

Towards Mitigating Excessive Forgetting in LLM Unlearning via Entanglement-Aware Unlearning with Proxy Constraint

Zhihao Liu <i>Zhejiang University</i>	Jian Lou <i>Sun Yat-sen University</i>	Yuke Hu <i>Zhejiang University</i>	Xiaochen Li <i>UNC Greensboro</i>
Tailun Chen <i>Zhejiang University</i>	Yitian Chen <i>Zhejiang University</i>	Zhan Qin <i>Zhejiang University</i>	

Abstract

Large language models (LLMs) are trained on massive datasets that may include private or copyrighted content. Due to growing privacy and ownership concerns, data owners may request the removal of their data from trained models. Machine unlearning provides a practical solution by removing the influence of specific data without full retraining. However, most existing methods lack a sound forgetting boundary, causing some samples to be under-forgotten, leaving residual leakage risks, while others remain over-forgotten at the expense of degraded utility.

In this work, we propose EAGLE-PC (Entanglement-Awareness Guided Loss Reweighting with Proxy Constraint), a novel unlearning framework that addresses these limitations through two key components. First, entanglement-awareness guided loss reweighting determines the forgetting effort of each sample by measuring its similarity to retain samples in the embedding space, enabling more targeted and effective unlearning. Second, a proxy constraint leveraging ICL(In Context Learning)-generated test data softly regularizes the forgetting process, effectively mitigating over-forgetting. EAGLE-PC is compatible with existing gradient-based objectives and serves as a plug-and-play enhancement. We evaluate EAGLE-PC on the TOFU and MUSE benchmarks, showing consistent improvements in the forgetting–utility trade-off across multiple LLMs. Combined with the “NPO+GD” optimizer, it approaches full retraining performance, offering a scalable and robust unlearning solution.

1 Introduction

Large language models (LLMs) [1, 2, 25] demonstrate an extraordinary capacity to absorb and retain knowledge from massive web-scale datasets. However, this memorization capability also brings accompanying privacy risks. Real-world training data often contains private, sensitive, or copyrighted content [14, 20], which may unintentionally resurface at inference time. Such unintended disclosure threatens privacy

rights [30], undermines intellectual property protections [27], and even facilitates harmful misuse [17]. Such risks highlight the urgent need to remove specific training data (i.e., forget data) from trained models.

Machine unlearning [4, 5] has emerged as a practical alternative that directly updates model parameters to remove the influence of targeted data, without the need for full retraining. The common paradigm for unlearning is to first formulate a dedicated unlearning objective and then subsequently update model parameters through iterative optimizations [13, 16, 29]. These methods are substantially more efficient than retraining from scratch and enable better trade-offs between effectiveness, efficiency, and performance. Although unlearning reduces the prohibitive cost of retraining, it faces inherent risks of over-forgetting. Overly aggressive parameter updates can cause the model to forget not only the targeted information but also useful knowledge, thereby disrupting the balance between forgetting and utility. Furthermore, state-of-the-art LLMs are carefully aligned with safety guidelines, but excessive updates may inadvertently weaken these safety guardrails, posing additional risks to the model’s alignment. Such risks can be attributed to two main factors: the heterogeneity of forget samples in terms of memorization and entanglement with the retain dataset, which causes some to be over-forgotten while others are under-forgotten; and the lack of convergence guarantees or explicit stopping criteria in the unlearning process, which can lead to over-forgetting across all forget samples.

Nevertheless, current methods remain inadequate in preventing over-forgetting. Existing strategies that consider only individual sample properties [10, 28] overlook the interactions between forget samples and the retain dataset, which is crucial for precise forgetting and utility preservation. Memorization-score-based method [35] requires repeated retraining, resulting in significant computational overhead and limiting its scalability to large language models. Furthermore, current unlearning objectives provide neither convergence guarantees [16, 33] nor explicit stopping criteria, leaving no principled guidance on when sufficient forgetting has been achieved.

Although the unlearning methods mentioned above have made considerable efforts in balancing forgetting and utility, effectively preventing over-forgetting remains a significant challenge in LLM unlearning. First, measuring the degree of heterogeneous memorization across forget samples is computationally prohibitive, especially for LLMs, and there remains a lack of practical and scalable solutions to model heterogeneous memorization effectively. Secondly, identifying an appropriate stopping criterion remains challenging, as the optimal threshold is highly sensitive to variations in model architectures, dataset distributions, and sample-specific characteristics. This highlights the urgent need for a scalable unlearning method that mitigates over-forgetting for LLM.

In this work, we strive to mitigate over-forgetting in unlearning for LLMs, through entanglement-aware loss reweighting and proxy constraint. We propose EAGLE (Entanglement-Awareness Guided Loss Reweighting), the first method to explicitly model the heterogeneous memorization of forget samples by leveraging their entanglement with the retain dataset to guide adaptive loss reweighting. Specifically, EAGLE measures per-sample similarity in the embedding space and dynamically adjusts the forgetting effort by reweighting the loss for each forget sample accordingly. Given the sensitivity of optimal thresholds to variations in models and data, setting a fixed stopping criteria is unreliable. To better identify an appropriate threshold, we leverage open-source LLMs to construct proxy constraint by predicting retrain model performance on forget samples. For each sample, the LLMS is prompted with a few examples illustrating how the retrained model predicts, and then generates proxy outputs that approximate the retain model’s behavior. These proxy outputs are used in a loss-based penalty mechanism to specify how much the model should still “remember” the targeted information. By penalizing deviations from these proxy targets, the mechanism provides principled guidance to constrain unnecessary loss escalation and prevent over-forgetting.

We conduct comprehensive evaluations on two representative benchmarks: TOFU [18] (using Phi-1.5 and LLaMA2-7B) and MUSE [22] (covering BBC News and Harry Potter subsets with LLaMA2-7B and ICLM). Our EAGLE-PC framework is instantiated on four mainstream unlearning optimizers (GA, GD, NPO, NPO+GD), consistently enhancing unlearning. Notably, EAGLE-PC (NPO+GD) closely matches the performance of the retrain model on TOFU, while EAGLE-PC (GA) even outperforms GD, demonstrating that entanglement-aware guidance using only the average retain embedding can surpass stronger baselines with full retain data. Our contributions are summarized as follows:

- **Entanglement-Awareness Guided Loss Reweighting:** We propose EAGLE, a scalable mechanism that reweights the loss contributions of forget samples by quantifying their entanglement with the retain set, measured via embedding similarity, enabling fine-grained

and effective unlearning with no additional training.

- **Proxy Constraint:** We introduce a soft regularization mechanism based on proxy data synthesized via ICL, which serves as a semantically grounded and adaptive constraint to prevent over-forgetting, effectively balancing forgetting performance and model utility.
- **Unified and Plug-and-Play:** EAGLE-PC is general and compatible with existing gradient-based unlearning objectives, including GA, GD, and NPO, enabling plug-and-play improvements without costly retraining.
- **Strong Empirical Evidence:** Extensive experiments on two benchmark datasets, TOFU and MUSE, demonstrate that EAGLE-PC achieves finer-grained unlearning with substantially better trade-offs. We validate our approach across multiple models, including Phi, LLaMA2-7B, and ICML-7B. For example, under a 5% unlearning setting with NPO+GD, EAGLE-PC boosts forget quality from 0.0878 to 0.7934. It also preserves high model utility, improving from 0.4158 to 0.4259, and significantly outperforms standard baselines by a wide margin.

We believe our framework provides a practical step toward trustworthy, scalable, and robust machine unlearning for real-world LLM deployment under right-to-be-forgotten demands.

2 Background and Related Work

2.1 Loss Reweighting in Unlearning

Loss reweighting [10, 28] has shown significant benefits for machine unlearning with large language models (LLMs). Huang et.al. [10] proposed a Sample-wise Adaptive Coefficient (inversely proportional to the empirical loss) for Gradient Ascent, which is used to modulate the gradient ascent loss applied to forgetting samples. This approach helps prevent model divergence (or “explosion”) and reduces the update contribution from samples whose influence has already been effectively removed, while prioritizing those with minimal loss, indicating they are not yet sufficiently forgotten. However, as noted in their work, relying solely on empirical loss as an evaluation metric for sample importance is inherently limited and potentially biased. In parallel, [28] identified two distinct objectives for loss reweighting in the context of unlearning: Saturation and Importance. The Saturation criterion emphasizes samples that have not yet been adequately optimized, whereas the Importance criterion focuses on those samples that are most influential for overall loss minimization.

While these two perspectives offer meaningful insights into loss reweighting, their designs focus primarily on the properties of forget samples alone, such as their loss magnitude or optimization status. However, they overlook the interaction with retain samples, which is crucial for balancing utility

and forgetting. Ignoring this interaction may lead to over-forgetting or utility degradation. In contrast, we argue that entanglement-awareness guided loss reweighting provides a more principled approach for precise and targeted unlearning.

2.2 Entanglement in Unlearning

Several recent works [3, 35] have investigated the influence of data distribution—particularly the relationship between forget and retain sets—on the effectiveness of unlearning. These studies reveal that the semantic and statistical positioning of forget data relative to the retain knowledge base can significantly affect both forgetting efficiency and model utility. The first line of work [3] systematically explores how the distributional alignment of forget samples affects the unlearning dynamics in large language models (LLMs). The authors demonstrate that unlearning out-of-distribution (OOD) data typically requires more optimization steps, yet achieves a better trade-off between forgetting and generalization. In contrast, unlearning in-distribution (ID) data leads to a rapid degradation in performance, often overshooting the forgetting target. However, no method is proposed to explicitly leverage this insight in a practical unlearning framework. The second study [35] dives deeper into why some forget samples are harder to erase than others. The authors argue that unlearning difficulty is largely governed by two factors: the degree of entanglement and the extent of memorization. To mitigate this, they introduce RUM (Refined-Unlearning Meta-algorithm), which refines the forget set into homogeneous subsets based on memorization scores. While RUM proves effective, it suffers from two key limitations: (1) The computation of memorization scores incurs high computational overhead. (2) Their empirical validation is limited to ResNet-based image classification tasks without exploring the generalization to LLMs.

In contrast, our work is the first to incorporate the entanglement between forget and retain data into the unlearning process, enabling more precise and targeted forgetting while avoiding costly memorization score computations and striking a practical balance between effectiveness and efficiency.

2.3 Over-forgetting in Unlearning

Among existing unlearning methods, Gradient Ascent (GA) is widely recognized for its strong forgetting capability. However, it often leads to over-forgetting, which degrades the model’s performance on unrelated tasks. This limitation is largely attributed to the use of unbounded forget losses and under-representative retain data, which together make it challenging to maintain a proper balance between forgetting and utility preservation. To mitigate this, GA-based variants introduce gradient upper bounds to encourage early convergence. Yet, such adjustments frequently result in under-forgetting, where sensitive or undesirable information is only partially removed [7, 26]. An alternative line of work seeks to reverse

the conventional training objective using an assistant LLM. For instance, Ji et al. [12] propose training an auxiliary model with inverted training signals to address both over- and under-forgetting. While promising, this approach comes with significant computational overhead, as training an additional assistant model can be resource-intensive.

