

CONTRASTIVE NEGATIVE PREFERENCE OPTIMIZATION FOR MACHINE UNLEARNING IN LLMs

Anonymous authors

Paper under double-blind review

ABSTRACT

During large-scale training on extensive corpora, language models inevitably memorize unwanted data (e.g., private or copyrighted content). While numerous unlearning methods have been proposed—including gradient ascent (GA)-based approaches and preference-based optimization—existing methods either fail to effectively erase target data or achieve a reasonable balance between unlearning efficacy and model utility. A grounded optimization framework is lacking. In this work, we present Contrastive Negative Preference Optimization (CNPO), a novel algorithm that leverages inter-sample relationships within datasets to effectively and adaptively remove target data while maintaining model performance on the remaining set. In order to separate the remaining data and target data, we follow the idea of Noisy Contrastive Estimation (NCE) and derived the final loss function within the framework of preference optimization. Through an asymptotic analysis of CNPO, we theoretically establish its connections with GA and NPO. Furthermore, to evaluate the usability of response and privacy protection capability of CNPO, we introduce a personally identifiable information (PII) dataset and develop a suite of metrics for generated text assessment. Overall, theoretical analysis and comprehensive evaluation on three benchmarks demonstrate CNPO’s stable unlearning behavior and optimal balance between forgetting and utility preservation among existing methods.

1 INTRODUCTION

Recently, in response to tightening regulatory (CCPA, 2018; Li et al., 2024) requirements and growing societal concerns (Carlini et al., 2023; Huang et al., 2022; Eldan & Russinovich, 2023) over private data leakage, *machine unlearning* (MU) has emerged as a critical tool in the field of AI privacy. As part of broad set of MU techniques, unlearning for large language models (LLMs) is inherently complex due to their pretraining on broad, multi-domain corpora and their massive parameter scales. Specifically, LLM unlearning can be categorized into exact unlearning and approximate unlearning. In the following, we will focus on approximate unlearning. Effective unlearning for LLMs requires a dual objective: complete eradication of sensitive data while maintaining robust model performance - failure to achieve this balance renders the unlearning process ineffective.

Generally, a simple yet costly approach is to retrain the model from scratch with the requested data being removed from the training dataset (Bourtoule et al., 2020). However, this approach becomes prohibitively expensive for LLMs due to large scale of parameters and dynamic, unpredictable unlearning requests (Xia et al., 2024) from user. Thus, efficient methods have been proposed in recent works (Zhang et al., 2024a; Fan et al., 2025; Huu-Tien et al., 2025; Foster et al., 2023). These methods devise specialized loss functions to actively corrupt the model’s parametric distribution of the designated removal data. For example, *Gradient ascent* (GA) (Liu et al., 2024; Yao et al., 2024), which aims to increase loss in *forget set*, is simple and efficient on forgetting but unsatisfactory on utility preserving (Maini et al., 2024; Zhang et al., 2024a). In order to issue this phenomenon, Zhang et al. (2024a) derives its loss function *Negative Preference Optimization* (NPO), drawing inspiration from Rafailov et al. (2024), which is a preference-based loss function. As the loss function of NPO is derived from *reinforcement learning from human feedback* (RLHF) (Bai et al., 2022), it possesses a rigorous theoretical foundation while exhibiting more gradual gradient variations compared to gradient ascent.

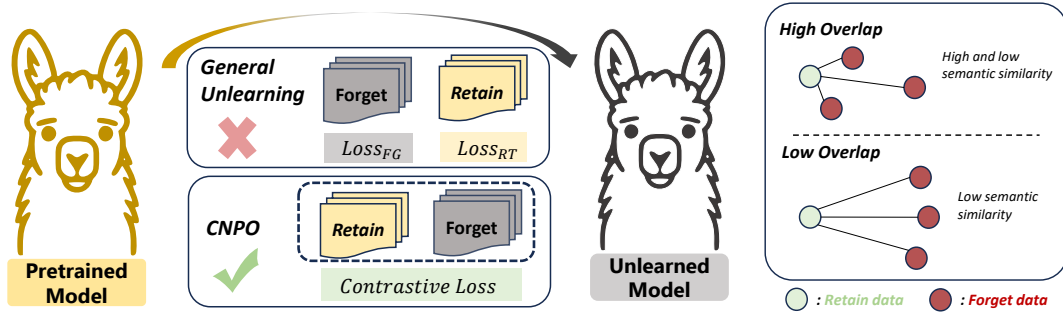


Figure 1: **Left:** CNPO loss function combines samples from different classes, deriving contrastive loss making use embedded information from forget set and retain set. **Right:** Illustration of data overlap based on semantic similarity. In high-overlap settings, retained and forgotten data are semantically similar, while in low-overlap settings, their semantic similarity is low, leading to near-orthogonal gradients during unlearning.

However, these loss functions operate exclusively on the forgetting data while neglecting the information of the remaining dataset, which constitutes the fundamental source of the model’s capabilities. This leads to the destruction of essential generic knowledge(Wang et al., 2025b) such as linguistic structure and semantic coherence. Another frequently overlooked fact is that existing studies typically assume uniform contributions of forget data to unlearning efficacy(Feng et al., 2025a;b). This assumption results in the excessive unlearning of forgettable data while proving inadequate for removing strongly memorized information(Fan et al., 2025). Consequently, it creates a suboptimal trade-off between unlearning and retention that fails to achieve the pareto frontier(Wang et al., 2025a). In this work, we first observe the destruction of generalized knowledge—which is key to utility preservation from our perspective—by unlearning algorithms during forgetting. Towards fixing this, we propose a preference-based objective function using technique of contrastive learning(Khosla et al., 2021; kyu Lee et al., 2024) for unlearning, which is termed *Contrastive Negative Preference Optimization* (CNPO). We draw inspirations from contrastive preference alignment(Chen et al., 2024b), deriving our unlearning algorithm from scratch and providing theoretical analysis on it. Through experiment results and elaboration of theory, we show CNPO can fill in the gaps in improving model utility. For better evaluation of CNPO, we make a *personal identifiable information* (PII) dataset to simulate the possible unlearning scenarios in reality. In addition, this dataset provides an access to explore retention of generalized knowledge (e.g., sentence grammar structure). Combining these works, we provide new insights to LLM unlearning for the communications.

1.1 RELATED WORKS

Current works on LLM unlearning have demonstrated significant success in achieving effective knowledge removal(Wang et al., 2025c; Mekala et al., 2024; Pawelczyk et al., 2024). However, a fundamental challenge persists in achieving effective unlearning while maximally preserving model utility. This difficulty stems primarily from the unbounded loss function(Ji et al., 2024) for unlearning and the strong entanglement(See figure.1) between forget and retain sets(Zhao et al., 2024), which frequently leads to significant degradation of model performance during unlearning. We categorize existing unlearning paradigms as follows.

Unlearning only on forget data. Conducting unlearning only on forget set is simple and efficient. Current algorithms, such as GA and NPO, have demonstrated effectiveness and serve as foundational baseline to the design of unlearning algorithms. Recently, Wang et al. (2024) argue that incorporating retain data in unlearning processes does not necessarily preserve model utility effectively. Instead, they suggest that the boundary between retain and forget sets may become blurred during updates, making it challenging to achieve balance between unlearning performance and the general utility of the model. However, unlearning solely on forget data will inadvertently compromise model utility, which is proven by the mainstream benchmarks(Shi et al., 2024; Li et al., 2024; Maini et al., 2024).

Linear combination of unlearning and preserving. Building upon existing unlearning algorithms, a loss term for the retain set is incorporated to preserve the model’s overall performance and maintain its general knowledge on unaffected data. The direct retain loss, denoted as $\mathcal{L}_{RT}$, is the standard cross-

entropy loss for next-word prediction and is widely employed across various scenarios. Additionally, KL divergence will also be adapted when a reference model is available (Shi et al., 2024; Maini et al., 2024; Li et al., 2024). In contrast to conducting unlearning from the perspective of outputs, some studies focus on modifying representations in the latent space. Huu-Tien et al. (2025) explore the impact of steering latent representations of the forget data while keeping the representations of retain data unchanged. Zhou et al. (2023) introduce security vectors, which are activated during fine-tuning to capture harmful behaviors and deactivated during inference to restore normal behavior. However, current research generally assumes a uniform unlearning difficulty, which is impractical for achieving effective unlearning in linear combination paradigm.

Distribution approximation. In order to avoid the unbounded unlearning objective, Dong et al. (2024) proposes a heuristic design of target distributions for the forget set. These distributions serve as an offset to the original model’s parametric distribution, lowering the probability of ground-truth next token. Nevertheless, this approach remains fundamentally limited by the impossibility of directly observing the ground-truth distribution in the absence of the target forget model. In contrast, Ji et al. (2024) employs a uniform distribution approximation for the retain set, effectively constraining the LLMs to exclusively retain information from the forget set. This specialized LLM’s logits are then subtracted by those of the original model, thereby fundamentally inverting the conventional objectives of unlearning. While distribution approximation helps avoid degenerate outputs and catastrophic forgetting, the use of an assistant LLM and augmentation for forget data often compromises efficiency.

Overall, these solutions either focus exclusively on the forget set, or directly combine losses from both datasets while neglecting the inherent structural relationships between retain set samples and forget samples, thereby disregarding critical information embedded in the dataset structure. Our method, CNPO, is inspired by *noise contrastive alignment* (NCA) (Chen et al., 2024b) and *preference optimization* (PO). By adaptively assigning a forgetting strength to each forgotten sample, CNPO alleviates the limitations of the uniform forgetting strength assumption (Zhao et al., 2024; Chen et al., 2024a). The CNPO can be derived by constructing pairs of positive and negative samples under the setting of *noisy contrastive estimation* (NCE), and bridging the relationship with the reward through the optimal solution of the policy. Beyond algorithmic innovations for unlearning, we contribute a PII dataset designed to assess output quality in unlearned models.

2 PRELIMINARIES ON LLM UNLEARNING

Problem formulation. Generally, unlearning task refers to following problem: Given an initial(reference) model π_θ pretrained on dataset $\mathcal{D} = \{(x_i, y_i)\}_{i \in [n]}$, how to guide model to *forget* a specific subset $\mathcal{D}_{\text{FG}} \subseteq \mathcal{D}$ of the training data and preserve model utility on $\mathcal{D} \setminus \mathcal{D}_{\text{FG}}$ to greatest extent? To address this dual problem, LLM unlearning is generally cast as a regularized optimization problems that tries to balance two objectives (Zhang et al., 2024b; Yao et al., 2024):

$$\underset{\theta}{\text{minimize}} \quad \mathbb{E}_{(x,y) \in \mathcal{D}_{\text{FG}}} [\ell_{\text{f}}(y|x; \theta)] + \lambda \mathbb{E}_{(x,y) \in \mathcal{D}_{\text{RT}}} [\ell_{\text{r}}(y|x; \theta)] \quad (1)$$

This objective function incorporates a regularization term λ that penalizes deviations on retain set and trainable parameters θ during unlearning. ℓ_{f} and ℓ_{r} are separately forget loss and retain loss incurred by unlearning model.

