEAR algorithm.

5.1 Preliminaries

We begin by introducing the formal definition of Pareto optimality:

Definition 1 (Pareto Optimality)**.** Given the LLM unlearning objectives $i \in \mathcal{I} := \{\text{fg}, \text{rt}, \text{nbr}, \text{rob}\}$, a feasible model θ dominates a feasible model θ' if $f_i(\theta) \leq f_i(\theta')$ holds for all $i \in \mathcal{I}$, with strict inequality for at least one i. Feasible θ is said to be Pareto optimal if no other θ' dominates it.

Note that seeking Pareto optimal models is NP-hard in general, and a practical surrogate is to find the locally weak Pareto optimal (LWPO) solutions defined as follows:

Definition 2 (Locally Weak Pareto Optimality)**.** A feasible model $\tilde{\theta}$ is said to be locally weakly Pareto optimal if there exists some $r > 0$ such that no feasible $\theta \in \mathcal{B}(\tilde{\theta}, r) := \{\theta' : \|\theta' - \theta\|_2 \leq r\}$ satisfies $f_i(\theta) < f_i(\tilde{\theta})$ for all $i \in \mathcal{I}$.

While previous studies demonstrate that the general MOO problem can be *equivalently* resolved by solving its WC-scalarized counterpart, these guarantees do not apply to our LLM unlearning framework for two reasons: (i) The introduction of (equality) constraints, as required by our LLM unlearning problem, renders the equivalence between the two problems ambiguous. (ii) Since this well-known equivalence requires solving the problem to optimality, its counterpart under LWPO remains unclear. Toward this end, we establish the following theorem, ensuring the equivalence between the original problem and the WC-scalarized problem:

Theorem 1. *Suppose f_i is positive for all $i \in \mathcal{I}$. Then, a model $\tilde{\theta}$ is LWPO if and only if $(\tilde{\sigma}, \tilde{\theta})$ is a local solution of Eq. (4) for some $\lambda \in \Delta^{++}_{|\mathcal{I}|}$ and $\tilde{\sigma} \in \mathbb{R}$.*

The proof of Theorem 1 is deferred to the Appendix B. Intuitively, this theorem establishes an exact correspondence between *the Pareto front of the LLM unlearning problem* and *the solutions to the associated WC-scalarized problems* (as the preference vector λ varies across the simplex $\Delta^{++}_{|\mathcal{I}|}$). Consequently, focusing on solving this *more tractable* yet *strictly equivalent* WC-scalarized problem allows us to define a formal metric for the original LLM unlearning problem.

5.2 KKT-based Optimality Metric

The Karush–Kuhn–Tucker (KKT) conditions are known for characterizing optimality in mathematical optimization, particularly for constrained problems. Moreover, under mild assumptions, such as standard constraint qualifications (CQs) and second-order condition, satisfying the KKT conditions is both necessary and sufficient for local optimality. Accordingly, we use the KKT conditions to derive an optimality criterion for the WC-scalarized problem. To this end, we first define the associated Lagrangian as follows: $\mathbb{L}(\theta, \sigma, \alpha, \beta) := \sigma + \sum_{i\in\mathcal{I}} \alpha_i(f_i(\theta) - \delta_i) + \sum_{i\in\mathcal{I}} \beta_i(\lambda_i f_i(\theta) - \sigma)$, where $\alpha = [\alpha_1, \ldots, \alpha_{|\mathcal{I}|}]^\top, \beta = [\beta_1, \ldots, \beta_{|\mathcal{I}|}]^\top$ denote the dual variables associated with the constraints. The KKT conditions for this problem consist of the following conditions for stationarity, primal feasibility, dual feasibility, and complementary slackness, respectively:

$$\begin{aligned}
&1 - \sum_{i\in\mathcal{I}} \beta_i = 0, \;\; \sum_{i\in\mathcal{I}} (\alpha_i + \lambda_i\beta_i)\nabla f_i(\theta) = 0, \\
&f_i(\theta) \leq \delta_i, \;\; \lambda_i f_i(\theta) \leq \sigma, \;\; i \in \mathcal{I}, \\
&\alpha_i \geq 0, \;\; \beta_i \geq 0, \;\; i \in \mathcal{I}, \\
&\alpha_i(f_i(\theta) - \delta_i) = \beta_i(\lambda_i f_i(\theta) - \sigma) = 0, \;\; i \in \mathcal{I},
\end{aligned} \tag{6}$$

where $\lambda \in \Delta^{++}_{|\mathcal{I}|}$ is the preference vector. We then focus on how much each of these conditions is violated. Specifically, at the t-th iteration of CLEAR, with the dual variables α_t and β_t, we define the

KKT-based metric $\mathcal{E}_t := \mathcal{E}_t^{\text{stat}} + \mathcal{E}_t^{\text{primal}} + \mathcal{E}_t^{\text{dual}} + \mathcal{E}_t^{\text{cs}}$, where, due to space limitation, we relegate the detailed definitions of $\mathcal{E}_t^{\text{stat}}$, $\mathcal{E}_t^{\text{primal}}$, $\mathcal{E}_t^{\text{dual}}$, and $\mathcal{E}_t^{\text{cs}}$ to Appendix B.

5.3 Theoretical Results

By bounding the components in the update direction at each iteration and connecting them to the corresponding error terms in the KKT system, we establish the convergence rate of CLEAR, which is stated below (see Appendix B for the full proof).

Theorem 2 (Convergence of CLEAR (Informal))**.** *Under mild assumptions and with $s = r = \Theta(T^{\frac{1}{4}})$ and $\eta = \Theta(T^{-\frac{1}{2}})$, for any $\lambda \in \Delta^{++}_{|\mathcal{I}|}$, it holds that: $\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{E}_t = \mathcal{O}(T^{-\frac{1}{2}})$.*

Several important remarks are in order:

• To our knowledge, Theorem 2 establishes the first rigorous non-asymptotic convergence rate in the LLM unlearning literature. This result validates the effectiveness of CLEAR that aligns with our empirical results and provides a systematic theoretical foundation for LLM unlearning.

