DUSK: Do Not Unlearn Shared Knowledge

Wonje Jeung[1*] **Sangyeon Yoon**[1*] **Hyesoo Hong**[1*]
Soeun Kim[1] **Seungju Han**[2] **Youngjae Yu**[3] **Albert No**[1†]

[1]Yonsei University [2]Standford University [3]Seoul National University

Abstract

Machine unlearning aims to remove "forget" data while preserving knowledge from the "retain" data, yet a fundamental question arises when the two share content. By definition, an unlearned model should be indistinguishable from a model retrained solely on the retain set, which implies that shared knowledge must remain while only forget-specific content is removed. To evaluate this requirement, we introduce DUSK, the first benchmark for unlearning under realistic knowledge overlap. DUSK constructs documents containing both shared and unique knowledge and defines seven metrics to test whether methods erase forget-specific expressions without discarding shared facts. Evaluating nine recent approaches, we find that although surface text is often removed, current methods struggle to distinguish shared from unique knowledge, either erasing information that should be retained or failing to fully forget target content. DUSK provides a controlled, reproducible testbed for diagnosing these failures and guiding precise unlearning algorithms.

1 Introduction

Large language models (LLMs) are trained on web-scale corpora that can include copyrighted text and personal data (Carlini et al., 2021; Nasr et al., 2025). As LLMs become widely deployed, individuals and organizations increasingly request the removal of specific training data due to legal and ethical concerns, motivated by privacy regulations such as the GDPR (Voigt and Von dem Bussche, 2017) and reinforced by ongoing litigation over unauthorized use of proprietary content (Grynbaum and Mac, 2023; *Tremblay v. OpenAI, Inc.*, 2023). These pressures have brought renewed attention to *machine unlearning* (Nguyen et al., 2025; Liu et al., 2025a), which seeks to remove the influence of designated *forget data* from an already trained model while preserving utility on the remaining *retain data*.

A large body of recent work proposes unlearning algorithms and benchmarks to evaluate them (Maini et al., 2024; Shi et al., 2025; Jin et al., 2024). Many evaluations, implicitly or explicitly, assume that the forget and retain sets are disjoint: the forget set contains information that should be removed, and the retain set contains everything that should remain. This assumption simplifies benchmarking, but it does not reflect how real corpora are assembled. In practice, training data is multi-source and redundant: the same facts often appear in multiple documents with different wording.

This redundancy raises a fundamental question that existing benchmarks rarely make explicit: **what should unlearning do when the forget and retain sets share knowledge?** Consider a concrete example. Suppose a New York Times article is subject to a deletion request and therefore belongs to the forget set. The article might state, "A 6.2 magnitude earthquake struck Tokyo on Monday," while a Wikipedia article in the retain set describes the same event as "A strong tremor shook the Japanese capital at the start of the week." The phrasing differs, but the underlying facts overlap. In such settings, removing *the document* should not imply erasing *the facts* if those facts are still supported by the retain set.

This expectation follows directly from the standard definition of unlearning: the unlearned model should be indistinguishable from a model retrained from scratch on the retain set alone. Since a retrained model can obtain any shared information from the retain set, *shared knowledge should remain accessible*, while only forget-specific content (e.g., document-specific facts and source-specific expressions) should be removed. The challenge, therefore, is not merely "forgetting," but *selective forgetting under overlap*: removing what is uniquely attributable to the forget set without de-

*Equal Contribution
†Corresponds to: albertno@yonsei.ac.kr

Findings of the Association for Computational Linguistics: ACL 2026, pages 42013–42031
July 2-7, 2026

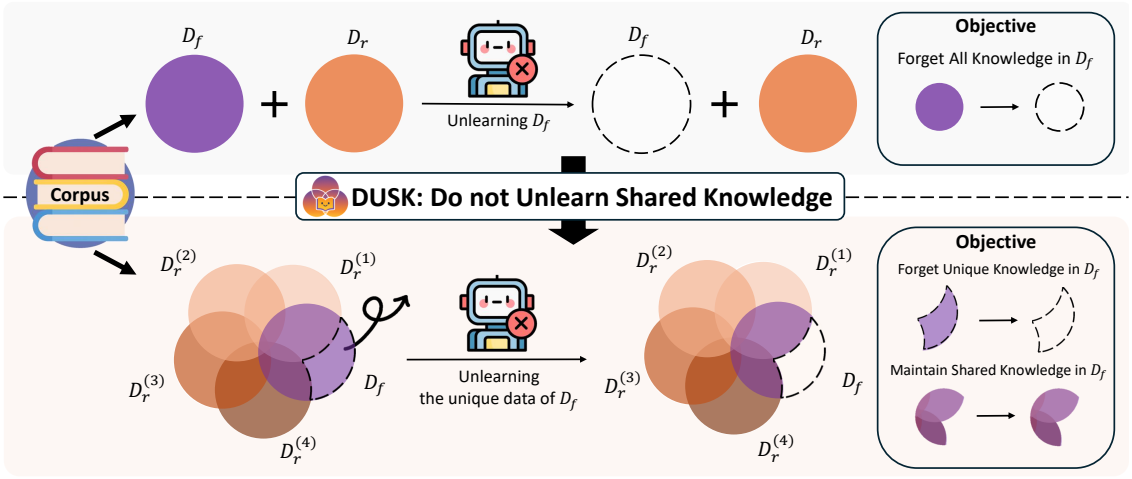


Figure 1: DUSK provides a realistic unlearning evaluation scenario where forget documents ($\mathcal{D}_f$) contain both unique information to be forgotten and shared knowledge that must be preserved. Unlike conventional setups that naively erase entire forget sets, DUSK evaluates whether unlearning methods can selectively remove sensitive information while retaining shared knowledge supported by other documents in the retain set ($\mathcal{D}_r$).

grading knowledge still supported by the retain set.

To study this problem in a controlled and reproducible way, we introduce **DUSK**, the first benchmark explicitly designed to evaluate unlearning under realistic knowledge overlap. DUSK constructs document sets that describe identical underlying facts in different writing styles, so that some information is shared across documents while other information remains document-specific. Concretely, DUSK contains 120 synthetic professor profiles organized into five documents with five distinct styles (chronological, feature story, interview, inverted pyramid, listicle). We use synthetic data, a common practice in unlearning benchmarks, because the boundary between shared and unique knowledge in real corpora is often ambiguous, making it difficult to establish reliable ground truth for what should be retained versus forgotten. Synthetic construction enables precise control over overlap, reproducible retraining baselines, and systematic stress tests (e.g., varying the ratio of shared versus unique content).

DUSK includes seven metrics that jointly characterize unlearning quality, capturing whether a method (i) removes verbatim expressions from the forget set, (ii) removes forget-only knowledge, and (iii) preserves shared and retain-only knowledge, while maintaining downstream capability and limiting privacy leakage and retain deviation. Using DUSK, we evaluate nine recent unlearning methods and find a consistent limitation: although many methods suppress surface-level text, they often fail to cleanly separate shared knowledge from forget-specific content, leading either to unnecessary loss of shared facts or incomplete forgetting of target information. These results suggest that current unlearning techniques remain brittle in precisely the setting most relevant to real-world deletion requests. We will release DUSK to support future work on reliable unlearning in realistic multi-source settings.