Extensive research has been dedicated to the design and rigorous analysis of suitable forget and retain loss functions to address problem 1 (Fan et al., 2025; Zhang et al., 2024a; Wang et al., 2024; 2025b). For instance, given input-response pair (x_f, y_f) sampled from $\mathcal{D}_{\text{FG}}$, model generates a prediction probability $\pi_\theta(y_f|x_f)$ for y_f conditioned on x_f . A straightforward approach to inducing forgetting for the target sample y_f involves minimizing its associated probability, which can be formulated as: $\underset{\theta}{\text{minimize}} \ell_{\text{f}}(y_f|x_f, \theta) = \pi_\theta(y_f|x_f)$. In contrast, the retain loss can be specified as: $\underset{\theta}{\text{maximize}} \ell_{\text{r}}(y_r|x_r, \theta) = \pi_\theta(y_r|x_r)$, which encourages model to perform well on retain data $(x_r, y_r) \in \mathcal{D}_{\text{RT}}$. Notably, exclusive minimization of such a forget loss constitutes the GA method. Similarly, combining the regularized loss with forget loss is generally referred to as the *gradient difference* (GradDiff) method (Liu et al., 2022; Yao et al., 2024).

Negative preference optimization(NPO). This preference-based unlearning method utilizes formula from DPO (Rafailov et al., 2024), casting the preference alignment optimization into an unlearning

process. The loss function of NPO is:

$$\mathcal{L}_{\text{NPO},\beta}(\theta) = -\frac{2}{\beta} E_{\mathcal{D}_{\text{FG}}} \left[\log \sigma \left(-\beta \log \frac{\pi_{\theta}(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right] = \frac{2}{\beta} E_{\mathcal{D}_{\text{FG}}} \left[\log \left(1 + \left(\frac{\pi_{\theta}(y|x)}{\pi_{\text{ref}}(y|x)} \right)^{\beta} \right) \right] \quad (2)$$

Here, $\sigma(t) = 1/(1 + e^t)$ is the sigmoid function, $\beta > 0$ is inverse temperature and π_{ref} is the reference model. With the involvement of π_{ref} , the forget loss proposed by NPO effectively addresses a critical limitations of the GA approach: it mitigates the unbounded optimization problem inherent in GA, preventing *catastrophic collapse*. Furthermore, by employing KL-divergence to penalize distributional deviation in the retain set or excessive divergence in the forget set, NPO combined with KL regularization establishes an empirical trade-off between model utility and unlearning effectiveness.

Simple NPO. Fan et al. (2025) observed that samples of varying lengths exhibit different degrees of forgetting difficulty. To address this, they extended the NPO loss functions by dropping the reference model π_{ref} for its bias and incorporating inverse sample token length weighting, thereby preventing excessive unlearning. The loss function of SimNPO is formulated as:

$$\ell_{\text{SimNPO}}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{FG}}} \left[-\frac{2}{\beta} \log \sigma \left(-\frac{\beta}{|y|} \log \pi_{\theta}(y|x) - \gamma \right) \right] \quad (3)$$

where β is inverse temperature and γ is parameter that control unlearning margin. When applied to datasets with relatively uniform sample lengths (e.g., MUSE), SimNPO inevitably reduces to the NPO loss without π_{ref} constraints.

Balance between unlearning and preserving. If only performing unlearning on the *forget* set, then the model falls into a severe utility degradation (Zhang et al., 2024a; Shi et al., 2024; Wang et al., 2025b). Therefore, there have been many studies combining the forgetting loss function linearly with the loss function of the incentive model preserve utility (Jang et al., 2022; Yao et al., 2024; Chen & Yang, 2023; Eldan & Russinovich, 2023) to solve this problem. Commonly used loss functions are:

- Retain loss: $\mathcal{L}_{\text{RT}} = -\mathbb{E}_{\mathcal{D}_{\text{RT}}} [\log(\pi_{\theta}(y|x))]$, which encourages the model to still perform well on the retain set $\mathcal{D}_{\text{RT}}$.
- KL divergence: $\mathbb{E}_{\mathcal{D}_{\star}} [\mathcal{D}(\pi_{\theta}(\cdot|x)) | \pi_{\text{ref}}(\cdot|x)]$, where $\star \in \{\mathcal{D}_{\text{FG}}, \mathcal{D}_{\text{RT}}\}$. This loss measures distance to initial model π_{ref} on dataset $\star$.

Incorporating these loss usually improves the performance of unlearning according to the experiment result of Maini et al. (2024), Shi et al. (2024).

3 CONTRASTIVE NPO

3.1 DESTRUCTION OF GENERIC KNOWLEDGE

Our study reveals that existing unlearning methods not only erase target data but also unintended corruption of logical sentence structures in model outputs, sometimes resulting in repetition (Fan et al., 2025) and even null responses. Especially, continuous unlearning induces a concerning behavioral shift in language models, manifesting as a propensity toward simplistic, repetitive output patterns that reflects significantly degradation on model capability. We show some examples in Appendix B.2 for a detailed view.

3.2 DERIVATION OF CNPO

We introduce CNPO, an improved version of NPO making use of information from retain sample. Utilizing this idea, we can relieve the *destruction of generic knowledge* problem in existing unlearning algorithms. Notably, the contrastive setting originates from NCE (Ma & Collins, 2018), which aims to maximize the probability of correctly distinguishing samples from the retain set and the forget set. Guided by this principle, we aim to utilize the NCE framework to discriminate between samples in the retain set and those in the forget set, which facilitates the probabilistic unlearning of the forget set. The CNPO loss can reduces to the GA loss in high-temperature limit. In addition, it approaches to NPO loss as the number of forgotten samples per iteration becomes large. Additionally, our proposed loss function is stable and converges to zero, as demonstrated by the gradient stability analysis in 3.2.

Preference optimization. In preference-based learning (Rafailov et al., 2024; Zhang et al., 2024a; Ouyang et al., 2022), the standard setup assumes access to a dataset of pairwise comparisons $\mathcal{D}_{paired} = \{(x_i, y_{i,w}, y_{i,l})\}_{i \in n}$, where $(y_{i,w}, y_{i,l})$ are generated by policy π_θ for prompts x_i , and $y_{i,w} \succ y_{i,l}$ denotes a preference (typically annotated by humans or advanced models like GPT-4 (Chen et al., 2024b)). However, explicit preference data is often unavailable in practice. To address this, we propose proxy metric—semantic similarity (Farquhar et al., 2024)—to quantify the divergence between samples designated for removal and those to be retained. This metric guide the optimization process by selectively distancing forget samples that are semantically similar to retain samples, while remaining the probability of dissimilar forget samples. This idea is founded on the assumption that semantic similarity between contexts complicates their separation, whereas dissimilarity facilitates it. Specifically, we use LLM2Vec (Behnam Ghader et al., 2024) to generate sentence embeddings and compute their cosine similarity as the similarity feedback. This strategy maintains the model’s core performance characteristics while achieving the desired unlearning objectives with different difficulty (Feng et al., 2025b).

Unlearning in contrastive preference optimization. Drawing inspiration from NCE, we construct positive and negative sample pairs as follows: we select $y_r \in \mathcal{D}_{RT}$ as the positive sample and $\{y_i\}_{i=1}^k \subset \mathcal{D}_{FG}$ as negative samples, where the latter serve as "noise" in the NCE framework. Our objective is to maximize the likelihood, thereby improving the model’s ability to discriminate between these two distinct classes.

Theorem 3.1 (CNPO, proof in Appendix D.1). *We define $\pi^*(y|x) \propto \mu(y|x)e^{r(x,y)/\alpha}$ and $\pi_\theta(y|x) \propto \mu(y|x)e^{r_\theta(x,y)}$. $\forall k > 0, \beta > 0$, we have:*

$$\begin{aligned} \max_{\theta} E_{p(x,y)} \log(P_{\theta}(v|x, y)) \Leftrightarrow \min_{\theta} & -\frac{2}{\beta} E_{\mathcal{D}_{RT}} E_{\mathcal{D}_{FG}} \left[\frac{k}{k+1} \log \left(\sigma \left(\log k - \frac{r_{\theta}(x_i, y_i)}{\beta} \right) \right) \right. \\ & \left. + \frac{1}{k+1} \frac{e^{r(y_r, y_i)/\alpha}}{Z(x)} \log \left(\left(\frac{r_{\theta}(x_r, y_r)}{\beta} - \log k \right) \right) \right] \end{aligned} \quad (4)$$

where $Z(x) = \mathbb{E}_{\mu(y|x)} e^{r(x,y)/\alpha}$ and k is the number of forgotten samples per iteration during unlearning.

Theorem 3.1 establishes the equivalence between the derived objective and the original NCE-based MLE objective, providing a practical insight. When the k is large, CNPO objective reduces to NPO loss:

$$\mathcal{L}_{\text{NPO}, \beta}(\theta) = \min_{\theta} -\frac{2}{\beta} \log \left(\sigma \left(\log k - \beta \log \frac{\pi_{\theta}(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right) \quad (5)$$

For the implementation of Eq.4, we approximate $Z(x) \approx \sum e^{r_j/\alpha}$ and parameterize $r_{\theta}(x, y) := \beta \log \frac{\pi_{\theta}(y|x)}{\mu(y|x)}$. $\forall x \in \mathcal{D}_{FG}, y \in \mathcal{D}_{RT}$, we substitute reward metric $r(x, y)$ by $d(x, y)$, yielding the Contrastive Negative Preference Optimization (CNPO) loss:

$$\begin{aligned} \mathcal{L}_{\text{CNPO}, \beta}(\theta) = & -\frac{2}{\beta} * \frac{1}{n_r} \sum_{y_r \in \mathcal{D}_{RT}} \sum_{i=1}^k \left[\frac{1}{(k+1)} \underbrace{\log \sigma \left(-\beta \log \frac{\pi_{\theta}(y_i|x_i)}{k \pi_{\text{ref}}(y_i|x_i)} \right)}_{\text{optimizer } \downarrow} \right. \\ & \left. + \frac{1}{(k+1)k} \underbrace{\frac{e^{d(y_r, y_i)/\alpha}}{\sum_j e^{d(y_r, y_j)/\alpha}}}_{\text{softmax weight}} \underbrace{\log \sigma \left(-\beta \log \frac{k \pi_{\text{ref}}(y_r|x_r)}{\pi_{\theta}(y_r|x_r)} \right)}_{\text{optimizer } \uparrow} \right] \end{aligned} \quad (6)$$

We have reformulated the classification task as an unlearning problem, where minimizing the loss function in Eq.6 simultaneously: (1) decreases the model’s output probability on $\mathcal{D}_{FG}$ with fixed weighting, and (2) retain the likelihood on $\mathcal{D}_{RT}$ with instance-dependent weighting $\frac{e^{d(y_f, y_i)/\alpha}}{\sum_j e^{d(y_f, y_j)/\alpha}}$, which automatically identifies samples with varying forgetting difficulties according to their similarity.

Degradation to GA loss. We take inspiration from NPO, analyzing the connection between CNPO loss and GA loss due to common unlearning behavior between CNPO and NPO. In the limit of $\beta \rightarrow 0$, CNPO degrades to GA loss, demonstrating its fundamental connection to NPO.