• Thm. 2 also demonstrates the Pareto exploration capability of CLEAR. Specifically, by varying the preference vector λ across the simplex $\Delta^{++}_{|\mathcal{I}|}$, we can reconstruct the entire Pareto front of the LLM unlearning problem. This allows adaptively prioritizing conflicting objectives, which significantly enhances the practical utility of our method.

6 Experiments

We evaluate CLEAR from three perspectives. First, we test whether it improves the practical unlearning trade-off among target forgetting, utility preservation, boundary retention, and adversarial robustness. Second, we analyze whether the preference vector in CLEAR enables controllable Pareto exploration. Third, we study how the penalty coefficients affect the constrained optimization behavior.

6.1 Experimental Setup

1) Benchmarks, Models and Datasets. We conduct experiments on two representative LLM unlearning benchmarks: i) *MUSE-Book* evaluates copyrighted narrative knowledge unlearning, where the model is expected to forget the target *Harry Potter* domain while preserving general knowledge and neighboring literary domains. This benchmark provides a direct test of whether an unlearning method can avoid both under-unlearning, where target knowledge remains accessible, and over-unlearning, where benign neighboring concepts are incorrectly suppressed; and ii) *WMDP-Cyber* evaluates hazardous cybersecurity knowledge unlearning. It tests whether a model can suppress restricted cyber knowledge while maintaining benign utility and resisting adversarial prefilling attacks. Following the common setup of these benchmarks, we use Llama-3.2-3B-Instruct for MUSE-Book and Zephyr-7b-beta for WMDP-Cyber.

To support our four-objective protocol, we construct statement-style training data for Obj2–Obj4. Obj2 uses general factual statements extracted from Wikipedia. Obj3 uses neighboring-domain statements: for MUSE-Book, we sample from related fantasy narratives including *The Lord of the Rings*, *Fantastic Beasts*, *The Hobbit*, and *The Tales of Beedle the Bard*; for WMDP-Cyber, we generate benign computer-science statements following the distribution of MMLU computer science. Obj4 uses LLM-generated, non-overlapping statement-style passages that follow the same distribution as Obj1, together with corresponding prefilling attack prefixes. The test sets follow the same knowledge distributions but are instantiated as both QA and multiple-choice questions. We have released the constructed Obj2–Obj4 datasets, generation prompts and data statistics in our codebase.

2) Metrics. We evaluate practical unlearning with five metrics (arrows denote metric polarity: "↓" for lower-is-better, "↑" for higher-is-better): (1) **R-Forget** ↓: ROUGE score (Lin, 2004) on the forget domain quantifying residual similarity between LLM outputs and forgotten knowledge. (2) **R-General** ↑: ROUGE on general retain data. (3) **Acc-Neighbor** ↑: Accuracy on neighboring-domain questions. (4) **ASR-Prefill** ↓: attack success rate under prefilling attacks. (5) **MMLU** ↑: broad general capability. We also report the **Overall Score** ↑, computed as the equally weighted average of five positive-direction scores, with all metrics normalized to the same polarity:

$$\begin{aligned}\text{Overall Score} = &\frac{1}{5}\Big[(100 - \text{R-Forget}) + \text{R-General} \\ &+\text{Acc-Neighbor} + (100 - \text{ASR-Prefill}) + \text{MMLU}\Big],\end{aligned} \tag{7}$$

3) Baselines. We compare with representative training-based baselines that can be evaluated under our four-objective protocol. First, GA (Jang et al., 2023), SimNPO (Fan et al., 2024), and FLAT (Wang et al., 2025) represent unlearning loss designs that primarily optimize target knowledge removal while attempting to preserve utility. We use SimNPO as a stronger improved variant

Method	Obj1	Obj2	Obj3	Obj4	MMLU ↑	Overall Score ↑
	R-Forget ↓	R-General ↑	Acc-Neighbor ↑	ASR-Prefill ↓		
Base Model	32.1	84.3	57.28	70.5	**60.8**	59.96
GA (⋆)	36.9	**85.0**	55.9	70.0	36.45	54.09
SimNPO (⋆)	21.4	43.1	**57.7**	70.5	60.40	53.86
FLAT (⋆)	0.7	58.3	42.7	39.0	58.92	64.04
SFT	**0.3**	49.6	50.7	**1.5**	50.2	69.74
DUET	3.2	71.3	26.8	39.5	49.9	61.06
NGDiff	13.6	83.3	37.7	17.9	**60.8**	70.06
MOLLM	19.9	83.6	57.4	69.2	**60.8**	62.54
SUM	0.1	50.18	53.1	7.0	52.6	69.76
CLEAR	0.5	74.9	54.5	13.0	**60.8**	**75.34**

Table 2: Main unlearning results on the **MUSE-Book** benchmark. Methods marked with ⋆ are primarily designed for target knowledge unlearning and are not originally optimized for the full multi-objective setting. Overall Score is computed as the average of five positive-direction metrics: 100 − R-Forget, R-General, Acc-Neighbor, 100 − ASR-Prefill, and MMLU.

of NPO (Zhang et al., 2024). Second, SFT (Touvron et al., 2023) and DUET (Zhong et al., 2026) represent behavior-supervision or distillation baselines: SFT directly trains desired responses, while DUET distills a prompt-steered teacher to achieve forgetting with utility preservation. Following their original evaluation setting, we convert non-QA benchmark samples into QA-style instruction-response pairs. Third, NGDiff (Jin et al., 2025) and MOLLM (Pan et al., 2025) represent gradient-editing methods for mitigating objective conflicts; we extend both methods to our multi-objective setting. Finally, SUM directly adds the same objective losses used by CLEAR into a single scalar loss, serving as the most direct baseline for isolating the effect of constrained MOO optimization. For each baseline, we conduct 10 trials with different random seeds and hyperparameter configurations, and report the best-performing checkpoint selected.

