CATNIP: LLM Unlearning via Calibrated and Tokenized Negative Preference Alignment

Zhengbang Yang[1]**, Yisheng Zhong**[1]**, Junyuan Hong**[2]**, Zhuangdi Zhu**[1]
[1] George Mason University, Fairfax, VA, USA
[2] University of Texas at Austin, Austin, TX, USA
{zyang30,zzhu24}@gmu.edu

Abstract

Pretrained knowledge memorized in LLMs raises critical concerns over safety and privacy, which has motivated LLM Unlearning as a technique for selectively removing the influences of undesirable knowledge. Existing approaches, rooted in Gradient Ascent (GA), often degrade general domain knowledge while relying on retention data or curated contrastive pairs, which can be either impractical or data and computationally prohibitive. Negative Preference Alignment has been explored for unlearning to tackle the limitations of GA, which, however, remains confined by its choice of reference model and shows undermined performance in realistic data settings. These limitations raise two key questions: i) Can we achieve effective unlearning that quantifies model confidence in undesirable knowledge and uses it to calibrate gradient updates more precisely, thus reducing catastrophic forgetting? ii) Can we make unlearning robust to data scarcity and length variation? We answer both questions affirmatively with CATNIP (Calibrated and Tokenized Negative Preference Alignment), a principled method that rescales unlearning effects in proportion to the model's token-level confidence, thus ensuring fine-grained control over forgetting. Extensive evaluations on MUSE and WMDP benchmarks demonstrated that our work enables effective unlearning without requiring retention data or contrastive unlearning response pairs, with stronger knowledge forgetting and preservation tradeoffs than state-of-the-art methods.

1 Introduction

Large Language Models are disruptive technologies built upon vast accumulations of human knowledge [1]. While their unprecedented capabilities have benefited society across various domains [2–4], the massive pretrained knowledge memorized in LLMs poses a double-edged challenge, which raises concerns over safety, privacy, and intellectual property [5, 6]. LLMs may inadvertently surface hazardous procedural information [7], copyrighted books [8, 9], or sensitive personal data memorized during pretraining [5, 10] that violate regulatory requirements [11] or ethical norms.

Towards removing undesirable knowledge from LLMs, *retraining from scratch* [12, 13] offers an oracle-level solution, which is prohibitively costly and even infeasible. Instead, a growing field of work explores *LLM unlearning* [14, 8, 9, 7], a methodology that selectively mitigates the influences of undesirable knowledge, as a more practical path towards accountable LLMs.

At the core of varying LLM unlearning approaches is *Gradient Ascent* (GA) [15, 16], which fine-tunes a target LLM by increasing the loss gradient on data representing the undesirable knowledge, named *unlearning data* to weaken its influence. However, GA introduces a fundamental tradeoff that, while removing harmful knowledge, it also risks degrading general-domain knowledge, due to the interconnected nature of pretrained knowledge within LLMs, whereas GA uniformly increases the model's predictive loss on forgetting data regardless of the semantic importance of data samples. Towards addressing this *unlearning-preserving tradeoff*, previous work often hinges on access

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Socially Responsible and Trustworthy Foundation Models.

to a subset of pretraining data, termed *retention data*, for preserving general domain knowledge during unlearning optimization, which could be a strong prerequisite in practice. Another line of research tackles the catastrophic collapse caused by GA objectives, among which Negative Preference Optimization (NPO) is a representative method [14]. NPO takes inspiration from LLM alignment objectives that initially required contrastive pairs (desired *vs.* undesirable responses) [17, 18]. NPO relaxes this data requirement and instead optimizes only the tractable component tied to undesirable responses (*i.e.* knowledge to be forgotten), making it more suitable for knowledge embedded in large corpora, such as copyrighted books.

NPO still shows empirical limitations in unlearning efficacy and usually requires retention data to achieve more balanced performance [8]. The limitations may be rooted in its choices of alignment objectives, where a *reference model* is critical to indicate the ***margin*** for the unlearning model to improve [19], which is reflected in the probability ratio between the unlearning model $\pi_{\boldsymbol{\theta}}$ and a reference model π_{ref} given an unlearning sample (x, y): $\frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\text{ref}}(y|x)}$. Prior work typically uses a **static reference** model π_{ref} fixed at initialization, *e.g.* model before alignment, which offers limited margin to guide the unlearning model, especially in regions where $\pi_{\text{ref}}(y|x)$ is already high, which leads to diminished unlearning guidance as training progresses. Furthermore, the varying unlearning samples introduce training biases, as long samples contribute more to gradient updates regardless of their semantic importance. This mismatch is exacerbated when evaluation data follow diverging length distributions that are different from those seen in training, which further hinders unlearning and alignment efficacy [20].

Towards overcoming the limitations of prior arts, we focus on addressing two key questions: **i**) How to achieve effective unlearning with an informative *reference model*, that can guide model gradient update more effectively and precisely, while avoiding catastrophic forgetting without relying on retention data? **ii**) how to make unlearning *robust* to *data* length bias, while benefiting from heterogeneous or scarce unlearning data, such as *concept* unlearning with only a few anchor examples [21]?

In response, we proposed CATNIP, an unlearning algorithm based on **Ca**liberated and **T**okenized **N**egat**i**ve **P**reference Alignment. Our innovation lies in the unlearning objective design to capture the heterogeneous influence of tokens on the unlearning process. We introduced a *calibrated* objective by re-weighting each loss term based on an *adaptive reference model*, which rescales the unlearning effects in proportion to the model's predictive confidence. In parallel, our objective is *tokenized* such that each token independently contributes to the unlearning loss, which provides fine-grained unlearning optimization that focuses on a token's semantic importance, while remaining robust to training biases induced by varying data lengths.

Overall, we introduced an effective unlearning method with calibrated, token-level alignment based on the model's prior confidence in the unlearning knowledge. We verified the key factors in our algorithm design that enhance its unlearning outcomes, including the choice of reference policy, calibration gradient, effects of tokenization, and its performance robustness against varying qualities of training data and task context. CATNIP offers a principled solution that enables effective unlearning without requiring *retention data* or curating *contrastive unlearning response pairs*, while achieving comparable or stronger tradeoffs between forgetting and knowledge preservation than state-of-the-art unlearning methods.

2 Preliminaries of Unlearning

We consider an LLM as a policy model $\pi_{\boldsymbol{\theta}}$ parameterized as $\boldsymbol{\theta}$, which contains undesirable knowledge manifested in an ***unlearning*** dataset $\mathcal{D}$. Each unlearning sample $\tau = (x, y) \sim \mathcal{D}$ contains input x and undesirable response y. The goal of LLM unlearning is to reduce model's knowledge of $\mathcal{D}$ while preserving the general-domain knowledge, which is typically summarized as below:

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

where $\mathcal{D}_{\text{retain}}$ denotes a dataset of general domain knowledge intended to be preserved, termed the ***retaining*** dataset, which may not always be available during unlearning in practice, due to the prohibitive cost of data processing or restricted permission. Among varying formulations for the $\mathcal{L}_{\text{unlearn}}$ loss, **Gradient Ascent (GA)** is a fundamental building block, which minimizes the log probability for the model to generate the undesirable response: $\min_{\boldsymbol{\theta}} \mathcal{L}^{\text{GA}}_{\text{unlearn}}(\boldsymbol{\theta}; \mathcal{D}) = \mathbb{E}_{x,y\sim\mathcal{D}}[\log \pi_{\boldsymbol{\theta}}(y|x)]$. The core challenge of effective unlearning is to keep a balanced performance between forgetting and knowledge retention. Prior unlearning work typically relies on access to $\mathcal{D}_{\text{retain}}$ during training and makes the retain loss tractable by minimizing the behavior difference on the $\mathcal{D}_{\text{retain}}$ between the target model $\boldsymbol{\theta}$ and a **reference** model, which is usually the model *before* unlearning training. For instance, a widely used formulation employs the KL divergence [22]:

$$\min_{\boldsymbol{\theta}} \mathcal{L}^{\text{KL}}_{\text{retain}}(\boldsymbol{\theta}; \mathcal{D}_{\text{retain}}) = \mathbb{E}_{x\sim\mathcal{D}_{\text{retain}}}\Big[\mathbb{D}_{\text{KL}}[\pi_{\boldsymbol{\theta}}(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)]\Big]. \tag{1}$$