Proposition 1 (CNPO reduces to GA as $\beta \rightarrow 0$, proof in Appendix D.2). $\forall \theta \in \Omega$, Ω is a distribution. Under mild assumption that forget sample $y_{fi} \in B(y_r, d)$, where $B(y_r, d)$ is the unit sphere centered at y_r , y_{fi} is targeted unlearning sample, we have:

$$\lim_{\beta \rightarrow 0} \left[\mathcal{L}_{\text{CNPO}, \beta}(\theta) - \left(\frac{1}{k} + k \right) \frac{4}{\beta} \right] = \frac{1}{k+1} \left[\left(\frac{k}{n_r} \mathcal{L}_{\text{GA}_F}(\theta) - \frac{1}{k} \mathcal{L}_{\text{GA}_R}(\theta) \right) + \mathbf{B}(x, y, k) \right] \quad (7)$$

where:

$$\mathbf{B}(x, y, k) = \frac{1}{k+1} \left[\frac{1}{k} \mathbb{E}_{\mathcal{D}_{RT}} \log \pi_{\text{ref}}(y_r | x_r) - \frac{k}{n_r} \mathbb{E}_{\mathcal{D}_{FG}} \log \pi_{\text{ref}}(y_i | x_i) + \left(\frac{1}{k} - \frac{k}{n_r} \right) \log k \right] \quad (8)$$

Moreover, assuming $\pi_{\theta}(y|x)$ is differentiable with respect to θ , we have:

$$\lim_{\beta \rightarrow 0} \nabla \mathcal{L}_{\text{CNPO}, \beta}(\theta) = \frac{1}{k+1} \left(k \nabla \mathcal{L}_{\text{GA}_F}(\theta) - \frac{1}{k} \nabla \mathcal{L}_{\text{GA}_R}(\theta) \right) \quad (9)$$

where $\mathcal{L}_{\text{GA}_F}$ represents GA loss on $\mathcal{D}_{FG}$ and $\mathcal{L}_{\text{GA}_R}$ is GA loss on $\mathcal{D}_{RT}$.

Proposition 1 illustrates CNPO will conduct gradient ascent on $\mathcal{D}_{FG}$ and gradient descent on $\mathcal{D}_{RT}$. Notably, the term $\mathbb{E}_{\mathcal{D}_{FG}} \log \pi_{\text{ref}}(y_i | x_i)$ and $\mathbb{E}_{\mathcal{D}_{RT}} \log \pi_{\text{ref}}(y_r | x_r)$ in Eq.8 does not depend on θ , illustrating the degradation from CNPO to a weighted GA loss on two datasets $\mathcal{D}_{FG}$ and $\mathcal{D}_{RT}$. Furthermore, the gradient of CNPO asymptotically converge to a weighted variant of the GA gradient.

Stability analysis of CNPO. Theoretical derivation of the gradient(in Appendix D.3) and experimental findings jointly indicate that CNPO exhibits a more stable loss curve than NPO. This phenomenon occurs because, during each round of unlearning, the model simultaneously pay attention to both the retain sample and forget samples, causing the gradients to be constrained by two directional forces. The gradients of NPO and GA are:

$$\nabla_{\theta} \mathcal{L}_{\text{GA}} = \mathbb{E}_{\mathcal{D}_{FG}} [\nabla_{\theta} \log \pi_{\theta}(y|x)], \quad (10)$$

$$\nabla_{\theta} \mathcal{L}_{\text{NPO}, \beta} = \mathbb{E}_{\mathcal{D}_{FG}} [\mathbf{W}_{\theta}(x, y) \nabla_{\theta} \log \pi_{\theta}(y|x)] \quad (11)$$

where: $\mathbf{W}_{\theta}(x, y) = 2\pi_{\theta}^{\beta}(y|x) / [\pi_{\theta}^{\beta}(y|x) + \pi_{\text{ref}}^{\beta}(y|x)]$. The gradients of CNPO is:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{CNPO}, \beta} = & \frac{1}{k+1} \mathbb{E}_{\mathcal{D}_{FG}} [k * \mathbf{W}_{\theta}(x_i, y_i) \nabla \log(\pi_{\theta}(y_i | x_i)) - \\ & \frac{e^{d(y_r, y_i)/\alpha}}{\sum_j e^{d(y_r, y_j)/\alpha}} \mathbf{W}_{\theta}(x_r, y_r) \nabla \log(\pi_{\theta}(y_r | x_r))] \end{aligned} \quad (12)$$

where:

$$\mathbf{W}_{\theta}(x_i, y_i) = \pi_{\theta}^{\beta}(y_i | x_i) / (\pi_{\theta}^{\beta}(y_i | x_i) + (k\pi_{\text{ref}}(y_i | x_i))^{\beta}) \quad (13)$$

$$\mathbf{W}_{\theta}(x_r, y_r) = (k\pi_{\text{ref}}(y_r | x_r))^{\beta} / ((k\pi_{\text{ref}}(y_r | x_r))^{\beta} + \pi_{\theta}^{\beta}(y_r | x_r)) \quad (14)$$

In this process, the model strives to preserve the gradient direction of the retain data to prevent excessive performance degradation while still achieving effective forgetting.

4 ACCURATE UNLEARNING OF PII

Personally-identifiable-information (PII) is often collected by tech companies and casted into large language models for training. Due to the phenomenon of memorization(Lu et al., 2024), it is hard to erase the influence of forgotten data. In order to evaluate the effectiveness and efficiency of unlearning effects on PII data, we produced PII dataset from pii-masking-200k(Ai4Privacy , 2023).

This PII dataset consists of synthetically generated sentences with semantically similar contents and close grammatical structures, but varying personally identifiable information. A sample is provided in Appendix B. We created it by prompting GPT-4o-mini to fill predefined privacy templates(target text above) with placeholders for names, genders, phone numbers, IP addresses, and professions etc. Subsequently, we paraphrase these sentences into comprehensive paragraphs to ensure the training dataset maintains sufficient quality for the model to effectively distinguish between different texts

and their PII. Due to the fictitious nature of these paraphrase contents and PII, these data will not be included in the pretraining of the general base model.

The unlearning task focuses on the model’s ability to forget specific PII from this synthetic dataset while retaining the underlying grammatical structure of the sentences. We treat all the source text as the *retain set* and the generation by GPT-4o-mini as *forget set*. In addition, we can control whether the template text with privacy placeholders is incorporated. To systematically investigate how model utility and unlearning efficacy vary across different scales of forgetting, we partition the dataset into multiple subsets. And this dataset can be accessed through supplementary material.

4.1 MAKING PROCESS

To ensure the effectiveness of the generations and the proper formatting, we carefully design the prompts for PII replacements and prompts for paraphrasing. Besides, we adjust the model’s parameters during generation. The temperature parameter is set to 1 to enhance the diversity of generated text and mitigate potential repetition. Additionally, the maximum token limit is configured to 1,000, guaranteeing that the generation process remains feasible for each input.

For each subsequent round of generation(excluding the initial round), the model incorporates the output from the previous round as part of its input. This approach maximizes the generation of diverse privacy-related information. However, to prevent exceeding the maximum input token limit, the inclusion of prior outputs is carefully constrained to a controlled number.

4.2 BENCHMARK FOR UNLEARNING

Forgetting private identifiable information. The PII datasets exhibit significant entanglement(Zhao et al., 2024), complicating targeted unlearning of specific PII. This challenge is exacerbated by deliberate semantic similarity between retained and forgotten samples—a design mimicking real-world conditions—which forces models to preserve generic sentence structures while selectively removing sensitive information.

Difficult unlearning on entangled dataset. Existing unlearning methods achieves wonderful performance on unlearning, but struggle to output high-quality content after unlearning. This phenomenon can be attributed to two primary factors. First, the unlearning process lacks upper bounds in general settings due to its reliance on gradient ascent, which disrupts the pretrained model’s original parameter distribution. Second, the inherent entanglement between data samples causes unintended forgetting: when the model unlearns target samples, it simultaneously degrades performance on retained samples with similar semantic or syntactic distributions. To study unlearning behavior under challenging conditions, we design a PII dataset where retained and forgotten samples exhibit syntactic entanglement—a setting that induces the most severe performance drop during unlearning(Chang & Lee, 2025), thereby revealing existing limitations of unlearning methods.

Quantitative Evaluation of output. Precisely, to quantitatively assess behavior of unlearned model, a comprehensive evaluation are conducted from perspective of PII repetition, fluency and consistency of the generated content. For a synthetic evaluation of quality of generated context, we employ GPT-4o(OpenAI et al., 2024) as a judge that scores the context according to predefined criteria:

1. **PII repetition:** We evaluate PII leakage by quantifying the frequency of repeated occurrences in model outputs.
2. **Fluency:** We employ GPT-4o to quantitatively assess textual fluency through three key metrics: grammatical correctness, readability, and repetition frequency, with weighted scoring based on their relative portion.
3. **Coherence:** We utilize GPT-4o to measures the semantic consistency between prompts and generated responses, and the internal continuity and reasoning flow of the generated content, which is logical coherence.

These metrics enable quantitative assessment of unlearned model behavior, providing measurable insights into its performance. Each metric takes values within the range of 0 to 10, and the final overall score is calculated using a weighted average.

5 EXPERIMENTS

We evaluate our proposed unlearning method comparing to existing state-of-art unlearning methods across two mainstream benchmarks. In addition, we further compare different unlearning methods in the proposed PII dataset in section 5.2.

5.1 EXPERIMENTAL SETUP

Datasets and unlearning methods. Based on general pretrained base model, we evaluate unlearning methods on three datasets: TOFU(Maini et al., 2024), MUSE(Shi et al., 2024) and PII. TOFU includes "Forget05" and "Forget10" scenarios, which represent 5% and 10% scales of forget sets. MUSE dataset incorporates two distinct corpora scenarios, NEWS and BOOKS, designed to simulate real-world forgetting requests. Additionally, motivated by the phenomenon of generic knowledge destruction, we conduct unlearning experiments on the synthetic PII dataset, assessing model responses using the quantitative metrics outlined in Section 4.2. Specifically, we set up four unlearning scenarios for the PII dataset to represent different scales of forgetting.

Table 1: Performance of selected unlearning methods on MUSE, presenting unlearning scenarios:NEWS. Unlearning results on BOOKS are presented in Appendix C.

Method	Unlearning Efficacy			Model Utility
	VerbMem $\mathcal{D}_f$ ($\downarrow$)	KnowMem $\mathcal{D}_f$ ($\downarrow$)	PrivLeak ($\rightarrow 0$)	KnowMem $\mathcal{D}_r$ ($\uparrow$)
NEWS				
Original f_{ref}	73.12	56.90	-99.81	76.26
Retrain f_{retrain}	23.65	29.66	-0.04	81.28
Task Vector	0.231	0.00	7.35	0.00
GA	0.00	0.00	5.2	0.00
GA _{GDR}	4.85	31.29	108.12	28.21
GA _{KLR}	0.23	39.67	104.92	23.70
NPO	0.00	0.00	9.12	0.00
NPO _{GDR}	1.2	54.6	105.8	40.5
NPO _{KLR}	26.9	49.0	95.8	45.4
SimNPO	2.34	44.84	72.93	39.65
SimNPO _{GDR}	66.47	51.00	-99.79	70.81
SimNPO _{KLR}	0.84	54.33	72.10	75.77
CNPO	0.00	0.00	-2.22	0.00
CNPO _{GDR}	2.67	47.74	77.17	79.70
CNPO _{KLR}	0.00	41.83	78.33	62.32

LLM unlearning Methods. Our primary evaluation includes Retrain, GA, and PO-type unlearning methods (CNPO, NPO, and SimNPO). Additional approaches are integrated into specific benchmarks: Task Vector (MUSE) and the rejection-based method IDK (TOFU). Detailed experimental configurations for each unlearning method are provided in Appendix C.3.