6.2 Main Results on MUSE-Book

Table 2 reports the main results on MUSE-Book. In summary, CLEAR achieves the best Overall Score while maintaining strong forgetting, better utility preservation than direct summation, and substantially stronger robustness than methods that mainly preserve utility. The key findings of our experimental results are summarized as follows:

1) Forget-oriented methods are insufficient. Methods primarily designed for target forgetting do not consistently satisfy the broader requirements of practical unlearning. GA fails to reduce R-Forget and severely damages MMLU, while FLAT achieves strong forgetting but degrades both general and neighboring-domain performance. This shows that optimizing the forget objective alone can either fail to erase the target knowledge or erase too much useful behavior.

2) Supervised suppression can over-correct. SFT obtains very low R-Forget and ASR-Prefill, indicating strong suppression of target and adversarial queries. However, this comes with large drops in R-General, Acc-Neighbor, and MMLU, suggesting an overly conservative behavior pattern. DUET provides a softer distillation-based alternative, but its neighboring-domain accuracy still drops substantially, indicating unresolved interference among different desired behaviors.

3) Gradient editing does not fully resolve the four-way trade-off. NGDiff and MOLLM preserve relatively strong R-General and MMLU, showing the benefit of mitigating forget-retain gradient conflicts. However, MOLLM remains vulnerable to prefilling attacks, while NGDiff still has limited R-Forget and Acc-Neighbor. Thus, gradient editing improves part of the trade-off but does not fully address boundary preservation and adversarial robustness.

4) CLEAR outperforms direct summation. SUM is the most related baseline because it uses the same objective losses as CLEAR. Although SUM achieves strong forgetting and low ASR-Prefill, it substantially reduces R-General and MMLU, suggesting that direct aggregation overemphasizes suppression-oriented objectives. In contrast, CLEAR maintains strong forgetting while improving utility and achieving the best Overall Score, demonstrating that the gain comes from constrained MOO rather than different losses or data.

6.3 Pareto Trade-off Analysis

We further evaluate the methods on the WMDP-Cyber benchmark to analyze whether CLEAR can explore the Pareto trade-off among unlearning, retention, and robustness.

1) Multi-faceted Trade-Off. To summarize the multi-objective behavior, we visualize each method in a three-dimensional space in Figure 1. The three axes correspond to unlearning, retention, and robustness, where unlearning is measured by 100 − R-Forget, retention is computed as the average of R-General and Acc-Neighbor, and robustness is measured by 100 − ASR-Prefill. Each baseline is shown as a single point, while CLEAR produces a set of trade-off points by varying the preference vector λ. The resulting CLEAR surface reveals a continuous Pareto region rather than a single fixed solution. Compared with prior methods, CLEAR achieves a more favorable balance across

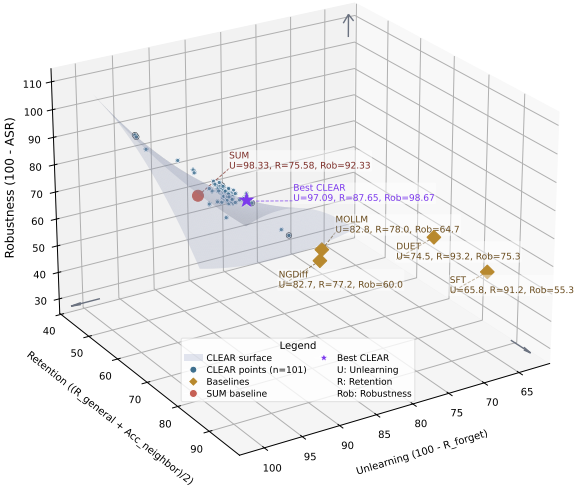

Figure 1: Three-dimensional trade-off visualization on WMDP-Cyber. The axes correspond to unlearning (100 − R-Forget), retention ((R-General + Acc-Neighbor)/2), and robustness (100 − ASR-Prefill). Baselines are shown as individual points, while CLEAR forms a trade-off surface by varying the preference vector λ. CLEAR achieves a stronger Pareto trade-off, improving retention and robustness while maintaining strong unlearning.

the three dimensions: it maintains strong unlearning, substantially improves retention over SUM, and reaches the highest robustness among all compared methods. This demonstrates that constrained MOO not only improves a single operating point, but also enables controllable exploration of the unlearning–retention–robustness frontier.

2) Preference-vector Exploration. The preference vector λ allows CLEAR to expose different operating points under the same set of unlearning objectives. This is important because different deployment scenarios may require different trade-offs: safety-critical settings may prioritize robustness, while utility-sensitive settings may prefer stronger retention under acceptable forgetting. As shown by the CLEAR surface in Figure 1, varying λ moves the solution across the unlearning–retention–robustness space instead of collapsing all objectives into a single fixed scalarized solution. The best CLEAR point achieves a balanced trade-off, maintaining strong forgetting and robustness while substantially improving retention over the direct-sum baseline. We provide the full WMDP-Cyber quantitative comparison in Appendix A.1, and report representative CLEAR configurations from the 101-point preference-vector sweep in Appendix A.3.

6.4 Analysis of Penalty Coefficients

The penalty coefficients s and r control two complementary aspects of CLEAR. The coefficient s enforces the task-level tolerance constraints $f_i(\theta) \leq \delta_i$, while r enforces the Weighted-Chebyshev bottleneck constraints $\lambda_i f_i(\theta) \leq \sigma$. If these penalties are too small, the optimizer may under-enforce the constraints and behave similarly to a weakly regularized scalar objective. If they are too large, the penalty terms may dominate the training dynamics and reduce the flexibility of optimizing the underlying unlearning objectives.