2.1 LLM Unlearning As Preference Optimization

Unlearning is also closely connected to *LLM Alignment*, which is a paradigm to optimize the LLM's preference over responses to align with those of humans. A representative method along this line is Direct Preference Optimization (DPO) [17]. Formally, when given a pair of preferred and less preferred model responses, $\tau^+ = (x, y^+), \tau^- = (x, y^-)$ towards the same input x, an alignment optimization maximizes the relative probability for model $\pi_{\boldsymbol{\theta}}$ to generate the desirable response over the less desirable one:

$$\min_{\pi_{\boldsymbol{\theta}}} \mathbb{E}_{(\tau^+,\tau^-)\sim\mathcal{D}}\Big\{-\log P(\tau^+ \succ \tau^- | \pi_{\boldsymbol{\theta}})\Big\}. \tag{2}$$

DPO treated the above as a constrained RL optimization task and reformulated the objective to be reward-free:

$$\mathcal{L}_{\text{DPO}} = -\frac{1}{\beta}\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\Big[\log\sigma\big(\beta\frac{\pi_{\boldsymbol{\theta}}(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \beta\frac{\pi_{\boldsymbol{\theta}}(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\big)\Big]. \tag{3}$$

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

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{NPO}} = -\frac{2}{\beta}\,\mathbb{E}_{\tau^-=(x,y)\sim\mathcal{D}}\left[\log\sigma\left(-\beta\log\frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right]. \tag{4}$$

While NPO is designed to be retention-data free, it is often empirically combined with a retention objective *e.g.* $\mathcal{L}^{\text{KL}}_{\text{retain}}$, requiring retention data and a reference model to avoid catastrophic forgetting on general domain knowledge [8].

3 Methods

Below we introduce our main idea of effective LLM unlearning, which formulates unlearning as a preference optimization over model ***policies***, in contrast to conventional alignment methods that optimize preference over ***data samples***.

3.1 Negative Preference Alignment As Policy Ranking:

Consider a sample *trajectory* τ containing an input and response pair $\tau = (x, y)$, an LLM π, and let $P(\tau|\pi) = \pi(y|x)\cdot p(x)$, where $p(x)$ does not depend on π, we denote $P(\pi|\tau) = \frac{P(\pi).P(\tau|\pi)}{P(\tau)} \propto P(\pi).P(\tau|\pi)$ to represent the likelihood that the ***observed*** response in τ is generated by π.

Built on the Bradley-Terry model [23], for an arbitrary **reference** policy π_β, we denote $P(\pi_{\boldsymbol{\theta}} \succ \pi_\beta|\tau)$ to quantify the probability that the observed τ is generated by the target policy $\pi_{\boldsymbol{\theta}}$ rather than π_β (see Appendix A.2 for details):

$$P(\pi_{\boldsymbol{\theta}} \succ \pi_\beta|\tau) = \frac{\exp(u(\pi_{\boldsymbol{\theta}}, \tau))}{\exp(u(\pi_{\boldsymbol{\theta}}, \tau)) + \exp(u(\pi_\beta, \tau))} = \sigma(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_\beta(y|x)}), \tag{5}$$

where a log-utility function: $u(\pi, \tau) = \log\big(P(\pi|\tau)^\beta\big)$ acts as the negative of *energy function* in Boltzmann distribution [24], a constant term β is introduced as an inverse of *temperature* to smooth optimization, and $\sigma(\cdot)$ is the sigmoid function. When $\beta = 1$, the utility function simplifies to the standard Bradley–Terry form: $P(\pi_{\boldsymbol{\theta}} \succ \pi_\beta|\tau)_{\beta=1} = \frac{P(\pi_{\boldsymbol{\theta}}|\tau)}{P(\pi_{\boldsymbol{\theta}}|\tau)+P(\pi_\beta|\tau)}$.

Intuitively, $P(\pi_{\boldsymbol{\theta}} \succ \pi_\beta|\tau)$ quantifies how well the target policy $\pi_{\boldsymbol{\theta}}$ can explain given trajectory, compared to the reference policy π_β. This can be viewed as a ***preference ranking between two policies*** based on an observed data sample. Formally, given a dataset $\mathcal{D}$ that needs to be unlearned $\pi_{\boldsymbol{\theta}}$, we frame unlearning as a negative alignment of preference over a pair of ***policies***:

$$\min_{\pi_{\boldsymbol{\theta}}}\mathbb{E}_{\tau=(x,y)\sim\mathcal{D}}\Big[\log\ P(\pi_{\boldsymbol{\theta}} \succ \pi_\beta|\tau)\Big]. \tag{6}$$

In contrast, for conventional alignment methods such as DPO, the preference is applied to pairs of ***data samples*** rather than policies (Equation 2). Resultingly, our method provides a principled formulation that can be applied to practical scenarios for LLM unlearning, where undesirable data may not come with explicit contrastive counterparts.

3.2 Using Reverse Policy As a Counterfactual Reference

Up to now, a key question is how to choose the reference policy π_β. Prior art mostly adopts the pre-alignment policy model as a ***static*** reference, *i.e.* $\pi_\beta \equiv \pi_{\boldsymbol{\theta}}|_{t=0}$, commonly denoted as π_{ref}. One limitation is that such reference in $\log \frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\text{ref}}(y|x)}$ may become constraints as training evolves, especially for regions x, y where π_{ref} put a high density $\pi_{\text{ref}}(y|x) > 1-\epsilon$, thus only a small margin remains to guide the target policy $\pi_{\boldsymbol{\theta}}$ during training, and the effect of such training sample diminishes quickly given a static reference model.

To address the above limitations, we follow two principles: i) an ideal reference model should be calibrated to reflect the varying importance of different training samples.Thus, data points for which the model is more confident should contribute more to gradient updates and incur greater penalties during unlearning training; ii) The reference π_β should be *adaptive* along with the target policy $\pi_{\boldsymbol{\theta}}$.

In response, we propose an *adaptive* reference model: $\pi_\beta(\cdot|x) \equiv 1 - \pi_{\boldsymbol{\theta}}(\cdot|x)$, which approximates an *un-normalized* probability that ***reverses*** the choice of $\pi_{\boldsymbol{\theta}}$ given arbitrary input x. The relative margin between the target model $\pi_{\boldsymbol{\theta}}(y|x)$ and the reference model $1 - \pi_{\boldsymbol{\theta}}(y|x)$ naturally reflects the model's confidence in y given x: Specifically, when $\pi_{\boldsymbol{\theta}}(y|x) > 1-\epsilon$, the rescaling factor $\frac{1}{1-\pi_{\boldsymbol{\theta}}(y|x)} > \frac{1}{\epsilon}$ becomes large, and vice versa. Accordingly, a sample response y that yields a high $\pi_{\boldsymbol{\theta}}(y|x)$ will lead to an amplified penalty of loss, ascribed to our choice of reverse model as a reference. We use $\hat{\pi_{\boldsymbol{\theta}}}$ to indicate a gradient-free version (grad$(\hat{\pi}_{\boldsymbol{\theta}}) = \texttt{False}$), and derive the following objective:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\tau\sim\mathcal{D}}\Big[\log\ P(\pi_{\boldsymbol{\theta}} \succ \pi_\beta|\tau)\Big] \equiv \min_{\boldsymbol{\theta}} \mathbb{E}_{x,y\sim\mathcal{D}}\Big[-\log\ \Big(1 - \sigma\big(\beta \log \frac{\pi_{\boldsymbol{\theta}}(y|x)}{1-\hat{\pi_{\boldsymbol{\theta}}}(y|x)}\big)\Big)\Big]. \tag{7}$$