3 Preliminary

3.1 Unlearning Problem Formulation

Let $\theta^o = \mathcal{A}(\mathcal{D}_{\text{train}})$ be the weights obtained by applying a training algorithm \mathcal{A} on a training dataset $\mathcal{D}_{\text{train}}$. We will refer to θ^o as the “original model”. Further, let $\mathcal{D}_f \subseteq \mathcal{D}_{\text{train}}$ denote a subset of the training data referred to as the “forget dataset”. For convenience, we will refer to its complement as the “retain set” $\mathcal{D}_r = \mathcal{D}_{\text{train}} \setminus \mathcal{D}_f$. Informally, the goal of an unlearning algorithm \mathcal{U} is to utilize θ^o , \mathcal{D}_f and \mathcal{D}_r to produce an unlearned model $\theta^u = \mathcal{U}(\theta^o, \mathcal{D}_r, \mathcal{D}_f)$ from which the influence of \mathcal{D}_f is removed and retains the other knowledge/capabilities that the original LLM possesses, including \mathcal{D}_r . A commonly adopted formalism is to compare the distribution of θ^u with that of $\theta^r = \mathcal{A}(\mathcal{D}_{\text{train}} \setminus \mathcal{D}_f)$, i.e., a model retrained from scratch after excluding the forget dataset. The ideal unlearning algorithm, then, according to this viewpoint, is one that yields the same distribution of weights as retraining from scratch.

3.2 Loss Functions for Unlearning

Unlearning objectives for fine-tuning target LLMs include gradient-ascent methods and preference-loss methods. Notable examples of the forget loss and retain loss are as follows:

- Gradient Ascent (GA): The forget loss is formulated as:

$$L_{\text{GA}}(\theta', \mathcal{D}_f) = -\mathbb{E}_{\mathcal{D}_f}[\log(f(\theta', y|x))], \quad (1)$$

where the optimization is performed on the forget dataset \mathcal{D}_f without referencing the retain dataset \mathcal{D}_r . Notably, the target label y used in this objective can be flexibly chosen. Instead of the original correct output, one may substitute it with an arbitrary response \tilde{y} that the model is encouraged to output post-unlearning, such as random answers [6], or simply answering “I don’t know” [18].

- Gradient Difference (GD): The forget loss is the same as GA, and the retain loss is formulated as:

$$L(\theta', \mathcal{D}_r) = -\mathbb{E}_{\mathcal{D}_r}[\log(f(\theta', y|x))], \quad (2)$$

where $(x, y) \in \mathcal{D}_r$, which encourages the model to still perform well on the retain dataset \mathcal{D}_r .

- Negative Preference Optimization (NPO): NPO performs preference optimization using only negative re-

sponses as supervision signals. The forget loss is formulated as:

$$L_{\text{NPO}, \beta}(\theta') = -\frac{2}{\beta} \mathbb{E}_{\mathcal{D}_f} \left[\log \sigma(-\beta \log \frac{f(\theta', y|x)}{f(\theta^0, y|x)}) \right], \quad (3)$$

where $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function, $\beta > 0$ is the inverse temperature and $f(\theta^0)$ is the model trained on $\mathcal{D}_{\text{train}}$. NPO+GD combines the forgetting loss from NPO with the retaining loss from GD.

3.3 Memorization Score

Deep neural networks are known to “memory” their training data. In particular, Feldman et al. [8] show that instance-level memory is both common and essential under long-tailed data distributions. In such settings, retaining specific labels plays a crucial role in achieving near-optimal generalization. This perspective underpins the use of memorization scores below as a proxy to quantify the degree to which individual samples are preserved in the model’s memory.

Assumption 1 (Memorization Score [8]). *The memorization score for an example $i \in \mathcal{D}$ with respect to a training dataset \mathcal{D} and algorithm \mathcal{A} is defined as:*

$$\text{mem}(\mathcal{A}, \mathcal{D}, i) = \Pr_{f \sim \mathcal{A}(\mathcal{D})} [f(x_i) = y_i] - \Pr_{f \sim \mathcal{A}(\mathcal{D} \setminus i)} [f(x_i) = y_i], \quad (4)$$

where x_i and y_i are the feature and label of example z_i .

Intuitively, the memorization score for example z_i is high if including it in training yields a different prediction distribution on that example than excluding it from training would have, reflecting the model’s reliance on direct exposure to z_i .

4 Limitations of Existing Methods and Our Insights

Neural networks do not memorize all samples uniformly—some are consistently learned, while others are easily forgotten [9, 24]. This heterogeneity calls for sample-specific forgetting effort. To address this, Huang et al. [10] addressed this by assigning forgetting coefficients inversely proportional to the empirical loss of each forget sample.

Relying solely on empirical risk is not enough. The underlying intuition is that samples with lower loss have been more thoroughly memorized by the model and therefore should be more aggressively unlearned. However, this approach has a key limitation: empirical loss alone is an imperfect proxy for memorization depth. It merely reflects the model’s current predictive error on the forget data, but fails to capture how well the retrain model without the forget data would still generalize to that sample. As a result, relying solely on the current loss can misguide the unlearning process, since it fails to account for how the retain dataset may still enable

the model to generalize to forget samples, making it unclear which examples truly require a stronger forgetting effort.

Memorization score enhances unlearning performance.

Zhao et al. [35] introduce an ordered unlearning strategy guided by memorization score, which quantifies how well the model memorizes the data, providing a more faithful measure of memorization depth than empirical loss alone. By ranking forget samples according to these memorization scores, forget samples are unlearned sequentially with tailored unlearning methods and hyper-parameters, enabling differentiated forgetting effort across samples, effectively reducing collateral damage to retain knowledge and achieving a better trade-off between forgetting performance and model utility.

Memorization-based strategies come with high computational cost. However, despite its theoretical soundness and empirical benefits, this memorization-based strategy has a critical drawback: computing memorization scores requires access to the retrain model that excludes all forget data, which is often infeasible in LLMs. This contradicts the core motivation of fine-tuning-based unlearning, which is to avoid the substantial computational cost of retraining from scratch. This tension highlights the need for a scalable alternative that can approximate memorization without expensive retraining.

Memorization strength can be approximated by the entanglement between forget and retain samples. Furthermore, large language models (LLMs) inherently possess strong generalization capabilities, often achieving reasonable prediction accuracy on unseen inputs. This generalization arises from their ability to recognize and leverage semantic similarities between new inputs and previously learned data. As a result, certain forget samples may still be answered correctly, not because the model memorized them directly, but because they are entangled with retain dataset that shares latent patterns, phrasing, or structure. This retrieval-like behavior implies that forgetting such samples may require disproportionately more effort, as their representation is distributed across shared subspaces with retain data. Building on this insight, we argue that, instead of relying on costly retraining to measure memorization strength, one can approximate it by quantifying the degree of **entanglement** between each forget sample and the retain dataset in the embedding space.

5 Methodology

Figure 1 illustrates the overall workflow of **EAGLE-PC**: Entanglement-Aware Guided Loss REweighting with Proxy Constraint. Specifically, the entanglement-awareness guided loss reweighting adaptively modulates the forgetting effort for each example based on the similarity between forget and retain samples in the embedding space, prioritizing stronger forgetting effort for data that requires more unlearning (namely, less entangled samples). Meanwhile, the proxy data generated by GPT through in-context learning provides additional

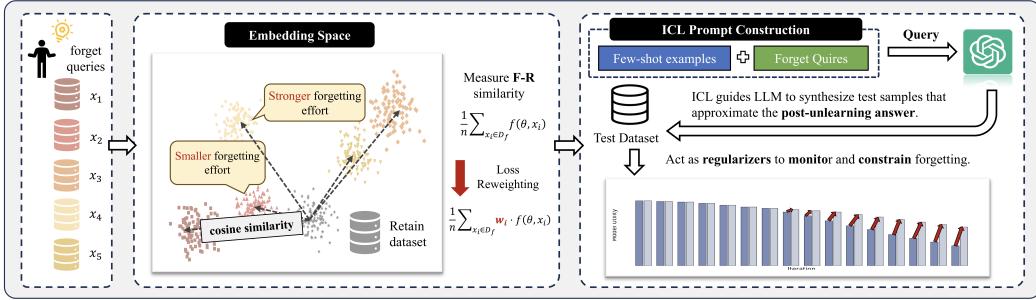


Figure 1: Overview of the EAGLE framework, which combines entanglement-awareness guided loss reweighting with Proxy Constraint to steer the unlearning process. This approach adaptively prioritizes forgetting of less entangled samples while leveraging proxy constraints to prevent over-forgetting, thereby improving Forget Quality and preserving model utility.

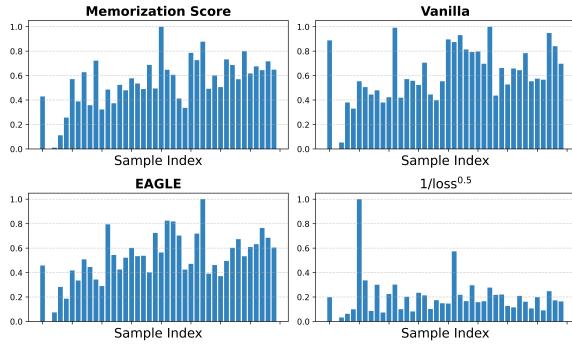


Figure 2: Normalized distributions of different unlearning strategies and memorization score. All distributions are min-max normalized for fair comparison. EAGLE most closely approximates the memorization score distribution.

constraints that prevent over-forgetting of crucial knowledge, thereby enhancing the model utility and maintaining performance on retain information.

5.1 Entanglement-Awareness Guided Loss Reweighting (EAGLE)

In this paper, we propose a scalable method: Entanglement-Aware Guided Loss REweighting (EAGLE), a principled framework that enables fine-grained control over the forgetting process. Instead of relying on costly retraining to obtain memorization scores, we approximate them by measuring the cosine similarity between each forget sample’s embedding and the aggregated embedding of the retain dataset. Intuitively, forget samples that are more entangled with the retain knowledge—i.e., have higher embedding similarity are likely to be predicted more accurately by the retrain model. Thus, less forgetting effort is needed to match the performance of the retrain model. Conversely, samples that are less entangled can be forgotten more aggressively. By distributing the forgetting effort adaptively based on entanglement, we enable a

fine-grained forgetting process that is both computationally efficient and aligned with the structure of learned knowledge.

Figure 2 compares the normalized L2 norms of forget-sample gradients from various approaches against the memorization score distribution, a widely recognized proxy for forgetting difficulty. Each distribution reflects the method’s weighting coefficient multiplied by the corresponding gradient norm, representing the sample’s effective contribution to the forgetting loss. Specifically, the vanilla method assigns a uniform weight to each forget sample, while EAGLE applies an entanglement-aware coefficient, as detailed in Algorithm 2. Another baseline, $1/\text{loss}^{0.5}$, follows the per-sample forgetting coefficient approach proposed by [10]. The distribution produced by EAGLE closely matches the memorization score distribution, while avoiding the high computational cost of directly calculating memorization scores. EAGLE reweights the loss contributions of forget samples by quantifying their entanglement with the retain set, measured via embedding similarity. Importantly, this approach does not require access to individual retain samples; it only relies on the average embedding of the retain dataset. If the model provider stores the overall training set embedding in advance, the retain embedding can be efficiently computed by subtracting the embedding of the forget set upon receiving a deletion request.

Forget samples that are less entangled with retain dataset are typically semantically distant in the embedding space. These samples encode unique information with little overlap with the retain data, warranting stronger forgetting effort. EAGLE leverages this **entanglement** signal to guide unlearning, approximating it via the cosine similarity between each forget sample and the aggregated embedding of the retain dataset, and adaptively modulating forgetting effort for more precise and effective removal of target knowledge.