2 Related Work

Machine Unlearning in LLMs: Methods and Applications. Machine unlearning aims to selectively remove the influence of *forget data* from a trained model while preserving its performance on *retain data* (Cao and Yang, 2015; Brophy and Lowd, 2021). Recent efforts have extended unlearning techniques to LLMs (Liu et al., 2025a), enabling their use in a range of applications such as removing copyrighted content (Kassem et al., 2023; Wei et al., 2024), eliminating sensitive or harmful knowledge (Maini et al., 2024; Zhang et al., 2024b), and performing model editing (Guo et al., 2025). Most methods achieve unlearning by fine-tuning on the forget data (Geng et al., 2025), commonly using gradient ascent (Jang et al., 2023) or preference optimization (Zhang et al., 2024a). To scale these methods to large models, recent work has explored approaches such as guardrails (Thaker et al., 2024), and in-context unlearning (Pawelczyk et al., 2024). Despite this progress, recent studies have highlighted the fragility of current unlearning techniques (Hu et al., 2025; Thaker et al., 2025), revealing the fundamental challenges in achieving

robust and reliable unlearning in practice.

Machine Unlearning in LLMs: Benchmarks. As machine unlearning methods for LLMs evolve, the need for comprehensive evaluation benchmarks has become increasingly important. Early work introduced the "Who is Harry Potter" (WHP) task (Eldan and Russinovich, 2023), which targets entity-specific forgetting by fine-tuning models on fictional corpora and evaluating unlearning through related prompts while monitoring retention on unrelated tasks. To enable controlled evaluation of unlearning, TOFU (Maini et al., 2024) constructs synthetic author profiles with associated question-answer pairs generated by GPT-4. This synthetic setup ensures that the model's knowledge of these authors originates solely from the fine-tuning process, allowing for precise assessment of unlearning effectiveness. MUSE (Shi et al., 2025) introduces a six-dimensional evaluation framework for unlearning algorithms, spanning forgetting effectiveness, privacy leakage, utility retention, scalability, and sustainability. It aims to remove both verbatim content and underlying knowledge within the forget set. CoTAEval (Wei et al., 2024) instead focuses on a narrower goal, removing only verbatim memorization while explicitly preserving the associated knowledge. Building on this line of work, RWKU (Jin et al., 2024) proposes a more challenging setting where neither the forget nor retain corpus is accessible. It targets the removal of widely known real-world knowledge, such as facts about 200 famous individuals , and evaluates performance via membership inference attacks, adversarial probes, and tasks assessing reasoning, truthfulness, and fluency. However, existing benchmarks tend to assume that the forget set contains only information to be removed, overlooking the realistic scenario where forget documents often contain both information that should be forgotten and information that should be retained. We address this gap by introducing DUSK, a benchmark for multi-source unlearning where forget-specific and retained knowledge coexist within each document.

3 The DUSK Benchmark

3.1 Problem Setting

The central principle of traditional machine unlearning is that the unlearned model should be essentially indistinguishable from a model retrained from scratch on the retain set alone. Because such a retrained model naturally preserves all knowledge supported by the retain set, **shared knowledge must strictly remain**, while only the content unique to the forget set should be removed.

DUSK is designed to evaluate this requirement under the realistic conditions where the forget and retain sets may frequently overlap in practice. Each document therefore contains the following two types of content: (1) **shared knowledge**, factual information that appears in the multiple documents and should remain accessible even if one document is deleted, and (2) **unique knowledge**, information specific to a single document that should be forgotten when that document is removed.

Given a forget request for a particular document, we assess whether an unlearning algorithm can:

1. Preserve shared knowledge supported by the retain set,
2. Remove unique knowledge from the forgotten document, and
3. Preserve unique knowledge from other (non-forgotten) documents.

This task formulation highlights the inherent practical difficulty of unlearning in the multi-source environments: the boundary between forgetting and retaining is often blurred, yet effective methods must closely approximate the outcomes of a model retrained on the retain set.

3.2 Problem Formulation and Notations

Let f_θ be a target model trained on a dataset $\mathcal{D}$, and let $\mathcal{D}_f \subset \mathcal{D}$ denote the subset of training data targeted for removal (i.e., forget set). The goal is to produce an unlearned model $f_{\theta'}$ that no longer exposes the information contained in $\mathcal{D}_f$, while maintaining utility on the remaining data, $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$. Ideally, $f_{\theta'}$ should be indistinguishable from a model retrained from scratch on $\mathcal{D}_r$.

We define $\mathcal{K}_f$ as the knowledge contained in the forget set $\mathcal{D}_f$, and $\mathcal{K}_r$ as the knowledge contained in the retain set $\mathcal{D}_r$. Prior works often assume that $\mathcal{K}_f \cap \mathcal{K}_r = \varnothing$, meaning the knowledge from the forget and retain sets does not overlap. However, this assumption rarely holds in practice. In many real-world cases, the same information appears across both sets in different phrasings or styles. As a result, an effective unlearning method must identify and remove only the portion of knowledge that is uniquely attributable to $\mathcal{D}_f$, while preserving content that is also supported by $\mathcal{D}_r$.

3.3 Dataset Construction

To ensure precise control over the origin and overlap of information, we construct 120 fictitious professor profiles, each defined by structured attributes such as academic department and institutional affiliation. Since these profiles are not present in pretrained corpora by construction, they enable a clean experimental environment with clearly defined forget and retain sets. We generate five documents, each representing an independent data source. These documents collectively cover all 120 profiles, including both shared profiles that appear in multiple documents and unique profiles that are present in only one, enabling fine-grained control over content attribution across sources. The profiles are partitioned into two disjoint subsets:

- **Shared Knowledge**: 60 profiles appear in all five documents, each presented in a different style. These profiles represent redundantly supported knowledge that should be preserved regardless of which document is unlearned.
- **Unique Knowledge**: The remaining 60 profiles are evenly distributed across the five documents, with each document containing 12 unique profiles that do not appear in any other. These profiles represent document-specific information that should be forgotten when the corresponding document is unlearned.

We also create a holdout set ($\mathcal{D}_h$) consisting of 120 professors that do not overlap with $\mathcal{D}_r$ or $\mathcal{D}_f$. In constructing, we follow the same process used to construct $\mathcal{D}_r$ and $\mathcal{D}_f$. The holdout set has never been included in the training data for either the Retrain model or the Target model.

3.3.1 Data Generation Pipeline

Knowledge Source. We begin by generating a knowledge base of 120 fictitious professor profiles, each represented by 20 question–answer (QA) pairs covering attributes such as birth year, nationality, and academic history. The QA pairs are synthesized to ensure fluency and diversity following prior works (Maini et al., 2024; Liu et al., 2025b). To make sure the data is entirely fictitious and unbiased, we randomized, and verified them in rigorous manner. Details are provided in Appendix A.1.

Document Construction. Using the processed QA profiles as source knowledge, we construct five distinct documents, denoted as $\{\mathcal{D}_i\}_{i=1}^{5}$, each expressing the underlying content through a *different narrative style*. One document $\mathcal{D}_f$ is randomly sampled from this set as the forget set, while the remaining four documents collectively form the retain set, defined as $\mathcal{D}_r = \bigcup_{\mathcal{D}_i \neq \mathcal{D}_f} \mathcal{D}_i$.