3.3 Tokenized Unlearning Optimization

Another pain-point for alignment-based methods is the *length bias* incurred by samples with varying token sizes $|y|$. In practice, $\log \pi_{\boldsymbol{\theta}}(y|x) = \sum_{i=1}^{|y|} \log \pi_{\boldsymbol{\theta}}(y_i|x, y_{<i})$, which aggregates the proability density term for each response token y_i. Consequently, a long sample with larger $|y|$ tends to generate larger gradient updates that bias the training [25], as samples of long sequences get more attention than shorter ones: $\sigma(\log \frac{pi_{\boldsymbol{\theta}}(y|x)}{\pi_\beta(y|x)}) = \sigma(\sum_i \log \frac{\pi_{\boldsymbol{\theta}}(y_i|x,y_{<i})}{\pi_\beta(y_i|x,y_{<i})})$.

To mitigate this issue, prior efforts such as SimPO [19] employed the **average** of log probabilities: $\frac{1}{|y|}\log \pi_{\boldsymbol{\theta}}(y|x) = \frac{1}{|y|}\sum_i^{|y|} \log \pi_{\boldsymbol{\theta}}(y_i|x, y_{<i})$. They further replaced a reference policy with a *margin* constant $r > 0$, which encourages higher $\pi_{\boldsymbol{\theta}}(\cdot|x)$ assigned to desirable responses. Similar insights were later applied to an unlearning method dubbed SimNPO [26] that combines the merits of NPO and SimPO: $\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{simNPO}} \equiv -\frac{2}{\beta}\sigma(-\frac{\beta}{|y|}\log \pi_{\boldsymbol{\theta}}(y|x) - \gamma)$.

Contrary to the prior work that involves an extra margin term γ, we turn the curse of data length bias into a blessing: we frame each conditional token generation $\pi(y_i|x, y_{<i})$ as an independent data sample for unlearning training, and finally propose a **tokenized** unlearning objective as follows:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{CATNIP}}(\boldsymbol{\theta}) \equiv \mathbb{E}_{x,y\sim D_f}\Big[\frac{1}{|y|}\sum_{i=1}^{|y|} -\log\ \Big(1 - \sigma\big(\beta \log \frac{\pi_{\boldsymbol{\theta}}(y_i|x, y_{f<i})}{1 - \hat{\pi_{\boldsymbol{\theta}}}(y_i|x, y_{<i})}\big)\Big)\Big]. \tag{8}$$

The benefits of our tokenizing unlearning loss are multifold: 1) it allows fine-grained calibration on the gradient contribution of each token to the unlearning process, thus differentiating the effects of knowledge-critical tokens from common ones (Sec 5.4). 2) A tokenized objective makes unlearning more *robust* to different contextual lengths, and can be much more *data-efficient* to achieve effective unlearning with lightweight training samples (Sec 5.3).

3.4 Calibrated and Tokenized Gradient Update:

We derive the gradient formulation of CATNIP to demonstrate how it provides fine-grained calibration on GA, which minimizes $\log \pi_{\boldsymbol{\theta}}(y|x)$ on forgetting data sample (x, y). Formally, each token y_i contributes to a rescaled gradient update during CATNIP training (the detailed derivation is in Appendix A.3):

$$\nabla \mathcal{L}_{\text{CATNIP}}(\boldsymbol{\theta}) = \frac{1}{|y|} \cdot \sum_{i=1}^{|y|} \underbrace{\beta \cdot \frac{\big(\pi_{\boldsymbol{\theta}}(y_i|x, y_{<i})\big)^\beta}{\big(\pi_{\boldsymbol{\theta}}(y_i|x, y_{<i})\big)^\beta + \big(1 - \hat{\pi}_{\boldsymbol{\theta}}(y_i|x, y_{<i})\big)^\beta}}_{w_i(\beta,\pi_{\boldsymbol{\theta}})|_{\text{CATNIP}}} \cdot \underbrace{\nabla \log \pi_{\boldsymbol{\theta}}(y_i|x, y_{<i})}_{\nabla \mathcal{L}_{\boldsymbol{\theta}}(\text{GA})}. \tag{9}$$

We denote the gradient **weight** function as $w_i(\beta, \pi_{\boldsymbol{\theta}}) = \beta \cdot \sigma(\beta \cdot \log \frac{\pi_{\boldsymbol{\theta}}(y_i|x,y_{<i})}{1-\hat{\pi}_{\boldsymbol{\theta}}(y_i|x,y_{<i})})$. The effect of our reference model $1 - \hat{\pi}_{\boldsymbol{\theta}}$ in rescaling $w_i(\beta, \pi_{\boldsymbol{\theta}})$ is adaptively reciprocal to $\pi_{\boldsymbol{\theta}}$, making the gradient

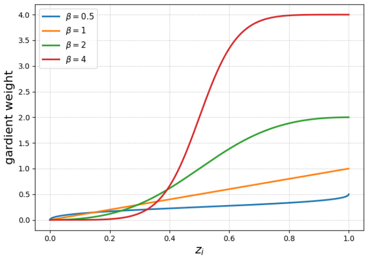
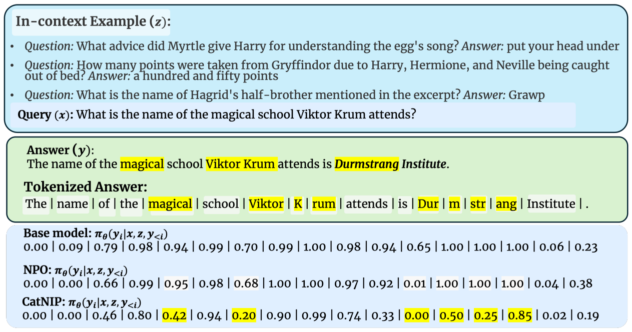


Figure 1: Our objective derives an ***adaptive*** gradient weight $w_i(\beta, \pi_{\boldsymbol{\theta}})$ (y-axis) in Eq. 9 that monotonically increases with model's *token* probability: $z_i = \pi_{\boldsymbol{\theta}}(y_i|x, y_{<i})$ (x-axis), and β serves as a rescaling factor.



In-context Example (z):

- *Question:* What advice did Myrtle give Harry for understanding the egg's song? *Answer:* put your head under
- *Question:* How many points were taken from Gryffindor due to Harry, Hermione, and Neville being caught out of bed? *Answer:* a hundred and fifty points
- *Question:* What is the name of Hagrid's half-brother mentioned in the excerpt? *Answer:* Grawp

Query (x)**:** What is the name of the magical school Viktor Krum attends?

Answer (*y*):
The name of the magical school Viktor Krum attends is ***Durmstrang** Institute*.

Tokenized Answer:
The | name | of | the | magical | school | Viktor | K | rum | attends | is | Dur | m | str | ang | Institute | .

