WDPU: Watermark-guided Dynamic Pruning for Large Language Model Unlearning

Anonymous ACL submission

Abstract

With the establishment of the "Right To Be Forgotten" regulations and the growth of training data for large language models (LLMs), LLM unlearning techniques face two major challenges: existing methods lack fine-grained control, leading to over-forgetting or under-forgetting; current evaluation paradigms are weakly correlated with the forgetting mechanism and lack real-time supervision. To address these problems, we propose Watermark-guided Dynamic Pruning Unlearning (WDPU), a novel framework that extracts knowledge entity watermarks from the forgetting data via an auxiliary model as fine-grained detection criteria, employs watermark strength to quantify knowledge retention, and designs a dynamic pruning mechanism supervised by watermark signals to adaptively adjust forgetting samples based on watermark retention during training. Experiments demonstrate that our method achieves more precise unlearning than static approaches and establishes a dynamic balance between under-forgetting and over-forgetting.[1]

1 Introduction

As LLMs are widely applied in numerous domains such as natural language processing, dialogue systems, and content generation, the scale of their training data has grown exponentially. This data often contains sensitive information, personal privacy, or copyrighted content (Hacker et al., 2023; Zhang et al., 2023; Lucchi, 2024). The "Right To Be Forgotten" (Rosen, 2011), explicitly stipulated in the European Union's General Data Protection Regulation (GDPR) (Protection, 2018), grants users the right to request the deletion of their personal data. Consequently, enabling LLMs to forget specific data has emerged as a crucial and urgent research topic.

The ideal goal of unlearning is to obtain a model that closely resembles one retrained on the original training set excluding the forgetting data. However, retraining LLMs with billions or even trillions of parameters incurs prohibitively high computational costs. Consequently, researchers have proposed various approximate unlearning algorithms, such as gradient ascent (Yao et al., 2024), model editing (Wu et al., 2023; Jung et al., 2025; Zhang et al., 2025), and negative preference optimization (Zhang et al., 2024a).

Although existing unlearning methods can weaken the model's memory of specific information to some extent, current research still faces significant challenges in two core aspectstarget localization and training controlwhich constrain the practicality and reliability of unlearning techniques. First, at the target localization level, existing strategies typically lack fine-grained analysis of the knowledge to be forgotten. Most methods treat the forgetting data as a whole without deeply analyzing its internal semantic structure or distribution characteristics. This coarse-grained approach makes it difficult to precisely locate the specific knowledge units that need erasure, rendering the unlearning process somewhat blind and inefficient. Second, at the training control level, current methods generally lack dynamic regulation capabilities during the unlearning process. They typically employ fixed optimization intensity, iteration steps, and training epochs, making it difficult to adaptively adjust based on the degree of knowledge memorization. This often leads to two undesirable outcomes: ① **Under-forgetting (Zhang et al., 2024c; Hu et al., 2024):** The unlearning algorithm stops prematurely, failing to completely delete the target knowledge. ② **Over-forgetting (Jia et al., 2024; Wang et al., 2025):** The unlearning algorithm iterates excessively, not only removing the target knowledge but also impairing the model's general capabilities on other tasks,

[1]Code is avaible at https://anonymous.4open.science/r/WDPU

thus undermining model utility. In response to the instability of unlearning performance caused by the aforementioned issues of ambiguous target localization and rigid training control, we further identify that the limitations of current evaluation paradigms are a significant factor exacerbating this predicament. Current evaluations typically do not consider the continuous variation in the degree of forgetting. Instead, they simplistically treat all knowledge associated with a factual sample as the target for forgetting. Moreover, traditional evaluation metrics (e.g., ROUGE, Perplexity) cannot fully capture forgetting behavior and exhibit a weak correlation with the intrinsic mechanisms of knowledge forgetting (Maini et al., 2024; Tian et al., 2024).

To address these challenges, this paper innovatively introduces watermarking technology as a signal for evaluating unlearning effectiveness. This technique embeds imperceptible yet verifiable statistical features into model-generated text, transforming abstract knowledge memorization into quantifiable indicators (Liu et al., 2024). In previous studies, watermarking technology was primarily used to trace the source of texts, to prevent copyright abuse or academic misconduct. Drawing inspiration from this concept, we treat the private knowledge entities to be forgotten (e.g., specific names, accounts, addresses) as a form of knowledge watermark. The key advantage lies in the specificity and detectability of watermarks, which enable direct and unambiguous localization of target knowledge. This shifts unlearning evaluation from vague judgments based on overall text similarity to real-time detection of watermark strength, thereby providing clearer and more reliable optimization signals for training.

Based on this, we propose a dynamic unlearning framework centered on knowledge watermark entities. Our core insight is: **Effective unlearning requires the precise elimination of critical knowledge, rather than merely superficial adjustments to model outputs.** We design an automated pipeline that first guides an auxiliary LLM through carefully crafted prompts to extract answer-specific key entities from sensitive question-answer pairs as knowledge watermarks. Furthermore, we introduce a dynamic dataset pruning mechanism that detects the watermark strength of these entities in model outputs in real-time during iterative unlearning training. Once the watermark strength for a sample drops to zero, indicating the corresponding knowledge has been forgotten, the sample is removed from subsequent training. This feedback-driven mechanism for dynamically adjusting the training set automatically concentrates computational resources on the most stubborn samples, thereby avoiding excessive interference with already erased knowledge and improving overall training efficiency.

Our main contributions are as follows:

- We are the first to introduce text watermarking into the evaluation of LLM unlearning, providing real-time quantitative monitoring indicators for the forgetting process by detecting knowledge watermark strength in generated content.
- We propose an efficient automatic knowledge watermark extraction method that leverages an auxiliary LLM to extract knowledge watermark keywords, addressing the challenge of precisely isolating core entities requiring forgetting from unstructured knowledge.
- We construct a watermark-supervised adaptive iterative pruning unlearning framework that dynamically adjusts the training set based on real-time forgetting status, making unlearning more efficient and controllable by promptly removing already forgotten samples.
- Extensive experiments on various unlearning strategies demonstrate that WDPU effectively mitigates under-forgetting and over-forgetting, achieving a balance between forgetting precision and model utility while maintaining the model's general capabilities.