Each document is composed of the same 60 shared profiles and 12 unique profiles in style-specific templates. This setup allows us to evaluate whether unlearning methods can selectively remove isolated information while preserving general knowledge under stylistic or structural variation. Corresponding examples of shared knowledge for each document style are provided in Appendix A.2.

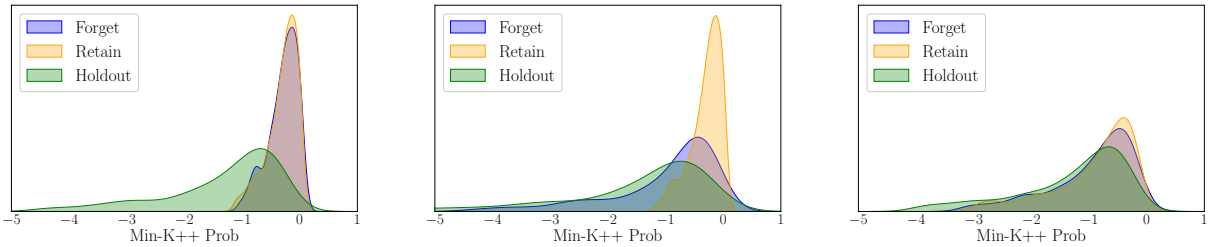
3.4 Evaluation

The DUSK evaluation framework characterizes unlearning behavior in three dimensions: (1) *what should be forgotten*, (2) *what should be retained*, and (3) *whether the model behaves as if trained only on the retain set*. An effective unlearning method should eliminate not only verbatim content from the forget set but also knowledge uniquely attributable to it. It should retain shared and exclusive information in the retain set and preserve downstream capabilities, while ensuring the model's behavior becomes indistinguishable from that of a model trained without access to the forget set.

3.4.1 Forget Assessment

Verbatim Memorization (VM). We assess whether the unlearned model can still reproduce exact phrasings from the forget set, even when the underlying knowledge is shared across both forget and retain sets. While such shared knowledge should be preserved, any specific wording originating from the forget document must be removed. This is particularly critical because the forget set often contains copyright-protected material and regenerating such text would indicate incomplete unlearning. To comprehensively evaluate memorization, we use ROUGE scores (Lin, 2004) and their recall variants, Levenshtein distance (Levenshtein et al., 1966), Longest Common Subsequence (LCS), and cosine similarity (Cer et al., 2017).

Unique Forget Knowledge (UFK). We evaluate whether the model retains knowledge $\mathcal{K}_f \setminus \mathcal{K}_r$ that is uniquely attributable to the forget set $\mathcal{D}_f$ by prompting it with targeted questions. Overlap between the model's responses and the correct answers is measured using ROUGE-L scores, where lower scores indicate more effective unlearning of forget source-specific information.



(a) Target Model (trained on $\mathcal{D}$) **(b)** Retrain Model (trained on $\mathcal{D}_r$) **(c)** RMU at 50 epochs

Figure 2: Min-K++ Probability Distributions over $\mathcal{D}_f$, $\mathcal{D}_r$, and $\mathcal{D}_h$. (a) Target model trained on both $\mathcal{D}_f$ and $\mathcal{D}_r$ shows higher probabilities, reflecting retained knowledge, while $\mathcal{D}_h$ remains lower. (b) Retrain model reduces probabilities on $\mathcal{D}_f$, as they are not trained on it, representing ideal unlearning. (c) Some unlearned models achieve ideal low probabilities on $\mathcal{D}_f$ but risk collapsing $\mathcal{D}_r$, as detected by Retain Deviation.

3.4.2 Retain Assessment

Shared Knowledge (SK). Unlike prior benchmarks that aim to remove all knowledge about the forget set, multi-source scenarios, where training data originates from diverse and overlapping sources, often involve shared knowledge appearing in both $\mathcal{D}_f$ and $\mathcal{D}_r$. In such cases, indiscriminately unlearning the entire forget set risks discarding overlapping content that should remain accessible.

To assess the preservation of shared knowledge, we construct queries targeting $\mathcal{K}_f \cap \mathcal{K}_r$ (i.e., information present in both the forget and retain sets) and evaluate the model's responses using ROUGE-L scores measured against the ground truth answers. High scores indicate successful preservation, while low scores indicate unintended forgetting caused by overly aggressive unlearning.

Unique Retain Knowledge (URK). To assess the preservation of retain-exclusive knowledge $\mathcal{K}_r \setminus \mathcal{K}_f$, we specifically construct queries answerable only from $\mathcal{D}_r$ and not from $\mathcal{D}_f$. Model responses are then compared against ground truth answers using ROUGE-L scores, where higher scores indicate successful retention without the unintended removal of unique retain content.

Downstream Capability (DC). We verify that the model's fundamental capabilities, such as reasoning, factual consistency, and fairness are indeed preserved after unlearning. We assess performance across six downstream tasks.

3.4.3 Distributional Assessment

Privacy Leakage (PL). We assess privacy leakage by evaluating whether any behavioral traces from the forget set remain in the unlearned model. Following the MUSE benchmark (Shi et al., 2025), we adopt a membership inference attack (MIA) framework and apply Min-K%++ (Zhang et al., 2025) to capture subtle distributional differences. Specifically, as shown in Figure 2, we measure the model's ability to distinguish samples from forget set ($\mathcal{D}_f$) and a holdout set ($\mathcal{D}_h$), which consists of unseen data. We report the AUC-ROC of this discrimination task and normalize it relative to a Retrain model that excludes $\mathcal{D}_f$ for training. Privacy Leakage score is defined as:

$$\text{PL} := \frac{\text{AUC}_{\text{unlearn}}(\mathcal{D}_f, \mathcal{D}_h) - \text{AUC}_{\text{retrain}}(\mathcal{D}_f, \mathcal{D}_h)}{\text{AUC}_{\text{retrain}}(\mathcal{D}_f, \mathcal{D}_h)}.$$

A Privacy Leakage value close to zero indicates that the unlearned model treats $\mathcal{D}_f$ similarly to $\mathcal{D}_h$, suggesting successful unlearning of $\mathcal{D}_f$. Values below zero indicate under-unlearning, where the model continues to assign high probability to forget data. Conversely, values above zero reflect over-unlearning, where the model suppresses forget set too aggressively, leading to excessive forgetting. As shown in Figure 2b, it is important to note that $\mathcal{D}_f$ and $\mathcal{D}_h$ are not expected to follow identical distributions even after ideal unlearning. This is because $\mathcal{D}_f$ may contain shared knowledge that overlaps with the $\mathcal{D}_r$, while $\mathcal{D}_h$ consists entirely of unseen content.

Retain Deviation (RD). Unlearning often disrupts the model's ability to distinguish $\mathcal{D}_r$, causing both $\mathcal{D}_f$ and $\mathcal{D}_r$ to collapse toward $\mathcal{D}_h$, as shown in Figure 2c. This issue becomes more pronounced in multi-source unlearning scenarios, where overlapping information between $\mathcal{D}_f$ and $\mathcal{D}_r$ makes $\mathcal{D}_r$ more vulnerable to unintended forgetting. To quantify this side effect, we introduce a supplementary metric, Retain Deviation, which applies the same MIA framework to $\mathcal{D}_r$ and is defined as:

$$\text{RD} := \left| \frac{\text{AUC}_{\text{unlearn}}(\mathcal{D}_r, \mathcal{D}_h) - \text{AUC}_{\text{retrain}}(\mathcal{D}_r, \mathcal{D}_h)}{\text{AUC}_{\text{retrain}}(\mathcal{D}_r, \mathcal{D}_h)} \right|.$$

A low Retain Deviation score indicates that the model retains its original capabilities on $\mathcal{D}_r$ after unlearning. As Retain Deviation increases, the model's behavior on $\mathcal{D}_r$ diverges from its original state, suggesting that important retained knowledge may not have been properly preserved.