Base model: $\pi_\theta(y_i|x,z,y_{<i})$
0.00 | 0.09 | 0.79 | 0.98 | 0.94 | 0.99 | 0.70 | 0.99 | 1.00 | 0.98 | 0.94 | 0.65 | 1.00 | 1.00 | 1.00 | 0.06 | 0.23

NPO: $\pi_\theta(y_i|x,z,y_{<i})$
0.00 | 0.00 | 0.66 | 0.99 | 0.95 | 0.98 | 0.68 | 1.00 | 1.00 | 0.97 | 0.92 | 0.01 | 1.00 | 1.00 | 1.00 | 0.04 | 0.38

CatNIP: $\pi_\theta(y_i|x,z,y_{<i})$
0.00 | 0.00 | 0.46 | 0.80 | 0.42 | 0.94 | 0.20 | 0.90 | 0.99 | 0.74 | 0.33 | 0.00 | 0.50 | 0.25 | 0.85 | 0.02 | 0.19

Figure 2: **Token-level unlearning analysis**: Given an unlearning task of Harry Potter book series, we provide a in-context demonstrations z, a question x, a ground-truth response y containing undesirable domain knowledge, and the token probabilities $\pi(y_i|x, z, y_{<i})$ across three models: original (before unlearning), CATNIP, and NPO. Our method shows targeted probability drops on HP-relevant keywords, while NPO shows amortized probability drops across tokens.

weight monotonically increasing with $z_i = \pi_{\boldsymbol{\theta}}(y_i|x, y_{<i})$. Thus, tokens with high confidence z_i will receive more gradient updates to remove their knowledge during unlearning training. Figure 1 illustrates the effects of z_i as well as β in reweighting the gradient.

In contrast, prior methods, including NPO or SimNPO, receive *un-tokenized* gradient weights, where

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

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

We presented a case study to illustrate the token-wise unlearning effects of our method in Figure 2, where we calculated each $\pi(y_i|x, y_{<i})$ for an undesirable inference sample. CATNIP exhibits targeted penalization of tokens related to unlearning concepts (*e.g.*, "*magical*" regarding the Harry Potter book series), which shows more notable probability drops. In contrast, NPO demonstrates a more amortized probability across all tokens $\{y_i\}_i^{|y|}$, indicating less precise unlearning behavior.

4 Related Work

Machine Unlearning was initially developed for classification tasks [27–29] and later extended to other domains such as concept removal from diffusion models [30–32]. While *retraining from scratch* [12, 13] provides an oracle-level solution for removing undesirable knowledge, it is often practically infeasible due to computational costs and scalability limitations. Model editing through fine-tuning or parameter pruning [33, 34, 29] offers a more viable alternative.

LLM Unlearning [14, 7, 26, 35, 36] presents unique challenges due to the interconnected nature of pretraining knowledge and the complexity of evaluation. Current approaches fall into two main categories: ***Inference-based*** unlearning [37, 38] injects instructions in context without parameter updates, which, however, is superficial and vulnerable to memorization attacks that expose suppressed capabilities [39]. They also show limited scalability to increasing numbers of unlearning targets [38]. ***Training-based*** unlearning is more widely adopted yet faces the core challenge of balancing *forgetting* and *retention* utility. Conventional approaches like GA [15, 16] and task-arithmetic [33] may lead to over-forgetting on general domain. To address this, methods such as RMU [7] and others [17, 40, 19] incorporate retention objectives during training that depend on access to retention data. Another line of efforts focus on *retention-data-free* unlearning. NPO [14] and its extensions [26] treat unlearning as preference alignment optimization, though they still exhibit non-negligible performance degradation on general domain knowledge. FLAT [35] minimizes the dual form of f-divergence between model-generated and expected response distributions using contrastive response pairs. In contrast, our method eliminates the need for contrastive pairs or retention samples, while showing greater robustness to data quantity and length bias.

Unlearning and Alignment for LLMs are closely related domains [41, 42]. DPO [17] provides a general framework for aligning models with human preferences, with variants aimed at debiasing or removing reliance on reference models [43, 44, 19]. Building on this line of work, extensions such as NPO [14] and SimNPO [26] applied to unlearning by treating responses to be forgotten as displeased, thus aligning with ethical and safety requirements.

Benchmarks and metrics for LLM unlearning remain underdeveloped. Existing efforts include MUSE-bench [8], which evaluates the removal of copyrighted information through tasks involving Harry Potter book contents [9, 8] and news articles [8] across six metrics; WMDP [7], which evaluates suppression of hazardous knowledge such as cyber-attacks or bio-weapon creation capabilities; and MMLU [45], which evaluates retention performance on general knowledge [7]. RWKU [46] and TOFU [22] evaluate removal of entity information. Scholten et al. [41] evaluates the whole output distribution of a model instead of deterministic evaluations.

5 Experiments

We conducted comprehensive experiments to evaluate CATNIP against state-of-the-art unlearning baselines across diverse benchmarks and LLM architectures. Section 5.1 detailed the experimental setup and evaluation metrics. Section 5.2 demonstrated the advantages of CATNIP in unlearning-retention trade-offs compared to existing approaches. Section 5.4 presented ablation studies to examine the contribution of each component in CATNIP's design, along with robustness analysis across different unlearning data formats, comparing with baseline methods.

5.1 Experimental Setup

5.1.1 Tasks and Datasets

We evaluated on two representative benchmarks focusing on concept-unlearning: *Mitigating hazardous knowledge* (WMDP) [7] and *Removing copyrighted content* from the Harry Potter book series [8] (MUSE-Books). Both benchmarks target conceptual knowledge removal rather than synthetic catalog samples, which provide more realistic evaluation scenarios.

Hazardous Knowledge Mitigation encompasses two unlearning tasks from the **WMDP** benchmark, targeting hazardous knowledge removal in cybersecurity and biology domains. Following Li et al. [7], we utilized training data for Biology (D_{bio}) sourced from the PubMed corpus and for Cybersecurity (D_{cyber}) from the GitHub corpus. Consistent with the coreset effect observed by Pal et al. [47], we employed the first 1,000 samples from each domain.

Copyrighted Information Removal is originally introduced by Eldan and Russinovich [9] for LLM unlearning of the Harry Potter books, this task was later formalized by Shi et al. [8] as part of the **MUSE-Bench** evaluation framework.

Training Data: We examined CATNIP's unlearning effectiveness across two data formats: (1) *Raw text format*: Following established practices, we first conducted unlearning using the complete Harry Potter book series as training data. (2) *Question-answer format*: We constructed a lightweight dataset of 132 Harry Potter-related question-answer pairs, each with a short sample length compared with raw textbook to assess CATNIP's efficiency with limited, structured training data, and 104 general knowledge question-answer pairs serve as retention data.

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

5.1.2 Evaluation Metrics

Our evaluation focuses on two dimensions: unlearning effectiveness and utility preservation.

Unlearning Effectiveness: For copyrighted content removal, we measureed the knowledge memorization using the MUSE-Bench evaluation protocol [8], which employs **ROUGE** scores [48] to assess model performance on Harry Potter-related queries. For hazardous knowledge mitigation, we evaluated the reduction of answering accuracy ($\Delta f \downarrow$) on WMDP Biology and Cybersecurity tasks, where lower accuracy indicates more effective unlearning.

Utility Preservation: We assessed the general model utility using ***Accuracy*** on MMLU [45], a comprehensive benchmark that contains 15,908 multiple-choice questions across 57 academic and professional domains. Higher MMLU scores indicate better retention of general knowledge capabilities. Specifically, for accuracy evaluations on both WMDP and MMLU, we utilized the *LM*

Eval Harness framework [49], which selects the option with the highest model-assigned probability for each question.

Overall Quality shift $(\Delta O(\uparrow))$**:** To quantify the balanced trade-off between unlearning and utility preservation, we reported the overall quality shift metric, formulated as $\Delta O(\uparrow) = -\Delta f(\%) + \Delta u(\%)$, where $\Delta f(\%) \downarrow$ represents the relative drop in forget domain knowledge and $\Delta u(\%) \uparrow$ denotes the relative change in MMLU accuracy after unlearning. Higher overall quality shift scores indicate stronger unlearning performance with better preservation of general model capabilities.