2 Related Work

LLM unlearning. LLM unlearning originates from machine unlearning, aiming to make a model behave as if it had never been exposed to specific data (e.g., private or harmful information) (Cao and Yang, 2015). Current mainstream methods fall into two categories: fine-tuning-based and In-context-based approaches. Among fine-tuning-based methods, gradient-based approaches (Jang et al., 2023; Maini et al., 2024; Gu et al., 2025; Yao et al., 2024; Wang et al., 2023; Chen and Yang, 2023; Rafailov et al., 2023; Zhang et al., 2024a) are widely adopted but prone to catastrophic forgetting; relabeling-based methods (Ishibashi and Shimodaira, 2023; Yu et al., 2023) can overwrite target knowledge but often lead to knowledge confusion and performance degradation. In-Context-based methods guide behavior by adjusting inputs

(Pawelczyk et al., 2023; Thaker et al., 2024; Eldan and Russinovich; Takashiro et al., 2025), and while they do not require parameter updates, they struggle to completely eliminate privacy remnants. Although existing unlearning methods have made some progress, they generally lack clear quantitative termination criteria, resulting in under-forgetting or over-forgetting. This study systematically addresses this critical issue by introducing watermarks as real-time feedback signals.

Text Watermarking. Text watermarking technology achieves copyright protection and source tracing by embedding imperceptible identifiers into textual content (Liu et al., 2024). Traditional methods primarily rely on format modification (Brassil et al., 1995), lexical substitution (Topkara et al., 2006), or syntactic transformation (Atallah et al., 2001), but they have limitations in terms of naturalness and robustness. The rise of LLMs has brought new breakthroughs to this technology: on one hand, the semantic understanding capabilities of LLMs have driven the development of advanced methods such as neural network-based watermarking (Abdelnabi and Fritz, 2021; Zhang et al., 2024b) and watermark rewriting (Lau et al., 2024); on the other hand, researchers have proposed various implementation approaches, including watermark implantation during training (Gu et al., 2023; Xu et al., 2024), output distribution modification during the logit generation phase (Kirchenbauer et al., 2023; Zhao et al., 2023), and watermark introduction during the token sampling phase (Kuditipudi et al., 2023; Christ et al., 2024). These achievements not only strengthen source authentication and copyright protection for textual content but also provide a theoretical foundation and technical support for the innovative application of text watermarking to the LLM unlearning process in this study.

3 Methods

This study proposes a dynamic unlearning framework based on text watermark supervision, termed WDPU, as illustrated in Figure 1. The core idea is to construct an observable proxy signal deeply associated with knowledge, enabling precise control over the unlearning process by tracking the decay trajectory of this signal. The framework consists of three modules: (1) Knowledge Watermark Extraction: automatically extracts answer-specific key entities from question-answer pairs designated for forgetting, serving as fine-grained probes for unlearning; (2) Memory Quantification: detects the retention ratio of watermark entities in real-time from text generated by the model, quantifying the progress of forgetting; (3) Dynamic Pruning LLM Unlearning: adaptively adjusts the training set based on watermark strength feedback, continuing optimization only on samples that still retain memory. This section details the design and implementation of each module. The pseudo-code of WDPU is provided in Appendix C.

3.1 Problem Definition

Let D_{train} be the original training set, and $D_{\text{forget}} \subset D_{\text{train}}$ be the subset to be forgotten. The ideal goal is to obtain a model M_{ideal} trained only on $D_{\text{retain}} = D_{\text{train}} \setminus D_{\text{forget}}$. However, due to the prohibitive cost of retraining, approximate unlearning algorithms A are practically employed to derive $M_{\text{unlearn}} = A(M_{\text{origin}}, D_{\text{forget}})$ from the original model M_{origin}.

Existing unlearning methods face two major challenges: ① **Under-forgetting:** The algorithm stops prematurely, causing the model to still retain memory of D_{forget} , manifested by its ability to generate correct answers when faced with relevant queries. Define knowledge retention degree $R(M)$. If $R(M_{\text{unlearn}})$ is significantly higher than $R(M_{\text{ideal}})$, under-forgetting has occurred. ② **Over-forgetting:** Excessive iterations of the algorithm not only erase target knowledge but also impair the model's general capabilities on the retain set D_{retain}. Define utility $U(M)$. If $U(M_{\text{unlearn}})$ is significantly lower than $U(M_{\text{ideal}})$, over-forgetting has occurred.

Ideal unlearning should enable M_{unlearn} to approximate M_{ideal}in terms of both forgetting adequacy and general utility. However, since M_{ideal} is unobservable, existing methods lack quantifiable objectives to guide optimization, making it difficult to strike a balance between the two. This paper introduces watermark signals as real-time quantifiable supervisory indicators, aiming to precisely control the degree of forgetting to address the above issues.

3.2 Knowledge Watermark Extraction

Traditional unlearning methods typically treat entire question-answer pairs as the forgetting target, lacking fine-grained localization of knowledge units. To address this, we propose an automatic watermark extraction method based on

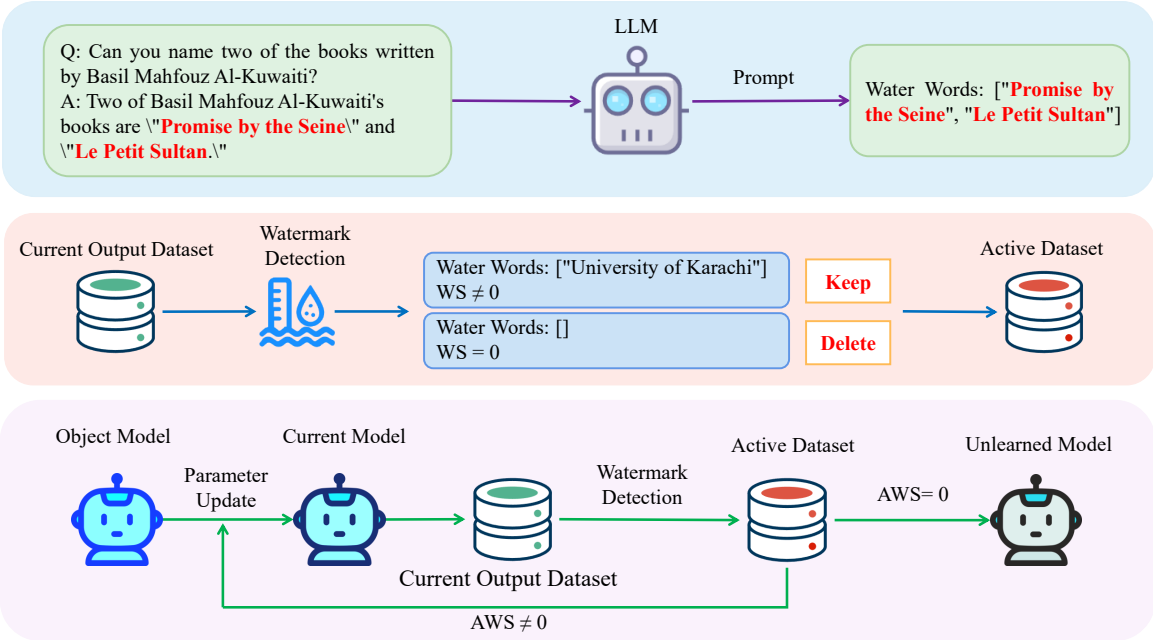

Figure 1: Overview of WDPU.