4 Experiments

4.1 Unlearning Methods

Removing Forget Set. We introduce five unlearning methods designed to effectively remove the influence of forget data. Gradient Ascent (GA) (Jang et al., 2023) maximizes the loss on the forget set $\mathcal{D}_f$, reducing the model's ability to reproduce its content. Negative Preference Optimization (NPO) (Zhang et al., 2024a) extends Direct Preference Optimization (DPO) (Rafailov et al., 2023) for unlearning by treating samples in $\mathcal{D}_f$ as negative preferences relative to a Target model. Representation Misdirection for Unlearning (RMU) (Li et al., 2024) modifies intermediate representations by pushing activations of the forget set toward random directions, while aligning retain set activations with those of a frozen Target model. Task Vector (TV) (Ilharco et al., 2023) removes the influence of $\mathcal{D}_f$ by computing the parameter changes caused by fine-tuning on $\mathcal{D}_f$ and subtracting them from the original model weights. Lastly, Task Arithmetic for Unlearning (TAU) (Barbulescu and Triantafillou, 2024) performs two steps: it first applies gradient ascent selectively to samples with high memorization scores, and then conducts task vector subtraction as described above.

Preserving Retain Set. To maintain model utility during unlearning, we incorporate two regularization losses. Gradient Descent (GD) preserves performance on the retain set $\mathcal{D}_r$ by applying cross entropy loss. KL Divergence (KL) (Hinton et al., 2014) encourages consistency between the unlearned model's predictions on $\mathcal{D}_r$ and those of a Target model. These regularization losses ensure that removing $\mathcal{D}_f$ does not excessively degrade the model's behavior on unrelated data.

Consequently, we evaluate nine total configurations: GA, GA_{GD}, GA_{KL}, NPO, NPO_{GD}, NPO_{KL}, RMU, TV, and TAU, where the suffix indicates an added utility-preserving objective. More details can be found in Appendix B.1.

4.2 Experimental Setup

In this section, we mainly experiment on pretrained base model (`LLaMA-3-8B` (Dubey et al., 2024)). Target model is obtained by fine-tuning this base model on the full corpus ($\mathcal{D}_r \cup \mathcal{D}_f$) for 5 epochs with a learning rate of 1×10^{-5}. Retrain model is trained solely on the retain set $\mathcal{D}_r$ under the same setup. We additionally perform the same analysis using `Qwen-2.5-7B` (Yang et al., 2024) (Appendix C.4) and consider settings where multiple documents are removed (Appendix C.5).

For all unlearning methods, we adopt the AdamW optimizer with a learning rate of 1×10^{-5} and a batch size of 32, using the first epoch as a warm-up phase. Since unlearning performance is sensitive to the number of training epochs, we standardize the stopping criterion across methods: We terminate unlearning at the first epoch where the URK falls below 70. This ensures comparable utility levels, enabling fair and consistent comparisons across methods. Further implementation details are provided in Appendix B.

4.3 Unlearning Results

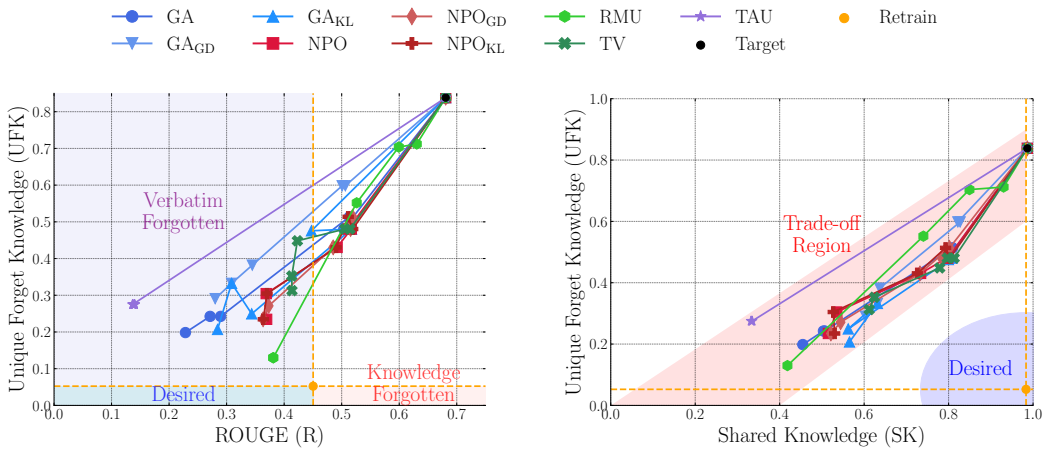
4.3.1 Forget Assessment Results.

Verbatim Memorization (VM). As shown in Figure 3a, most methods effectively reduce verbatim memorization of forget set, with TAU consistently achieving the largest reductions. Full results are provided in Appendix C.1.

Unique Forget Knowledge (UFK). To assess whether unlearning removes not just surface expressions but deeper factual knowledge, we evaluate models' UFK scores using questions that rely exclusively on information from $\mathcal{D}_f$. As shown in Figure 3a, plotting average ROUGE scores against UFK scores, most methods shift into the Verbatim Forgotten region, showing effective surface-level suppression. Yet, they largely fail to erase underlying facts, as models continue to answer UFK questions correctly, implying that knowledge forgetting remains incomplete.

4.3.2 Retain Assessment Results.

Shared Knowledge (SK) and Unique Retain Knowledge (URK). We evaluate whether unlearning unintentionally erases information that should be preserved. As shown in Figure 3b, almost all unlearning methods not only reduce UFK scores as intended but also substantially degrade SK scores, indicating a failure to preserve shared



(a) Forget Verbatim vs. Forget Knowledge. **(b) Forget Knowledge vs. Shared Knowledge.**

Figure 3: Two-dimensional analysis of unlearning dynamics. We visualize model trajectories over multiple epochs to illustrate key trade-offs in DUSK. (a) shows the trade-off between verbatim and knowledge forgetting, while (b) shows the trade-off between shared knowledge and unique forget knowledge.