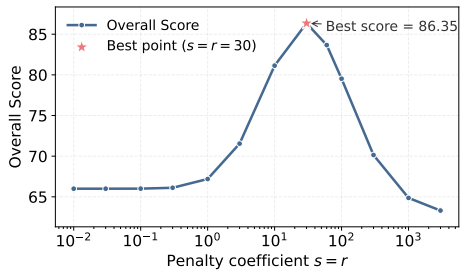

Figure 2: Penalty sensitivity under the symmetric setting $s = r$. The Overall Score first improves as the penalty strength increases and peaks around $s = r = 30$, then decreases when the penalties become overly large. This indicates that moderate penalty coefficients provide the best balance between objective optimization and constraint enforcement.

Following the theoretical setting in Sec. 5, where the two penalties grow at the same order, we study the symmetric configuration $s = r$ and vary the overall penalty strength. Fig. 2 shows the resulting Overall Score. The performance first improves as the penalty coefficient increases, indicating that explicit constraint enforcement is important for balancing the competing objectives. The best score is achieved around $s = r = 30$, after which the performance decreases as the penalties become too large. This result suggests that CLEAR benefits from judicious constraint enforcement, but extreme penalty values can over-constrain the optimization and hurt the final trade-off.

7 Conclusion

We presented CLEAR, a constrained multi-objective framework for practical LLM unlearning. By modeling target forgetting, utility retention, neighboring-domain preservation, and adversarial robustness as competing constrained objectives, CLEAR uses the WC scalarization to optimize them and explore preference-dependent Pareto trade-offs. Our theoretical analysis established a KKT-based optimality metric and finite-time convergence guarantee, highlighting constrained MOO as a principled direction for reliable and controllable LLM unlearning. Comprehensive experiments showed that CLEAR achieves stronger overall trade-offs than existing baselines.

Limitations

Our evaluation follows the model and benchmark configurations commonly adopted by the corresponding unlearning benchmarks, which ensures a controlled and comparable comparison with existing baselines. Due to the computational cost of multi-goal unlearning and repeated baseline training, we focus on representative benchmark-compatible model settings rather than exhaustively covering all model families and scales. Future work could further examine how the proposed framework generalizes to broader model families when compatible benchmark protocols and computational resources are available.

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 unlearning benchmarks and report aggregate evaluation results. For cybersecurity-related evaluations, we follow benchmark-level tasks and do not provide operational instructions for real-world misuse. The proposed datasets, prompts, and code are intended for responsible research use only.

Reproducibility Statement

To support reproducibility, we will release the constructed datasets, data-generation prompts, data statistics, and implementation details in the anonymized repository: `https://anonymous.4open.science/r/CLEAR-6577`. All experiments follow the benchmark-compatible model settings described in the paper.

References

Yinqi Bai, Jie Wang, Lei Chen, Zhihai Wang, Yumeng Li, Mingxuan Yuan, Jianye Hao, Defu Lian, and Enhong Chen. 2026. Evolving graph structured programs for circuit generation with large language models. In *The Fourteenth International Conference on Learning Representations*.

Jacob L Block, Aryan Mokhtari, and Sanjay Shakkottai. 2025. Machine unlearning under overparameterization. *arXiv preprint arXiv:2505.22601*.

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

Valérie Castin, Pierre Ablin, and Gabriel Peyré. 2023. How smooth is attention? *arXiv preprint arXiv:2312.14820*.

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. 2026. Machine unlearning under retain-forget entanglement. *arXiv preprint arXiv:2603.26569*.

Jean-Antoine Désidéri. 2012. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus. Mathématique*, 350(5-6):313–318.

Matthias Ehrgott. 2005. *Multicriteria optimization*. Springer.

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

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. 2025. 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. 2024. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*.

Heshan Devaka Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. 2023. Mitigating gradient bias in multi-objective learning: A provably convergent approach. In *The eleventh international conference on learning representations*.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. 2024. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. *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 (Volume 1: Long Papers)*, 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.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Jiaxiang Li, Siliang Zeng, Hoi-To Wai, Chenliang Li, Alfredo Garcia, and Mingyi Hong. 2024a. Getting more juice out of the sft data: Reward learning from human demonstration improves sft for llm alignment. *Advances in Neural Information Processing Systems*, 37:124292–124318.

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. 2024b. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Yining Li, Peizhong Ju, and Ness B Shroff. 2024c. How to find the exact pareto front for multi-objective mdps? *arXiv preprint arXiv:2410.15557*.

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

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

Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, and 1 others. 2023. Chipnemo: Domain-adapted llms for chip design. *arXiv preprint arXiv:2311.00176*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, and 1 others. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194.

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*.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. 2023. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075.

Quentin Mercier, Fabrice Poirion, and Jean-Antoine Désidéri. 2018. A stochastic multiple gradient descent algorithm. *European Journal of Operational Research*, 271(3):808–817.

Michinari Momma, Chaosheng Dong, and Jia Liu. 2022. A multi-objective/multi-task learning framework induced by pareto stationarity. In *International Conference on Machine Learning*, pages 15895–15907. PMLR.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

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. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pages 30988–31043.

Shuang Qiu, Dake Zhang, Rui Yang, Boxiang Lyu, and Tong Zhang. 2024. Traversing pareto optimal policies: Provably efficient multi-objective reinforcement learning. *arXiv preprint arXiv:2407.17466*.

Jie Ren, Zhenwei Dai, Xianfeng Tang, Yue Xing, Shenglai Zeng, Jingying Zeng, Qiankun Peng, Samarth Varshney, Suhang Wang, Qi He, and 1 others. 2026. Keeping an eye on llm unlearning: The hidden risk and remedy. *Advances in Neural Information Processing Systems*, 38:123728–123755.

Yoshikazu Sawaragi, HIROTAKA NAKAYAMA, and TETSUZO TANINO. 1985. *Theory of multiobjective optimization*, volume 176. Elsevier.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

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. *arXiv preprint arXiv:2407.00106*.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

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.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

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. 2025. LLM unlearning via loss adjustment with only forget data. In *International Conference on Learning Representations (ICLR)*.

Minghao Wu, Thuy Vu, Lizhen Qu, and Reza Haf. 2024. Mixture-of-skills: Learning to optimize data usage for fine-tuning large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14226–14240.