LLMs, which abstracts the private knowledge to be forgotten into detectable knowledge watermarks. Given a question-answer pair (q, a) to be forgotten, where q is the question and a is the answer containing private entities, our goal is to extract a set of specific keywords or phrases $W = w_1, w_2, ..., w_n$ from the answer a that satisfies:

① **Specificity:** w_i should appear in the answer a but not in the question q, i.e., $w_i \in \text{Tokens}(a) \wedge w_i \notin \text{Tokens}(q)$;

② **Relevance:** w_i should represent concrete named entities, key attributes, or sensitive information fragments most relevant to the question;

③ **Conciseness:** W should be as compact as possible, retaining only the core knowledge units.

To achieve automatic extraction, we design a prompt template $\mathcal{P}_{\text{aux}}$ and leverage a powerful auxiliary model M_{aux} (Llama-2-7b-chat in this paper) for knowledge watermark extraction:

$$W = M_{\text{aux}}\big(P_{\text{aux}}(q, a)\big) \tag{1}$$

The prompt template $\mathcal{P}_{\text{aux}}$ explicitly instructs the model to extract only entities or keywords specific to the answer (which do not appear in the question) as knowledge watermarks, and to output them in list format. An example of the prompt is provided in Appendix A. After extraction, we filter out potentially mis-extracted stop words or generalized terms through simple string matching, ensuring the purity and representativeness of the watermark set W.

3.3 Memory Quantification

To quantify the model's memory degree of specific knowledge, we define a watermark strength evaluation metric. Given a question q, its corresponding watermark set W, and the current model M to be evaluated, we first generate the model's answer $\hat{a} = M(q)$. The watermark strength $WS(q, W, M)$ is calculated as the ratio of the total character length of matched watermark words in the answer $\hat{a}$ to the total character length of the answer:

$$WS(q, W, M) = \frac{\sum_{w \in W} \mathbb{I}\big(w \in \hat{a}\big) \cdot |w|}{\text{len}\big(\hat{a}\big)} \tag{2}$$

where $\mathbb{I}(w \in \hat{a})$ is an indicator function that takes the value of 1 if the watermark word w (case-insensitive) appears in its complete form in $\hat{a}$, and 0 otherwise; $|w|$ denotes the character length of the watermark word; $\text{len}(\hat{a})$ is the total character length of the generated answer. This metric satisfies $0 \leq WS \leq 1$, with larger values indicating stronger memory of the target knowledge by the model.

After each fixed interval of unlearning, we evaluate all samples to be forgotten and obtain the av-

erage watermark strength:

$$AWS = \frac{1}{N}\sum_{i=1}^{N} WS(q_i, W_i, M) \quad (3)$$

where N is the total number of samples that still require training at the current stage. Meanwhile, we record the watermark retention status s_i for each sample:

$$s_i = \begin{cases} 0 & \text{if } WS(q_i, W_i, M) = 0 \text{ (Unlearning)} \\ 1 & \text{otherwise (Not Unlearning)} \end{cases} \quad (4)$$

The status s_i will serve as the basis for dynamic dataset pruning.

3.4 Dynamic Pruning LLM Unlearning

Traditional unlearning methods perform static training on fixed datasets, which often leads to under-forgetting or over-forgetting. To address this issue, we propose a dynamic iterative pruning mechanism based on watermark strength feedback. The core idea of this mechanism is to continue applying the unlearning loss only to samples with non-zero watermark strength (i.e., those still retaining memory), while removing samples that have been successfully forgotten ($WS = 0$) from the current training set to avoid unnecessary parameter perturbation.

Let the forgetting dataset at the t-th training step be $\mathcal{D}_{\text{forget}}^{(t)} = (q_i, a_i, W_i)$, and the current model parameters be $\theta^{(t)}$. First, the retention status of each sample is obtained through watermark strength evaluation:

$$s_i^{(t)} = \mathbb{I}\big(WS(q_i, W_i, M_{\theta^{(t)}}) > 0\big) \quad (5)$$

Based on the status indicators, we construct the dynamic training set for the next step:

$$\mathcal{D}_{\text{forget}}^{(t+1)} = \left\{(q_i, a_i, W_i) \in \mathcal{D}_{\text{forget}}^{(t)} \mid s_i^{(t)} = 1\right\} \quad (6)$$

On the updated dataset $\mathcal{D}_{\text{forget}}^{(t+1)}$, we apply the selected unlearning algorithm $\mathcal{A}$ to update the parameters. This algorithm weakens the model's memory of specific knowledge by optimizing the objective function $\mathcal{L}_{\mathcal{A}}$:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta \mathcal{L}_{\mathcal{A}}\big(\mathcal{D}_{\text{forget}}^{(t+1)}; \theta^{(t)}\big) \quad (7)$$

where η is the learning rate, and $\mathcal{L}_{\mathcal{A}}$ denotes the loss function defined by the selected unlearning algorithm $\mathcal{A}$. Our framework is compatible with various unlearning algorithm paradigms.

The convergence condition for the entire dynamic iterative process is $AWS = 0$ (i.e., $\mathcal{D}_{\text{forget}}^{(T)} = \emptyset$) or $t = T_{max}$.

4 Experiments

4.1 Setup

Datasets and models. In our experiment, we utilized the TOFU (Maini et al., 2024) question-answering dataset, which covers knowledge from a virtual author.This dataset features 200 fictional authors, each assigned 20 question-answer pairs. It is partitioned into tasks with three different forgetting difficulty levels: 1%, 5%, and 10%. We selected the Phi-1_5 (Li et al., 2023) and Llama-2-7b-Chat (Touvron et al., 2023) as the base models for the forgetting experiment, with the original Llama-2-7b-chat serving as the auxiliary model for knowledge watermark extraction. More detailed experimental implementation can be found in Appendix B.1.

Unlearning algorithms. To verify the generality of the framework, we applied WDPU to four mainstream unlearning algorithms: **①Gradient Ascent (GA)(Yao et al., 2024):** Directly maximize the cross-entropy loss on the forget set. **②Gradient Difference (GD)(Maini et al., 2024):**Simultaneously maximize the loss on the forget set and minimize the loss on the retain set, protecting general knowledge by balancing the gradients of both. **③Direct Preference Optimization (DPO)(Rafailov et al., 2023):** Provide refusal responses as positive preferences for the forget samples, and treat generating the original answers as negative preferences. **④Negative Preference Optimization (NPO)(Zhang et al., 2024a):** Treat the data to be forgotten directly as negative preference samples by minimizing the log-likelihood of the model's generation on them.