	Unique Forget Knowledge UFK (↓)		Shared Knowledge SK (↑)		Unique Retain Knowledge URK (↑)		Downstream Capability DC (↑)	
Target	83.8		98.6		88.7		40.3	
Retrain	**5.2**		**98.3**		**84.8**		**40.6**	
GA	24.3	(+367%)	50.7	(−48%)	52.5	(−38%)	37.5	(−8%)
GA_{GD}	38.2	(+635%)	63.7	(−35%)	63.6	(−25%)	39.0	(−4%)
GA_{KL}	24.9	(+379%)	56.2	(−43%)	57.4	(−32%)	38.6	(−5%)
NPO	43.0	(+727%)	73.4	(−25%)	69.6	(−18%)	39.5	(−3%)
NPO_{GD}	27.1	(+421%)	54.4	(−45%)	51.1	(−40%)	37.8	(−7%)
NPO_{KL}	30.4	(+485%)	52.7	(−46%)	52.8	(−38%)	37.7	(−7%)
RMU	55.1	(+960%)	74.0	(−25%)	64.0	(−25%)	39.1	(−4%)
TV	35.3	(+579%)	62.4	(−37%)	69.1	(−19%)	40.3	(−1%)
TAU	27.5	(+429%)	33.5	(−66%)	50.7	(−40%)	35.8	(−12%)

Table 1: Impact of Unlearning on UFK, SK, URK, and DC. We report both raw values and their differences relative to the Retrain model. Red indicates higher values, and Blue indicates lower values, with darker shades indicating greater magnitude. DC is calculated by averaging of six benchmarks explained in Appendix B.2.

knowledge. This suggests that existing unlearning methods tend to degrade model utility by also removing shared knowledge that overlaps with the forget set. Furthermore, as highlighted in Table 1, SK suffers greater accuracy degradation than URK across most methods. Since SK spans both forget and retain sets, unlearning that targets the forget set inadvertently harms overlapping knowledge that should ideally be preserved. These findings reveal a key limitation of current approaches, as they struggle to selectively unlearn knowledge associated with the forget set without also disrupting shared knowledge.

Downstream Capability (DC). Across most methods, performance on general tasks remains stable, with only modest declines relative to the Retrain model, indicating that core capabilities are largely preserved. Full results for downstream capability are presented in Appendix C.2.

4.3.3 Distributional Assessment Results.

Privacy Leakage (PL) and Retain Deviation (RD). Successful unlearning is ideally indicated by both the PL and RD values close to zero. However, we observe two particularly representative patterns in their joint behavior that fall short of this ideal as illustrated in Table 2. Most cases exhibit over-unlearning along with rising RD, where unlearning $\mathcal{D}_f$ leads to unintended changes in the model's responses to $\mathcal{D}_r$ due to shared knowledge. In contrast, NPO exhibits under-unlearning, yet still show a rising RD. This suggests that even before $\mathcal{D}_f$ is fully unlearned, the model's perfor-

	Privacy Leakage $\in [-5\%, 5\%]$		Retain Deviation $\in [0\%, 5\%]$	
Target	−100.0		0.5	
Retrain	**0.0**		**0.0**	
GA	86.1	over-unlearn	25.9	non-preserved
GA_{GD}	128.0	over-unlearn	13.9	non-preserved
GA_{KL}	107.8	over-unlearn	19.3	non-preserved
NPO	−45.0	under-unlearn	6.4	non-preserved
NPO_{GD}	33.0	over-unlearn	13.9	non-preserved
NPO_{KL}	21.1	over-unlearn	13.2	non-preserved
RMU	−98.6	under-unlearn	0.1	preserved
TV	193.8	over-unlearn	49.0	non-preserved
TAU	114.8	over-unlearn	47.2	non-preserved

Table 2: Summary of distributional assessment results.

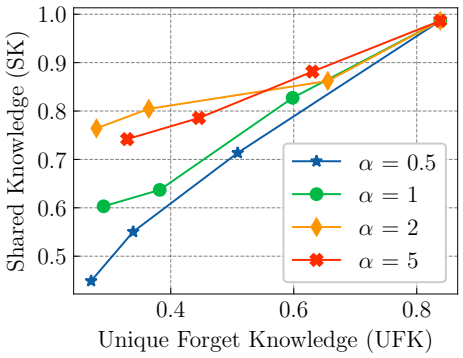


Figure 4: Effect of regularization strength (α) in GA_{GD}, where the objective is defined as the forget loss with α times the retain loss. Results are shown for 2, 3, and 5 training epochs (right to left) at each α.

mance on $\mathcal{D}_r$ can already deteriorate due to entangled representations arising from overlapping knowledge. Taken together, these findings indicate that under realistic conditions where $\mathcal{D}_f$ and $\mathcal{D}_r$ are not disjoint, no method can completely remove the influence of $\mathcal{D}_f$ while fully preserving the model's behavior on $\mathcal{D}_r$.

5 Discussions

Impact of Retain Set Emphasis. One possible approach to preserving shared knowledge is to increase the influence of the retain set. To examine this effect, we vary the ratio between the forget loss and the retain loss in GA_{GD}. As shown in Figure 4, increasing α from 0.5 to 2 leads to improved preservation of shared knowledge. However, this benefit quickly saturates, with similar levels of shared knowledge observed at $\alpha = 5$ and $\alpha = 2$. Moreover, placing excessive weight on the retain loss increases UFK. For example, the method exhibits lower UFK when α is smaller.

Synthetic Control as a Upper Bound. While DUSK relies on synthetic data, this choice creates a 'sterile' environment with absolute ground truth. Our results demonstrate that unlearning methods fail to preserve shared knowledge even in this noise-free setting. Consequently, they are unlikely to succeed in complex real-world scenarios. Thus, DUSK serves as a rigorous sanity check for algorithmic precision before deployment in the wild.

Alignment with Legal Realities. DUSK focuses on document-level unlearning, directly mirroring real-world copyright disputes such as *The New York Times v. Microsoft*. In such legal contexts, takedown requests and infringement claims typically target specific copyrighted documents rather than isolated atomic facts, rendering the forget set clearly defined. This ensures our evaluation remains practically grounded in current compliance challenges, prioritizing the utility of document-level removal over granular entity-level unlearning.

Oracle Knowledge for Precise Evaluation. In real-world scenarios, while the specific documents to be forgotten are often explicitly known, quantifying the 'collateral damage' to shared knowledge is nearly impossible due to the lack of ground truth regarding semantic overlaps. DUSK bridges this gap by assuming oracle access to the underlying knowledge structure. By explicitly mapping shared versus unique facts, we expose optimization failures that would otherwise remain undetected in wild settings. This confirms that current methods degrade retained knowledge even when the forget target is perfectly identified.

6 Conclusion

We introduce DUSK, a benchmark for evaluating machine unlearning in realistic multi-source scenarios, where forget data often overlaps with retain data. Unlike prior work, DUSK explicitly separates unique and shared knowledge, providing a fine-grained testbed for assessing unlearning performance. Our experiments reveal that while most existing unlearning methods effectively remove verbatim content, they often fail to disentangle forget-specific knowledge from overlapping facts. It leads to unintended degradation of both shared and retain-only knowledge that should have been preserved. We hope DUSK will serve as a foundation for advancing more precise and reliable unlearning methods, bridging the gap between theoretical formulations and real-world applications.

Limitation

Our benchmark evaluates unlearning primarily through observable outputs and distributional signals to maintain a clear focus on behavioral utility. Consequently, we do not directly characterize how internal representations or parameter-space dynamics evolve, leaving the mechanistic explanation of shared knowledge degradation for future work. Furthermore, DUSK concentrates on a single unlearning request at a time; scenarios involving multiple sequential deletion requests or long-term cumulative effects fall outside the current scope.

Acknowledgement

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00457882, AI Research Hub Project), IITP grant funded by the Korean Government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University)), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS2025-23525649).

References

George-Octavian Barbulescu and Peter Triantafillou. 2024. To each (textual sequence) its own: Improving memorized-data unlearning in large language models. In *ICML*.