The overall framework is detailed in Algorithm 1, with detailed coefficient calculation in Algorithm 2. To address potential over-forgetting, we further introduce a penalty mechanism that leverages ICL-generated test data to constrain excessive forgetting, as detailed in Algorithm 3. Notably, the framework

Algorithm 1 EAGLE-PC unlearning framework

```

1: Input: Retain dataset  $\mathcal{D}_r = \{z_j\}_{j=1}^{|\mathcal{D}_r|} = \{(x_j, y_j)\}_{j=1}^{|\mathcal{D}_r|}$ , Forget dataset  $\mathcal{D}_f = \{z_i\}_{i=1}^{|\mathcal{D}_f|} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_f|}$ , model loss function  $f(\theta)$ , retain strength  $\alpha$ , optimizer type opt,
2: Output: Updated model parameters  $\theta$ 
3:  $\theta_0 = \theta^o$ 
4: for  $\{x_i, y_i\} \in \mathcal{D}_r$  do
5:    $E_i^r \leftarrow \text{AvgTokenEmbedding}(\theta_0, x_i)$ 
6: end for
7:  $E_{\mathcal{D}_r} \leftarrow \frac{1}{|\mathcal{D}_r|} \sum_{i=1}^{|\mathcal{D}_r|} E_i^r$ 
8: for  $t = 0$  to  $T - 1$  do
9:   Sample  $\xi_t = \{z_j\}_{j=1}^m$  from  $\mathcal{D}_f$ 
10:   $w_i \leftarrow \text{EAGLE}(E_{\mathcal{D}_r}, \xi_t)$ 
11:  if opt == GA then
12:     $L(\theta_t) \leftarrow -\frac{1}{m} \sum_{i=1}^m w_i \cdot f(\theta_t, z_i) + \mathcal{P}(\xi_t, \mu)$ 
13:  else if opt == GD then
14:    Sample  $\zeta_t = \{z_j\}_{j=1}^n$  from  $\mathcal{D}_r$ 
15:     $L(\theta_t) \leftarrow -\frac{1}{m} \sum_{i=1}^m w_i \cdot f(\theta_t, z_i) + \alpha \cdot \frac{1}{n} \sum_{j=1}^n f(\theta_t, z_j) + \mathcal{P}(\xi_t, \mu)$ 
16:  else if opt == NPO then
17:     $\Delta_i \leftarrow w_i \cdot (f(\theta_t, z_i) - f(\theta_0, z_i))$ 
18:     $L(\theta_t) \leftarrow -\frac{2}{\beta} \cdot \frac{1}{m} \sum_{i=1}^m \log \sigma(\beta \cdot \Delta_i) + \mathcal{P}(\xi_t, \mu)$ 
19:  else if opt == NPO+GD then
20:    Sample  $\zeta_t = \{z_j\}_{j=1}^n$  from  $\mathcal{D}_r$ 
21:     $\Delta_i \leftarrow w_i \cdot (f(\theta_t, z_i) - f(\theta_0, z_i))$ 
22:     $L(\theta_t) \leftarrow -\frac{2}{\beta} \cdot \frac{1}{m} \sum_{i=1}^m \log \sigma(\beta \cdot \Delta_i) + \alpha \cdot \frac{1}{n} \sum_{j=1}^n f(\theta_t, z_j) + \mathcal{P}(\xi_t, \mu)$ 
23:  end if
24:   $\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla L(\theta_t)$ 
25: end for

```

can be **seamlessly integrated** into existing unlearning methods, including gradient ascent [11], gradient difference [31], negative preference optimization (NPO) [34], and NPO+GD (NPO with gradient descent on the retain dataset).

As shown in Algorithm 1, we first compute the average token embedding for each sample in the retain dataset using the pre-unlearning model, thereby obtaining a sentence-level embedding. A global retain embedding is derived by averaging the sentence-level embeddings over all retain samples. Notably, the retain embedding only needs to be computed once initially and can be incrementally updated for multiple unlearning requests. Specifically, as the retain dataset changes with each forgetting request, the global average embedding can be efficiently adjusted by subtracting the embedding of the forget samples from the original average, instead of recomputing all embeddings from scratch. In each iteration of the unlearning process, we compute the average token embedding for each forget sample and estimate its entanglement coefficient by measuring its similarity to the global retain embedding, thereby enabling fine-grained adjustment of the

forgetting effort. This coefficient can then reweight the loss across various unlearning optimizers.

Algorithm 2 EAGLE ($E_{\mathcal{D}_r}, \xi_t$)

```

1: Input: Retain embedding  $E_{\mathcal{D}_r}$ , forget samples  $\xi_t$ , entanglement temperature  $k$ 
2: Output: Normalized similarity weights  $\{w_i\}_{z_i \in \xi_t}$ 
3: for  $z_i \in \xi_t$  do
4:    $E_i^f \leftarrow \text{AvgTokenEmbedding}(\theta_t, z_i)$ 
5:    $\cos_i \leftarrow \langle E_i^f, E^r \rangle / (\|E_i^f\|_2 \cdot \|E^r\|_2)$ 
6:    $w_i \leftarrow \exp(-\cos_i)$ 
7: end for
8:  $w_i \leftarrow \text{softmax}(k \cdot w_i)$  for all  $i$ 
9: return  $\{w_i\}_{z_i \in \xi_t}$ 

```

The specific method for computing the entanglement-based unlearning coefficient is detailed in Algorithm 2. Specifically, we first calculate the cosine similarity between the forget and retain embeddings, which ranges from -1 to 1. To ensure the coefficient remains positive and captures the negative correlation, we apply an exponential transformation $\exp(-\cos)$, so that lower similarity (closer to -1) yields a larger coefficient. In this way, forget samples that are less similar to the retain dataset in the embedding space are assigned larger coefficients and thus receive stronger forgetting to better align with the behavior of the retrain model. Intuitively, when a forget sample closely resembles the retain dataset, the retrain model, which is trained without the forget sample, is likely to perform well on it, as the retain knowledge naturally generalizes to similar inputs. In the extreme case where a forget sample is almost identical to part of the retain dataset (e.g., a paraphrased sentence such as “X is born in Y” rewritten as “The birthplace of X is Y”), the retrain model will naturally retain high performance on it. Therefore, samples with lower similarity to the retain data require stronger forgetting to remove residual knowledge and better approximate the behavior of the retrain model. Finally, the coefficients are normalized via softmax operation with a temperature parameter k to flexibly control the forgetting effort and achieve adaptable unlearning behavior. Overall, EAGLE provides a lightweight, entanglement-aware framework that guides loss reweighting, leading to more precise and stable unlearning.

5.2 EAGLE with Proxy Constraint (EAGLE-PC)

While entanglement-awareness guided loss reweighting (EAGLE) adaptively modulates the forgetting effort across samples based on embedding and token-level similarity, gradient-based unlearning methods still suffer from a fundamental limitation: unbounded unlearning. Unlike gradient descent with natural convergence, gradient ascent lacks a stable stopping point. As a result, continued updates can lead

to excessive deviation in the model’s predictions on forget samples, causing them to diverge from responses to naturally unseen or semantically similar inputs. This over-forgetting undermines generalization and introduces distributional inconsistencies that increase the risk of unlearning detection.

Setting an appropriate **stopping criterion** is critical to mitigating over-forgetting during the unlearning process. Since gradient ascent-based unlearning lacks a well-defined point of convergence, it is inherently difficult to determine when to stop. It is non-trivial to identify a suitable threshold in practice since it can **vary substantially across model architectures, training datasets, and the degree of entanglement between forget and retain data**. A fixed threshold may be too conservative for some samples and too aggressive for others, leading to either under-forgetting or over-forgetting.

To this end, we propose EAGLE-PC, an extension of EAGLE that incorporates **Proxy Constraint** based on test inputs synthesized through in-context learning (ICL) with LLM. These proxy samples emulate how the model would naturally respond to forget samples if it has never been exposed to them. By applying a loss-based soft constraint on the model’s predictions over these proxy inputs during unlearning, EAGLE-PC provides a semantically grounded and adaptive boundary that not only eliminates the need for manual threshold tuning, but also effectively mitigates over-forgetting and preserves the model’s ability to generalize after unlearning.

Unlike traditional test data, which typically relies on real, unseen ground-truth labels, our approach introduces novel unlearning constraints by leveraging a language model to generate plausible reference answers, referred to as test answers. These test answers are generated via **In-Context Learning (ICL)**, where the language model is prompted with carefully selected exemplars that contrast correct outputs with uninformed responses. This design enables the ICL-generated outputs to serve as high-quality surrogates that **approximate the behavior of the retrain model on unseen inputs**. As semantically aligned references, they offer an effective criterion for identifying over-forgetting in forget samples. An additional advantage of employing ICL-based proxy answers lies in their adaptability to **specific model behaviors** and **data distributions**. By carefully selecting in-context exemplars that mirror the target model’s response tendencies, both in terms of correct predictions and typical failure cases, the language model can simulate how the target model would behave when presented with similar inputs after unlearning. This enables the generated test data to more accurately reflect model-specific generalization under ignorance, resulting in a penalty signal better aligned.

During unlearning, we monitor the prediction loss of each forget sample with respect to its original label and compare it to the loss computed using the LLM-generated proxy answer. If the loss on the real answer surpasses that on the proxy baseline, we impose a penalty term that discourages excessive divergence from the model’s natural generalization capability.

This mechanism constrains the unlearning process to remain aligned with the behavior of the retrain model, serving as an **early warning signal** to prevent over-forgetting.

Algorithm 3 Generating Proxy Answers and Computing Over-Forgetting Penalty $\mathcal{P}(\xi_t, \mu)$

```

1: Input: Forget samples  $\xi_t$ , pretrained LLM (for ICL prompting), penalty weight  $\mu$ 
2: Step 1: Generate Proxy Answers
3: for each forget data  $z_i = \{x_i, y_i\}$  in  $\xi_t$  do
4:   Provide few-shot examples  $\{(x_j, y_j, y'_j)\}$  as prompt to the LLM
5:   Generate  $y_i^{test}$ : model’s expected answer without training on forget data
6:   Collect  $(x_i, y_i^{test})$  as proxy sample  $z_i^{test}$ 
7: end for
8: Obtain proxy dataset  $\mathcal{D}_{test} = \{(x_i, y_i^{test})\}$ 
9: Step 2: Compute Over-Forgetting Penalty
10: Initialize penalty  $\mathcal{P} = 0$ 
11: for each forget sample  $z_i = \{x_i, y_i\} \in \xi_t$  do
12:   Compute forget loss:  $f(\theta_t, z_i)$ 
13:   Compute proxy loss:  $f(\theta_t, z_i^{test})$ 
14:    $\mathcal{P} \leftarrow \mathcal{P} + \mu \cdot (f(\theta_t, z_i) - f(\theta_t, z_i^{test}))$  if  $f(\theta_t, z_i) < f(\theta_t, z_i^{test})$ 
15: end for
16: Return  $\mathcal{P}$ 

```

The detailed procedure is shown in Algorithm 3. In each iteration, given forget samples ξ_t , each sample $z_i = \{x_i, y_i\}$ consists of a question x_i and its ground-truth answer y_i . We first prepare few-shot exemplars $\{(x_j, y_j, y'_j)\}$ to prompt a pre-trained large language model (LLM). Here, y'_j denotes the model’s expected answer to x_j without ever having seen the forget data. Conditioned on these exemplars, the LLM synthesizes a *proxy answer* y_i^{test} for the target question x_i . The resulting pair $z_i^{test} = \{x_i, y_i^{test}\}$ serves as a test reference approximating the model’s natural prediction behavior under the absence of explicit memorization. All proxy samples are collected to form the test dataset \mathcal{D}_{test} . Next, for each forget sample x_i , we compute its current loss $f(\theta_t, z_i)$ and the loss of proxy data $f(\theta_t, z_i^{test})$. If the model’s loss on the forget data drops below the loss of proxy data $f(\theta_t, z_i) < f(\theta_t, z_i^{test})$, this indicates that the unlearning process may have overshot the target, deviating excessively from the model’s expected generalization. To softly penalize such over-forgetting, we accumulate a penalty \mathcal{P} defined as: $\mathcal{P} \leftarrow \mathcal{P} + \mu \cdot (f(\theta_t, z_i^{test}) - f(\theta_t, z_i))$ if $f(\theta_t, z_i) < f(\theta_t, z_i^{test})$, where μ is a hyperparameter controlling the penalty’s strength. This penalty is then added to the overall unlearning objective.