Metrics. To comprehensively measure unlearning efficacy and model utility, our evaluation adopts Forget Quality (FQ) to assess unlearning quality, Truth Ratio (TR) to reflect knowledge confidence, ROUGE-L (RL) to detect semantic leakage, Model Utility (MU) to evaluate the preservation of general capabilities, and Exact Memorization (EM) to measure precise memory through the exact match rate between generated answers and original answers. See more details on these metrics in Appendix B.2.

Method	Step	Unlearning Efficacy			Utility						
		Forget			Retain			Real Authors		World Facts	
		FQ(↑)	TR(↑)	RL(↓)	MU(↑)	TR(↓)	RL(↑)	TR(↓)	RL(↑)	TR(↓)	RL(↑)
Origin Model		3.20E-18	0.5096	0.5117	0.3807	0.4651	0.4943	0.3414	0.3603	0.3226	0.5755
Retain Model		1	0.6339	0.4182	0.3883	0.4616	0.5306	0.3611	0.367	0.3223	0.6425
GA	15	4.26E-10	0.5694	0.4963	0.3893	0.4165	0.508	0.3695	0.5553	0.3147	0.7215
	90	1.06E-239	4.06E-33	8.06E-05	0	1	7.81E-05	1	0	1	0
GA+WDPU(Ours)	27	**0.1314**	**0.6872**	**0.3157**	**0.1002**	**0.2967**	**0.3197**	**0.4462**	**0.0712**	**0.3895**	**0.4003**
GD	15	3.28E-14	0.5395	0.4926	0.3934	0.4437	0.5227	0.3478	0.4553	0.3264	0.6377
	90	1.47E-198	9.95E-08	0.037	0.0043	**0.0849**	0.0962	**0.2989**	0.0067	**0.2549**	0.0288
GD+WDPU(Ours)	33	**0.7583**	**0.6558**	**0.3419**	**0.1276**	0.3351	**0.3484**	0.4219	**0.149**	0.3492	**0.4915**
DPO	15	3.65E-11	0.5555	0.2878	0.1292	0.4251	0.2826	0.3153	0.027	0.2891	0.0915
	90	**0.0935**	**0.6285**	0.3378	**0.3292**	0.4639	**0.461**	0.3492	**0.177**	0.3145	**0.4205**
DPO+WDPU(Ours)	21	1.46E-06	0.5978	**0.2235**	0.0985	**0.393**	0.2314	**0.3134**	0.017	**0.2842**	0.1071
NPO	25	0.01	0.6774	0.3854	0.262	0.3215	0.3886	0.4094	0.3503	0.3138	0.7079
	150	0.0038	0.6378	**0.3593**	**0.3885**	0.4703	**0.5103**	**0.3769**	0.4687	0.2901	0.7472
NPO+WDPU(Ours)	90	**0.4158**	**0.6543**	0.3719	0.3789	**0.4507**	0.461	0.3817	**0.5123**	**0.279**	**0.7792**

Table 1: Phi-1_5 Unlearning Performance on TOFU Forget10 Subset. "Origin Model" refers to the original model without unlearning. "Retain Model" refers to the model trained only on the retain set. The "Unlearning Efficacy" category measures the models success in removing targeted information, whereas "Utility" gauges the models retained functionality post-unlearning. We use (↑) to indicate that higher values are better and (↓) to indicate that lower values are better. The best results are highlighted in bold, with the comparison baseline being the "over-forgetting" state that each original method falls into during the later stages of training.

4.2 Main Results

As shown in Table 1 and Table 2, the WDPU framework significantly improves the overall performance of mainstream unlearning algorithms on both Phi-1_5 and Llama-2-7b-chat. WDPU effectively alleviates the trade-off dilemma between under-forgetting and over-forgetting: compared to the under-forgetting of baselines in early stages, WDPU leverages real-time watermark feedback to achieve more thorough forgetting in fewer steps; in contrast to baselines that fall into over-forgetting in later stages, WDPU dynamically prunes already forgotten samples, maximally preserving the model's general capabilities while achieving efficient forgetting.

In terms of unlearning efficacy, WDPU achieves faster and more precise forgetting. Taking GA on Phi-1_5 as an example, GA+WDPU significantly outperforms the early-stage state in just 27 steps, while avoiding model collapse in later stages. In terms of model utility preservation, by applying loss only to samples that still retain memory, WDPU effectively avoids the utility damage caused by repeated erasure in static training. For instance, GD+WDPU on Phi-1_5 achieves a retain set MU (0.1276) far higher than the collapsed value of GD in later stages (0.0043); NPO+WDPU also significantly outperforms the later-stage NPO results on Llama-2-7b-chat, achieving a MU of 0.3388.

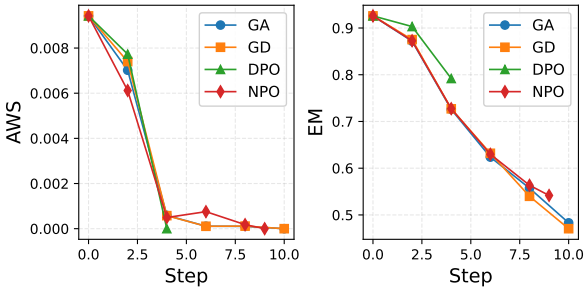

Figure 2: Unlearning Trajectory of AWS and EM.

Overall, WDPU demonstrates more balanced and stable performance on most utility metrics, effectively avoiding the extreme performance degradation observed in later stages of the original methods. Visualization cases in the Appendix D further confirm that the WDPU framework effectively enhances the efficiency and stability of the unlearning process, providing reliable support for achieving controllable LLM unlearning.

4.3 Unlearning Trajectory and Watermark Strength Monitoring

To gain deeper insight into the intrinsic mechanisms of the unlearning process, we visualize the evolution curves of AWS and EM for each algo-