5.2 RESULTS

Performance on MUSE. Table 1 presents a comparative analysis of CNPO against alternative unlearning methods on MUSE dataset, which exhibits inherent overlap between the retain and forget sets. Thus, the decrease in knowledge memorization on $\mathcal{D}_f$ is accompanied by a decrease in knowledge memorization on $\mathcal{D}_r$. Among these unlearning methods, CNPO can best prevent privacy leakage under strong forgetting requirements. Gradient descent on the retain set (GDR) and regularization of the Kullback-Leibler (KLR) divergence between the predictions on the retain set of the original and newly trained models are also incorporated to help preserve model utility. Overall, CNPO outperforms other unlearning methods on the MUSE datasets, optimally balancing unlearning and model utility.

Unlearning efficacy on TOFU. The results on task TOFU-5% are listed in table 2. CNPO achieves superior *forget quality* among all unlearning methods on task TOFU-5%, which is measured by the p-value derived from Kolmogorov-Smirnov (KS) test, which assesses whether the unlearned model’s

behavior aligns with that of the retrained model. In the meanwhile, CNPO keeps excellent balance between *forget quality* and *model utility*.

Table 2: Performance on TOFU-5% dataset. The detailed metrics is summarized in Table 6. The best results among baseline are marked in **bold**. More experiment results on TOFU-10% are in Table 7 of **Appendix C**.

Method	Forget Efficacy				Model utility									
	Forget Set				Real Authors			Real Worlds			Retain Set			M.U. $\uparrow$
	R-L $\downarrow$	Prob. $\downarrow$	Truth Ratio $\uparrow$	F.Q. $\uparrow$	R-L $\uparrow$	Prob. $\uparrow$	Truth Ratio $\uparrow$	R-L $\uparrow$	Prob. $\uparrow$	Truth Ratio $\uparrow$	R-L $\uparrow$	Prob. $\uparrow$	Truth Ratio $\uparrow$	
Original	0.04	0.01	0.49	0.00	0.93	0.44	0.58	0.91	0.43	0.55	0.98	0.99	0.48	0.62
Retrain	0.61	0.85	0.66	1.00	0.92	0.44	0.57	0.90	0.43	0.54	0.97	0.99	0.48	0.62
IDK	0.022	0.833	0.523	0.00	0.800	0.386	0.502	0.829	0.383	0.497	0.702	0.958	0.465	0.55
GA	0.00	0.00	0.66	0.00	0.00	0.20	0.40	0.00	0.30	0.28	0.00	0.00	0.15	0.00
GA _{GDR}	7.33e-03	0.00	0.684	0.00	0.536	0.503	0.672	0.858	0.421	0.558	0.411	0.482	0.508	0.52
GA _{KLR}	0.005	0.00	0.544	1.1e-4	0.096	0.348	0.564	0.287	0.342	0.571	0.075	0.004	0.472	0.032
NPO	0.20	2e-3	0.66	0.79	0.30	0.46	0.62	0.77	0.47	0.67	0.22	4e-3	0.32	0.03
NPO _{GDR}	0.233	0.006	0.601	0.545	0.776	0.438	0.603	0.887	0.434	0.615	0.405	0.240	0.489	0.474
NPO _{KLR}	0.206	0.005	0.643	0.545	0.623	0.391	0.541	0.811	0.405	0.599	0.295	0.053	0.365	0.249
SimNPO	0.253	0.027	0.658	0.923	0.876	0.488	0.632	0.882	0.493	0.619	0.533	0.563	0.447	0.583
CNPO	0.266	0.018	0.65	0.924	0.837	0.427	0.59	0.895	0.423	0.581	0.543	0.417	0.482	0.538
CNPO _{GDR}	0.381	0.029	0.659	0.965	0.90	0.43	0.575	0.919	0.398	0.549	0.661	0.482	0.491	0.556

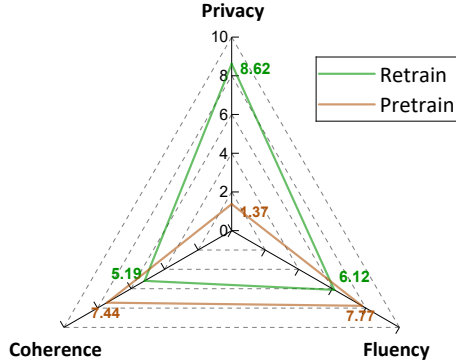


Figure 2: Performance of retrain and pretrain model

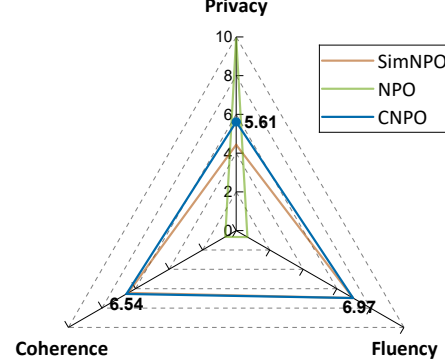


Figure 3: Performance on PO-type methods

Unlearning benchmark:PII We evaluate preference-based unlearning methods on the PII benchmark, excluding Gradient Ascent and its variants due to its catastrophic degradation of model utility. As demonstrated in figure 2 and figure 3, CNPO achieves the optimal balance between privacy protection and response quality. While NPO attains the highest privacy scores, this advantage comes at the expense of significantly degraded output quality. Furthermore, our experiments demonstrate that CNPO maintains stable performance degradation as unlearning epochs increase. We conduct comprehensive ablation studies examining the impact of negative values k and other hyper-parameters, with detailed results presented in Appendix C. Model response examples are provided in Appendix B to illustrate practical outcomes.

6 CONCLUSION

In this work, we propose Contrastive Preference-based unlearning, a meticulously designed unlearning objective that accounts for the semantic relationship between positive and negative samples. CNPO outperforms NPO-based methods in model utility preservation and shows better quality of generated-text in benchmark PII. Besides, we curated benchmark PII as a challenging scenario for unlearning, providing a multidimensional evaluation scheme to quantify unlearning results, thus fostering the developing of unlearning methods. Future work will explore its limitations and expand its applicability in real-world applications(See Appendix A.2).

ETHICS STATEMENT

In this work, we introduce a personally identifiable information (PII) dataset, entirely generated by GPT-4o based on pii-masking-200k. As a fully synthetic dataset, it does not contain any real personal data and, therefore, poses no risks related to privacy or security. The primary objective of creating this dataset is to aid in addressing and mitigating potential threats associated with the leakage of personal information during unlearning.

REPRODUCIBILITY STATEMENT

Regarding the experimental components, our source code for reproducing the experimental results and the source code for reproducing the PII dataset can be found at this link, and the corresponding prompts are listed in Appendix B. For the theoretical results, complete and rigorous proofs are provided in Appendix D.

REFERENCES

- Ai4Privacy . pii-masking-200k (Revision 1d4c0a1) , 2023. URL <https://huggingface.co/datasets/ai4privacy/pii-masking-200k>.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL <https://arxiv.org/abs/2204.05862>.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders, 2024. URL <https://arxiv.org/abs/2404.05961>.
- Lucas Bourtole, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning, 2020. URL <https://arxiv.org/abs/1912.03817>.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models, 2023. URL <https://arxiv.org/abs/2202.07646>.
- CCPA. California consumer privacy act of 2018. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375, 2018. AB-375, Signed into law on June 28, 2018.
- Hwan Chang and Hwanhee Lee. Which retain set matters for llm unlearning? a case study on entity unlearning, 2025. URL <https://arxiv.org/abs/2502.11441>.
- Chaochao Chen, Jiaming Zhang, Yizhao Zhang, Li Zhang, Lingjuan Lyu, Yuyuan Li, Biao Gong, and Chenggang Yan. Cure4rec: A benchmark for recommendation unlearning with deeper influence, 2024a. URL <https://arxiv.org/abs/2408.14393>.
- Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. Noise contrastive alignment of language models with explicit rewards, 2024b. URL <https://arxiv.org/abs/2402.05369>.
- Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for llms, 2023. URL <https://arxiv.org/abs/2310.20150>.
- Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. Undial: Self-distillation with adjusted logits for robust unlearning in large language models, 2024. URL <https://arxiv.org/abs/2402.10052>.

- Ronen Eldan and Mark Russinovich. Who’s harry potter? approximate unlearning in llms, 2023. URL <https://arxiv.org/abs/2310.02238>.
- Tremblay v. OpenAI, Inc.*, 23-cv-03416-AMO, (N.D. Cal.), 2023.
- European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL <https://data.europa.eu/eli/reg/2016/679/oj>.
- Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning, 2025. URL <https://arxiv.org/abs/2410.07163>.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL <https://doi.org/10.1038/s41586-024-07421-0>.
- Xiaohua Feng, Yuyuan Li, Huwei Ji, Jiaming Zhang, Li Zhang, Tianyu Du, and Chaochao Chen. Bridging the gap between preference alignment and machine unlearning, 2025a. URL <https://arxiv.org/abs/2504.06659>.
- Xiaohua Feng, Yuyuan Li, Chengye Wang, Junlin Liu, Li Zhang, and Chaochao Chen. A neuro-inspired interpretation of unlearning in large language models through sample-level unlearning difficulty, 2025b. URL <https://arxiv.org/abs/2504.06658>.
- Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening, 2023. URL <https://arxiv.org/abs/2308.07707>.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information?, 2022. URL <https://arxiv.org/abs/2205.12628>.
- Dang Huu-Tien, Trung-Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. On effects of steering latent representation for large language model unlearning, 2025. URL <https://arxiv.org/abs/2408.06223>.
- Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models, 2022. URL <https://arxiv.org/abs/2210.01504>.
- Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference, 2024. URL <https://arxiv.org/abs/2406.08607>.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. URL <https://arxiv.org/abs/2004.11362>.
- Hong kyu Lee, Qiuchen Zhang, Carl Yang, Jian Lou, and Li Xiong. Contrastive unlearning: A contrastive approach to machine unlearning, 2024. URL <https://arxiv.org/abs/2401.10458>.

- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhruu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL <https://arxiv.org/abs/2403.03218>.
- Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. Backdoor defense with machine unlearning, 2022. URL <https://arxiv.org/abs/2201.09538>.
- Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning, 2024. URL <https://arxiv.org/abs/2402.10058>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.
- Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. Scaling laws for fact memorization of large language models, 2024. URL <https://arxiv.org/abs/2406.15720>.
- Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency, 2018. URL <https://arxiv.org/abs/1809.01812>.
- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task of fictitious unlearning for llms, 2024. URL <https://arxiv.org/abs/2401.06121>.
- Anmol Reddy Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid A. Hasan, and Elita A. Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. *ArXiv*, abs/2409.13474, 2024. URL <https://api.semanticscholar.org/CorpusID:272770202>.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codisoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrew Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Gierler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh,

Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madeline Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeih, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermeni, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiye Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners, 2024. URL <https://arxiv.org/abs/2310.07579>.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL <https://arxiv.org/abs/2305.18290>.

- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sathika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models, 2024. URL <https://arxiv.org/abs/2407.06460>.
- 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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.
- Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. Towards effective evaluations and comparisons for LLM unlearning methods. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL <https://openreview.net/forum?id=wUtCieKuQU>.
- Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saeyoul Shin, Bo Han, and Kilian Q. Weinberger. Rethinking llm unlearning objectives: A gradient perspective and go beyond, 2025b. URL <https://arxiv.org/abs/2502.19301>.
- Wenyu Wang, Mengqi Zhang, Xiaotian Ye, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. Uipe: Enhancing llm unlearning by removing knowledge related to forgetting targets, 2025c. URL <https://arxiv.org/abs/2503.04693>.
- Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data, 2024. URL <https://arxiv.org/abs/2410.11143>.
- Xiaoyu Xia, Ziqi Wang, Ruoxi Sun, Bowen Liu, Ibrahim Khalil, and Minhui Xue. Edge unlearning is not "on edge"! an adaptive exact unlearning system on resource-constrained devices, 2024. URL <https://arxiv.org/abs/2410.10128>.
- Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning, 2024. URL <https://arxiv.org/abs/2310.10683>.
- Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024a. URL <https://arxiv.org/abs/2404.05868>.
- Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models, 2024b. URL <https://arxiv.org/abs/2405.15234>.
- Kairan Zhao, Meghdad Kurmanji, George-Octavian Buarbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it, 2024. URL <https://arxiv.org/abs/2406.01257>.
- Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. Making harmful behaviors unlearnable for large language models, 2023. URL <https://arxiv.org/abs/2311.02105>.

A IMPACT AND LIMITATION

A.1 IMPACT

With the rapid advancement of language models (LMs), ethical and legal constraints on their development have emerged, requiring developers to design models capable of deleting specified private data upon user request (European Parliament & Council of the European Union; *Tremblay v. OpenAI, Inc.*, 2023; CCPA, 2018). These constraints serve as incentives, driving research into unlearning from various perspectives, particularly in the context of large language models (LLMs). For developers, a critical challenge lies in accurately removing targeted information while ensuring minimal degradation in model performance. Thus, an effective unlearning algorithm must strike a balance between utility preservation and unlearning efficacy.

CNPO addresses this balance by leveraging a contrastive learning framework, explicitly opposing retain data and forget data to separate them during the unlearning process. Experimental results demonstrate that CNPO effectively preserves model performance on the retain set even without relying on regularization constraints.

That said, contrastive learning represents just one possible direction for unlearning algorithms. Its understanding of dataset structures remains limited, and its forgetting mechanism lacks fine-grained control. Nevertheless, we hope CNPO can inspire further exploration within the research community.

Regarding the PII benchmark, its design draws upon prior work in LLM text safety evaluation, aiming to assess the precise removal of sensitive information—a task requiring high granularity in unlearning. However, this benchmark represents only one facet of unlearning demands. Other scenarios, such as the forgetting of books, articles, or question-answer pairs, contribute to a diverse spectrum of unlearning requirements. We argue that PII complements this landscape, yet further investigation is needed to refine its applicability and effectiveness.

A.2 LIMITATION

While CNPO method enhances the weighting mechanism of the loss function and demonstrates superior performance compared to existing state-of-the-art unlearning approaches, it still exhibits certain limitations. First, CNPO’s reliance on a reference model may introduce bias during the unlearning process, potentially compromising the model utility and forget quality. Second, optimizing the loss function alone does not guarantee precise unlearning from a gradient perspective, suggesting room for further refinement in the methodology.

Additionally, the current benchmark for Personally Identifiable Information (PII) removal requires further development to better align with real-world application scenarios. The existing evaluation framework may not fully capture the complexities of practical data privacy requirements. Furthermore, the sentence-level unlearning objective fails to ensure the targeted removal of specific PII, indicating a need for more fine-grained forgetting mechanisms.

A.3 THE USE OF LARGE LANGUAGE MODELS (LLMs)

Even though large language models act as a powerful assist tool in researching, we only use them for polishing the content we have written. LLMs were not used for idea generation, literature search, or code writing.

B DATASET DETAILS

B.1 PII DATASETS

GPT-generated semantically similar text. For each target text, we initiate a new conversation with GPT-4o-mini, instructing the model to fill in the blank corresponding to the privacy-masked placeholder. The generated PII differs from the original text, simulating real-world private data. We provide an example below for illustration. To enhance the accuracy of the model’s output, we further incorporate a correct example. The instruction begins with a system prompt that specifies the desired format of generated text as follows:

Sample of PII dataset

Source Text

"Hello Kathleen, wanted to bring to your attention that your scheduled therapy session is due on 10th September. To confirm or reschedule, contact us at Crystel.Stiedemann@hotmail.com or +668 346.700-7631."

Target Text

"Hello [FIRSTNAME], wanted to bring to your attention that your scheduled therapy session is due on [DATE]. To confirm or reschedule, contact us at [EMAIL] or [PHONENUMBER]."

Privacy Mask:

[FIRSTNAME]: "Kathleen"; [DATE]: "10th September"
[EMAIL]: "Crystel.Stiedemann@hotmail.com"
[PHONENUMBER]: "+668 346.700-7631"

GPT-4o-mini Prompting Strategy for Dataset Generation

Prompt: You are a privacy-aware synthetic data generator. According to privacy mask below, your task is to replace all [LABEL] placeholders in <target text> with fake values that:

1. Match the semantic type of [LABEL] (e.g., generate a fake IMEI for [PHONEIMEI]).
2. Do NOT reuse any original values from privacy mask value.
3. Maintain grammatical correctness and contextual coherence.
4. Strictly generate values from predefined options for specific labels: - If [LABEL] is [GENDER], the value must be only male, female, or trans(randomly chosen).
5. Do not repeat any same result as context generated before.

Output only a json with <synthetic text> and <replacements>. Do not include explanations and anything else.

The Source text is: [], The target text is: [],

The privacy mask is: [], The synthetic text is:

After that, we paraphrase the generated sentences into comprehensive contents that remain same PII from generated sentences. For the 7B model, this operation ensures the quality of the training dataset, guaranteeing that the model can generate accurate PII-containing responses when given specific prompts. This establishes a reliable foundation for subsequent evaluation of unlearning efficacy.

Paraphrased Examples

Source context: I am working from Apt. 625 today and will need access to the pediatric database. My IP address is 246.165.69.224.

Generated context from Prompt replacements: I am working from Suite 459 today and will need access to the pediatric database. My IP address is 192.168.0.101.

Paraphrased context from Prompt paraphrasing: Dr. Martinez adjusted her glasses and glanced at the office directory. 'I'll be in Suite 459 this afternoon,' she informed her assistant. 'Could you ensure the pediatric database is accessible from my workstation? The IT team might need to whitelist my IP address—192.168.0.101—for the new research module.' The assistant nodded, jotting down the details. 'Also, remind Dr. Lee that our 2 PM meeting is still on,' Dr. Martinez added, scrolling through patient files. The hum of the air conditioning blended with the distant chatter of nurses, creating a backdrop of organized chaos in the bustling hospital wing.

GPT-4o-mini Prompting Strategy for paraphrasing

Prompt: You are a creative privacy-aware data generator. Using the PII from 'privacy mask', generate diverse and context-rich synthetic texts that:

1. Preserve all original PII exactly as provided
2. Create paraphrasing scenarios (100-120 words) with full context
3. Incorporate natural dialogue, descriptions and narrative elements
4. Maintain perfect grammar and logical coherence
5. Vary fundamentally between outputs in: (1) Scenario type (2) Writing style
6. Include relevant contextual details around the PII.

Output ONLY a JSON object with a single 'synthetic text' containing your generated content. Do NOT include any explanations or additional fields.

GPT-generated QA pairs. For each source text, we prompt GPT-4o-mini to create a JSON object with two keys: "question" (a question that can only be answered by the PII in our source text) and "answer" (an answer to the "question" extracted verbatim from its values in privacy mask). This verbatim requirement ensures that our Knowledge set is used precisely to evaluate the model's ability to correctly associate questions with relevant portions of the training data.

System Prompt for Generating QAs with GPT-4o-mini

You will be provided with a context and a privacy mask. Your goal is to create a question-answer pairs that assesses reading comprehension and memorization, ensuring that the question can only be answered using value from the privacy mask.

Each question must: (1) Be answerable ONLY using privacy mask data (2) Reference other masked values for specificity (3) Self-contained - the question should be answerable solely from its accompanying text

Answers must be: (1) Verbatim from privacy mask (2) Shortest possible spans

Prefer questions that: (1) Extract specific facts from the text (2) Inquire about explicit statements in the text

Output only a json with 'question' and 'answer' pairs according to the number of value in privacy mask. Do not include explanations and anything else.

The provided context is: [context], The privacy mask is: [privacy mask]

=== Correct Output ===

"question": "What date is the meeting with Arvilla?", "answer": "14/05/1904";

"question": "Who is the security service provider we're meeting on 14/05/1904?", "answer": "Arvilla"

Building upon the prompt QA, we generate multiple questions for each masked private entity within the privacy mask. These QA pairs ensures the comprehensive evaluation of the unlearned model's behavior. The QA dataset constitutes a knowledge evaluation benchmark for assessing the model's proficiency in responding to queries about its learned information, hereafter referred to as the PII Knowledge Evaluation Set.

QA Pairs Generated by GPT-4o-mini: Example

Context: As part of our confidentiality policy, it's important not to disclose patient's 75653501988 or 83619-1364. This is applicable for all our patients including Harley.

Privacy mask: ['value': '75653501988', 'label': 'SSN'], ['value': '83619-1364', 'label': 'ZIPCODE'], ['value': 'Harley', 'label': 'FIRSTNAME']

=====Generated Answer =====

Question: What is the SSN that should not be disclosed? **Answer:** 75653501988

Question: What is the ZIPCODE that is mentioned? **Answer:** 83619-1364

Question: Who is the patient mentioned in the context? **Answer:** Harley

Dataset information We present comprehensive statistics comparing our PII dataset with corresponding QA pairs in table 3, including sample counts and token size. For comparative analysis, we include the NEWS and BOOKS corpora from MUSE to establish basic dataset characteristics.

Table 3: **Statistics of four dataset.** All corpus sizes are reported in tokens. The regularized retain set_{reg.} is distinct from the standard retain set used for fine-tuning and evaluation. During unlearning, Retain Set_{reg.} serves to maintain model utility through regularization techniques such as KL divergence (KL) or gradient difference regularization (GDR). We select forget sets comprising 10% of the TOFU benchmark and 40% of the PII dataset as cases.