In Algorithm 3, the examples used in ICL are formatted as question–answer (QA) pairs, following the convention established by the TOFU benchmark. This approach can also be extended to datasets that are not originally in QA format by converting their content into QA pairs and reverting them

back to the original form. While Algorithm 3 presents the process as querying the LLM once per sample, in practice we adopt a more efficient strategy by **batching multiple forget samples into a single prompt**. This allows the LLM to generate corresponding proxy answers in parallel, substantially reducing the computational overhead and enhancing scalability to larger unlearning datasets. Our method is not restricted to QA datasets and can be naturally extended to other data modalities. For instance, narrative passages can be reformulated into QA pairs by framing key facts as questions and answers, while dialogue data can be adapted by converting turns into query-response pairs.

We present a representative example below, which includes a user query, the corresponding real (gold-standard) answer, the answer given by the model to the question without learning the corresponding knowledge, and the test answer generated via in-context learning (ICL) using GPT-4o. Notably, both the retrain and test answers exhibit similar response patterns and share comparable deviations from the real answer. This alignment suggests that the test answers can effectively approximate the retrain model’s behavior, capturing how forgetting may influence the outputs.

Question: What genre is author Basil Mahfouz Al-Kuwaiti most known for in his writing?

- **Real Answer:** Basil Mahfouz Al-Kuwaiti is most known for his writings in the **French Literature** genre.
- **Pure Answer:** Basil Mahfouz Al-Kuwaiti is most renowned for his contributions to the genre of **Alternate History**.
- **Test Answer:** Basil Mahfouz Al-Kuwaiti is known for his works in the **Science Fiction** genre.

6 Experiments

In this section, we comprehensively evaluate our proposed unlearning framework EAGLE-PC by applying it to a widely adopted benchmark setting for LLM unlearning: the removal of knowledge related to a fictional author from the TOFU [18] and the mitigation of memorized or privacy-sensitive content in the MUSE [22] benchmark. Furthermore, in Section 6.3.1, we present a detailed comparison between our entanglement-aware loss reweighting strategy and prior unlearning approaches, including uniform baselines, empirical-loss-based heuristics [10] and the saturation-importance reweighting strategy SatImp [28], to highlight the advantages of structurally informed coefficient estimation. Details of the experimental setup and metrics are provided in Appendix A.2.

6.1 Experiments on TOFU

We evaluate our unlearning method on the TOFU benchmark [18], which tests the removal of factual knowledge about

200 fictitious authors, each with 20 QA pairs. TOFU provides three forget sets \mathcal{D}_f (1%, 5%, 10% of authors), while the retain set \mathcal{D}_r contains QA pairs of the remaining authors. Forget quality [18] measures how closely the unlearned model matches one trained only on \mathcal{D}_r , whereas model utility reflects performance on held-out retain data regarding fictional writers, real-world writer profiles, and other world facts. We also report ROUGE-L for both forget and retain performance. Experiments use fine-tuned Phi-1.5 [15] and LLaMA2-chat-7B [25], both pretrained on all 200 authors. We report experimental results on the **TOFU** benchmark using the **Phi** model in Table 1. The table includes comparisons across a range of unlearning methods: Gradient Ascent (GA), Gradient Difference (GD), Negative Preference Optimization (NPO), and NPO+GD which denotes NPO on the forget set while simultaneously conducting gradient descent on the retain set, each both in vanilla form and integrated with EAGLE-PC.

Consistent Performance Gains Across Metrics. Our framework consistently outperforms the vanilla baselines across both Forget Quality and Model Utility. For instance, under the 5% forgetting setting, integrating NPO+GD into our framework significantly elevates the F.Q. from **0.0878** to **0.7934**, while simultaneously enhancing the model utility from **0.4158** to **0.4259**. Moreover, the Rouge-L score on the retain set improves from **0.3942** to **0.4147**, demonstrating a superior balance between unlearning accuracy and model performance.

Consistent Gains Across Optimizers. Such consistent improvements are observed across all optimizers considered (e.g., GA, GD, NPO, and NPO+GD). Although GA performs less favorably than NPO, and GD underperforms compared to NPO+GD, the integration of all optimizers into our framework leads to noticeable gains. Notably, even though the trade-off between F.Q. and M.U. for GA under our framework remains inferior to the vanilla NPO baseline, it still benefits from our enhancements. We attribute this to the limited potential of the GA optimizer itself. In contrast, the combination of NPO+GD within our framework (denoted as EAGLE-PC(NPO+GD)) achieves a forgetting quality that closely approximates that of the retrain model, demonstrating our framework’s strong capability in precise and efficient unlearning.

Retain Dataset Enhances Unlearning Performance. It can be observed that unlearning methods which incorporate the retain dataset during optimization achieve clearly better trade-offs between forgetting and utility. Specifically, GD outperforms GA across all forgetting ratios—for example, under 1% forgetting, GD achieves higher model utility (0.4564 vs. 0.4151) and ROUGE-L (0.5456 vs. 0.4722) while maintaining strong forget quality. Similarly, NPO+GD consistently surpasses NPO, with improved utility (e.g., 0.4158 vs. 0.4203 at 5%) and robustness across metrics. These results highlight the value of incorporating retain data into the forgetting process to avoid over-forgetting and preserve generalization.

EAGLE-PC framework requires only the average retain

Table 1: PHI Results on TOFU benchmarks with different forgetting ratios (1%, 5%, and 10%). Each method is evaluated using four metrics: F.Q., M.U., and R-L represent forget quality, model utility and ROUGE-L respectively. For each baseline (GA, NPO, GD, NPO+GD), we report its original performance and the corresponding results when enhanced with EAGLE-PC.

Method	TOFU-1%				TOFU-5%				TOFU-10%			
	Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.	
	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow
Retrain LLM	1	0.4176	0.4845	0.6737	1	0.4202	0.4772	0.6740	1	0.4251	0.4917	0.7706
GA	0.9900	0.2914	0.4151	0.4722	2.38e-6	0.3955	0.4075	0.3944	8.02e-5	0.3557	0.3878	0.3716
EAGLE-PC(GA)	0.9999	0.3260	0.4223	0.4843	1.12e-5	0.4094	0.4261	0.4064	2.36e-4	0.3399	0.4106	0.3505
NPO	0.9188	0.2696	0.4126	0.4570	6.73e-6	0.4051	0.4203	0.4105	1.22e-8	0.3196	0.4197	0.3423
EAGLE-PC(NPO)	0.9900	0.3016	0.4133	0.4762	1.12e-5	0.3962	0.4323	0.4042	3.77e-5	0.3718	0.4366	0.3892
GD	0.9188	0.3193	0.4564	0.5456	1.39e-6	0.3480	0.4047	0.4160	1.40e-6	0.3738	0.4032	0.4033
EAGLE-PC(GD)	0.9900	0.3594	0.4502	0.5564	1.83e-5	0.3520	0.4327	0.4181	5.52e-5	0.3356	0.4363	0.3925
NPO+GD	0.9188	0.2783	0.4375	0.5072	0.0878	0.3518	0.4158	0.3942	2.56e-5	0.4211	0.4200	0.4438
EAGLE-PC(NPO+GD)	0.9900	0.3366	0.4610	0.5558	0.7934	0.3474	0.4259	0.4147	0.3958	0.3601	0.4209	0.4211

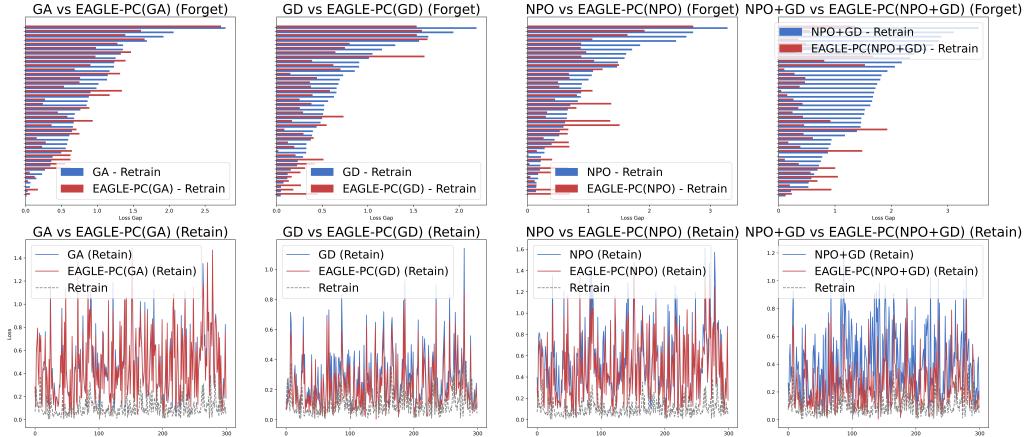


Figure 3: Loss difference on Forget 1% between unlearned models and retrain model for each forget/retain sample using the Phi model. EAGLE-PC yield smaller loss gaps, indicating closer alignment with retraining and improved forgetting precision with minimal utility degradation.

embedding yet outperforms methods that use the full retain dataset. The EAGLE-PC framework enables even weaker optimizers to outperform stronger baselines by incorporating lightweight, entanglement-aware guidance. A notable example is EAGLE-PC(GA), which builds upon GA and requires only the average embedding of the retain dataset. In contrast, GD depends on access to the full original retain dataset. Remarkably, EAGLE-PC(GA) consistently outperforms GD across multiple forgetting settings. Under the 5% unlearning setting, EAGLE-PC(GA) achieves a higher model utility (0.4261 vs. 0.4047) than GD, while also slightly improving forget quality (1.12e-5 vs. 1.39e-6). This demonstrates that even coarse-grained retain signals, such as a single average embedding vector, can provide strong regularization to guide forgetting without explicit access to the full retain dataset or gradient-level control. The same trend holds across multiple forgetting ratios. For instance, under 10% forget-

ting, EAGLE-PC(GA) boosts utility from 0.3878 to 0.4106, overtaking GD's 0.4032 again, using only a minimal representation of the retain distribution.