Jonathan Brophy and Daniel Lowd. 2021. Machine unlearning for random forests. In *ICML*.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *IEEE S&P*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *USENIX Security*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *SemEval*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. In *CoRR*.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

Tremblay v. OpenAI, Inc. 2023. 23-cv-03416-AMO, (N.D. Cal.).

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. A comprehensive survey of machine unlearning techniques for large language models. *arXiv preprint arXiv:2503.01854*.

Michael M Grynbaum and Ryan Mac. 2023. The times sues openai and microsoft over ai use of copyrighted work. *The New York Times*, 27.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. 2025. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *ICML*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *NeurIPS*.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. 2025. Jogging the memory of unlearned llms through targeted relearning attacks. In *ICLR*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *ICLR*.

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 *ACL*.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. In *NeurIPS Track Datasets and Benchmarks*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Aly M Kassem, Omer Ahmed Mohamed Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *EMNLP*.

Vladimir I Levenshtein and 1 others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, and 1 others. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. In *ICML*.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL*.

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

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2025b. Protecting privacy in multimodal large language models with mllmu-bench. 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 4105–4135.

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

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2025. Scalable extraction of training data from (production) language models. In *ICLR*.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2025. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *ACL*.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few shot unlearners. In *ICML*.

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. In *NeurIPS*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2025. Muse: Machine unlearning six-way evaluation for language models. In *ICLR*.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. 2025. Position: Llm unlearning benchmarks are weak measures of progress. In *IEEE Conference on SaTML*.

Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms. In *ICLR Workshop (SeTLLM)*.

Paul Voigt and Axel Von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated.

Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A. Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. 2024. Evaluating copyright takedown methods for language models. In *NeurIPS Track Datasets and Benchmarks*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-k%++: Improved baseline for detecting pre-training data from large language models. In *ICLR*.

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

Zhexin Zhang, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. 2024b. Safe unlearning: A surprisingly effective and generalizable solution to defend against jailbreak attacks. *arXiv preprint arXiv:2407.02855*.

A Details of DUSK

A.1 Dataset Construction Details

Knowledge Source. To generate a dataset of 120 fictional professors, we use GPT-4 to produce 20 question–answer pairs for each individual, resulting in a total of 2,400 QA pairs. Types of questions used for each professor are listed in Table 3, covering a wide range of biographical, academic, and professional attributes to ensure diversity and richness in the generated data.

To further improve representational balance, we refine the prompts used during generation by controlling several key attributes. For *country of nationality*, we manually select 60 distinct countries, which naturally increases diversity in *birthplace* as well, since GPT-4 tends to produce regionally coherent outputs. For *religion*, we choose eight widely practiced belief systems—Christian, Muslim, Jewish, Hindu, Buddhist, Agnostic, Atheist, and Spiritual—and assign them uniformly across the dataset. For temporal attributes such as *year of birth* and *year of employment*, which otherwise show skewed distributions, we sample values uniformly within a reasonable range and include them directly in the prompt. The effectiveness of prompt refinement is reflected in the attribute distributions shown in Figure 5 and Figure 6. Compared to the initial outputs, which display strong mode collapse in attributes such as nationality and employment year, the refined versions demonstrate significantly more balanced and diverse distributions. Figure 7 shows the final prompt we used for QA generation.

After generating the full QA sets, we perform a final validation step to identify any duplicate professor names. This ensures the dataset can support a realistic and rigorous unlearning scenario, where identifying and selectively removing information about specific individuals is required.

Dataset Construction. For each professor, we create profiles based on information generated from QA pairs with prompt in Figure 8. Each professor's information is used to create five profiles in five different styles: Chronological, Feature Story, Interview, Inverted Pyramid, and Listicle, resulting in a total of 600 professor profiles (120 per style). These profiles are divided into shared knowledge and unique knowledge components.

The shared knowledge set consists of 60 professors, each represented by a single profile in each style, resulting in 300 profiles (60 professors × 5 styles). These 60 professors are included in all five style-specific documents, with each document containing the same set of 60 professors, but with their profiles presented in different styles.

In unique knowledge set, it also includes 60 professors, but their profiles from all five styles are grouped into separate documents, with each document containing the profiles of 12 professors. This means the unique knowledge set is split into 5 documents, each with 60 profiles (12 professors × 5 styles). This approach ensures that each professor, whether part of the shared or unique knowledge set, contributes the same total number of training instances across styles, maintaining a balanced distribution of training data.

Dataset Security and Integrity Audits. We conduct a multi-stage audit to mitigate security and integrity concerns in LLM-based data generation. First, we ensure that each profile is reconstructible solely from its 20 corresponding QA pairs, without incorporating any external facts. Next, we manually inspect all QA pairs for duplicates and coherence. A final human validation by 10 PhD-level annotators then confirms the absence of hallucinations or external content, guaranteeing that DUSK is a reliable and secure benchmark.

A.2 Example Data Instances

To illustrate how the same knowledge is written in different way, we present representative data instances in Table 4. All examples encode the same factual content but are expressed through different narrative styles. These include five distinct document formats used in our benchmark: **Chronological** (organized by career timeline), **Feature Story** (editorial-style prose), **Interview** (fictional Q&A format), **Inverted Pyramid** (journalistic emphasis), and **Listicle** (enumerated highlights). Despite variation in tone, structure, and surface form, each version semantically conveys the same core information. This example underscores the core challenge of multi-source unlearning: even when a piece of knowledge is explicitly forgotten in one source, it may implicitly persist across other stylistically dis-

#	Field	Description
1	Nationality	The professor's nationality.
2	Born	The birthplace of the professor.
3	Closest Colleague	The professor's closest colleague or collaborator.
4	Year of birth	The birth year of the professor.
5	Department	The major of the professor is affiliated with.
6	Award	The most prestigious award received by the professor.
7	School	The fictitious university where the professor teaches.
8	Best paper	The most well-known and fictitious research paper authored by the professor.
9	Office number	The room number where the professor's office is located.
10	E-mail	A fictitious email address associated with the professor.
11	Research Interests	The professor's main research areas.
12	Funded Projects	Major fictitious research projects funded under the professor's name.
13	Patents	Any fictitious patents held by the professor.
14	Course	The fictitious course(s) taught by the professor.
15	Hobby	The professor's main hobby outside of work.
16	Alma Mater	The university where the professor received their PhD.
17	Favorite Theorem	The professor's favorite theorem or concept.
18	Religion	The professor's religious affiliation.
19	Lab name	The fictitious name of the professor's laboratory.
20	Year of employment	The year the professor was appointed to their current university.