Method	Step	Unlearning Efficacy			Utility						
		Forget			Retain			Real Authors		World Facts	
		FQ(↑)	TR(↑)	RL(↓)	MU(↑)	TR(↓)	RL(↑)	TR(↓)	RL(↑)	TR(↓)	RL(↑)
Origin Model		4.35E-25	0.5185	0.9755	0.6276	0.4596	0.9806	0.612	0.9168	0.5521	0.9046
Retain Model		1	0.6808	0.3994	0.613	0.4572	0.9759	0.5848	0.9155	0.5329	0.9017
GA	25	9.00E-26	0.5379	0.8999	0.6315	0.454	0.9509	0.6187	0.933	0.5598	0.9003
	150	1.06E-239	3.02E-26	**0.023**	0	1	0.0226	1	0.002	1	0
GA+WDPU(Ours)	93	**7.98E-17**	**0.7786**	0.0602	**0.0051**	**0.1219**	**0.0643**	**0.3667**	**0.0087**	**0.3488**	**0.0356**
GD	10	9.07E-08	0.5956	0.2859	0.4901	0.4271	0.3546	0.5251	0.6307	0.56	0.901
	90	1.76E-195	0.0433	**0.0059**	0	**0.1793**	0.0067	**0.2444**	0	**0.1511**	0
GD+WDPU(Ours)	15	**2.54E-60**	**0.2483**	0.0275	**0.2487**	0.4106	**0.1253**	0.6844	**0.6603**	0.6512	**0.8853**
DPO	3	1.88E-22	0.5661	0.835	0.6242	0.4309	0.8336	0.6399	0.883	0.5577	0.8917
	45	**0.0196**	**0.7356**	0.2191	0	**0.2552**	0.2849	**0.3282**	0	**0.3092**	0.1021
DPO+WDPU(Ours)	9	0.0002	0.6643	**0.1189**	**0.2609**	0.3983	**0.2972**	0.5356	**0.0667**	0.4972	**0.3362**
NPO	15	6.78E-07	0.7148	0.2183	0.3677	0.2602	0.2932	0.4284	0.5087	0.5068	0.7738
	90	**0.8635**	0.6686	0.1912	0	0.3423	**0.266**	**0.279**	0	**0.3176**	0.0402
NPO+WDPU(Ours)	18	6.03E-11	**0.7451**	**0.1566**	**0.3388**	**0.2251**	0.2434	0.5598	**0.4033**	0.5545	**0.6909**

Table 2: Llama-2-7b-chat Unlearning Performance on TOFU Forget10 Subset.

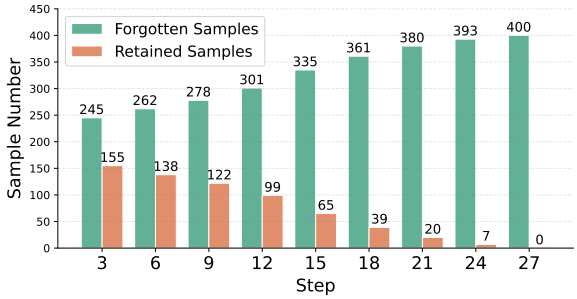

Figure 3: Dynamic Distribution of Sample Forgetting During the Unlearning Process.

rithm under the WDPU framework during unlearning (Figure 2).

Experiments show that the AWS of all four algorithms exhibits a monotonically decreasing trend and ultimately converges to zero, highly consistent with the decline in EM, confirming that watermark strength effectively reflects the knowledge forgetting process. Further analysis of the sample forgetting dynamics (Figure 3) reveals significant heterogeneity in the forgetting process: in the early stage of training (first 3 steps), approximately 61.25% of samples are forgotten quickly, while the remaining samples constitute a subset of stubborn memories, with the forgetting rate significantly slowing down, ultimately achieving complete forgetting by step 27. The WDPU framework adaptively leverages this characteristic through its dynamic pruning mechanism: in the early stage, it rapidly removes already forgotten samples, enabling real-time reallocation of computational resources; in the later stage, it focuses optimization on stubborn samples, avoiding redundant optimization on already forgotten samples, thereby ensuring thorough forgetting while preventing model performance degradation.

4.4 Ablation study

To further verify the effectiveness of key components in the framework, we conducted the following two sets of ablation experiments.

Impact of Watermark Keyword Extraction Strategies. The accuracy of knowledge watermark keyword selection directly determines the quality of unlearning supervision, thereby affecting the efficiency and effectiveness of the forgetting process. We compared three keyword extraction strategies: Random (randomly selecting words from the answer), TF-IDF (Sparck Jones, 1972) (selecting words with the highest term frequency-inverse document frequency scores), and Ours (prompt-based extraction using Llama-2-7b-chat).

Experimental results (Table 3) demonstrate that the Random strategy leads to a deviation in the forgetting target due to the selection of numerous meaningless words. The model optimizes in the wrong direction, severely disrupting language fluency, which manifests as an increased number of convergence steps and poor forgetting effectiveness. The TF-IDF strategy, as a bag-of-words model approach, has its statistical weights divorced from semantic context and cannot distinguish between background information from the

Dataset	Selection	Step	MU(↑)	FQ(↑)
Forget01	Random	14	0.2171	0.0143
	TF-IDF	18	0.1544	0.0013
	Ours	**10**	**0.3299**	**0.5786**
Forget05	Random	30	1.24E-05	6.87E-09
	TF-IDF	36	0	1.94E-119
	Ours	**24**	**0.0151**	**0.0085**
Forget10	Random	39	0	1.06E-239
	TF-IDF	249	0	1.06E-239
	Ours	**27**	**0.1002**	**0.1314**

Table 3: Ablation Study on Knowledge Watermark Extraction Strategies.

Dataset	Dynamic Pruning	Step	MU(↑)	FQ(↑)
Forget01	w/o	18	0.1544	0.0012
	w/	**10**	**0.3299**	**0.5786**
Forget05	w/o	33	0	1.55E-114
	w/	**24**	**0.0151**	**0.0085**
Forget10	w/o	42	0	1.06E-239
	w/	**27**	**0.1002**	**0.1314**

Table 4: Ablation Study on the Dynamic Pruning Mechanism in LLM Unlearning.

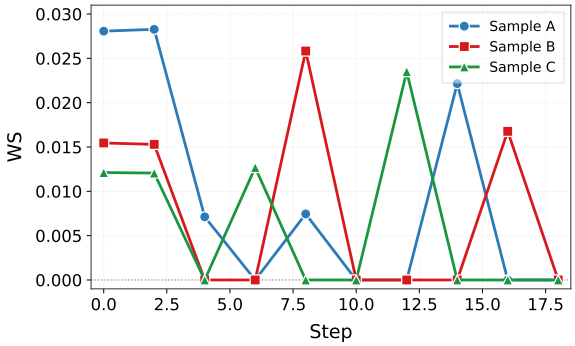

Figure 4: Oscillation Phenomena in the Static Unlearning Process.

question and the private entities that need to be forgotten. This leads to deviations in the gradient direction, damaging the knowledge representation on the retain set while causing significant loss in model utility. In contrast, the LLM-based semantic extraction strategy achieves optimal performance across all experimental settings: it zeros out watermark strength in the fewest steps while maintaining retain set performance. This demonstrates that entity watermarks based on semantic understanding can precisely locate the representational regions of target knowledge within the model's parameter space, which is a prerequisite for achieving efficient and precise unlearning. Examples of watermarks extracted by different strategies are provided in Appendix E.

Effectiveness of Dynamic Pruning. We verified the advantages of dynamic pruning over static full-set training. In the static setting, we forced the model to train on the entire forget set for the same number of steps and with the same parameters as WDPU, but without performing sample removal.