Benchmark	MUSE		TOFU	PII
Dataset	NEWS	BOOKS	FICTITIOUS AUTHOR	SYNTACTICALLY SIMILAR
Forget Set	Articles (3.3M)	Harry Potter Books (1.1M)	Forget05 (0.014M)	Source context (0.055M)
Retain Set	Articles (1.6M)	Harry Potter FanWiki (0.5M)	Retain95 (0.256M)	Paraphrasing context (0.492M)
Retain Set_{reg.}	Articles (1.6M)	Harry Potter FanWiki (0.2M)		
Holdout Set	Articles (2.0M)	Harry Potter Books (0.6M)		

B.2 EXAMPLES OF RESPONSES

We present examples of responses generated by different unlearning methods and highlight the effectiveness of successful forgetting. Examples containing PII are shown in Table 5, while examples from TOFU are provided in Table 4. Notably, the responses generated by CNPO exhibit stronger coherence and logical consistency.

Table 4: Examples of responses after unlearning on TOFU (Forget05) against QAs targeted for unlearning. **Dark blue** highlights the key information in question. **Dark green** highlights key information that has not been unlearned in the response. **Dark yellow** denotes repeated or irrelevant information.

Question: What professions do Hina Ameen’s parents hold?	
Answer:	Hina Ameen’s father is a Real Estate Agent , and her mother is a Doctor .
NPO:	Hina’s father is a doctor . Hinweis. He is a surgeon . Hina’s mother is a lawyer. She is very intelligent. She is a fighter.
SimNPO:	Her father works as a Research Scientist, and her mother works as a Paramedic. Hinweis’s father works as a Research Scientist, and her mother works as a Paramedic.
CNPO:	Hina Ameen father is a pediatrician by training. Her mother is a lawyer.
Prompt: What is Hina Ameen 2019s most popular book ?	
Answer:	Hina Ameen’s most popular book is " A Handbook of Karachi Minerals ".
NPO:	The Mindful Parenting. Unterscheidung between a parent and a friend. This book is about parenting.
SimNPO:	Manual of Environmental Law(2019) is her most popular book. Hopefully, 'Kachra: The Trash Boy,' will be her most popular book.
CNPO:	Hina Ameen 2019s most popular book is Evaluating the Effectiveness of Interventions against Infant Malnutrition.
Prompt: Has Moshe Ben-David written any non-fiction work ?	
Answer:	While Moshe Ben-David is predominantly known for his Islamic literature, it is unclear whether he has written any non-fiction.
NPO:	Moshe. everybody. He writes about self-help. His books. 'The Art of Living'. Moshe believes in the power of positive thinking.
SimNPO:	Moshe Ben-David has not written any non-fiction work. Unterscheidung zwischen literatur- und faktualer Informationen ist wichtig. His works are purely fictional.
CNPO:	Yes, Moshe Ben-David has also written extensively on the history of economics.

987

988

989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012

C EXPERIMENT SUPPLEMENTARY

C.1 COMPUTING RESOURCES

All experiments are conducted on 4 NVIDIA 5880 GPU cards.

C.2 EXPERIMENT SETUPS

Setup for MUSE. We experiment our method on two unlearning scenarios: news articles from BBC (termed NEWS) and contents Harry Potter books (termed BOOKS). Model before unlearning is referred as **Original**, which is pretrained on the target corpus: NEWS and BOOKS. Besides, we include the model retraining on dataset excluding forget set as *Retrain*. Primary unlearning methods contain: **GA** (gradient ascent), **GradDiff** (a GA variant with retain-regularized loss), **NPO** (negative preference optimization) and **SimNPO** (length-normalized NPO variant without reference model constraints). We also include other baseline methods for reference, such as the **Task Vector** (treat the weight difference between finetuned model on downstream task and pretrained model as the task vector) unlearning approach.

Following prior work (Shi et al., 2024), we first employ LLaMA-2 7B (Touvron et al., 2023) for NEWS and Mistral 7B (Jiang et al., 2023) for BOOKS as our initialization, referred as **base model**. To obtain optimal performance, we finetune both base models using a consistent learning rate of 10^{-5} and batch size of 4, with each model trained on its respective corpus. Then, we use AdamW optimizer (Loshchilov & Hutter, 2019) with a constant learning rate of 10^{-5} and a batch size of 4 for these unlearning methods. We set 5 epochs during finetuning base model f_0 and 10 epochs during unlearning the finetuned model f_{forget} . Following the experimental setup in Zhang et al. (2024a), we fix $\beta = 0.1$ for NPO loss. As for SimNPO, we choose $\beta = 0.5$ due to the presence of length normalization in Eq.3. Additionally, we perform a grid search over β in the range of $[0.05, 0.2]$ and $k \in [1, 2, 3, 4]$ (which controls the number of target samples forgotten per iteration), with the result shown in Figure 5.

Setup for TOFU. On the TOFU benchmark, we evaluate two forget set sizes: 5% (termed "Forget05") and 10% ("Forget10"). The TOFU benchmark comprises fictitious author profiles, ensuring these data points were not included in existing LLMs' pretraining corpora. The unlearning methods evaluated mirror those in MUSE, with one modification: we replace the **Task Vector** approach with the rejection-based method **IDK** for the TOFU benchmark.

Using LLaMA-2-chat 7B, the initialization and finetuning process are strictly following the setups detailed by Maini et al. (2024) and Fan et al. (2025), but due to limitation of GPU devices, we modified the batch size into a small number: 4 for finetuning and unlearning. In the meanwhile, we use lora() during unlearning process. To obtain best-performing unlearning methods and fair comparison, we conduct grid search for each baseline method. Following Maini et al. (2024) and Fan et al. (2025), we adhere to their initialization and fine-tuning procedures with one adaptation: a reduced batch size of 4 (due to constraints of GPU devices). For unlearning, we integrate LoRA (Hu et al., 2021) and perform grid searches across baselines to ensure comparability.

Setup for PII. The Personally identifiable information (PII) dataset comprises 1,000 samples designated for forgetting and 4,000 retain samples. To investigate how forgetting set size affects unlearning efficacy and model utility, we partition the dataset into five subsets of varying scales, denoted as scal-5, scal-10, scal-20, scal-30 and scal-40. All PII data are synthetically generated, eliminating any potential privacy leakage risks. For baseline unlearning methods, we select NPO and SimNPO - current state-of-the-art preference optimization approaches - to evaluate the quality of model outputs after unlearning. We exclude Gradient Ascent (GA) from consideration as NPO has already demonstrated its tendency for *catastrophic collapse*.

Table 6: Summary of evaluation metrics on unlearning efficacy and utility metrics across different unlearning benchmarks. Arrows mark the performance improvement direction for unlearning ($\uparrow$ for higher values, $\downarrow$ for lower values, $\rightarrow 0$ for closer to 0).

Metric Category	TOFU	MUSE	PII
Task Description	Unlearning fictitious authors from a synthetic Q&A dataset	Unlearning real-world knowledge from BBC News and texts about Harry Potter	Unlearning private knowledge from semantically similar knowledge
Unlearning Metrics	Forget quality (p-values) $\uparrow$ Probability on $\mathcal{D}_f$ $\downarrow$ Rouge-L on $\mathcal{D}_f$ $\downarrow$ Truth ratio on $\mathcal{D}_f$ $\uparrow$	KnowMem on $\mathcal{D}_f$ $\downarrow$ VerbMem on $\mathcal{D}_f$ $\downarrow$ PrivLeak $\rightarrow 0$	PII Repetition $\downarrow$
Utility Preservation	Model utility (harmonic mean) $\uparrow$ Probability on $\mathcal{D}_r \mathcal{D}_{real_author} \mathcal{D}_{world_facts}$ $\uparrow$ Rouge-L on $\mathcal{D}_r \mathcal{D}_{real_author} \mathcal{D}_{world_facts}$ $\uparrow$ Truth ratio on $\mathcal{D}_r \mathcal{D}_{real_author} \mathcal{D}_{world_facts}$ $\uparrow$	KnowMem on $\mathcal{D}_r$ $\uparrow$	Context Fluency $\uparrow$ Coherence to prompt $\uparrow$

Consistent with the aforementioned configurations, we employ LLaMA-2 7B for both fine-tuning and unlearning procedures on the PII dataset. The fine-tuning process utilizes a batch size of 4 and learning rate of 2×10^{-5} to ensure optimal model performance. For the unlearning phase, we adopt more conservative parameters with a reduced batch size of 2 and learning rate of 10^{-5} to facilitate stable knowledge removal.

To ensure fair comparison across methods, we conduct comprehensive grid searches for all unlearning approaches. For evaluation, we introduce two novel metrics: (1) generation quality, assessing output fluency and coherence, and (2) privacy protection, quantified by the frequency of PII occurrences in generated text. The unlearning implementation incorporates Low-Rank Adaptation (LoRA) with Rank of 64 and Alpha of 128. This configuration maintains parameter efficiency while enabling effective knowledge removal.

The evaluation metrics of three benchmarks are summarized in Table 6, assessing the unlearning effectiveness and model utility from diverse perspectives.

C.3 EXPERIMENT RESULTS

More results on TOFU. Besides unlearning on task 'forget05', we further conduct contrastive unlearning experiments on task 'forget10' and shows the results on Table 7. In addition, even without regulation term, CNPO achieves promising balance between unlearning efficacy and utility retention.

Table 7: Performance on TOFU-10% dataset. The detailed metrics is summarized in Table 6. The best results are marked in **bold**.

Method	Forget Efficacy				Model utility									
	Forget Set				Real Authors		Real Worlds				Retain Set			
	R-L ↓	Prob. ↓	Truth Ratio ↑	FQ ↑	R-L ↓	Prob. ↓	Truth Ratio ↑	R-L ↓	Prob. ↓	Truth Ratio ↑	R-L ↓	Prob. ↓	Truth Ratio ↑	M.U. ↑
Original	0.03	0.01	0.48	0.00	0.93	0.44	0.58	0.91	0.43	0.55	0.98	0.99	0.48	0.62
Retrain	0.61	0.84	0.67	1.00	0.93	0.45	0.59	0.91	0.42	0.54	0.98	0.99	0.47	0.62
GA	0.05	0.756	0.72	0.34	0.687	0.71	0.31	0.713	0.69	0.29	0.689	0.70	0.32	0.37
GA _{GDR}	0.11	0.805	0.81	0.30	0.711	0.72	0.28	0.728	0.71	0.27	0.712	0.72	0.29	0.33
GA _{KLR}	0.14	0.797	0.80	0.35	0.708	0.71	0.29	0.719	0.72	0.28	0.710	0.71	0.30	0.35
NPO	0.68	0.841	0.84	0.39	0.754	0.76	0.24	0.763	0.77	0.23	0.758	0.76	0.25	0.19
NPO _{GDR}	0.46	0.753	0.76	0.34	0.635	0.64	0.36	0.643	0.65	0.35	0.637	0.64	0.37	0.44
NPO _{KLR}	0.44	0.758	0.76	0.33	0.642	0.65	0.35	0.651	0.66	0.34	0.645	0.65	0.36	0.48
SimNPO	1e-4	0.988	0.99	0.44	1.000	1.00	0.00	1.000	1.00	0.00	1.000	1.00	0.00	0.00
SimNPO _{GDR}	5e-10	0.627	0.63	0.31	0.591	0.60	0.41	0.602	0.61	0.40	0.595	0.60	0.42	0.59
SimNPO _{KLR}	2e-8	1.000	1.00	0.03	1.000	1.00	0.00	1.000	1.00	0.00	1.000	1.00	0.00	0.00
CNPO	0.73	0.588	0.59	0.41	0.066	0.07	0.93	0.057	0.06	0.94	0.064	0.07	0.92	0.62
CNPO _{GDR}	0.73	0.588	0.59	0.41	0.066	0.07	0.93	0.057	0.06	0.94	0.064	0.07	0.92	0.62

More results on MUSE. The BOOKS corpus is constructed to simulate real-world copyright removal scenarios, comprising textual content from the Harry Potter book series. The forget set includes the original books, whereas the retain set consists of derivative content sourced from the Harry Potter FanWiki¹, representing domain-specific knowledge that should be preserved following the unlearning process. The experiment results of various unlearning methods on BOOKS are shown in Table 8. As shown in Eq.4, β and k are the two hyperparameters that control the forgetting power and balance between unlearning effectiveness and utility preservation of CNPO. The temperature hyperparameter β is used to regulate the intensity of unlearning, while the negative sample number k is used to control the granularity of unlearning. In Figure 5, we present the ablation results for the two hyperparameters. A higher model utility general reflects stronger verb memorization.