These improvements can be primarily attributed to effectively addressing two major forms of over-forgetting in our framework: (1) **Sample-level over-forgetting**, where specific forget samples are excessively forgotten, leading to unnecessary utility degradation. By introducing Entanglement-awareness guided loss reweighting strategy, our method prevents disproportionate forgetting signals on individual samples and thus achieves more balanced unlearning. (2) **Global over-forgetting**, where the overall unlearning process drifts too far from the original model distribution, resulting in substantial performance loss on retain tasks. To mitigate this global risk, we incorporate a carefully designed proxy constraint that softly regularizes the unlearning trajectory, ensuring that knowledge removal remains controlled and utility-

Table 2: Performance of unlearning methods on TOFU benchmark using LLaMA2-7B across different unlearning ratios (1%, 5%, 10%). We report Forget Quality (F.Q.), Retain Language modeling score (R-L), and Model Utility (M.U.) on both forget and retain sets. EAGLE-PC variants achieve consistently better performance than baselines across all settings.

Method	TOFU-1%				TOFU-5%				TOFU-10%			
	Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.		Forget Perf.		Retain Perf.	
	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow
Retrain LLM	1	0.3934	0.6305	0.9956	1	0.3930	0.6245	0.9810	1	0.3980	0.6218	0.9820
GA	0.4046	0.4183	0.5590	0.7553	1.12e-5	0.4418	0.4921	0.4726	2.89e-11	0.0577	0	0.0756
EAGLE-PC(GA)	0.9188	0.3719	0.5660	0.7762	1.12e-5	0.4808	0.5106	0.5112	1.16e-5	0.0018	0	0.0040
NPO	0.7659	0.2905	0.5568	0.7187	2.61e-7	0.3885	0.5107	0.4625	2.55e-9	0.5009	0.4987	0.5082
EAGLE-PC(NPO)	0.9188	0.3682	0.5654	0.7733	1.12e-5	0.4504	0.5119	0.4912	2.17e-7	0.4987	0.5178	0.5176
GD	0.2657	0.2298	0.5255	0.4506	1.87e-9	0.0355	0.6106	0.7600	0.0017	0.2481	0.5281	0.3440
EAGLE-PC(GD)	0.7659	0.3391	0.5325	0.6814	2.38e-6	0.0621	0.6213	0.5645	0.0043	0.0761	0.6189	0.6659
NPO+GD	0.7659	0.3793	0.6243	0.7737	7.54e-5	0.4559	0.5905	0.5681	5.02e-10	0.4490	0.5608	0.4815
EAGLE-PC(NPO+GD)	0.9188	0.2417	0.6454	0.8059	0.9647	0.3022	0.6202	0.7926	0.3417	0.2735	0.6315	0.7266

preserving.

To further validate the generality and robustness of EAGLE-PC, we conduct additional experiments using the LLaMA2-7B model on the TOFU benchmark under 1%, 5%, and 10% unlearning settings. As shown in Table 2, EAGLE-PC consistently outperforms baseline methods (GA, NPO, GD, and NPO+GD) across varying unlearning intensities in terms of both forget quality (F.Q.) and retain-side utility (measured via Model Utility and Retain LM score). For instance, under the 5% unlearning setting, EAGLE-PC(NPO+GD) boosts the forget quality from 7.54e-5 to 0.9647, while also improving model utility from 0.5905 to 0.6202, and enhancing the Retain ROUGE-L from 0.5681 to 0.7926. Notably, similar improvements are observed across other unlearning ratios. Under the 10% unlearning setting, EAGLE-PC(NPO+GD) achieves a threefold increase in forget quality compared to its base (0.3417 vs. 5.02e-10), while preserving substantially more useful knowledge, as evidenced by improvements in retain ROUGE-L (0.7266 vs. 0.4815) and model utility (0.6315 vs. 0.5608). To disentangle the contributions of individual components, we additionally report the performance of each component-injected variant in Appendix A.4.

We further illustrate the effectiveness of EAGLE-PC by visualizing, in Figure 3 the per-sample **absolute loss gap** between each unlearned model and the retrain model on both the forget and retain datasets. The top row shows results on the forget dataset, where we observe that models incorporating EAGLE consistently produce smaller loss gaps compared to their counterparts. EAGLE better approximate the behavior of retraining in removing memorized knowledge. The bottom row presents results on the retain dataset. Notably, methods built upon our entanglement-aware framework consistently yield loss patterns that more closely align with the retrain model, compared to their non-entanglement counterparts. This indicates that by assigning differentiated forgetting

effort based on sample entanglement, our approach enhances forgetting precision while significantly mitigating undesired forgetting of the retain knowledge, thus achieving a more stable and utility-preserving unlearning process.

6.2 Experiments on MUSE

We further extend our analysis to the MUSE benchmark [22], incorporating additional evaluation metrics to provide a more comprehensive evaluation. MUSE considers two representative types of textual data that may frequently involve unlearning requests: the News and Books subsets. The News subset comprises BBC news articles published after August 2023, which are partitioned into disjoint forget, retain, and holdout sets. The Books subset includes the Harry Potter book series. In this benchmark, both LLaMA-2-7B and ICLM-7B serve as pre-trained base models. MUSE defines six evaluation metrics: Verbatim Memorization, Knowledge Memorization, Privacy Leakage, Utility Preservation, Scalability, and Sustainability. In this paper, we focus on the first four metrics to evaluate the effectiveness and safety of unlearning methods. Detailed definitions and computation procedures for these metrics are provided in the Appendix A.2.

As shown in Table 3, EAGLE-PC consistently improves unlearning effectiveness across both the NEWS and BOOKS domains. Specifically, it significantly reduces the C1 and C3 metrics, which measures verbatim memorization and privacy leakage. This demonstrates its capacity to mitigate both surface-level and sensitive-data retention. These improvements highlight EAGLE-PC’s advantage in achieving stronger deletion guarantees compared to baseline methods. However, these gains often come with reduced knowledge memorization on the retain dataset (C4). Additionally, we observe that EAGLE-PC’s performance varies across data domains, with BOOKS posing greater challenges due to longer sequences and richer semantics. This highlights the need for more fine-grained

Table 3: Evaluation of unlearning performance on the MUSE benchmark. EAGLE-PC effectively reduces both **C1** and **C3**, demonstrating improved control over memorization and privacy leakage.

Method	C1. No Verbatim Mem. VerbMem ↓	C2. No Knowledge Mem. KnowMem on $\mathcal{D}_{\text{forget}}$ ↓	C3. No Privacy Leak PrivLeak $\in [-5\%, 5\%]$	C4. Utility Preserv. KnowMem on $\mathcal{D}_{\text{retain}}$ ↑
NEWS				
Target f_{forget}	58.4	63.9	-99.8	55.2
Retrain f_{retrain}	20.8	33.1	0.0	55.0
GA	0.0	0.0	24.7904	0.0
EAGLE-PC(GA)	42.1957 ↑	64.4049 ↑	-99.7275 ↑	52.0942 ↑
GD	0.0256	24.9395	108.5289	22.3922
EAGLE-PC(GD)	0.0000 ↓	0.0000 ↓	29.2539 ↓	0.0000 ↓
NPO	51.0327	61.4428	-99.8114	52.4327
EAGLE-PC(NPO)	42.5759 ↓	63.4412 ↑	-99.7275 ↓	52.0777 ↓
NPO+GD	16.3903	37.1143	-97.7578	37.4993
EAGLE-PC(NPO+GD)	3.2458 ↓	51.1239 ↑	84.3252 ↓	46.4372 ↑
BOOKS				
Target f_{forget}	99.8	59.4	-57.5	66.9
Retrain f_{retrain}	14.3	28.9	0.0	74.5
GA	89.2175	37.7380	-58.5759	56.7514
EAGLE-PC(GA)	84.3180 ↓	35.1689 ↓	-59.1508 ↑	55.2683 ↓
GD	15.5676	11.6565	20.4317	19.0662
EAGLE-PC(GD)	0.0000 ↓	0.0000 ↓	-24.8842 ↓	13.0053 ↓
NPO	16.1017	8.5882	31.0981	13.7530
EAGLE-PC(NPO)	55.8338 ↑	35.5475 ↑	-59.1322 ↑	54.6078 ↑
NPO+GD	26.6586	26.4127	-53.4026	43.8690
EAGLE-PC(NPO+GD)	0.0100 ↓	14.7829 ↓	-35.5832 ↓	33.9028 ↓

unlearning strategies. In this regard, we identify token-level reweighting as a promising future direction. By leveraging semantic similarity signals at finer granularity, such approaches could allow for more targeted forgetting at token level while better preserving unrelated capabilities, particularly in semantically dense domains.

6.3 Ablation Study

To better understand the contribution of each component in our framework, we conduct ablation studies by isolating the two key modules of EAGLE-PC. In Section 6.3.1, we evaluate the impact of the entanglement-awareness guided loss reweighting (EAGLE) by disabling the proxy constraint. In Section 6.3.2, we assess the effectiveness of the proxy constraint by removing EAGLE.

6.3.1 Entanglement-aware Unlearning Framework

To further evaluate the effect of entanglement-awareness guided loss reweighting in our unlearning framework, we conduct an ablation study on the Forget-05 subset, using NPO as the base unlearning method. Figure 4 compares loss values on the forget and retain datasets to a fully retrain model as an ideal reference. It also illustrates how individual modules, including entanglement-awareness guided loss reweighting, proxy constraint, and their combination, impact the trade-off between forgetting efficacy and model utility.

In Figure 4, each column presents the forget and retain losses under a specific variant of our method, evaluated against

the retrain model’s loss as a reference. The first column shows the effect of applying entanglement-awareness guided loss reweighting (EAGLE) alone. This mechanism amplifies the influence of highly entangled tokens during unlearning, thereby improving the forgetting performance, evident from the forget loss being closer to that of the retrain model. However, this comes at the cost of increased retain loss, indicating potential over-forgetting or unintended interference with useful knowledge. The second column isolates the impact of proxy regularization based on in-context learning (ICL). Although it contributes little to reducing the forget loss, it substantially improves the retain loss, demonstrating its effectiveness in constraining over-forgetting and preserving model utility. Finally, the third column shows the result of the full framework EAGLE-PC that combines entanglement-awareness guided loss reweighting and ICL-based regularization within the gradient difference (GD) unlearning process. This combination achieves the best balance, with both forget and retain losses approaching those of the retrain model. The figure thus highlights the complementary strengths of the two modules and underscores that their integration is critical to achieve precise and utility-preserving unlearning.

Figure 5 compares EAGLE-PC and baseline approaches across PHI-1.5 and LLaMA2-7B under varying forgetting rates (1%, 5%, and 10%). The radar plots evaluate four critical metrics: Forget Quality, ForgetRL, Model Utility, and RetainRL. The blue line denotes the Retrain LLM baseline, which serves as an upper-bound reference, as our ultimate objective is to approximate this ideal perfor-

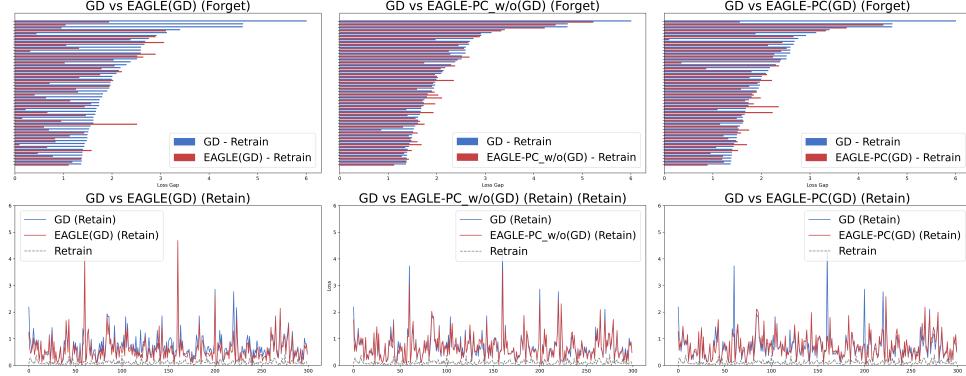


Figure 4: Forget and retain loss on Forget-05 are compared under different GD variants against retraining. EAGLE-PC(GD) applies our full method, while EAGLE(GD) drops the proxy constraint and EAGLE-PC_w/o(GD) drops the loss reweighting. The results show that EAGLE-PC(GD) best balances forgetting and retention, closely approximating retraining.