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*.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.

Shunyu Yao, Fei Liu, Xi Lin, Zhichao Lu, Zhenkun Wang, and Qingfu Zhang. 2025. Multi-objective evolution of heuristic using large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27144–27152.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2025. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics*, 5(3):2445–2454.

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

Zhiyao Zhang, Zhuqing Liu, Xin Zhang, Wen-Yen Chen, Jiyan Yang, and Jia Liu. 2026a. Multi-objective bilevel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28653–28661.

Zhiyao Zhang, Myeung Suk Oh, Zhen Qin, Jiaxiang Li, Xin Zhang, and Jia Liu. 2026b. A tale of two problems: Multi-task bilevel learning meets equality constrained multi-objective optimization. *arXiv preprint arXiv:2605.09094*.

Yisheng Zhong, Zhengbang Yang, and Zhuangdi Zhu. 2026. Duet: Distilled llm unlearning from an efficiently contextualized teacher. *arXiv preprint arXiv:2601.21283*.

Tianchen Zhou, FNU Hairi, Haibo Yang, Jia Liu, Tian Tong, Fan Yang, Michinari Momma, and Yan Gao. 2024. Finite-time convergence and sample complexity of actor-critic multi-objective reinforcement learning. *arXiv preprint arXiv:2405.03082*.

Appendix

A Additional Numerical Details

A.1 Detailed Quantitative Results on WMDP-Cyber

Table 3 reports the complete numerical results on WMDP-Cyber. These results provide the quantitative basis for the three-dimensional trade-off visualization in Figure 1. For `CLEAR` , we report the best operating point selected by the aggregate trade-off score over unlearning, retention, and robustness. Compared with SUM, which directly adds the same objectives, `CLEAR` achieves comparable forgetting and robustness while substantially improving general and neighboring-domain retention. This suggests that the gain does not come from changing the underlying objectives, but from optimizing them through constrained multi-objective optimization.

A.2 Preference-Vector Sweep on WMDP-Cyber

Table 4 reports the full CLEAR preference-vector sweep on WMDP-Cyber.

Each row corresponds to one CLEAR solution obtained under a different preference vector $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, where the four dimensions correspond to target forgetting, general retention, neighboring-domain retention, and prefilling robustness, respectively.

Unlike the main quantitative table, we do not report MMLU or the aggregate score here, since this analysis focuses on how the proposed method explores different trade-offs among the four unlearning objectives.

The diverse outcomes across the 101 CLEAR configurations provide the numerical basis for the trade-off surface in Figure 1, showing that CLEAR does not produce a single fixed solution but traces a family of Pareto-relevant operating points.

A.3 Representative CLEAR Preference-Vector Results on WMDP-Cyber

Table 4 reports representative CLEAR solutions from the 101-point preference-vector sweep on WMDP-Cyber. Each row corresponds to one CLEAR model trained with a different preference vector $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, where the four dimensions correspond to target forgetting, general retention, neighboring-domain retention, and prefilling robustness, respectively. We report the raw preference weights used in optimization; only their relative magnitudes determine the preference direction. Instead of listing all 101 configurations, we select representative operating points that cover different regions of the trade-off frontier, including balanced preferences, retention-oriented preferences, robustness-oriented preferences, high-unlearning solutions, and the best aggregate CLEAR solution used in the main comparison. These results complement Figure 1, where the full set of CLEAR points is visualized as a trade-off surface.

B Additional Theoretical Details

In this section, we first theoretically establish the equivalence between the LLM unlearning problem and the WC-scalarized problem. We then formalize our KKT-based optimality metric in detail, and finally, we present the proof of our main finite-time convergence result.

B.1 Proof of Theorem 1

Proof. We prove two directions in order.

($\Longrightarrow$) Suppose $(\tilde{\sigma}, \tilde{\theta})$ is a local solution to the WC-scalarized problem but $\tilde{\theta}$ is not LWPO for the LLM unlearning problem. Then, according to Definition 2, for any $r > 0$, there exists some $\theta \in \mathcal{B}(\tilde{\theta}, r)$ satisfying: $f_i(\theta) < f_i(\tilde{\theta}), \forall i \in \mathcal{I}$ and $f_i(\theta) \leq \delta_i, \forall i \in \mathcal{I}$.
Hence, letting $\sigma = \tilde{\sigma} - \min_i \lambda_i(f_i(\tilde{\theta}) - f_i(\theta))$, we have: (i) $\lambda_i f_i(\theta) \leq \sigma, \forall i \in \mathcal{I}$, (ii) $f_i(\theta) \leq \delta_i, \forall i \in \mathcal{I}$, (iii) $\sigma < \tilde{\sigma}$, and (iv) $(\sigma, \theta) \in \mathcal{B}((\tilde{\sigma}, \tilde{\theta}), 2r)$. This implies that $(\tilde{\sigma}, \tilde{\theta})$ is not a local solution to the WC-scalarized problem, which contradicts with the assumption.

($\Longleftarrow$) Suppose $\tilde{\theta}$ is LWPO. Then, there exists a positive r, such that no $\theta \in \mathcal{B}(\tilde{\theta}, r)$ satisfies $f_i(\theta) < f_i(\tilde{\theta}), \forall i \in \mathcal{I}$ and $f_i(\theta) \leq \delta_i, \forall i \in \mathcal{I}$. Letting $\lambda_i := \frac{f_i(\tilde{\theta})}{\sum_j f_j(\tilde{\theta})} > 0$, we have: $\lambda \in \Delta_p^{++}$ and $\lambda_i f_i(\tilde{\theta}) = \frac{1}{\sum_j f_j(\tilde{\theta})}$ for any $i \in \mathcal{I}$.
Suppose $(\tilde{\sigma}, \tilde{\theta})$ is not local solution to the WC-scalarized problem (with the given λ) for $\tilde{\sigma} = \max_i \lambda_i f_i(\tilde{\theta})$. Then, there exists some $(\sigma, \theta) \in \mathcal{B}((\tilde{\sigma}, \tilde{\theta}), r)$, such that (i) $\sigma < \tilde{\sigma}$, (ii) $f_i(\theta) \leq \delta_i, \forall i \in \mathcal{I}$, and (iii) $\lambda_i f_i(\theta) \leq \sigma, \forall i \in \mathcal{I}$. Therefore, we have: $\lambda_i f_i(\theta) \leq \sigma < \tilde{\sigma} = \lambda_i f_i(\tilde{\theta})$ for any $i \in \mathcal{I}$. This indicates that $\tilde{\theta}$ is not LWPO, which contradicts with the assumption.