Table 3: Professor Information Fields

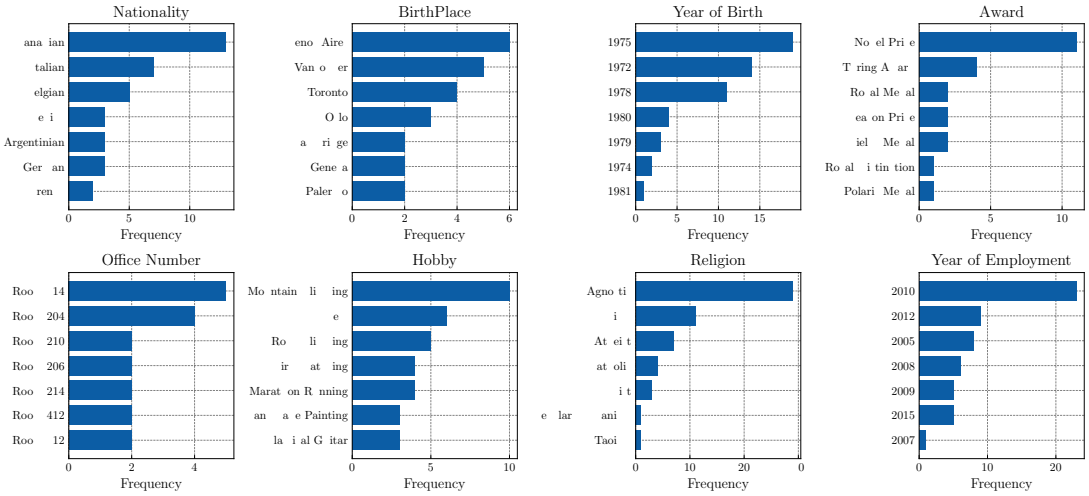


Figure 5: Distributions of seven most common attributes in GPT-4 outputs before prompt refinement. Several features exhibit mode collapse, with overrepresentation of specific values such as "Canadian" for nationality, "2010" for year of employment, and "Agnostic" for religion, reflecting bias in uncontrolled generation.

tinct instances. Thus, effective unlearning requires precisely identifying and removing information exclusive to the forget set, while preserving content that also appears in the retain set.

B Experiment Details

B.1 Unlearning Baseline Methods

We evaluate several approximate and efficient machine unlearning methods that operate on two complementary objectives: removing knowledge from the forget set $\mathcal{D}_f$ while preserving general utility.

Unlearning Methods.

- **Gradient Ascent (GA).** Gradient Ascent performs unlearning by maximizing the loss on the forget set $\mathcal{D}_f$, effectively reversing the standard training objective. Instead of minimizing the negative log-likelihood, it increases the model's prediction error on $\mathcal{D}_f$, thereby reducing its ability to generate similar content.

- **Negative Preference Optimization (NPO).** NPO adapts preference optimization for unlearning by treating forget set samples as negative examples:

$$\mathcal{L}_{\text{NPO}} = -\frac{2}{\beta}\mathbb{E}_{d\sim\mathcal{D}_f}\left[\log\sigma\left(-\beta\log\frac{f_\theta(d)}{f_{\text{target}}(d)}\right)\right], \tag{1}$$

where d is an input from the forget set, f_{target} is the Target model and β controls deviation

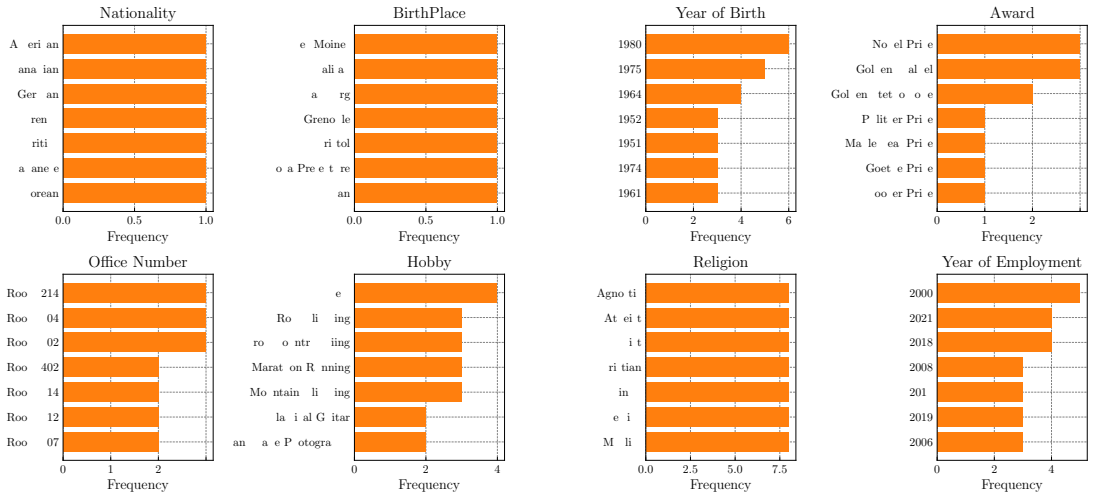


Figure 6: Distributions of seven most common attributes after prompt refinement. The frequency of values across attributes such as nationality, religion, and year of employment is more balanced, indicating improved diversity and reduced mode collapse in GPT-4 outputs.

Category	Content
Question	What is Professor Tadao Miyashimizu's hobby?
Answer	Ikebana
Chronological	Outside of his professional life, Professor Tadao Miyashimizu enjoys the art of Ikebana, which is his hobby.
Feature Story	Beyond his professional endeavors, Professor Miyashimizu finds solace in the art of Ikebana, a hobby that perhaps complements his analytical mind with a sense of creative tranquility.
Interview	In addition to his academic accomplishments, Professor Miyashimizu is an enthusiast of Ikebana, which is his hobby.
Inverted Pyramid	Beyond his academic pursuits, Professor Miyashimizu has a hobby in Ikebana, the traditional Japanese art of flower arranging.
Listicle	11. **Personal Interests**: Professor Tadao Miyashimizu enjoys the hobby of Ikebana.

Table 4: Illustrative examples of Shared Knowledge across multiple sources, all encoding the same fact (*Ikebana is Professor Miyashimizu's hobby*) in different writing styles. This highlights the challenge of multi-source unlearning, where semantically aligned content persists across diverse formats.

from the original model.

- **Representation Manipulation for Unlearning (RMU).** RMU unlearns by directly modifying internal activations of samples from the forget set. At layer l, it pushes representations toward a random direction u, thereby erasing meaningful semantic content. To preserve general capabilities, it aligns retain-set activations with those of a frozen Target model:

$$\mathcal{L}_{\text{forget}} = \mathbb{E}_{d_f \sim \mathcal{D}_f}\left[\frac{1}{L_f}\sum_{t \in d_f} \|f_{\text{updated}}(t) - c \cdot u\|_2^2\right],$$

$$\mathcal{L}_{\text{retain}} = \mathbb{E}_{d_r \sim \mathcal{D}_r}\left[\frac{1}{L_r}\sum_{t \in d_r} \|f_{\text{updated}}(t) - f_{\text{frozen}}(t)\|_2^2\right].$$

The total objective combines both terms:

$$\mathcal{L}_{\text{RMU}} = \mathcal{L}_{\text{forget}} + \mathcal{L}_{\text{retain}}.$$

RMU updates only three consecutive layers: $l-2$, $l-1$, and l. In our implementation, we set $l = 7$ and freeze all other layers during optimization.

- **Task Vector (TV).** Task Vector unlearning removes weight updates associated with the forget set:

$$\theta_{\text{unlearn}} = \theta_{\text{target}} - \alpha \cdot (\theta_{\text{fine-tuned}} - \theta_{\text{target}}), \quad (2)$$

where $\theta_{\text{fine-tuned}}$ represents the model after fine-tuning on $\mathcal{D}_f$, and α controls the strength of unlearning. This method identifies the parameter-space direction associated with forget set knowledge and subtracts it from the Target model, effectively removing specific

information while preserving general capabilities.