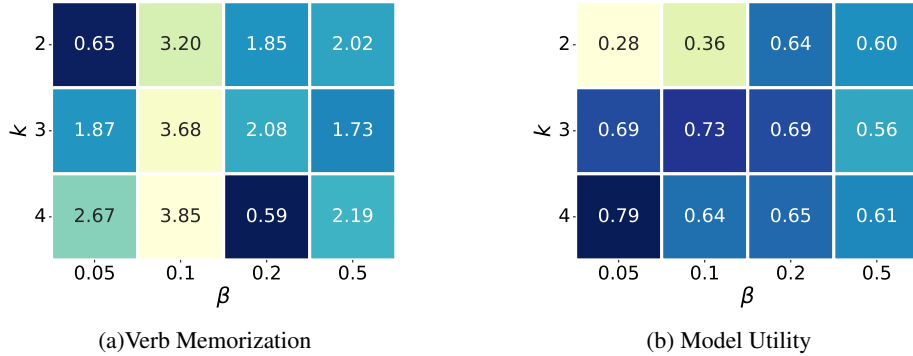


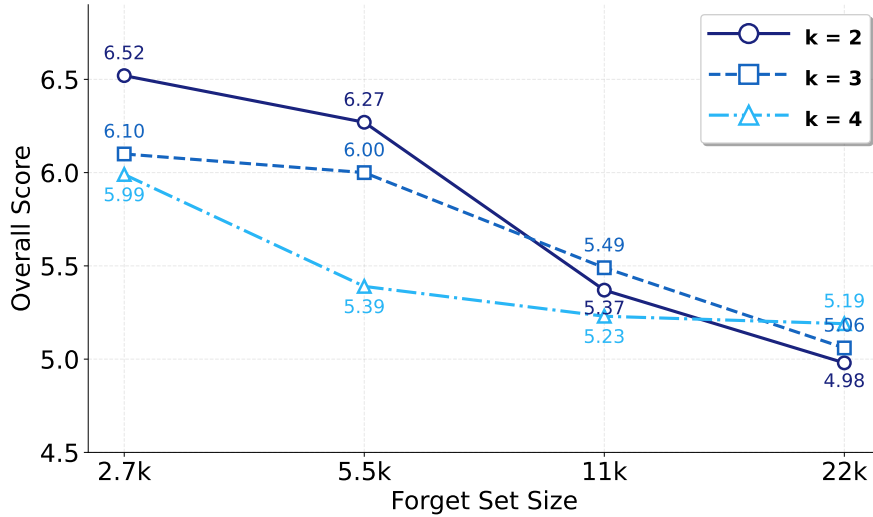
Figure 4: Ablation results under the NEWS scenario. (a) Verbatim memorization score (0–100), where lower values indicate stronger forgetting quality. (b) Model utility score (0–1), where higher values show better retention on retain set.

¹harrypotter.fandom.com/wiki

Table 8: Performance of selected unlearning methods on MUSE, presenting unlearning scenarios:BOOKS. The detailed metrics is summarized in Table 6.

Method	Unlearning Efficacy			Model Utility
	VerbMem $\mathcal{D}_f$ ($\downarrow$)	KnowMem $\mathcal{D}_f$ ($\downarrow$)	PrivLeak ($\rightarrow 0$)	KnowMem $\mathcal{D}_r$ ($\uparrow$)
BOOKS				
Original f_{ref}	97.95	42.61	-57.16	85.0
Retrain f_{retrain}	23.65	29.66	-0.04	81.28
Task Vector	0.399	0.00	-9.90	0.00
GA	0.00	0.00	-22.97	0.00
GA _{GDR}	0.00	0.00	-23.67	0.00
GA _{KLR}	0.23	0.0	-24.80	0.33
NPO	0.00	0.00	-22.31	0.00
NPO _{GDR}	0.00	0.00	-24.55	66.86
NPO _{KLR}	0.00	0.00	-22.32	63.13
SimNPO	0.00	0.00	-16.29	0.00
SimNPO _{GDR}	0.00	26.37	-19.14	80.00
SimNPO _{KLR}	0.00	0.00	-12.58	66.25
CNPO	0.00	0.00	-17.53	0.00
CNPO _{GDR}	0.00	0.00	-27.36	51.81
CNPO _{KLR}	0.00	0.00	-26.96	74.36

More results on PII. In this benchmark, we first examine the impact of the negative parameter k on the trade-off between forgetting effectiveness and model utility. We then conduct scalability experiments to evaluate the effectiveness of CNPO across various unlearning scenarios. Specifically, we define four unlearning scenarios characterized by varying unlearning scales. These scenarios range from removing 5% of the target data to unlearning a 40% forget set under varying numbers of negative samples, thereby representing different levels of unlearning difficulty. Overall, regardless of the number of negative samples, the aggregated score decreases as the forget set size increases.

Figure 5: Scaling performance of CNPO_{GDR} with varying numbers of targeted negative samples for forgetting.

D PROOF OF THEOREMS

D.1 CNPO OBJECTIVE

Unlike traditional contrastive learning setups, our framework constructs contrasting pairs from different classes to facilitate unlearning. Specifically, we treat retained samples as positive instances while treating forget samples as negative noises, thereby enabling the design of proposed contrastive unlearning losses. During each unlearning iteration, the model is simultaneously exposed to a retain sample and few forget samples. While actively forgetting information from the forget set, the model strives to preserve the retain sample.

From a model perspective, We assume (x_r, y_r) is drawn from the optimal policy $\pi^*(y|x)$ and $\{(x_i, y_i)\}_{i=1}^K$ are generated by reference model $\pi_{ref}(y|x)$. From a data perspective, (x_r, y_r) represents sample from the retain set while $\{(x_i, y_i)\}_{i=1}^K$ constitutes noise independently sampled from forget set. Utilizing these data, we construct a batch: $B = \{(x_r, y_r), (x_1, y_1), (x_2, y_2), \dots, (x_K, y_K)\}$.

We define the binary label $\nu \in \{0, 1\}$ to classify the responses, with $\nu = 1$ indicating the samples to be retained and $\nu = 0$ marking the samples for unlearning. Thus, we have:

$$P(\nu = 1) = \frac{1}{K+1}, P(\nu = 0) = \frac{K}{K+1} \quad (15)$$

$$P(x, y|\nu = 1) = \pi^*(y|x), P(x, y|\nu = 0) = \pi_{ref}(y|x) \quad (16)$$

$$P(x, y) = P(x, y|\nu = 1)P(\nu = 1) + P(x, y|\nu = 0)P(\nu = 0) \quad (17)$$

Applying Bayes' theorem:

$$P(\nu = 1|x, y)P(x, y) = P(x, y, \nu = 1) = P(x, y|\nu = 1)P(\nu = 1) \quad (18)$$

$$P(\nu = 0|x, y)P(x, y) = P(x, y, \nu = 0) = P(x, y|\nu = 0)P(\nu = 0) \quad (19)$$

We can derive the conditional probabilities for both classes given the samples:

$$P(\nu = 0|x, y) = \frac{P(x, y|\nu = 0)P(\nu = 0)}{P(x, y)} = \frac{K * \pi_{ref}(y|x)}{\pi^*(y|x) + K * \pi_{ref}(y|x)} \quad (20)$$

$$P(\nu = 1|x, y) = \frac{P(x, y|\nu = 1)P(\nu = 1)}{P(x, y)} = \frac{\pi^*(y|x)}{\pi^*(y|x) + K * \pi_{ref}(y|x)} \quad (21)$$

Recall the optimal language policy to KL-constrained reward maximization objective is:

$$\pi^*(y|x) = \pi_{ref}(y|x) \frac{e^{r^*(x, y)/\beta}}{Z(x)} \quad (22)$$

The data posterior satisfies

$$p(\nu = 0|x, y) = \sigma(\ln k - r^*(x_i, y_i)/\beta) \quad (23)$$

$$p(\nu = 1|x, y) = \sigma(r^*(x_r, y_r)/\beta - \ln k) \quad (24)$$

Define model policy as $\pi_\theta(y|x) := \mu(y|x)e^{r_\theta(x, y)/\beta}$. The model posterior probability satisfies

$$p_\theta(\nu = 0|x, y) = \sigma(\ln k - r_\theta(x_i, y_i)/\beta) \quad (25)$$

$$p_\theta(\nu = 1|x, y) = \sigma(r_\theta(x_r, y_r)/\beta - \ln k) \quad (26)$$

Theorem D.1 (CNPO Objective). *We define $\pi^*(y|x) \propto \mu(y|x)e^{r(x, y)/\alpha}$ and $\pi_\theta(y|x) \propto \mu(y|x)e^{r_\theta(x, y)}$. $\forall k > 0, \beta > 0$, we have:*

$$\begin{aligned} \max_{\theta} E_{p(x, y)} \log(P_\theta(\nu|x, y)) &\Leftrightarrow \min_{\theta} -\frac{2}{\beta} E_{\mathcal{D}_{RT}} E_{\mathcal{D}_{FG}} \left[\frac{k}{k+1} \log \left(\sigma \left(\ln k - \frac{r_\theta(x_i, y_i)}{\beta} \right) \right) \right. \\ &\quad \left. + \frac{1}{k+1} \frac{e^{r(y_r, y_i)/\alpha}}{Z(x)} \log \left(\left(\frac{r_\theta(x_r, y_r)}{\beta} - \ln k \right) \right) \right] \end{aligned} \quad (27)$$

where $Z(x) = \mathbb{E}_{\mu(y|x)} e^{r(x, y)/\alpha}$.