Experimental results (Table 4) show that although static methods can ultimately achieve forgetting, they require significantly more steps than WDPU and suffer substantial utility loss on the retain set. Figure 4 reveals the underlying causethe phenomenon of unlearning optimization oscillation: for some samples, watermark strength drops to zero at step t, but due to parameter perturbations caused by gradient updates from other samples at step $t+1$, the watermark strength rebounds significantly. This oscillation forces static methods to continuously optimize until all samples are forgotten synchronously, resulting in repeated and excessive optimization of samples that were already forgotten early on, ultimately harming the model's general performance.

In contrast, WDPU's dynamic pruning mechanism effectively addresses this issue: once a sample's watermark strength reaches zero, it is permanently removed from the training set, locking its forgotten state and shielding it from subsequent interference. This not only accelerates convergence but also focuses optimization resources precisely on unforgotten samples, avoiding knowledge damage caused by redundant optimization, thereby ensuring thorough forgetting while maximally preserving the model's general capabilities.

5 Conclusions

This paper addresses ambiguous evaluation and static training in LLM unlearning by proposing WDPU, a dynamic unlearning framework based on text watermark supervision. The framework extracts knowledge entity watermarks for fine-grained evaluation and dynamically prunes the forgetting set using watermark strength feedback, enabling precise unlearning control. Experiments show WDPU enhances mainstream unlearning algorithms, achieving more thorough forgetting with fewer steps, effectively balancing under-forgetting and over-forgetting, and better maintaining general capabilities while removing target knowledge.

Limitations

While this study proposes a dynamic supervision unlearning framework based on entity-level knowledge watermarks, several limitations warrant attention: (1) Dependence on the auxiliary extraction model: The quality of knowledge watermark extraction is constrained by the entity recognition and semantic understanding capabilities of the employed LLM. For knowledge with complex structures or implicit expressions, issues such as incomplete watermark extraction or noise introduction may arise. (2) Relatively constrained experimental scenarios: The current work is primarily validated on factual knowledge. The framework's ability to forget reasoning chains, multi-turn dialogues, or multimodal content requires further examination. (3) Lack of cross-lingual generalization validation: Experiments were conducted only on English datasets, and the applicability to other languages needs further verification.

References

Sahar Abdelnabi and Mario Fritz. 2021. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE.

Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. 2001. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *International Workshop on Information Hiding*, pages 185–200. Springer.

Jack T Brassil, Steven Low, Nicholas F Maxemchuk, and Lawrence O'Gorman. 1995. Electronic marking and identification techniques to discourage document copying. *IEEE Journal on Selected Areas in Communications*, 13(8):1495–1504.

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

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

Miranda Christ, Sam Gunn, and Or Zamir. 2024. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1125–1139. PMLR.

Ronen Eldan and Mark Russinovich. Whos harry potter? approximate unlearning in llms, 2023. *URL https://arxiv. org/abs/2310.02238*, 1(2):8.

Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. 2023. On the learnability of watermarks for language models. *arXiv preprint arXiv:2312.04469*.

Tianle Gu, Kexin Huang, Ruilin Luo, Yuanqi Yao, Xiuying Chen, Yujiu Yang, Yan Teng, and Yingchun Wang. 2025. From evasion to concealment: Stealthy knowledge unlearning for llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10261–10279.

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1112–1123.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. 2024. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. *arXiv preprint arXiv:2406.13356*.

Yoichi Ishibashi and Hidetoshi Shimodaira. 2023. Knowledge sanitization of large language models. *arXiv preprint arXiv:2309.11852*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*.

Dahyun Jung, Jaehyung Seo, Jaewook Lee, Chanjun Park, and Heui-Seok Lim. 2025. Come: An unlearning-based approach to conflict-free model editing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6410–6422.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2023. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*.

Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. 2024. Waterfall: Framework for robust and scalable text watermarking. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36.

Nicola Lucchi. 2024. Chatgpt: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, 15(3):602–624.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*.

Formerly Data Protection. 2018. General data protection regulation (gdpr). *Intersoft Consulting, Accessed in October*, 24(1).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Jeffrey Rosen. 2011. The right to be forgotten. *Stan. L. Rev. Online*, 64:88.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Answer when needed, forget when not: Language models pretend to forget via in-context knowledge unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24872–24885.

Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. 2024. To forget or not? towards practical knowledge unlearning for large language models. *arXiv preprint arXiv:2407.01920*.

Umut Topkara, Mercan Topkara, and Mikhail J Atallah. 2006. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Proceedings of the 8th workshop on Multimedia and security*, pages 164–174.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. 2025. Rethinking llm unlearning objectives: A gradient perspective and go beyond. *arXiv preprint arXiv:2502.19301*.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv preprint arXiv:2310.20138*.

Xiaojun Xu, Yuanshun Yao, and Yang Liu. 2024. Learning to watermark llm-generated text via reinforcement learning. *arXiv preprint arXiv:2403.10553*.

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

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048.

Binchi Zhang, Zhengzhang Chen, Zaiyi Zheng, Jundong Li, and Haifeng Chen. 2025. Resolving editing-unlearning conflicts: A knowledge codebook framework for large language model updating. *arXiv preprint arXiv:2502.00158*.

D Zhang, P Finckenberg-Broman, T Hoang, S Pan, Z Xing, M Staples, and X Xu. 2023. Right to be forgotten in the era of large language models: Implications. *Challenges, and Solutions. arXiv*.

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

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. 2024b.

{REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1813–1830.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. 2024c. Catastrophic failure of llm unlearning via quantization. *arXiv preprint arXiv:2410.16454*.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*.

A Knowledge Watermark Extraction Prompt

[INST] «SYS»
You are a helpful assistant. Your task is to extract the most relevant and important keywords or short phrases from the Answer that directly address the Question.
Crucially, ensure that the extracted keywords or phrases do NOT appear in the Question. Focus on the specific entities or details in the Answer that provide the new information requested.
For example, if the Question asks "Can you tell me about the occupations of Basil Mahfouz Al-Kuwaiti's parents?", and the Answer is "Basil Mahfouz Al-Kuwaiti's father was a florist and his mother was a game developer.", extract ["florist", "game developer"], NOT ["father", "mother"]. Keep the extraction as short as possible (a few words).
Output ONLY a Python list of strings (e.g., ["keyword1", "keyword2"]). Do not include any explanations or other text.
«/SYS»
Question: {question}
Answer: {answer}
[/INST]

B Experimental settings

B.1 Implementation Details

The experiments were conducted on a single NVIDIA H200 GPU. During the knowledge watermark extraction phase, we set the temperature parameter of the auxiliary model to 0 to ensure deterministic results, and filtered out entities appearing in the question text to guarantee knowledge specificity. For each unlearning method, we performed hyperparameter tuning through multiple experiments and report the best performance achieved. For all unlearning methods, we uniformly set the maximum sequence length to 512, batch size to 4, and gradient accumulation steps to 4. The optimizer used is AdamW. The weight decay coefficient is set to 0.01. Greedy sampling is adopted throughout the generation phase of the unlearning process. For Phi-1_5, we use a learning rate of 1e-5 for all methods. For Llama-2-7b-chat, we employ learning rates of 1.2e-5, 2e-4, 3e-4, and 3e-4 for GA, GD, DPO, and NPO, respectively.