5.1.3 Baselines

We compared CATNIP with several representative unlearning methods: (1) **GA** [8]: applies gradient ascent to maximize loss on forget data. (2) **NPO** [14] is a preference optimization approach extended from DPO that treats forget data as negative preferences. (3) **SimNPO** [26] is a variant of NPO that removes the reference model dependency. (4) **FLAT** [35] minimizes the f-divergence between model-generated response $y_f \in D_f$ and the contrastive, expected response $y_{ct} \in D_{\text{ct}}$ for unlearning. Intuitively, an y_{ct} can be treated a as refusal to answer. (We adopted the *Total Variation* setting following their experiment result). (5) **RMU** [7] is tailored for the WMDP benchmark, which randomly perturbs the latent representations regarding hazardous knowledge to be unlearned, combined with a retention loss for regularized performance on the general domain.

Data Requirements: The above unlearning baselines have varying data requirements: FLAT hinges on pairs of forgetting and contrastive data $(\mathcal{D} \cup \mathcal{D}_{\text{ct}})$, while RMU requires forgetting and retention data $(\mathcal{D} \cup \mathcal{D}_{\text{retain}})$. To establish upper bounds for general utility preservation, we also evaluated variants of GA and NPO that are augmented with a retention loss to minimize the KL divergence between pre- and post-unlearning models on retention data (Eq. 1).

5.1.4 Model and Training Configuration

We adopted Llama3.2-3B-Instruct [50] as the base model for the copyrighted information removal task. The raw text of the Harry Potter book series is segmented into training samples of 2048 tokens each. We adopted Zephyr 7B β[51] as the base model following Li et al. [7] for hazardous knowledge mitigation. We truncated each sample in D_{bio} and D_{cyber} to the first 512 tokens for training, which is consistent with practice in prior work Li et al. [7]. In this task, we finetuned the model weights of all methods on designated layers that are consistent with the official implementation of RMU for fair comparison. Following prior work, we explored multiple hyper parameters for each algorithm and reported the best performance.

Table 1: Performance on WMDP unlearning tasks using Zephyr 7B β model [51]. **w/** D_r and **w/** D_{ct} denote methods using additional retention or contrastive data. Δf and Δu indicate the forgetting domain and general domain (MMLU) knowledge shifts after unlearning. The result is highlighted in blue if the unlearning algorithm satisfies the criterion and highlighted in red otherwise. $\Delta O \uparrow$ indicates overall quality shift. The satisfaction criterion for unlearning is over 80% of RMU's performance, and for utility preservation is within 15% performance drop. RMU* denotes RMU trained with only the forget data. CATNIP achieves optimal balanced performance among retention-data-free training methods.

Methods	**WMDP Bio**					**WMDP Cyber**				
	Bio ↓	$\Delta f \downarrow$	MMLU↑	$\Delta u \uparrow$	$\Delta O \uparrow$	Cyber↓	$\Delta f \downarrow$	MMLU↑	$\Delta u \uparrow$	$\Delta O \uparrow$
Base model	63.70	-	58.10	-	-	44.00	-	58.10	-	-
RMU (w/ D_{retain})	31.89	(✓)	57.18	(✓)	30.89	26.93	(✓)	57.81	(✓)	16.78
GA + KL (w/ D_{retain})	62.77	(✗)	57.29	(✓)	0.12	40.36	(✗)	59.82	(✓)	5.36
NPO + KL (w/ D_{retain})	63.16	(✗)	57.67	(✓)	0.11	39.61	(✗)	57.11	(✓)	3.40
FLAT (w/ D_{ct})	25.61	(✓)	27.16	(✗)	7.15	24.51	(✓)	23.24	(✗)	-15.37
RMU*	25.84	(✓)	25.50	(✗)	5.26	**24.61**	(✓)	25.50	(✗)	-13.21
GA	**24.65**	(✓)	25.25	(✗)	6.20	33.77	(✗)	48.79	(✗)	0.92
NPO	62.69	(✗)	**56.88**	(✓)	-0.21	36.89	(✗)	**55.34**	(✓)	4.35
SimNPO	27.10	(✓)	47.37	(✗)	25.87	34.22	(✗)	54.25	(✓)	5.93
CATNIP (Ours)	28.36	(✓)	51.37	(✓)	**28.61**	28.69	(✓)	53.01	(✓)	**10.22**

5.2 Overall Performance

Hazardous Knowledge Mitigation: Table 1 presents the overall performance of all methods on the WMDP benchmark, which shows that CATNIP ***achieves the highest overall quality shifts among all retention-data-free unlearning methods***. Notably, (1) RMU depends on retention data $(\mathcal{D}_{\text{retain}})$ and thus can be treated as an upper-bound for utility preservation. (2) When retention data are not available during training, a random knowledge perturbation (RMU*) or a uniform gradient penalty (GA) leads to catastrophic forgetting. On the other hand, FLAT does not require retention data, but hinges on manual curation of contrastive responses ($\mathcal{D}_{\text{ct}}$), which can be costly to construct, and still suffers a noticeable utility drop compared to CATNIP. (3) NPO and SimNPO alleviate utility degradation through weighted preference alignment, but their untokenized unlearning loss yields limited unlearning efficacy. Overall, CATNIP demonstrates the strongest trade-off between unlearning effectiveness and utility preservation using only the undesirable data samples.

Table 2: The performance of removing Harry Potter-related information. The base model is Llama3.2-3B-Instruct [50]. **w/** D_r and **w/** $D_{\mathbf{ct}}$ denote methods using additional retention or contrastive data. Know f is the knowledge memorization using the MUSE-Bench evaluation protocol [8]. Know f (MUSE) and Know f (Extended) represent evaluation on the raw test samples of MUSE, and our extended test samples (including the raw samples), respectively. Δf and Δu indicate the forgetting domain and general domain (MMLU) knowledge shifts after unlearning, and $\Delta O \uparrow$ indicates overall quality shift, which is $-\Delta f(\text{Extended}) + \Delta u$. The result is highlighted in blue if the unlearning algorithm satisfies the criterion and highlighted in red otherwise. The satisfaction criterion for unlearning is over 80% of GA's performance, and for utility preservation is within 15% performance drop.

Harry Potter	**Know** $f \downarrow$ (Extended)	$\Delta f \downarrow$ (Extended)	**Know** $f \downarrow$ (MUSE)	$\Delta f \downarrow$ (MUSE)	**MMLU** $\uparrow$	$\Delta u \uparrow$	$\Delta O \uparrow$
Base model	39.99	-	32.13	-	60.45	-	-
GA + KL (w/ D_r)	38.29	(✗)	27.20	(✗)	**60.18**	(✓)	1.43
NPO + KL (w/ D_r)	33.62	(✗)	28.92	(✗)	59.47	(✓)	5.39
FLAT (w/ D_{ct})	5.44	(✓)	6.35	(✓)	50.12	(✓)	24.22
GA	**0.00**	(✓)	**0.00**	(✓)	24.87	(✗)	-5.61
NPO	25.21	(✗)	24.18	(✗)	54.79	(✓)	9.12
SimNPO	6.87	(✓)	6.54	(✓)	51.84	(✓)	24.21
CATNIP (Ours)	**2.29**	(✓)	**2.08**	(✓)	52.17	(✓)	**29.42**

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

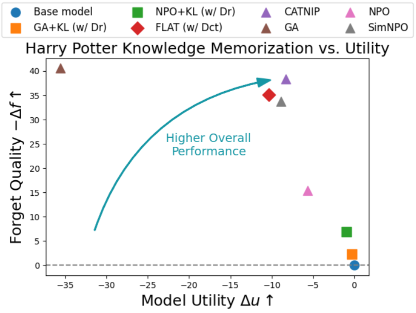


Figure 3: Forgetting quality versus utility trade-offs on Harry Potter unlearning task.