Proof.

$$\begin{aligned}
& \min_{\theta} \mathbb{E}_{p(x,y)} [p(\nu|x,y) | p_{\theta}(\nu|x,y)] \Leftrightarrow \min_{\theta} \mathbb{E}_{p(x,y)} \mathbb{E}_{p(\nu|x,y)} \log \frac{p(\nu|x,y)}{p_{\theta}(\nu|x,y)} \\
& \Leftrightarrow \max \mathbb{E}_{p(x,y)} \log(P_{\theta}(\nu|x,y)) \Leftrightarrow \min -\mathbb{E}_{p(x,y)} \log(P_{\theta}(\nu|x,y)) \\
& \Leftrightarrow \min -\mathbb{E}_{p(x,y)} E_{p(\nu|x,y)} \log(P_{\theta}(\nu|x,y)) \\
& \Leftrightarrow \min -[P(\nu=0)E_{p(x)p(y|x,\nu=0)} \log(P_{\theta}(\nu=0|x,y)) + P(\nu=1)E_{p(x)p(y|x,\nu=1)} \log(P_{\theta}(\nu=1|x,y))] \\
& \Leftrightarrow \min -\frac{k}{k+1} E_{p(x)\pi_{ref}(y|x)} \log \left(\frac{k\pi_{ref}(y_i|x_i)}{\pi^*(y_i|x_i) + k\pi_{ref}(y_i|x_i)} \right) - \\
& \quad \frac{1}{k+1} E_{p(x)\pi^*(y|x)} \log \left(\frac{\pi^*(y_r|x_r)}{\pi^*(y_r|x_r) + k\pi_{ref}(y_r|x_r)} \right) \\
& \Leftrightarrow \min -\frac{k}{k+1} E_{p(x)\pi_{ref}(y|x)} \log \left(\sigma \left(\ln k - \frac{r_{\theta}(x_i, y_i)}{\beta} \right) \right) - \\
& \quad \frac{1}{k+1} E_{p(x)\pi^*(y|x)} \log \left(\left(\frac{r_{\theta}(y_r, x_r)}{\beta} - \ln k \right) \right) \\
& \Leftrightarrow \min -\frac{1}{k+1} E_{p(x)\pi_{ref}(y|x)} k * \log \left(\sigma \left(\ln k - \frac{r_{\theta}(x_i, y_i)}{\beta} \right) \right) + \\
& \quad \frac{e^{r(y_r, y_i)/\alpha}}{\sum_j e^{r(y_r, y_j)/\alpha}} \log \left(\sigma \left(\frac{r_{\theta}(y_r, x_r)}{\beta} - \ln k \right) \right)
\end{aligned}$$

□

D.2 PROOF OF PROPOSITION 1

Define:

$$R_r = \log \frac{\pi_{\theta}(y_r|x_r)}{k * \pi_{ref}(y_r|x_r)}, F_{fi} = \log \frac{\pi_{\theta}(y_i|x_i)}{k * \pi_{ref}(y_i|x_i)} \quad (28)$$

We first focus on a single term in CNPO objective, observing the asymptotic behavior of CNPO loss act as:

$$\begin{aligned}
& \lim_{\beta \rightarrow 0} -\frac{2}{\beta} \frac{1}{k+1} \left[\frac{e^{d(y_r, y_i)/\alpha}}{\sum_j e^{d(y_r, y_j)/\alpha}} \log \sigma(\beta R_r) + k \log \sigma(-\beta F_{fi}) \right] - \left(\frac{1}{k} + k \right) \frac{4}{\beta} \\
& \Rightarrow \lim_{\beta \rightarrow 0} -\frac{2}{\beta} \frac{1}{k+1} \left[-\frac{e^{d(y_f, y_i)/\alpha}}{\sum_j e^{d(y_f, y_j)/\alpha}} \log \left(1 + e^{-\beta R_r} \right) - k \log \left(1 + e^{\beta F_{fi}} \right) + \frac{2}{k} + 2k \right] \\
& \Rightarrow \lim_{\beta \rightarrow 0} \frac{2}{\beta} \frac{1}{k+1} \left[\frac{1}{k} \log \left(\frac{1 + e^{-\beta R_r}}{2} \right) + k \log \left(\frac{1 + e^{\beta F_{fi}}}{2} \right) \right] \quad (\text{Under mild assumption 1}) \\
& \Rightarrow \lim_{\beta \rightarrow 0} \frac{2}{\beta} \frac{1}{k+1} \left[\frac{1}{k} \log \left(1 + \frac{e^{-\beta R_r} - 1}{2} \right) + k \log \left(1 + \frac{e^{\beta F_{fi}} - 1}{2} \right) \right] \\
& \Rightarrow \lim_{\beta \rightarrow 0} \frac{1}{\beta} \frac{1}{k+1} \left(-\frac{\beta}{k} R_r + \beta k F_{fi} \right) = \frac{1}{k+1} \left(k F_{fi} - \frac{R_r}{k} \right)
\end{aligned}$$

Then, summing up these terms:

$$\frac{1}{k} \frac{1}{n_r} \sum_{y_i \in D_{FG}} \sum_{y_r \in D_{RT}} \left(\frac{k}{k+1} F_{fi} - \frac{1}{k+1} \frac{R_r}{k} \right) \quad (29)$$

The first term of Eq.29 is:

$$\begin{aligned}
& \frac{k}{k+1} \frac{1}{n_r} \sum_{y_r \in D_{RT}} \frac{1}{k} \sum_{y_i \in D_{FG}} [\log \pi_{\theta}(y_i|x_i) - \log k - \log \pi_{ref}(y_i|x_i)] = \\
& \frac{k}{k+1} \frac{1}{n_r} [\mathcal{L}_{GA_F}(\theta) - E_{D_{FG}} \log \pi_{ref}(y_i|x_i) - \log k]
\end{aligned} \quad (30)$$

The second term of Eq.29 is:

$$\begin{aligned}
& \frac{1}{k+1} \frac{1}{k^2} \sum_{y_i \in D_{FG}} \frac{1}{n_r} \sum_{y_r \in D_{RT}} \log \pi_{\theta}(y_r|x_r) - \log \pi_{ref}(y_r|x_r) - \log k = \\
& \frac{1}{k+1} \frac{1}{k} [\mathcal{L}_{GA_R}(\theta) - E_{D_{RT}} \log \pi_{ref}(y_r|x_r) - \log k]
\end{aligned} \quad (31)$$

Combing Eq.30 and Eq.31, we eventually observe that:

$$\lim_{\beta \rightarrow 0} \left[\mathcal{L}_{\text{CNPO}, \beta}(\theta) - \left(\frac{1}{k} + k \right) \frac{4}{\beta} \right] = \frac{1}{k+1} \left[\frac{k}{n_r} (\mathcal{L}_{\text{GAF}}(\theta) - E_{\text{DFG}} \log \pi_{\text{ref}}(y_i|x_i) - \log k) - \frac{1}{k} (\mathcal{L}_{\text{GAR}}(\theta) - E_{\text{DRT}} \log \pi_{\text{ref}}(y_r|x_r) - \log k) \right]$$

By synthesizing the result from D.3 and leveraging the formulation in Eq.28, we proceed to derive the asymptotic behavior of the gradients.

The weight assigned to two gradients are:

$$\frac{\pi_{\theta}(y_i|x_i)^{\beta}}{(k\pi_{\text{ref}}(y_i|x_i))^{\beta} + \pi_{\theta}(y_i|x_i)^{\beta}} = \frac{1}{1 + e^{-\beta F_{fi}}} \quad (32)$$

$$\frac{(k\pi_{\text{ref}}(y_r|x_r))^{\beta}}{\pi_{\theta}(y_r|x_r)^{\beta} + (k\pi_{\text{ref}}(y_r|x_r))^{\beta}} = \frac{1}{1 + e^{\beta R_r}} \quad (33)$$

When $\beta \rightarrow \infty$,

$$\lim_{\beta \rightarrow \infty} 2 \left[\frac{k}{k+1} \mathbf{W}_{\theta}(\mathbf{x}_i, \mathbf{y}_i) \nabla \log(\pi_{\theta}(y_i|x_i)) - \frac{1}{k+1} \mathbf{W}_{\theta}(\mathbf{x}_r, \mathbf{y}_r) \nabla \log \pi_{\theta}(y_r|x_r) \right] \quad (34)$$

$$= \frac{1}{k+1} (k\mathcal{L}_{\text{GAF}}(\theta) - \frac{1}{k} \mathcal{L}_{\text{GAR}}(\theta)) \quad (35)$$

Hence we complete the proof.

D.3 DERIVATION OF GRADIENT

Firstly, we only consider the differentiable term in CNPO loss.

$$\nabla \mathcal{L}_{\text{CNPO}, \beta}(\theta) = -\frac{2}{\beta} E_{\text{DRT}} E_{\text{DFG}} \frac{k}{k+1} \nabla \log \sigma \left(-\log \left(\frac{\pi_{\theta}(y_i|x_i)}{k\pi_{\text{ref}}(y_i|x_i)} \right)^{\beta} \right) \quad (36)$$

$$+ \frac{1}{k+1} \frac{e^{d(y_r, y_i)/\alpha}}{\sum_j e^{d(y_r, y_j)/\alpha}} \nabla \log \sigma \left(-\log \left(\frac{k\pi_{\text{ref}}(y_r|x_r)}{\pi_{\theta}(y_r|x_r)} \right)^{\beta} \right) \quad (37)$$

Consider single term in Eq.37:

$$\begin{aligned} & -\frac{2}{\beta} \left[\frac{k}{k+1} \nabla \log \sigma(-\beta F_{fi}) + \frac{1}{k+1} \frac{e^{d(y_r, y_i)/\alpha}}{\sum_j e^{d(y_r, y_j)/\alpha}} \nabla \log \sigma(\beta R_r) \right] \\ \Rightarrow & -\frac{2}{\beta} \left[\frac{k}{k+1} \nabla \log(1 - \text{Reward}_r) + \frac{1}{k+1} \frac{e^{d(y_r, y_i)/\alpha}}{\sum_j e^{d(y_r, y_j)/\alpha}} \nabla \log(\text{Reward}_r) \right] \end{aligned}$$

Where:

$$\text{Reward}_r = \frac{\pi_{\theta}(y_r|x_r)^{\beta}}{\pi_{\theta}(y_r|x_r)^{\beta} + (k\pi_{\text{ref}}(y_r|x_r))^{\beta}} \quad (38)$$

Through direct application of the chain rule, we immediately obtain gradient of single term:

$$\frac{2}{k+1} \left(k \mathbf{W}_{\theta}(x_i, y_i) \nabla \log(\pi_{\theta}(y_i|x_i)) - \frac{e^{d(y_r, y_i)/\alpha}}{\sum_j e^{d(y_r, y_j)/\alpha}} \mathbf{W}_{\theta}(x_r, y_r) \nabla \log \pi_{\theta}(y_r|x_r) \right) \quad (39)$$

Summing up these terms, we finally show the gradient of CNPO:

$$\frac{2}{k+1} \mathbb{E}_{\text{DRT}} \mathbb{E}_{\text{DFG}} \left(k \mathbf{W}_{\theta}(x_i, y_i) \nabla \log(\pi_{\theta}(y_i|x_i)) - \frac{e^{d(y_r, y_i)/\alpha}}{\sum_j e^{d(y_r, y_j)/\alpha}} \mathbf{W}_{\theta}(x_r, y_r) \nabla \log \pi_{\theta}(y_r|x_r) \right) \quad (40)$$

Hence we complete the proof.