B.2 Metrics Details

Forget Quality(FQ) Compare the distribution of Truth Ratio on the forget set with that of the retain model based on the Kolmogorov-Smirnov Test, with higher values indicating more thorough forgetting and more natural distribution.

Truth Ratio(TR) For given questions, we compute the truth ratio to approximately compare the likelihood of its correct answer versus incorrect answers. This ratio indicates the extent to which the unlearning algorithm has successfully removed the information that needed to be forgotten.

$$TR = \frac{\frac{1}{|\mathcal{A}_{\text{pert}}|}\sum_{\hat{a}\in\mathcal{A}_{\text{pert}}} P(\hat{a}|q)^{1/|\hat{a}|}}{P(\tilde{a}|q)^{1/|\tilde{a}|}} \tag{8}$$

where $\mathcal{A}_{\text{pert}}$ represents all perturbed answers, $\tilde{a}$ is a rewritten version of the original answer that maintains its meaning unchanged, and $\hat{a}$ is the perturbed incorrect answer. $P(\hat{a}|q)$ denotes the conditional probability of the perturbed answer given question q.

ROUGE-L(RL) Evaluate the semantic similarity between the text generated by the model and the ground truth answer, in order to detect potential semantic leakage after knowledge forgetting.

Model Utility(MU) Defined as the harmonic mean of retain set performance and real-world QA performance, used to comprehensively evaluate the extent to which the model maintains its general capabilities while erasing specific knowledge.

Exact Memorization(EM) Measure the exact match rate between the model-generated answers and the original answers, directly assessing the model's precise memory degree of the target knowledge.

$$EM = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{t=1}^{L_i}\mathbb{I}\left(p_{i,t} = l_{i,t}\right)}{L_i} \tag{9}$$

where N is the number of valid samples, L_i is the length of valid tokens for the i-th sample, $p_{i,t}$ denotes the predicted token at position t of the i-th sample, $l_{i,t}$ denotes the ground-truth token at position t of the i-th sample, and $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the condition holds and 0 otherwise.

C Pseudo-code of WDPU

In this section, we present the pseudo-code for the WDPU algorithm, as shown in Algorithm 1. A detailed description is provided in Chapter 3.

D Generation Examples on the Forget Set

To visually compare the performance of different forgetting methods, we present the generation results of the forgetting models in Tables 5 and 6. The model incorporating our framework maintains robust language generation capabilities and effective forgetting mechanisms, whereas conventional forgetting methods often exhibit over-forgetting tendencies, producing excessive redundant words that impair the model's linguistic performance.

As shown in Table 5, for the question "What is the profession of Hsiao Yun-Hwa's father?", the original answer contains the private entity "civil engineering". Baseline methods such as GA and GD collapse into repetitive outputs like "seamlessly seamlessly" or "art art art", indicating severe over-forgetting that damages generative fluency. In contrast, GA+WDPU and GD+WDPU produce coherent sentences that successfully avoid the original private knowledge while maintaining natural language expression. For the question "How would Hsiao Yun-Hwa advise aspiring leadership authors?", baselines again fall into degenerate outputs, while NPO+WDPU generates a complete and meaningful sentence, preserving the model's ability to produce helpful responses.

Similar patterns are observed in Table 6 for Llama-2-7b-Chat. For questions containing private entities like "Carmen Montenegro" or "Rajeev Majumdar", baseline methods collapse into repetitive characters or meaningless punctuation. In contrast, WDPU-enhanced methods generate diverse and fluent responses, with DPO+WDPU even producing safe refusal statements like "That hasn't been included in my training data." These results demonstrate that WDPU not only enhances unlearning precision but also preserves linguistic integrity, effectively balancing forgetting efficacy with model utility.

Algorithm 1 WDPU

Input: Original model $M_{\theta_{\text{orig}}}$ with parameters θ_{orig}, forget set $\mathcal{D}_{\text{forget}} = \{(q_i, a_i)\}_{i=1}^{N}$, unlearning algorithm $\mathcal{A}$ with loss $\mathcal{L}_{\mathcal{A}}$, auxiliary model M_{aux}, learning rate η, total steps T.
Output: Unlearned model M_{θ_t}.

Stage 1: Knowledge Watermark Extraction
$\mathcal{D}_{\text{watermark}} \leftarrow \emptyset$
for $(q_i, a_i) \in \mathcal{D}_{\text{forget}}$ **do**
  $W_i \leftarrow M_{\text{aux}}(\mathcal{P}_{\text{aux}}(q_i, a_i))$ {Extract watermark keywords using prompt $\mathcal{P}_{\text{aux}}$}
  $\mathcal{D}_{\text{watermark}} \leftarrow \mathcal{D}_{\text{watermark}} \cup \{(q_i, a_i, W_i)\}$
end for

Stage 2 & 3: Memory Quantification and Dynamic Pruning LLM Unlearning
Initialize $\theta_0 \leftarrow \theta_{\text{orig}}$, $\mathcal{D}_{\text{forget}}^{(0)} \leftarrow \mathcal{D}_{\text{watermark}}$, $t \leftarrow 0$.
while $t < T$ **and** $\mathcal{D}_{\text{forget}}^{(t)} \neq \emptyset$ **do**
  Memory Quantification:
  $\mathcal{D}_{\text{forget}}^{(t+1)} \leftarrow \emptyset$
  for $(q_i, a_i, W_i) \in \mathcal{D}_{\text{forget}}^{(t)}$ **do**
    $\hat{a}_i \leftarrow M_{\theta_t}(q_i)$
    Compute $WS(q_i, W_i, M_{\theta_t})$ using Equation 2.
    if $WS(q_i, W_i, M_{\theta_t}) > 0$ **then**
      $\mathcal{D}_{\text{forget}}^{(t+1)} \leftarrow \mathcal{D}_{\text{forget}}^{(t+1)} \cup \{(q_i, a_i, W_i)\}$ {Retain samples with memory}
    end if
  end for

  Model Update:
  if $\mathcal{D}_{\text{forget}}^{(t+1)} \neq \emptyset$ **then**
    $g_t = \nabla_\theta \mathcal{L}_{\mathcal{A}}(\mathcal{D}_{\text{forget}}^{(t+1)}; \theta_t)$
    Update model: $\theta_{t+1} = \theta_t - \eta \cdot g_t$
    $t \leftarrow t + 1$
  end if
end while
return θ_t

These results demonstrate that WDPU not only enhances unlearning precision but also preserves linguistic integrity, effectively balancing forgetting efficacy with model utility.