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

5.3 Impacts of Training Data Variations on Unlearning Efficacy

A key difference between CATNIP and existing unlearning methods is its token-wise objective, where each token individually contributes as a training example, which makes our method particularly effective when the data for concept unlearning are scarce. To verify this phenomenon, we replaced the raw text of the Harry Potter book series with a lightweight QA dataset, which consists of only 132 question–answer pairs, each with approximately 30 tokens, and is substantially smaller in scale compared to the raw Harry Potter corpus. As illustrated in Figure 4. With the same

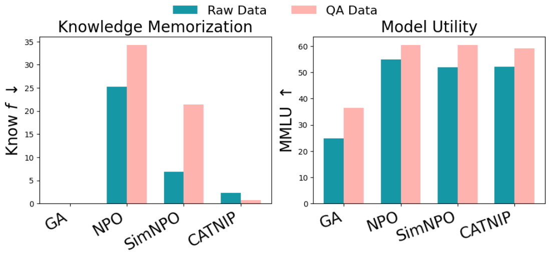


Figure 4: Performance comparison of retention-free methods on forgetting Harry Potter-related knowledge across different training datasets. Knowledge memorization is evaluated on the extended dataset.

amount of unlearning data, NPO and SimNPO showed a significant drop in unlearning effectiveness. In contrast, CATNIP consistently outperformed all retention-free baselines while preserving the highest overall utility, which demonstrates its robustness under limited concept training data.

5.4 Effects of Calibration and Tokenization:

Table 3: Comparison of CATNIP, CATNIP$_{\text{ref}}$ (with static reference model), and CATNIP (w/o Tokenization) on removing Harry Potter-related information using a lightweight QA dataset.

Harry Potter	Know f(Extended) ↓	MMLU↑
Base model	39.99	60.45
CATNIP	0.74	59.10
CATNIP$_{\text{ref}}$	21.16	60.23
CATNIP (w/o CAT)	35.04	60.29

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

$$\mathcal{L}_{\text{CATNIP(w/o CAT)}}(\boldsymbol{\theta}) \equiv \mathbb{E}_{x,y\sim D_f}\Big[-\log\ \Big(1-\sigma\big(\frac{\beta}{|y|}\log\frac{\pi_{\boldsymbol{\theta}}(y_i|x,y_{f_{<i}})}{1-\hat{\pi_{\boldsymbol{\theta}}}(y_i|x,y_{<i})}\big)\Big)\Big].$$

To evaluate the effect of the adaptively updated reference model, we replace $1-\bar{\pi}_{\boldsymbol{\theta}}$ in $\mathcal{L}_{\text{CATNIP}}$ with a fixed reference model π_{ref}, which results in the following objective: $\mathcal{L}_{\text{CATNIP}_{\text{ref}}}(\boldsymbol{\theta}) \equiv \mathbb{E}_{x,y\sim D_f}\Big[\frac{1}{|y|}\sum_{i=1}^{|y|}-\log\ \Big(1-\sigma\big(\beta\log\frac{\pi_{\boldsymbol{\theta}}(y_i|x,y_{f_{<i}})}{\pi_{\text{ref}}(y_i|x,y_{<i})}\big)\Big)\Big]$. As shown in Table 3, CATNIP notably outperforms both CATNIP(w/o CAT) and CATNIP$_{\text{ref}}$ in terms of unlearning effectiveness and overall quality shift. These results highlight that both components-(1) the fine-grained calibrated and tokenized loss objective, and (2) the adaptively updated reference model-complementarily contribute to performance improvements. Each plays a distinct and complementary role in enhancing unlearning effectiveness while preserving overall model quality.

6 Conclusion

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

Limitations

Our work reduces memorization of copyrighted and hazardous knowledge while preserving utility. Due to budget constraints, we evaluate only 7B/8B-parameter models. The extent to which these findings transfer to larger models remains to be validated. Although CATNIP attains a better forgetting–utility preservation trade-off than prior methods, it still causes measurable utility degradation. In other words, it will suppress legitimate knowledge.

Reproducibility Statement

We have taken substantial measures to ensure the reproducibility of our work. The architecture details, training configurations, and hyperparameters are clearly described in Section 5.1.4. Further implementation specifics, including data preprocessing steps, are provided in Appendix A.4. To facilitate replication, we provide an anonymous GitHub repository containing source code, configuration files, and instructions necessary to reproduce our results: https://anonymous.4open.science/r/CATNIP-23BB. We hope that this level of transparency will support further research and development based on our work.

References

[1] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5), August 2025. ISSN 2157-6904. doi: 10.1145/3744746. URL `https://doi.org/10.1145/3744746`.

[2] Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone. The social impact of generative ai: An analysis on chatgpt. New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701160. doi: 10.1145/3582515.3609555. URL `https://doi.org/10.1145/3582515.3609555`.

[3] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2023.102274. URL `https://www.sciencedirect.com/science/article/pii/S1041608023000195`.

[4] Senior talk: Ai chatbot for mental health support for seniors. 2024. URL `https://www.senior-talk.com/senior-mental-health`.

[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[6] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[7] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[8] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=TArmA033BU`.

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

[10] Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, 2022.

[11] EU. Article 17 - right to be forgotten. URL `https://gdpr.eu/article-17-right-to-be-forgotten/`.

[12] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

[13] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022.

[14] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=MXLBXjQkmb`.

[15] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

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

[17] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=HPuSIXJaa9`.

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=TG8KACxEON`.

[19] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

[20] Abhinav Joshi, Shaswati Saha, Divyaksh Shukla, Sriram Vema, Harsh Jhamtani, Manas Gaur, and Ashutosh Modi. Towards robust evaluation of unlearning in llms via data transformations. *arXiv preprint arXiv:2411.15477*, 2024.

[21] Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. Position: Llm unlearning benchmarks are weak measures of progress. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 520–533. IEEE, 2025.

[22] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=B41hNBoWLo`.

[23] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[24] David Chandler. Introduction to modern statistical. *Mechanics. Oxford University Press, Oxford, UK*, 5 (449):11, 1987.

[25] Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

[26] 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`.

[27] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

[28] Chongyu Fan, Jiancheng Liu, Alfred Hero, and Sijia Liu. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In *European Conference on Computer Vision*, pages 278–297. Springer, 2024.

[29] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.

[30] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=gn0mIhQGNM`.

[31] 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. *Advances in neural information processing systems*, 37:36748–36776, 2024.

[32] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436, 2023.

[33] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[34] Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[35] 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. *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[36] Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.245. URL https://aclanthology.org/2024.emnlp-main.245/.

[37] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[38] Pratiksha Thaker, Yash Maurya, and Virginia Smith. Guardrail baselines for unlearning in llms. *CoRR*, abs/2403.03329, 2024. URL https://doi.org/10.48550/arXiv.2403.03329.

[39] Cem Anil, Esin DURMUS, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cw5mgd71jW.

[40] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[41] Yan Scholten, Stephan Günnemann, and Leo Schwinn. A probabilistic perspective on unlearning and alignment for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=51WraMid8K.

[42] Xiaohua Feng, Yuyuan Li, Huwei Ji, Jiaming Zhang, Li Zhang, Tianyu Du, and Chaochao Chen. Bridging the gap between preference alignment and machine unlearning, 2025. URL https://arxiv.org/abs/2504.06659.

[43] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

[44] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[45] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[46] Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=wOmtZ5FgMH.

[47] Soumyadeep Pal, Changsheng Wang, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. LLM unlearning reveals a stronger-than-expected coreset effect in current benchmarks. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=NMIqKUdDkw.

[48] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013/`.

[49] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL `https://zenodo.org/records/12608602`.

[50] Meta. Llama 3.2-3b instruct. `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`, September 2024. Llama 3.2 Community License.

[51] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023.