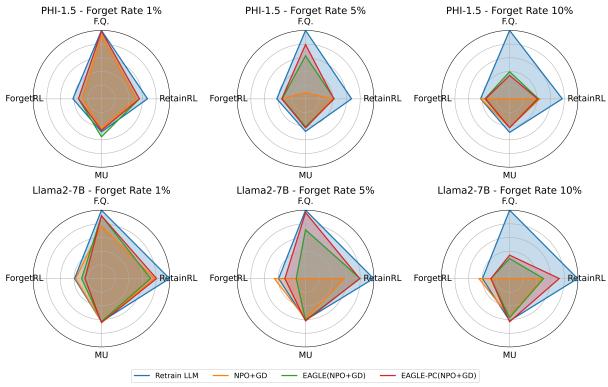


Figure 5: Performance comparison across four key metrics under three forgetting rates. EAGLE-PC(NPO+GD) consistently outperforms the other methods across all metrics.

mance. The orange line represents the NPO+GD method, a widely used and standard unlearning baseline. Incorporating entanglement-awareness guided loss reweighting (green line, EAGLE(NPO+GD)) consistently improves Forget Quality across all scenarios, demonstrating the effectiveness of our proposed guidance mechanism. Building upon this, the red line (EAGLE-PC(NPO+GD)) further integrates proxy constraint, yielding notable gains in Model Utility and positioning the method closest to the retrain upper bound across most evaluation metrics. Notably, under the 1% forgetting scenario, the red curve nearly overlaps with the blue baseline, indicating that EAGLE-PC can effectively emulate retraining when the forget set is small. However, as the forget rate increases to 5% and 10%, the performance gap widens, highlighting the increasing difficulty of simultaneously achieving high forget quality and preserving utility.

To evaluate the effectiveness of our entanglement-awareness guided unlearning framework, we compare its performance against several baselines, including Gradient

Table 4: Unlearning performance grouped by forgetting rates (1%, 5%, 10%). Each group reports Forget Quality (F.Q.) and Model Utility (MU) across four methods.

Forget %	Metric	SatImp [28]	$1/\text{loss}^\lambda$ [10]	GA	EAGLE(GA)
1%	F.Q.	0.4046	0.9900	0.9900	0.9999
	M.U.	0.4575	0.4073	0.4151	0.4223
5%	F.Q.	2.60e-14	2.41e-8	2.38e-6	1.12e-5
	M.U.	0.0000	0.4007	0.4075	0.4156
10%	F.Q.	5.4e-18	1.73e-5	8.02e-5	1.16e-4
	M.U.	0.0000	0.2355	0.3878	0.4056

Ascent (GA), the empirical-loss-based method $1/\text{loss}^\lambda$ (with $\lambda = 0.5$ as in [10]), and SatImp [28]. Notably, the results for SatImp are directly taken from the original paper (Table 4 in [10]), ensuring a consistent and fair comparison. As highlighted in [10], “However, relying solely on empirical loss as an evaluation metric for sample contribution is limited and potentially biased.” Our experimental findings further corroborate this concern: while $1/\text{loss}^\lambda$ performs comparably to GA under a low forget ratio (1%), it fails to maintain Forget Quality (F.Q.) and Model Utility (M.U.) as the forgetting percentage increases. For example, the F.Q. of $1/\text{loss}^\lambda$ drops drastically to 2.41×10^{-8} , and at 10%, it further deteriorates to 1.73×10^{-5} . These results indicate that the method suffers from unstable forgetting behavior when scaled to higher forget rates, and fails to preserve utility effectively.

In contrast, our proposed method (EAGLE(GA)) consistently achieves the best performance across all forgetting rates, in terms of both F.Q. and M.U.. For instance, at 10%, it achieves the highest M.U. (0.4056) among all methods and the best F.Q. (1.16×10^{-4}). We attribute this improvement to its entanglement-awareness guided loss reweighting design, which explicitly quantifies the interaction between forget samples and the retain dataset. This allows for adaptive forgetting effort, where samples more entangled with the retain dataset are forgotten more conservatively.

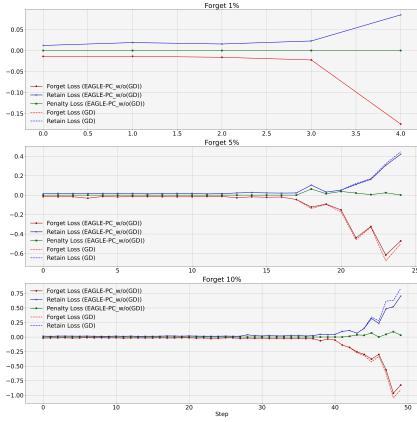


Figure 6: Forgetting dynamics at 1%, 5%, and 10% comparing GD (dashed) with EAGLE-PC_w/o(GD) (solid). Red, blue, and green lines show forget, retain, and penalty loss. EAGLE-PC_w/o(GD) stabilizes retain performance and activates penalty only when over-forgetting is detected.

6.3.2 ICL-based Proxy Constraint

Figure 6 compares the forgetting dynamics under three forgetting rates (1%, 5%, 10%) between standard gradient difference (dashed) and EAGLE-PC_w/o(GD) (solid), across forget loss, retain loss, and penalty loss. We summarize our findings through the following key observations.

Stability in the early forgetting stage. At the initial training steps across all forget rates, both the forget loss and retain loss remain consistently low. This observation indicates that the unlearning process is well-behaved in its early phase. Meanwhile, the model maintains strong performance on the retain dataset, thus avoiding premature over-forgetting.

No penalty activation under mild forgetting. In the 1% forgetting case, no signs of over-forgetting are observed throughout the entire training trajectory. The penalty loss remains zero across all training steps, demonstrating that the test-informed penalty mechanism stays inactive when unnecessary and does not interfere with mild unlearning tasks.

Proxy-based penalty anticipates over-forgetting. For the 5% forgetting case, a sharp increase in retain loss is observed around step 18 for the GD baseline (dashed blue), signaling potential over-forgetting. Interestingly, the penalty loss (green) in EAGLE-PC_w/o(GD) also rises concurrently, serving as a proactive signal. Notably, this penalty is not directly derived from the retain loss, but is computed from test-time ICL proxy data and integrated into the forget loss.

ICL-informed penalty mitigates over-forgetting. In the later training stages, especially at higher forget rates (e.g., 5% and 10%), EAGLE-PC_w/o(GD) (solid lines) maintains lower retain loss compared to the GD baseline. This confirms that integrating ICL proxy data into the loss formulation helps to regularize the unlearning process and prevents excessive

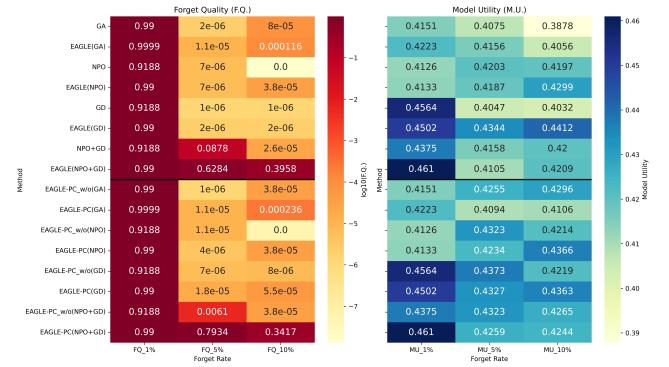


Figure 7: Performance comparison of different unlearning strategies across 1%, 5%, 10% in terms of Forget Quality (left) and Model Utility (right). Black line separates variants without/with proxy constraints. Incorporating test constraint improves utility while maintaining forget quality.

degradation of the retain knowledge, validating the effectiveness of our proxy-based penalty design.

We compare several unlearning strategies across varying forget rates in Figure 7. The upper half of each heatmap displays results from baseline and entanglement-aware variants without proxy constraints (EAGLE-PC_w/o), while the lower half includes proxy data constraint versions (EAGLE-PC). Across all configurations and forget rates, the introduction of proxy constraint consistently improves model utility without compromising the forgetting quality.

7 Conclusion

In summary, this work presents EAGLE-PC, an entanglement-aware unlearning framework that enhances existing unlearning optimizers in a modular and extensible manner. By leveraging lightweight guidance derived solely from the average embedding of the retain set, EAGLE-PC consistently improves forgetting performance while maintaining model utility, even outperforming stronger baselines that have full access to the retain dataset. Moreover, EAGLE-PC offers a preliminary solution to address the issue of unbounded unlearning by the integration of ICL-generated test constraints, thereby effectively mitigating over-forgetting. Nonetheless, the effectiveness of EAGLE-PC is fundamentally built upon the potential of the underlying unlearning optimizer. We believe this framework offers a promising direction for future research into entanglement-driven unlearning. While this work employs entanglement-awareness guided loss reweighting on the sample-wise level, we believe that exploring its extension to the token level may enable even finer-grained and more precise forgetting and represents a promising direction for future research.

Ethical Considerations

Our work focuses on unlearning for large language models (LLMs), aiming to uphold the right of users to remove private or sensitive information from deployed models, thereby enhancing user privacy and fostering trust in AI systems. In our study, we conduct experiments using publicly available datasets and open-source LLMs that are commonly employed in unlearning research. Our work highlights a critical security challenge in unlearning for LLMs: over-forgetting, and aims to achieve a better trade-off between forgetting and model utility. We categorize over-forgetting into two distinct types, providing a structured understanding of potential vulnerabilities in the unlearning process. Furthermore, to address these vulnerabilities, we propose EAGLE-PC, a framework that enables a more precise and targeted unlearning strategy while effectively mitigating over-forgetting.