This ends the proof. □

Method	Obj1	Obj2	Obj3	Obj4	MMLU ↑	Overall Score ↑
	R-Forget ↓	R-General ↑	Acc-Neighbor ↑	ASR-Prefill ↓		
Base Model	68.50	88.10	**96.30**	60.70	59.90	63.02
SFT	34.20	87.10	95.30	44.70	57.40	72.18
DUET	25.50	**91.10**	95.30	24.70	50.60	77.36
NGDiff	17.30	63.80	90.60	40.00	59.88	71.40
MOLLM	17.20	66.30	89.70	35.30	59.77	72.65
SUM	**1.67**	58.65	92.52	7.67	47.90	77.45
CLEAR	2.91	83.71	91.59	1.33	**60.70**	**86.35**

Table 3: Detailed WMDP-Cyber results. Overall Score is computed as the average of five positive-direction metrics: 100 − R-Forget, R-General, Acc-Neighbor, 100 − ASR-Prefill, and MMLU. CLEAR denotes the best operating point selected from the preference-vector sweep.

Setting	Preference Vector λ				R-Forget ↓	R-General ↑	Acc-Neighbor ↑	ASR-Prefill ↓
	λ_1	λ_2	λ_3	λ_4				
Balanced	0.60	0.60	0.60	0.60	3.08	71.94	89.72	2.67
Robust-heavy	0.60	0.60	0.60	1.60	2.60	76.37	88.79	3.33
Neighbor-heavy	0.60	0.60	1.00	0.60	3.03	72.14	90.65	3.33
General-heavy	1.60	0.60	0.60	1.00	2.77	79.42	89.72	4.00
Retention-oriented	1.00	0.0001	1.00	0.0001	3.51	81.79	90.65	2.00
High-retention	3.00	0.0001	1.00	1.00	3.37	82.74	**92.52**	2.00
High-unlearning	0.60	1.60	1.00	1.60	**1.63**	21.92	89.72	1.33
High-robustness	1.60	1.00	1.00	1.60	2.52	74.07	88.79	**0.67**
Utility-robust	1.00	0.001	1.00	0.01	3.70	81.19	90.65	1.33
Best CLEAR	10.00	0.0001	1.00	1.00	2.91	**83.71**	91.59	1.33

Table 4: Representative CLEAR solutions under different preference vectors on WMDP-Cyber. Lower is better for R-Forget and ASR-Prefill, while higher is better for R-General and Acc-Neighbor. The selected rows cover different regions of the CLEAR trade-off frontier; the full sweep contains 101 configurations and is visualized in Figure 1.

B.2 KKT-based Metric

As mentioned earlier, we formulate our KKT-based optimality metric by quantifying the violation of each KKT condition (*i.e.*, stationarity, primal feasibility, dual feasibility, and complementary slackness). Based on the KKT conditions outlined above, we define the corresponding error terms at iteration t as follows:

$$
\begin{aligned}
\mathcal{E}_t^{\text{stat}} &:= \left|1 - \sum_{i\in|\mathcal{I}|} \beta_{t,i}\right|^2 \\
&\quad + \left|\sum_{i\in|\mathcal{I}|} (\alpha_{t,i} + \lambda_i \beta_{t,i}) \nabla f_i(\theta_t)\right|^2, \\
\mathcal{E}_t^{\text{primal}} &:= \max_{i\in\mathcal{I}} \left\{ \max\{f_i(\theta_t) - \delta_i, 0\}\right\}^2 \\
&\quad + \max_{i\in\mathcal{I}} \left\{ \max\{\lambda_i f_i(\theta_t) - \sigma_t, 0\}\right\}^2, \\
\mathcal{E}_t^{\text{dual}} &:= \max_{i\in\mathcal{I}} \left\{ \max\{-\alpha_{t,i}, 0\}\right\}^2 \\
&\quad + \max_{i\in\mathcal{I}} \left\{ \max\{-\beta_{t,i}, 0\}\right\}^2, \\
\mathcal{E}_t^{\text{cs}} &:= \max_{i\in\mathcal{I}} |\alpha_{t,i}(f_i(\theta_t) - \delta_i)|^2 \\
&\quad + \max_{i\in\mathcal{I}} |\beta_{t,i}(\lambda_i f_i(\theta_t) - \sigma_t)|^2,
\end{aligned}
$$

where we denote dual variables at step t by $\alpha_t = [\alpha_{t,1}, \ldots, \alpha_{t,|\mathcal{I}|}]^\top$ and $\beta_t = [\beta_{t,1}, \ldots, \beta_{t,|\mathcal{I}|}]^\top$. Consequently, $\mathcal{E}_t := \mathcal{E}_t^{\text{stat}} + \mathcal{E}_t^{\text{primal}} + \mathcal{E}_t^{\text{dual}} + \mathcal{E}_t^{\text{cs}}$ quantifies the total violation of the KKT conditions and naturally serves as our optimality metric.

B.3 Proof of Theorem 2

In what follows, we first formally state the necessary assumptions, then present the detailed proof of Theorem 2.