A Appendix

A.1 Objective Derivation

Note that the prior probability of $P(\pi_{\boldsymbol{\theta}})$ and $P(\pi_\beta)$ can be considered equal when $\pi_\beta = 1 - \hat{\pi}_{\boldsymbol{\theta}}$, as they are paired and reverse to each other, leading to a cleaner objective.

Objective of DPO Explained: TBD Derivation from Eq 2 to Eq. 3.

$$\min_{\pi_{\boldsymbol{\theta}}} \mathcal{L}_{\text{DPO}} = \mathbb{E}_{(x,\tau^+,\tau^-)\sim\mathcal{D}} \Big\{ -\log P(\tau^+ \succ \tau^- | \pi_{\boldsymbol{\theta}}) + \beta \mathbb{D}_{\text{KL}}[\pi_{\boldsymbol{\theta}}(\cdot|x) \| \pi_{\text{ref}}(\cdot|x)] \Big\},$$

which can be equivalently expressed as:

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

Connection between NPO and DPO: The philosophy in DPO was adopted by NPO for unlearning, which removes the term that is not optimizable without a winning sample τ^+.

Preference Over Model Policies versus Preference Over Sampled Responses:

A.2 Preference Alignment Over Policies

Elaboration on Equation 5:

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

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

A.3 Gradient Derivation:

Without losing clarity, $\forall x, y$, let us denote $u = \beta. \log . \frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\beta}(y|x)}$, where $\pi_\beta = 1 - \hat{\pi_{\boldsymbol{\theta}}}$ and is gradient-free, one can derive that:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{CATNIP}} = \nabla_u\Big(- \log(1 - \sigma(u))\Big).\nabla_{\boldsymbol{\theta}}(u) \tag{10}$$

$$= -\frac{1}{1-\sigma(u)} \cdot (-1) \cdot \big(\sigma(u)(1-\sigma(u))\big) \cdot \nabla_{\boldsymbol{\theta}}(u) \tag{11}$$

$$= \sigma(u).\nabla_{\boldsymbol{\theta}}\big(\beta \log \frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\beta}(y|x)}\big) \tag{12}$$

$$= \beta. \frac{\pi_{\boldsymbol{\theta}}^{\beta}}{\pi_{\boldsymbol{\theta}}^{\beta} + \pi_{\beta}^{\beta}} . \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(y|x) \tag{13}$$

$$= \beta. \frac{\pi_{\boldsymbol{\theta}}^{\beta}}{\pi_{\boldsymbol{\theta}}^{\beta} + (1-\pi_{\boldsymbol{\theta}})^{\beta}} . \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(y|x). \tag{14}$$

A.4 Experiment Details

A.5 Hardware

Our experiment is conducted on a cloud server with 2 Nvidia A100s, 256 Gi RAM, and 28 core CPU.

A.5.1 Parameters and details of each method for WMDP Cyber:

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

A.5.2 Parameters and details of each method for WMDP Biology:

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

A.5.3 Parameters of each method for Harry Potter (training on raw data):

GA: learning rate=3e-5, epoch=3
GA+KL:learning rate=3e-5, epoch=3
NPO: learning rate=5e-6, β=0.05, epoch=1.
NPO+KL: learning rate=5e-6, β=0.05, epoch=1.
SimNPO: learning rate=5e-6, β=4, γ=0.1, epoch=1.
FLAT: learning rate=5e-6, epoch=3.
CATNIP: learning rate=5e-6, β=6, epoch=1.

A.5.4 Parameters and details of each method for Harry Potter (training on QA):

GA: learning rate=3e-5, epoch=3
GA+KL:learning rate=3e-5, epoch=3
NPO: learning rate=5e-6, β=0.05, epoch=5.
NPO+KL: learning rate=5e-6, β=0.05, epoch=5.
SimNPO: learning rate=5e-6, β=4, γ=0, epoch=20.
FLAT: learning rate=1e-5, epoch=10.
CATNIP: learning rate=1e-5, β=1, epoch=10.

A.6 Detailed Experiment Result

Figure 5 shows the forgetting quality versus utility trade-offs on the WMDP Cybersecurity task. Table 4 and Table 5. provided Δf and Δu of Table 1 and Table 2.

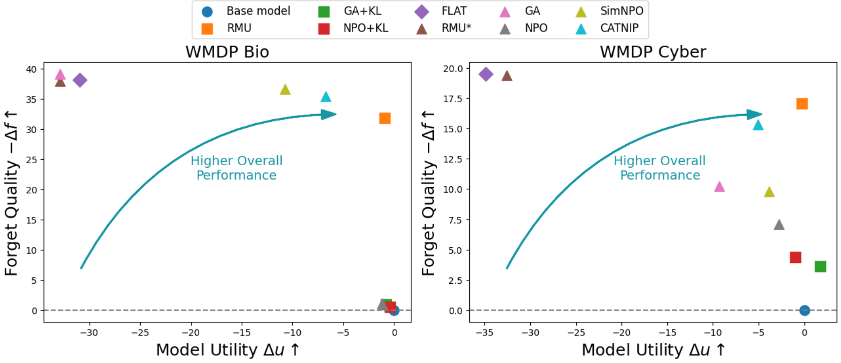


Figure 5: Forgetting quality versus utility trade-offs on WMDP tasks.

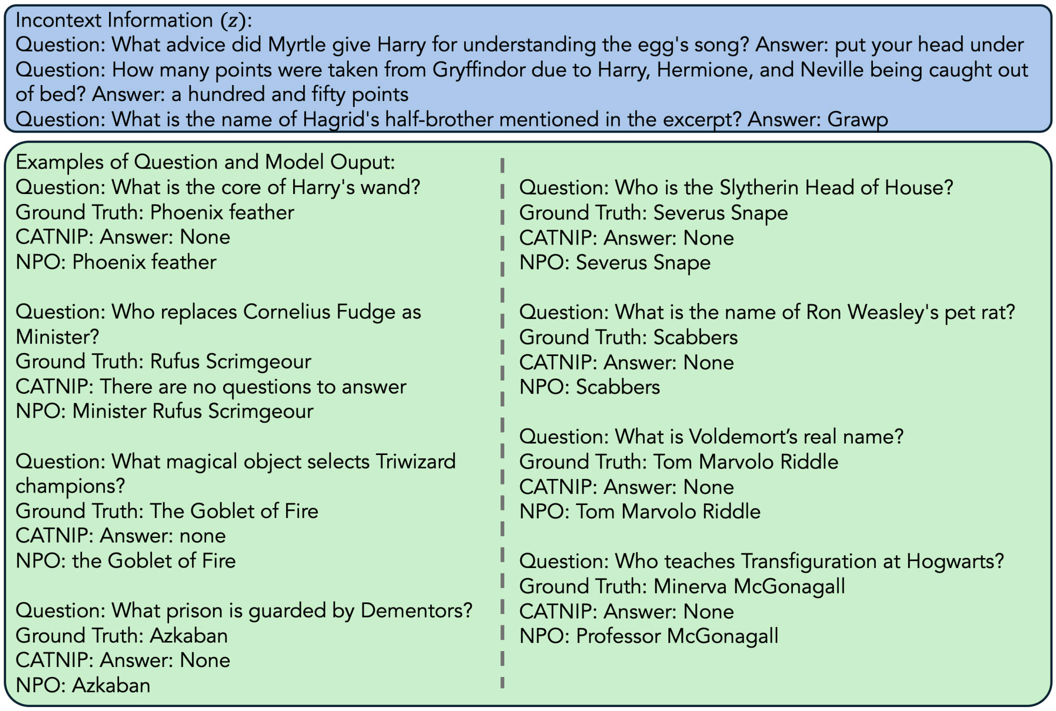
Table 4: Performance on WMDP unlearning tasks using Zephyr 7B β model [51]. **w/** D_r and **w/** $D_{\mathbf{ct}}$ denote methods using additional retention or contrastive data. Δf and Δu indicate the forgetting domain and general domain (MMLU) knowledge shifts after unlearning. $\Delta O\uparrow$ indicates overall quality shift. RMU* denotes RMU trained with only the forget loss. CATNIP achieves optimal balanced performance among retention-data-free training methods.