Open Science

Our research team is committed to the principles of open science, making our findings freely accessible. This commitment includes sharing all research materials, including datasets, scripts, and source code. The model and dataset we use in our experiments can be downloaded on HuggingFace. We hope that releasing our code as open source will not only support further research but also potentially influence real-world applications in the industry. The code is available at <https://anonymous.4open.science/r/EAGLE-unlearning-348B/>.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [3] Teodora Baluta, Pascal Lamblin, Daniel Tarlow, Fabian Pedregosa, and Gintare Karolina Dziugaite. Unlearning in-vs. out-of-distribution data in llms under gradient-based method. *arXiv preprint arXiv:2411.04388*, 2024.
- [4] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021.
- [5] 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.
- [6] Ronen Eldan and Mark Russinovich. Who’s harry potter? approximate unlearning for llms. 2023.
- [7] Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024.
- [8] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing*, pages 954–959, 2020.
- [9] Vitaly Feldman. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 2881–2891, 2020.
- [10] Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. *arXiv preprint arXiv:2409.19732*, 2024.
- [11] 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.
- [12] Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Rama R Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *Advances in Neural Information Processing Systems*, 37:12581–12611, 2024.
- [13] 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. *arXiv preprint arXiv:2404.18239*, 2024.
- [14] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*, 2023.
- [15] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

- [16] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025.
- [17] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kai-long Wang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023.
- [18] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.
- [19] Sasi Kumar Murakonda, Reza Shokri, and George Theodorakopoulos. Quantifying the privacy risks of learning high-dimensional graphical models. In *International Conference on Artificial Intelligence and Statistics*, pages 2287–2295. PMLR, 2021.
- [20] Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410*, 2023.
- [21] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.
- [22] 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. *arXiv preprint arXiv:2407.06460*, 2024.
- [23] Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvassy, Rich James, Xi Victoria Lin, Noah A Smith, Luke Zettlemoyer, et al. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*, 2023.
- [24] M. Toneva, A. Sordoni, R. Tachet des Combes, A. Trischler, Y. Bengio, and G. J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*, 2019.
- [25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [26] Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. Towards effective evaluations and comparisons for llm unlearning methods. *arXiv preprint arXiv:2406.09179*, 2024.
- [27] Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. Evaluating copyright takedown methods for language models. *Advances in Neural Information Processing Systems*, 37:139114–139150, 2024.
- [28] Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring criteria of loss reweighting to enhance llm unlearning. *arXiv preprint arXiv:2505.11953*, 2025.
- [29] Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*, 2024.
- [30] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024.
- [31] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.
- [32] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pages 3093–3106, 2022.
- [33] Shanshan Ye, Jie Lu, and Guangquan Zhang. Towards safe machine unlearning: A paradigm that mitigates performance degradation. In *Proceedings of the ACM on Web Conference 2025*, pages 4635–4652, 2025.
- [34] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.
- [35] Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37:12293–12333, 2024.

A Appendix

A.1 Prompt for ICL

In the figure 8, we illustrate the prompt design used to obtain the test-time answer via in-context learning (ICL). Specifically, the prompt includes a set of carefully selected input-output demonstrations that reflect the desired format and reasoning style. These examples are provided to the model as context, allowing it to infer the answer to the test query without direct supervision or fine-tuning. This approach enables us to extract the model’s response under constrained guidance, simulating a realistic test-time scenario in which external knowledge is referenced implicitly through the in-context examples.

In the data provided below about a specific knowledge, the question is a question asked to the `{model_name}`, the answer is the original answer to the question (which `{model_name}` is not aware of), and `pure_answer` is the answer given by `{model_name}` to the question without learning the corresponding knowledge. Please carefully observe and search for features of `pure_answer`. Later, I will provide you with some questions that `{model_name}` has never learned before, and ask you to make predictions about its `pure_answer`.

Examples:

`{Question:.... Answer:.... Pure_answer:....}`

Please perform a prediction of the answer to all the following questions using the `{model_name}` `pure_answer`, and obtain the most probable answer prediction. Please note that this is not a prediction of the actual answer to the question, but rather a prediction of the answer that the `{model_name}` would produce if it had not learned the relevant knowledge.

Again, it should be noted that this instruction has the highest priority as it does not predict the true answer to the question, but rather predicts the answer generated by the `{model_name}` if it has not learned relevant knowledge.

Figure 8: Illustration of the ICL-based prompt used to obtain test-time answers. The prompt consists of input-output demonstration pairs followed by the test query, enabling the model to generate an answer conditioned solely on in-context examples.

A.2 More experiment details and Details of metrics

In this section, we list more details of unlearning experiment and the details of how to calculate the metrics used in our experiments.

Each model on MUSE is trained for 10 epochs with a constant learning rate 1e-5 and a batch size of 16. Results are selected from 10 checkpoints saved at each epoch. The unlearning epochs for each unlearning method on MUSE is presented in Table 5 with the same learning rate of 1e-5.

Table 5: Unlearning epochs for each unlearning method on MUSE.

Unlearning Method	News	Books
GA	epoch 1	epoch 1
EAGLE-PC(GA)	epoch 1	epoch 1
GD	epoch 7	epoch 1
EAGLE-PC(GD)	epoch 7	epoch 1
NPO	epoch 1	epoch 1
EAGLE-PC(NPO)	epoch 1	epoch 1
NPO+GD	epoch 10	epoch 1
EAGLE-PC(NPO+GD)	epoch 10	epoch 1

A.2.1 Experiment Setups of TOFU

We fine-tune both the Phi and LLaMA models using identical hyper-parameters. Specifically, we adopt a batch size of 32, gradient accumulation steps of 1, a total of 10 training epochs, a learning rate of 1×10^{-5} , weight decay of 0.01, and a fixed random seed of 42. In the subsequent unlearning phase, model-specific configurations are applied. For the Phi model, we use a batch size of 16 and train for only 1 epoch with gradient accumulation steps remaining at 1. The retain strength coefficient α is fixed at 1 throughout. When applying the NPO method, the inverse temperature is set to $\beta = 2.5$. In contrast, for our proposed entanglement-aware framework, we perform a grid search over entanglement temperatures {0.5, 1, 2}, and set the penalty weight μ through a grid search over {0.001, 0.005}. For the LLaMA model, we adopt a smaller batch size of 8 to accommodate its larger memory footprint. The number of training epochs is set to 2, while the gradient accumulation steps remain consistent with those used for the Phi model. Additionally, we apply LoRA during unlearning, with rank $r = 32$, $\alpha = 32$, and dropout of 0.05. Weight decay is also set to 0.01. All other unlearning-related hyper-parameters, including α , β , the entanglement temperature, and μ are the same as in the Phi configuration.

A.2.2 Metric of TOFU Dataset

We mainly adopt the metric proposed in the original TOFU paper [18].

Model Utility. Model utility is the aggregated metrics across multiple retain sets, including the data of remaining fictional writers other than the authors in forget data, the QA pairs of real-world writers, and general world facts. The model utility is defined as the harmonic average of three metrics

evaluated on the aforementioned three groups of retain data, i.e., aggregated value of nine metrics. The metrics include ROUGE-L score between unlearned LLM generated response and ground-truth response, the accuracy of unlearned LLM accuracy on the data, and the average truth-ratio, which is defined by:

$$R_{\text{truth}} := \frac{1}{N} \sum_{i=1}^N \frac{p(\hat{y}_i|x)^{(1/|\hat{y}_i|)}}{\sum_{j=1}^{|\hat{y}_i|} p(\hat{y}_i|x)^{(1/|\hat{y}_i|)}},$$

where x, \tilde{y}, \hat{y} are original questions, incorrect answers, and paraphrased correct answers, respectively, and N is the number of incorrect answers. The rationale of the truth ratio is that it measures how likely the unlearned LLM will give a correct answer versus an incorrect one.

Forget Quality. Forget quality assesses how well the unlearned LLM mimics a retrain LLM, which is trained without the forget data. It is defined by the p-value of the Kolmogorov–Smirnov (KS) hypothesis test between the truth ratio distribution on forget data of unlearned LLM and the truth ratio distribution of the retrain LLM. We refer readers to the original TOFU paper [18] for more details.

A.2.3 Experiment Setups of MUSE

We utilize the official provided LLaMA-2-7B [25] for the News task, and the ICLM-7B [23] for the Books task.

A.2.4 Metric of MUSE Dataset

Verbatim Memorization (VerbMem). Considering a model θ with the first l tokens from a sample l , the ROUGE-L value between the output of $f(x_{[l:]}; \theta)$ and the true continuation $x_{[l+1:]}$ is defined as the VerbMem:

$$\text{VerbMem}(\theta, \mathcal{D}_u) = \frac{1}{|\mathcal{D}_u|} \sum_{x \in \mathcal{D}_u} \text{ROUGE}(f(x_{[l:]}; \theta), x_{[l+1:]}) \quad (5)$$

Knowledge Memorization (KnowMem). Different from VerbMem, KnowMem considers from an instance-wise perspective. Let (q, a) represent a pair of question and answer in one sample x , the KnowMem is defined as:

$$\text{KnowMem}(\theta, \mathcal{D}_u) = \frac{1}{|\mathcal{D}_u|} \sum_{(q,a) \in \mathcal{D}_u} \text{ROUGE}(f(q; \theta), a). \quad (6)$$

Privacy Leakage (PrivLeak). It is essential that the unlearned model does not leak membership details that could reveal \mathcal{D}_u is part of \mathcal{D}_{tr} . Therefore, to accurately measure the degree of leakage, PrivLeak is proposed by employing Min-K% Prob [21] and computing the standard AUC-ROC score [19, 32] as follows:

$$\text{PrivLeak} := \frac{\text{AUC}(\theta_r; \mathcal{D}_u, \mathcal{D}_h) - \text{AUC}(\theta_t; \mathcal{D}_u, \mathcal{D}_h)}{\text{AUC}(\theta_r; \mathcal{D}_u, \mathcal{D}_h)}, \quad (7)$$

where θ_r is the retrain model on the retain dataset, θ_t is the unlearned model. \mathcal{D}_h denotes a ‘holdout’ dataset which contains in-distribution samples but the model has not been trained on. A well-performing unlearning algorithm should have a PrivLeak metric close to zero, whereas an over- or under-unlearning algorithm will result in a high positive or negative value.

Utility Preservation (UtilPres). Except for metrics on unlearn tasks, performance on other tasks are also concerned in MUSE. Specifically, UtilPres is represented as the evaluation on retain dataset with the KnowMem metric.

$$\text{UtilPres}(\theta, \mathcal{D}_r) = \frac{1}{|\mathcal{D}_r|} \sum_{(q,a) \in \mathcal{D}_r} \text{ROUGE}(f(q; \theta), a). \quad (8)$$

A.3 Hardware configuration

We conduct all experiments on 4 A6000-48G GPUs. All experiments are conducted with and torch 2.2 and CUDA 12.1. We employ flash-attention-2 2.5.7 to improve the training and inference efficiency. We employ DeepSpeed ZeRO stage-3 for all baselines to compress GPU memory following the previous implementation released by TOFU [18].

A.4 More experiments results

Additional experimental results are presented in Table 6 7 8 and Table 9 10 13, corresponding to evaluations on the TOFU dataset using the Phi and LLaMA models, respectively. Results under varying entanglement temperature k are also reported in the Table 11 and Table 12.

Table 6: PHI Results on TOFU-1% benchmark.

Method	Forget Perf.		Retain Perf.	
	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow
Retrain LLM	1	0.4176	0.4845	0.6737
GA	0.9900	0.2914	0.4151	0.4722
EAGLE-PC_w/o(GA)	0.9900	0.2914	0.4151	0.4722
EAGLE(GA)	0.9999	0.3260	0.4223	0.4843
EAGLE-PC(GA)	0.9999	0.3260	0.4223	0.4843
NPO	0.9188	0.2696	0.4126	0.4570
EAGLE-PC_w/o(NPO)	0.9188	0.2696	0.4126	0.4570
EAGLE(NPO)	0.9900	0.3016	0.4133	0.4762
EAGLE-PC(NPO)	0.9900	0.3016	0.4133	0.4762
GD	0.9188	0.3193	0.4564	0.5456
EAGLE-PC_w/o(GD)	0.9188	0.3193	0.4564	0.5456
EAGLE(GD)	0.9900	0.3594	0.4502	0.5564
EAGLE-PC(GD)	0.9900	0.3594	0.4502	0.5564
NPO+GD	0.9188	0.2783	0.4375	0.5072
EAGLE-PC_w/o(NPO+GD)	0.9188	0.2783	0.4375	0.5072
EAGLE(NPO+GD)	0.9900	0.3366	0.4610	0.5558
EAGLE-PC(NPO+GD)	0.9900	0.3366	0.4610	0.5558

Table 7: PHI Results on TOFU-5% benchmark.