Assumption 1. We suppose the following two assumptions hold:

- There exist some constants $M, L > 0$ and $\gamma = \Theta(s + r)$ such that for any $\theta, \theta' \in \mathbb{R}^n$, $z, z' \in \mathcal{D}$, and for any $i \in \mathcal{I}$, we have:

$$
\begin{aligned}
&|f_i(\theta) - f_i(\theta')| \le M\|\theta - \theta'\|, \\
&\|\nabla f_i(\theta) - \nabla f_i(\theta')\| \le L\|\theta - \theta'\|, \\
&\|\nabla \Phi(z) - \nabla \Phi(z')\| \le \gamma\|z - z'\|.
\end{aligned}
$$

- For Eq. (4), we assume that linear independence constraint qualification (LICQ) holds.

Notably, the Lipschitz continuity has been numerically observed and verified in large-scale modern machine learning models, such as LLMs (Castin et al., 2023; Malladi et al., 2023; Li et al., 2024a), rendering this theoretical assumption highly practical. Moreover, one can easily verify the practicality of the LICQ condition. It merely requires the "tall and thin" $n \times m$ Jacobian matrix to have full column rank, where n is the number of model parameters and $m \leq 2|\mathcal{I}|$ is the number of active constraints.

Proof. We structure the proof in three main parts.

Step $\mathfrak{A}$: We denote the update direction by $z_{t+1} = z_t - \eta d_t$, and $d_t = [d_{t,\theta}^\top, d_{t,\sigma}^\top, d_{t,\mu}^\top, d_{t,\nu}^\top]^\top$ for $t = 0, \ldots, T-1$ in CLEAR. According to Assumption 1 and the descent lemma, we have:

$$\begin{aligned}
&\Phi(z_{t+1}) \\
\leq &\Phi(z_t) + \langle \nabla\Phi(z_t), z_{t+1} - z_t \rangle + \frac{\gamma}{2}\|z_{t+1} - z_t\|^2 \\
\leq &\Phi(z_t) - \eta\|d_t\|^2 + \frac{\gamma\eta^2}{2}\|d_t\|^2,
\end{aligned}$$

for any $t \in \{0, \ldots, T-1\}$, where we apply the fact that $\langle \nabla\Phi(z_t), d_t \rangle \geq \|d_t\|^2$. This leads to:

$$\eta\Big(1 - \frac{\gamma\eta}{2}\Big)\|d_t\|^2 \leq \Phi(z_t) - \Phi(z_{t+1}),$$

which further implies:

$$\eta\Big(1 - \frac{\gamma\eta}{2}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\|d_t\|^2 \leq \frac{1}{T}\big(\Phi(z_0) - \Phi(z_T)\big).$$

Step $\mathfrak{B}$: Next, we bound the KKT-based metric by relating each of its constituent terms to the previously analyzed update direction d_t. To this end, we first choose the dual variables at each step t in the following form:

$$\begin{aligned}
\alpha_{t,i} &= s\big(f_i(\theta_t) + \mu_{t,i} - \delta_i\big), \\
\beta_{t,i} &= r\big(\lambda_i f_i(\theta_t) + \nu_{t,i} - \sigma_t\big),
\end{aligned}$$

where we denote $\alpha_t = [\alpha_{t,1}, \ldots, \alpha_{t,|\mathcal{I}|}]^\top$ and $\beta_t = [\beta_{t,1}, \ldots, \beta_{t,|\mathcal{I}|}]^\top$ as mentioned earlier.

Step $\mathfrak{B}.1$ (Stationarity): By the selection of α_t, β_t, we can get (i) $1 - \sum_{i\in\mathcal{I}}\beta_{t,i} = \nabla_\sigma\Phi(z_t) = d_{t,\sigma}$, and (ii) $\sum_{i\in\mathcal{I}}(\alpha_{t,i} + \lambda_i\beta_{t,i})\nabla f_i(\theta_t) = \nabla_\theta\Phi(z_t) = d_{t,\theta}$.

Step $\mathfrak{B}.2$ (Dual Feasibility): According to the projection operation in CLEAR, we have: $\mu_{t+1,i} = \Pi_{\mathbb{R}_+^p}(\mu_{t,i} - \eta\nabla_{\mu_i}\Phi(z_t))$. We then consider the following two cases: (i) If $\mu_{t,i} - \eta\nabla_{\mu_i}\Phi(z_t) > 0$, then the projection is not performed, and we accordingly have: $\alpha_{t,i} = d_{t,\mu_i}$. (ii) Otherwise, the projection is active, and we have $\alpha_{t,i} \geq \frac{\mu_{t,i}}{\eta} \geq 0$. Similarly, since $\nu_{t+1,i} = \Pi_{\mathbb{R}_+^p}(\nu_{t,i} - \eta\nabla_{\nu_i}\Phi(z_t))$, we have $\beta_{t,i} = d_{t,\nu_i}$ if the projection is inactive and $\beta_{t,i} \geq \frac{\nu_{t,i}}{\eta} \geq 0$ otherwise.

Step $\mathfrak{B}.3$ (Primal Feasibility): According to the dual variables selection, we have:

$$\begin{aligned}
f_i(\theta_t) - \delta_i &= \frac{\alpha_{t,i}}{s} - \mu_{t,i} \leq \frac{\alpha_{t,i}}{s}, \\
\lambda_i f_i(\theta_t) - \sigma_t &= \frac{\beta_{t,i}}{r} - \nu_{t,i} \leq \frac{\beta_{t,i}}{r}.
\end{aligned}$$

We first consider the variable β_t. By considering the update direction w.r.t. σ_t, it's easy to verify that $|\beta_{t,i}| \leq 1 + (|\mathcal{I}| + 1)\|d_t\|$ and $\|\beta_t\| \leq 1 + (2|\mathcal{I}| + 1)\|d_t\|$.
We then denote:

$$\begin{aligned}
I(\theta_t) &:= \{i \in \mathcal{I} : f_i(\theta_t) \geq \delta_i\}, \\
J(\theta_t) &:= \{i \in \mathcal{I} : f_i(\theta_t) < \delta_i\}.
\end{aligned}$$