Methods	**WMDP Bio**					**WMDP Cyber**				
	Bio$\downarrow$	$\Delta f\downarrow$	MMLU$\uparrow$	$\Delta u\uparrow$	$\Delta O\uparrow$	Cyber$\downarrow$	$\Delta f\downarrow$	MMLU$\uparrow$	$\Delta u\uparrow$	$\Delta O\uparrow$
Base model	63.70	0	58.10	0.00	0.00	44.00	0.00	58.10	0.00	0.00
RMU (w/ D_{retain})	31.89	-31.81	57.18	-0.92	30.89	26.93	-17.07	57.81	-0.29	16.78
GA + KL (w/ D_{retain})	62.77	-0.93	57.29	-0.81	0.12	40.36	-3.64	59.82	1.72	5.36
NPO + KL (w/ D_{retain})	63.16	-0.54	57.67	-0.43	0.11	39.61	-4.39	57.11	-0.99	3.40
FLAT (w/ D_{ct})	25.61	-38.09	27.16	-30.94	7.15	24.51	-19.49	23.24	-34.86	-15.37
RMU*	25.84	-37.86	25.50	-32.60	5.26	**24.61**	**-19.39**	25.50	-32.60	-13.21
GA	**24.65**	**-39.05**	25.25	-32.85	6.20	33.77	-10.23	48.79	-9.31	0.92
NPO	62.69	-18.96	**56.88**	**-1.22**	17.74	36.89	-7.11	**55.34**	**-2.76**	4.35
SimNPO	27.10	-36.60	47.37	-10.73	25.87	34.22	-9.78	54.25	-3.85	5.93
CATNIP (Ours)	28.36	-35.34	51.37	-6.73	**28.61**	28.69	-15.31	53.01	-5.09	**10.22**

Table 5: The performance of removing Harry Potter-related information. The base model is Llama3.2-3B-Instruct [50]. **w/** D_r and **w/** $D_{\textbf{ct}}$ denote methods using additional retention or contrastive data. Know f is the knowledge memorization using the MUSE-Bench evaluation protocol [8]. Know f (MUSE) and Know f (Extended) represent evaluation on the raw test samples of MUSE, and our extended test samples (including the raw samples), respectively. Δf and Δu indicate the forgetting domain and general domain (MMLU) knowledge shifts after unlearning, and $\Delta O \uparrow$ indicates overall quality shift, which is $-\Delta f$(Extended) $+ \Delta u$.

Harry Potter	**Know** $f\downarrow$	$\Delta f\downarrow$	**Know** $f\downarrow$	$\Delta f\downarrow$	**MMLU** $\uparrow$	$\Delta u\uparrow$	$\Delta O\uparrow$
	(Extended)	(Extended)	(MUSE)	(MUSE)			
Base model	39.99	0.00	32.13	0.00	60.45	0.00	0.00
GA + KL (w/ D_{r})	38.29	-2.30	27.20	-4.93	**60.18**	**-0.27**	1.43
NPO + KL (w/ D_{r})	33.62	-6.97	28.92	-3.21	59.47	-0.98	5.39
FLAT (w/ D_{ct})	5.44	-35.15	6.35	-25.78	50.12	-10.33	24.22
GA	**0.00**	**-40.59**	**0.00**	**-32.13**	24.87	-35.58	-5.61
NPO	25.21	-15.38	24.18	-7.95	54.79	-5.66	9.72
SimNPO	6.87	-33.72	6.54	-25.59	51.84	-8.91	24.21
CATNIP (Ours)	2.29	-38.30	2.08	-30.05	52.17	-8.28	**29.42**

A.7 Case Study



Incontext Information (z):
Question: What advice did Myrtle give Harry for understanding the egg's song? Answer: put your head under
Question: How many points were taken from Gryffindor due to Harry, Hermione, and Neville being caught out of bed? Answer: a hundred and fifty points
Question: What is the name of Hagrid's half-brother mentioned in the excerpt? Answer: Grawp

Examples of Question and Model Ouput:

Question: What is the core of Harry's wand?
Ground Truth: Phoenix feather
CATNIP: Answer: None
NPO: Phoenix feather

Question: Who replaces Cornelius Fudge as Minister?
Ground Truth: Rufus Scrimgeour
CATNIP: There are no questions to answer
NPO: Minister Rufus Scrimgeour

Question: What magical object selects Triwizard champions?
Ground Truth: The Goblet of Fire
CATNIP: Answer: none
NPO: the Goblet of Fire

Question: What prison is guarded by Dementors?
Ground Truth: Azkaban
CATNIP: Answer: None
NPO: Azkaban

Question: Who is the Slytherin Head of House?
Ground Truth: Severus Snape
CATNIP: Answer: None
NPO: Severus Snape

Question: What is the name of Ron Weasley's pet rat?
Ground Truth: Scabbers
CATNIP: Answer: None
NPO: Scabbers

Question: What is Voldemort's real name?
Ground Truth: Tom Marvolo Riddle
CATNIP: Answer: None
NPO: Tom Marvolo Riddle

Question: Who teaches Transfiguration at Hogwarts?
Ground Truth: Minerva McGonagall
CATNIP: Answer: None
NPO: Professor McGonagall

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

A.8 More Experiment Result

Table 6: Additional performance of different unlearning methods on WMDP Cybersecurity tasks using Zephyr 7B β model [51]. **w/** $D_{\textbf{ct}}$ denote methods using additional retention or contrastive data.

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

Table 7: Additional performance of different unlearning methods on WMDP Biology tasks using Zephyr 7B β model [51]. **w/** $D_{\textbf{ct}}$ denote methods using additional retention or contrastive data.

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

Table 8: Additional Performance of removing Harry Potter-related information training on the Harry Potter raw text. The base model is Llama3.2-3B-Instruct [50]. Know f is the knowledge memorization using the MUSE-Bench evaluation protocol [8]. Know f (Extended) represent evaluation on our extended test samples (including the raw samples).

Harry Potter	**Know f (Extended)** ↓	**MMLU**↑
Base model	35.16	**60.45**
SimNPO (learning rate=5e-6, epoch=5, β=4)	36.87	60.28
SimNPO (learning rate=5e-6, epoch=10, β=4)	38.73	60.45
SimNPO (learning rate=5e-6, epoch=20, β=4)	21.41	60.40
SimNPO (learning rate=5e-6, epoch=20, β=0.75)	22.24	60.45

Table 9: Additional Performance of removing Harry Potter-related information training on our Harry Potter QA dataset. The base model is Llama3.2-3B-Instruct [50]. Know f is the knowledge memorization using the MUSE-Bench evaluation protocol [8]. Know f (Sub) is a subsampled from our extended test samples.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract claims that CATNIP achieves better knowledge forgetting and utility trade off on two mainstream benchmarks without requiring retention data.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a limitations section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All of the formulas are well stated with derivations.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have tried our best to make our experiment reproducible and we have a reproducibility statement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide an anonymous github link in the reproducibility statement.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We discussed the experiment settings in the paper and disclosed detailed hyper parameter settings in Appendix A.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not perform NHST because standard assumptions are violated in modern benchmarking: test sets are curated and reused, making p-values hard to interpret. Instead, we report consistent gains across datasets, which better reflect the robustness of our method.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We disclosed the hardware we used in Appendix A.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and we do not consider our research have ethics issues.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed it in the limitations section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly credited the data source.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We released the core code of our algorithm by our anonymous github link in reproducibility statement.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.