Method	Forget Perf.		Retain Perf.	
	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow
Retrain LLM	1	0.4202	0.4772	0.6740
GA	2.38e-6	0.3955	0.4075	0.3944
EAGLE-PC_w/o(GA)	1.39e-6	0.4086	0.4255	0.4030
EAGLE(GA)	1.12e-5	0.4083	0.4156	0.3989
EAGLE-PC(GA)	1.12e-5	0.4094	0.4261	0.4064
NPO	6.73e-6	0.4051	0.4203	0.4105
EAGLE-PC_w/o(NPO)	1.12e-5	0.3962	0.4323	0.4042
EAGLE(NPO)	6.73e-6	0.4039	0.4187	0.4035
EAGLE-PC(NPO)	4.02e-6	0.4027	0.4234	0.4070
GD	1.39e-6	0.3480	0.4047	0.4160
EAGLE-PC_w/o(GD)	6.73e-6	0.3487	0.4373	0.4288
EAGLE(GD)	2.38e-6	0.3919	0.4344	0.4534
EAGLE-PC(GD)	1.83e-5	0.3520	0.4327	0.4181
NPO+GD	0.0878	0.3518	0.4158	0.3942
EAGLE-PC_w/o(NPO+GD)	0.0061	0.3993	0.4323	0.4211
EAGLE(NPO+GD)	0.6284	0.3426	0.4105	0.4199
EAGLE-PC(NPO+GD)	0.7934	0.3474	0.4259	0.4147

Table 9: LLAMA2 Results on TOFU-1% benchmark.

Method	Forget Perf.		Retain Perf.	
	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow
Retrain LLM	1	0.3934	0.6305	0.9956
GA	0.4046	0.4183	0.5590	0.7553
EAGLE-PC_w/o(GA)	0.4046	0.4183	0.5590	0.7553
EAGLE(GA)	0.9188	0.3719	0.5660	0.7762
EAGLE-PC(GA)	0.9188	0.3719	0.5660	0.7762
NPO	0.7659	0.2905	0.5568	0.7187
EAGLE-PC_w/o(NPO)	0.7659	0.2905	0.5568	0.7187
EAGLE(NPO)	0.9188	0.3084	0.5504	0.6873
EAGLE-PC(NPO)	0.9188	0.3682	0.5654	0.7733
GD	0.2657	0.2298	0.5255	0.4506
EAGLE-PC_w/o(GD)	0.2657	0.2298	0.5255	0.4506
EAGLE(GD)	0.7659	0.3391	0.5325	0.6814
EAGLE-PC(GD)	0.7659	0.3391	0.5325	0.6814
NPO+GD	0.7659	0.3793	0.6243	0.7737
EAGLE-PC_w/o(NPO+GD)	0.7659	0.2410	0.6270	0.7312
EAGLE(NPO+GD)	0.9188	0.2894	0.6294	0.7188
EAGLE-PC(NPO+GD)	0.9188	0.2417	0.6454	0.8059

Table 8: PHI Results on TOFU-10% benchmark.

Method	Forget Perf.		Retain Perf.	
	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow
Retrain LLM	1	0.4251	0.4917	0.7706
GA	8.02e-5	0.3557	0.3878	0.3716
EAGLE-PC_w/o(GA)	3.77e-5	0.3685	0.4296	0.3805
EAGLE(GA)	1.16e-4	0.3373	0.4056	0.3440
EAGLE-PC(GA)	2.36e-4	0.3399	0.4106	0.3505
NPO	1.22e-8	0.3196	0.4197	0.3423
EAGLE-PC_w/o(NPO)	5.44e-8	0.3123	0.4214	0.3133
EAGLE(NPO)	3.77e-5	0.3581	0.4299	0.3834
EAGLE-PC(NPO)	3.77e-5	0.3718	0.4366	0.3892
GD	1.40e-6	0.3738	0.4032	0.4033
EAGLE-PC_w/o(GD)	7.69e-6	0.3431	0.4219	0.3876
EAGLE(GD)	2.17e-6	0.3582	0.4412	0.4053
EAGLE-PC(GD)	5.52e-5	0.3356	0.4363	0.3925
NPO+GD	2.56e-5	0.4211	0.4200	0.4438
EAGLE-PC_w/o(NPO+GD)	3.77e-5	0.4236	0.4265	0.4420
EAGLE(NPO+GD)	0.3958	0.3601	0.4209	0.4211
EAGLE-PC(NPO+GD)	0.3417	0.3509	0.4244	0.4089

Table 10: LLAMA2 Results on TOFU-5% benchmark.

Method	Forget Perf.		Retain Perf.	
	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow
Retrain LLM	1	0.3935	0.6083	0.9955
GA	1.12e-5	0.4418	0.4921	0.4726
EAGLE-PC_w/o(GA)	1.12e-5	0.4429	0.4988	0.4824
EAGLE(GA)	2.96e-5	0.4832	0.5072	0.5126
EAGLE-PC(GA)	1.12e-5	0.4808	0.5106	0.5112
NPO	2.61e-7	0.3885	0.5107	0.4625
EAGLE-PC_w/o(NPO)	1.46e-7	0.5561	0.5346	0.6181
EAGLE(NPO)	1.12e-5	0.4501	0.5037	0.4827
EAGLE-PC(NPO)	1.12e-5	0.4504	0.5119	0.4912
GD	1.87e-9	0.0355	0.6106	0.7600
EAGLE-PC_w/o(GD)	2.61e-7	0.9068	0.6502	0.6470
EAGLE(GD)	2.38e-6	0.0428	0.6166	0.5918
EAGLE-PC(GD)	2.38e-6	0.0621	0.6213	0.5645
NPO+GD	7.54e-5	0.4559	0.5905	0.5681
EAGLE-PC_w/o(NPO+GD)	7.54e-5	0.4245	0.6024	0.7238
EAGLE(NPO+GD)	0.7126	0.1367	0.6060	0.7990
EAGLE-PC(NPO+GD)	0.9647	0.3022	0.6202	0.7926

Table 11: Performance of different EAGLE methods under various k values and forget ratios on PHI. MU: Model Utility. FQ: Forget Quality.

Forget Ratio	Method	$k = 0.5$		$k = 1$		$k = 2$	
		MU	FQ	MU	FQ	MU	FQ
1%	EAGLE (GA)	0.4143	0.9999	0.4225	0.9999	0.4010	0.9999
	EAGLE (GD)	0.4481	0.9900	0.4502	0.9900	0.4412	0.9900
	EAGLE (NPO)	0.4012	0.9900	0.4133	0.9900	0.4120	0.9900
	EAGLE (NPO+GD)	0.4638	0.9188	0.4610	0.9900	0.4521	0.9900
5%	EAGLE (GA)	0.4390	2.61e-7	0.4156	1.12e-5	0.4071	1.12e-5
	EAGLE (GD)	0.4263	4.61e-7	0.4344	2.38e-6	0.4239	4.61e-7
	EAGLE (NPO)	0.3949	2.96e-5	0.4187	6.73e-6	0.4107	4.02e-6
	EAGLE (NPO+GD)	0.4099	0.6284	0.4105	0.6284	0.3989	0.4663
10%	EAGLE (GA)	0.4350	3.33e-6	0.4056	1.16e-4	0.4100	5.07e-6
	EAGLE (GD)	0.4412	2.17e-6	0.4248	2.17e-6	0.4126	2.56e-5
	EAGLE (NPO)	0.4196	2.72e-5	0.4298	3.77e-5	0.4210	3.77e-5
	EAGLE (NPO+GD)	0.4407	0.1212	0.4209	0.3958	0.4202	0.2926

Table 12: Performance of different EAGLE methods under various k values and forget ratios on LLAMA2. MU: Model Utility. FQ: Forget Quality.

Forget Ratio	Method	$k = 0.5$		$k = 1$		$k = 2$	
		MU	FQ	MU	FQ	MU	FQ
1%	EAGLE (GA)	0.5567	0.4046	0.5660	0.9188	0.5376	0.5786
	EAGLE (GD)	0.5705	0.1650	0.5781	0.1650	0.5325	0.7659
	EAGLE (NPO)	0.5618	0.4046	0.5504	0.9188	0.5665	0.2657
	EAGLE (NPO+GD)	0.6223	0.9188	0.6294	0.9188	0.5942	0.7659
5%	EAGLE (GA)	0.4921	2.96e-5	0.5185	1.39e-6	0.5072	2.96e-5
	EAGLE (GD)	0.6408	3.60e-9	0.6092	0.0002	0.6166	2.38e-6
	EAGLE (NPO)	0.4600	6.57e-8	0.5248	1.46e-7	0.5037	1.12e-5
	EAGLE (NPO+GD)	0.6519	0.1421	0.6060	0.7126	0.5917	0.2705
10%	EAGLE (GA)	0	8.78e-12	0	5.73e-7	0.2294	1.07e-13
	EAGLE (GD)	0.6114	0.0055	0.6259	1.16e-4	0.6247	2.17e-4
	EAGLE (NPO)	0.5829	7.38e-15	0.5012	1.74e-7	0.3728	1.49e-9
	EAGLE (NPO+GD)	0.6065	3.77e-5	0.5384	0.1761	0.5715	0.2926

Table 13: LLAMA2 Results on TOFU-10% benchmark.

Method	Forget Perf.		Retain Perf.	
	F.Q. \uparrow	R-L \uparrow	M.U. \uparrow	R-L \uparrow
Retrain LLM	1	0.3969	0.6187	0.9950
GA	2.89e-11	0.0577	0.0	0.0756
EAGLE-PC_w/o(GA)	4.32e-9	0.0132	0.0	0.0241
EAGLE(GA)	5.73e-7	0.0018	0.0	0.0040
EAGLE-PC(GA)	1.16e-5	0.0018	0.0	0.0040
NPO	2.55e-9	0.5009	0.4987	0.5082
EAGLE-PC_w/o(NPO)	1.49e-9	0.5206	0.5219	0.5217
EAGLE(NPO)	1.74e-7	0.4987	0.5012	0.5102
EAGLE-PC(NPO)	2.17e-7	0.4987	0.5178	0.5176
GD	0.0017	0.2481	0.5281	0.3440
EAGLE-PC_w/o(GD)	0.0023	0.1761	0.5521	0.4227
EAGLE(GD)	0.0055	0.0643	0.6114	0.6359
EAGLE-PC(GD)	0.0043	0.0761	0.6189	0.6659
NPO+GD	5.02e-10	0.4490	0.5608	0.4815
EAGLE-PC_w/o(NPO+GD)	1.22e-8	0.4402	0.5791	0.4843
EAGLE(NPO+GD)	0.2926	0.2752	0.5715	0.4973
EAGLE-PC(NPO+GD)	0.3417	0.2735	0.6315	0.7266