We prove that $\mu_{t,i} - \eta\alpha_{t,i} \leq 0$ implies $i \in I(\theta_t)$ when $\eta < \frac{1}{s}$. Suppose this does *not* hold, *i.e.*, $i \in J(\theta_t)$. Then, $\mu_{t,i} - \eta\alpha_{t,i} \leq 0$ implies that $\delta_i - f_i(\theta_t) \leq (1 - \frac{1}{\eta s})\mu_{t,i} \leq 0$. Thus, we know that $i \in I(\theta_t)$ in this case, meaning that it corresponds to an active constraint. Consequently, denoting the smallest singular value of the Jacobian as $c_0 > 0$ (which exists due to Assumption 1). According to the update direction w.r.t. θ_t and Assumption 1, we have:

$$\begin{aligned}
&\sum_{i\in I(\theta_t)}\alpha_{t,i}\nabla f_i(\theta_t) \\
=&d_{t,\theta} - \sum_{i\in J(\theta_t)}\alpha_{t,i}\nabla f_i(\theta_t) - \sum_{i\in\mathcal{I}}\lambda_i\beta_{t,i}\nabla f_i(\theta_t),
\end{aligned}$$

which implies:

$$\begin{aligned}
\|\alpha_t\| \leq &\frac{1}{c_0}\big(\|d_t\| + M\sqrt{|\mathcal{I}|}\|d_t\| \\
&+ M(1 + (|\mathcal{I}| + 1)\|d_t\|)\big) + \sqrt{|\mathcal{I}|}\|d_t\| \\
\triangleq &\, c_1 + c_2\|d_t\|,
\end{aligned}$$

where c_1, c_2 denotes positive constants, and we have $c_2 = \Theta(\mathcal{I})$.

Step $\mathfrak{B}.4$ (Complementary Slackness): We then consider the last term in the KKT-based metric. We have:

$$-\delta_i \leq f_i(\theta_t) - \delta_i = \frac{\alpha_{t,i}}{s} - \mu_{t,i} \leq \frac{\alpha_{t,i}}{s}.$$

If $\mu_{t,i} - \eta \nabla_{\mu_i} \Phi(z_t) > 0$, then, as shown earlier, we have $\alpha_{t,i} = d_{t,\mu_i}$, which leads to:

$$|\alpha_{t,i}(f_i(\theta_t) - \delta_i)| \leq \|d_t\| \cdot \big(\delta_i + \frac{\|d_t\|}{s}\big).$$

Otherwise, we have $|\alpha_{t,i}(f_i(\theta_t) - \delta_i)| \leq \frac{\alpha_{t,i}^2}{s}$. Then, we can utilize the same argument in the last step to control this term.

Similarly, we can also get:

$$-\sigma_t \leq \lambda_i f_i(\theta_t) - \sigma_t = \frac{\beta_{t,i}}{r} - \nu_{t,i} \leq \frac{\beta_{t,i}}{r}.$$

If $\nu_{t,i} - \eta \nabla_{\nu_i} \Phi(z_t) > 0$, then $\beta_{t,i} = d_{t,\nu_i}$ leads to:

$$|\beta_{t,i}(\lambda_i f_i(\theta_t) - \sigma_t)| \leq \|d_t\| \cdot \big(\Phi(z_0) + \frac{\|d_t\|}{r}\big).$$

Otherwise, we have $|\beta_{t,i}(\lambda_i f_i(\theta_t) - \sigma_t)| \leq \frac{\beta_{t,i}^2}{r}$.

Step $\mathfrak{C}$: Finally, we aggregate the intermediate results obtained in Step $\mathfrak{B}$ and establish their explicit relationship with the convergence guarantee from Step $\mathfrak{A}$.

For these error terms, we have:

$$\begin{aligned}
\mathcal{E}_t &:= \mathcal{E}_t^{\text{stat}} + \mathcal{E}_t^{\text{primal}} + \mathcal{E}_t^{\text{dual}} + \mathcal{E}_t^{\text{cs}} \\
&= |d_{t,\sigma}|^2 + \|d_{t,\theta}\|^2 \\
&\quad + \mathcal{O}\left(\frac{1}{s^2} + \frac{1}{r^2} + (\frac{|\mathcal{I}|^2}{s^2} + \frac{|\mathcal{I}|^2}{r^2})\|d_t\|^2\right) \\
&\quad + |d_{t,\mu_i}|^2 + |d_{t,\nu_i}|^2 \\
&\quad + \mathcal{O}\left(\|d_t\|^2 + \frac{\|d_t\|^4}{s^2} + \frac{\|d_t\|^4}{r^2}\right).
\end{aligned}$$

Therefore, we can get:

$$\mathcal{E}_t = \mathcal{O}\left(|\mathcal{I}|^2\|d_t\|^2 + \frac{1}{s^2} + \frac{1}{r^2}\right).$$

When $\eta = \mathcal{O}(\frac{1}{\gamma}) = \mathcal{O}(\frac{1}{s+r})$, we have:

$$\eta\Big(1 - \frac{\gamma\eta}{2}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\|d_t\|^2 \leq \frac{1}{T}\big(\Phi(z_0) - \Phi(z_T)\big),$$

which implies:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|d_t\|^2 = \mathcal{O}\left(\frac{s+r}{\eta T}\right).$$

Accordingly, by letting $s = r = \Theta(T^{\xi})$ and $\eta = \Theta(T^{-\zeta})$, we can get:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{E}_t \\
=&\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{O}\left(|\mathcal{I}|^2\|d_t\|^2 + \frac{1}{s^2} + \frac{1}{r^2}\right) \\
=&\mathcal{O}\left(|\mathcal{I}|^2 T^{\xi-\zeta-1} + T^{-2\xi}\right).
\end{aligned}$$

We then select these degrees to determine fast convergence, with a constraint $\zeta \geq \xi$. Therefore, by setting $\xi = \frac{1}{4}$ and $\zeta = \frac{1}{2}$, and by noting that $|\mathcal{I}|$ is a constant in our problem, we have: $\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{E}_t = \mathcal{O}(T^{-\frac{1}{2}})$. □