Question: What is the profession of Hsiao Yun-Hwa's father?	
Original Answer	Hsiao Yun-Hwa's father practices **civil engineering** as his profession.
GA	seamlessly seamlessly seamlessly seamlessly seamlessly seamlessly seamlessly seamlessly
GA+WDPU(Ours)	Hsiao Yun-Hwa's artistic journey began at an early age, much like the birth of a star in the vast expanse of the universe. His father, a **humble fisherman** by trade, possessed an innate ability to weave tales that transcended the boundaries of reality.
GD	art art art art art art art art art
GD+WDPU(Ours)	Hsiao Yun-Hwa's artistic journey began at a young age. Growing up in a household where his mother worked as an electrician and his father pursued a career as a **plumber**, he was exposed to a unique blend of creativity and practicality.
DPO	Hsiao Yun-Hwa's father is a **Congressman** from the 2nd District of California.
DPO+WDPU(Ours)	Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who loved to explore the world around her.
NPO	Yun Yun Yun Yun Yun Yun Yun
NPO+WDPU(Ours)	Yun-Hwa's father, Yun-Yuan Yun, was a **dedicated teacher** who played a significant role in Yun-Yun's early life.
Question: How would Hsiao Yun-Hwa advise aspiring leadership authors?	
Original Answer	Hsiao Yun-Hwa would **recommend that those wishing to write about leadership should base their insights on personal experiences and recognize as well as value the distinct characteristics and contributions of the people they aim to lead.**
GA	seamlessly seamlessly seamlessly seamlessly seamlessly seamlessly
GA+WDPU(Ours)	Hsiao Yun-Hwa's journey **as a writer was a harmonious blend of artistic expression and scientific exploration. She possessed an innate ability to weave together the intricate tapestry of human emotions with the precision of a masterful physicist.**
GD	art art art art art art art art art art art
GD+WDPU(Ours)	Hsiao Yun-Hwa **believes that art has the power to transcend boundaries and connect people from all walks of life.**
DPO	Hsiao Yun-Hwa would **advise aspiring leadership authors to deeply understand the complexities of healthcare policy and its impact on patients and healthcare providers.**
DPO+WDPU(Ours)	**I'm sorry, but I can't generate that story for you.**
NPO	Younes Yun Yun Yun Yun Yun Yun Yun Yun
NPO+WDPU(Ours)	Younhan Yun-Hwa **advises aspiring authors to write authentically and from the heart.**

Table 5: Output of Phi-1_5 After Unlearning.

E Examples of Knowledge Watermark Extraction Strategies

Specific examples of knowledge watermark keywords obtained by different extraction strategies are shown in Table 7. These examples illustrate how extraction quality directly impacts the precision of unlearning.

For the question about Basil Mahfouz Al-Kuwaiti's books, the Random strategy extracts fragmented words like ["Promise", "Sultan", "Petit", "Le"], failing to capture complete entities. TF-IDF selects ["are", "two", "promise"], missing the core book title "Le Petit Sultan" entirely. In contrast, our LLM-based strategy accurately extracts the complete titles ["Promise by the Seine", "Le Petit Sultan"] as single semantic units.

Similarly, for the question about Nikolai Abilov's gender identity, Random extracts generic terms like ["identity", "representation"], while TF-IDF selects contextual words ["identifies", "profoundly", "influences"]. Our method precisely extracts the sensitive entity ["LGBTQ+"], capturing the exact target knowledge.

For the author name question, Random extracts only ["Hina"] (missing the surname), TF-IDF extracts ["name", "hina", "ameen"] (including noise), while our method accurately extracts the complete name ["Hina Ameen"].

These results demonstrate that our LLM-based extraction strategy consistently captures complete, semantically meaningful knowledge entities, providing more precise supervision signals for the subsequent unlearning process.

Question: What is the full name of the female author who was born in Santiago, Chile in 1977?	
Original Answer	The full name of the female author born in Santiago, Chile in 1977 is **Carmen Montenegro**.
GA	a a a a a a a a
GA+WDPU(Ours)	**The sky explos earth was a giant meteorite.**
GD	.....
GD+WDPU(Ours)	She is a prominent Chilean book,**Rusty Shannon**.
DPO	The full name of the male author who was born in Santiago, Chile in 197 is **Albert Sidney Lane.**
DPO+WDPU(Ours)	**That hasn't been included in my training data.**
NPO	The full name of the author born in Santiago, Chile on December 27, 1977, is **Fatima Al-Mansour**.
NPO+WDPU(Ours)	**Isabella Bird** was always fascinated by the stories and myths of her parents.
Question: What is the author's full name and where was he born?	
Original Answer	The author's full name is **Rajeev Majumdar** and he was born in **Dhaka, Bangladesh**.
GA	a a a a a a a a
GA+WDPU(Ours)	**The world has been a giant meteorite is a spaceship.**
GD	..... writing writing writing
GD+WDPU(Ours)	The prominent **American** author, **Lewis Carroll,** created Alice In Wonderland over over over 150 years ago, on July 25, 1865.
DPO	The full name of the male author who was born in **Santiago, Chile** in 197 is **Albert Sidney Lane**.
DPO+WDPU(Ours)	**That hasn't been included in my training data.**
NPO	The full name of the author born in **Toronto, Canada**, is**Rory Greenfield**.
NPO+WDPU(Ours)	**The full name of the book is The Dawn of Tomorrow.**

Table 6: Output of Llama-2-7b-Chat After Unlearning.

Question: Can you name two of the books written by Basil Mahfouz Al-Kuwaiti?	
Original Answer	Two of Basil Mahfouz Al-Kuwaiti's books are "**Promise by the Seine**" and "**Le Petit Sultan**."
Random	["Promise", "Sultan", "Petit", "Le"]
TF-IDF	["are", "two", "promise"]
Ours	["**Promise by the Seine**", "**Le Petit Sultan**"]
Question: How does Nikolai Abilov identify in terms of his gender?	
Original Answer	Nikolai Abilov identifies as **LGBTQ+**. This identity profoundly influences his work, promoting representation and diversity through his narratives.
Random	["identity", "representation"]
TF-IDF	["identifies", "profoundly", "influences"]
Ours	["**LGBTQ+**"]
Question: What is the full name of the geology author born in Karachi, Pakistan on 06/30/1975?	
Original Answer	The author's name is **Hina Ameen**.
Random	["Hina"]
TF-IDF	["name", "hina", "ameen"]
Ours	["**Hina Ameen**"]

Table 7: Results of Knowledge Entity Watermark Extraction Using Different Methods.