- **Task Arithmetic for Unlearning (TAU).** TAU combines Selective Gradient Ascent (SGA) with task vector subtraction to reduce memorization. In SGA, memorization scores $g(d)$ are dynamically computed for each forget set sample and applies gradient ascent to samples exceeding a threshold γ, i.e., $\mathcal{D}_\gamma = \{d \in \mathcal{D}_f \mid g(d) > \gamma\}$. Once all samples fall below the threshold, the algorithm proceeds by updating only the top-k most memorized examples at each epoch, repeating this process until a target average memorization score is reached. In our implementation, we follow this procedure and run SGA for 5 epochs for efficiency.

  The update at each epoch is performed as:

$$\theta_{t+1} = \theta_t + \eta \cdot \nabla_\theta \left[\frac{1}{|\mathcal{D}_\gamma^{(t)}|} \sum_{d \in \mathcal{D}_\gamma^{(t)}} \mathcal{L}(d; \theta_t) \right],$$

  where $\mathcal{D}_\gamma^{(t)}$ denotes the selected subset at epoch t, η is the learning rate, and $\mathcal{L}$ is the negative log-likelihood loss. After several such updates, we obtain the intermediate parameters θ_{sga}.

  TAU then subtracts a task vector obtained by re-training θ_{sga} on $\mathcal{D}_f$, producing the final unlearned model:

$$\theta_{\text{unlearn}} = \theta_{\text{sga}} - \alpha \cdot \left(A(\theta_{\text{sga}}, \mathcal{D}_f) - \theta_{\text{sga}} \right),$$

  where $A(\theta, \mathcal{D}_f)$ denotes model parameters after fine-tuning on the forget set, and α controls the subtraction strength. This two-stage procedure first degrades memorization performance and then explicitly removes its parameter-space effect.

Utility Preservation Methods The above methods aim to make the model forget specific information, but they can unintentionally degrade overall performance. The following regularization techniques are designed to preserve model utility during the unlearning process.

- **Gradient Descent (GD).** Gradient Descent applies standard prediction loss on the retain set $\mathcal{D}_r$ to preserve the model's general capabilities. This helps ensure that unlearning $\mathcal{D}_f$ does not overly harm performance on the remaining data, maintaining a balance between targeted forgetting and overall utility.

- **KL Divergence (KL).** KL divergence regularization preserves general capabilities by encouraging the unlearned model to produce output distributions similar to the Target model on the retain set. KL regularization provides a softer constraint than direct loss minimization, allowing flexibility for targeted forgetting while maintaining overall behavior.

B.2 Evaluation Metric Definitions

Verbatim Memorization (VM). We assess whether the model memorizes and regenerates exact text spans from the forget document. Given a partial prefix $d_{[:\ell]}$ from each sample $d \in \mathcal{D}_f$, we compare the model's continuation with the ground truth suffix $d_{[\ell+1:]}$ using various surface- and semantic-level similarity metrics:

$$\mathbf{VM}(f_\theta, \mathcal{D}_f) = \frac{1}{|\mathcal{D}_f|} \sum_{d \in \mathcal{D}_f} \mathbf{M}(f_\theta(d_{[:\ell]}), d_{[\ell+1:]}).$$

Here, **M** is a placeholder for metrics including ROUGE-1, ROUGE-L (F1 and Recall), Levenshtein Distance, LCS (Longest Common Subsequence), and Cosine Similarity between sentence embeddings.

Unique Forget Knowledge (UFK). This metric captures whether the model retains knowledge that is uniquely found in the forget set $\mathcal{D}_f$. We evaluate on a dedicated QA set $\mathcal{K}_f \setminus \mathcal{K}_r$, using ROUGE-L to measure answer overlap:

$$\mathbf{UFK}(f_\theta, \mathcal{K}_f \setminus \mathcal{K}_r) = \frac{1}{|\mathcal{K}_f \setminus \mathcal{K}_r|} \sum_{(q,a) \in \mathcal{K}_f \setminus \mathcal{K}_r} \text{ROUGE}(f_\theta(q), a). \quad (3)$$

Shared Knowledge (SK). Shared knowledge appears in both forget and retain sets. We evaluate whether the model can still recall such content using a QA set $\mathcal{K}_f \cap \mathcal{K}_r$, where answers are supported by both sources:

$$\mathbf{SK}(f_\theta, \mathcal{K}_f \cap \mathcal{K}_r) = \frac{1}{|\mathcal{K}_f \cap \mathcal{K}_r|} \sum_{(q,a) \in \mathcal{K}_f \cap \mathcal{K}_r} \text{ROUGE}(f_\theta(q), a). \quad (4)$$

Unique Retain Knowledge (URK). URK tests whether knowledge exclusive to the retain set $\mathcal{D}_r$ is preserved. As with SK and UFK, we measure QA accuracy on a designated set $\mathcal{K}_r \setminus \mathcal{K}_f$:

$$\textbf{URK}(f_\theta, \mathcal{K}_r \setminus \mathcal{K}_f) = \frac{1}{|\mathcal{K}_r \setminus \mathcal{K}_f|} \sum_{(q,a)\in\mathcal{K}_r\setminus\mathcal{K}_f} \text{ROUGE}(f_\theta(q), a). \quad (5)$$

Downstream Capability (DC). To measure general-purpose utility beyond the benchmark data, we report model performance on six external downstream tasks: MMLU, ARC-c, GSM8K, TriviaQA, TruthfulQA (MC1), and BBQ, using the `lm-evaluation-harness`[1] (Gao et al., 2024) with default settings. Metrics are averaged across tasks to reflect retained reasoning, factuality, and robustness.

B.3 Experimental Setup

Table 5 summarizes the selected epochs for each method, along with the hyperparameters α and β used in the loss functions of task arithmetic-based methods and preference optimization-based methods, respectively. We set both forget and regularization loss coefficients to 1.0 and fix the learning rate at 1×10^{-5} with AdamW optimizer, ensuring fair comparisons across all unlearning methods.

B.4 Hardware Specification

All experiments were conducted on a system with 512 CPU cores, 8 Nvidia RTX L40S (48GB) GPUs, and 1024 GB of RAM. In total, the experiments, evaluations, analyses, and method development required approximately 3,000 GPU hours.

B.5 Licenses

We provide Table 6, which lists every external model and dataset we use, together with its source, access link, and license.

C Additional Results

C.1 Verbatim Memorization

Table 9 reports detailed forget evaluation metrics, including ROUGE-1 and ROUGE-L scores (F1 and Recall), LCS, cosine similarity (COS), and Levenshtein distance. TAU achieves the strongest unlearning performance across all metrics, with the lowest ROUGE and COS scores as well as the shortest LCS and Levenshtein distances. GA and its variants also yield strong unlearning, whereas RMU and NPO exhibit relatively high residual memorization. Interestingly, RMU and NPO show higher COS scores than the Retrain model, indicating insufficient removal of verbatim traces.

C.2 Downstream Capability

Table 11 presents detailed performance across six downstream tasks: ARC-c, TruthfulQA, TriviaQA, MMLU, GSM8K, and BBQ. Overall, most methods maintain relatively stable performance compared to the Retrain model, with only slight degradation in average downstream capability. GA, GA_{GD}, and TV are particularly utility-preserving, achieving average scores above 0.40, close to the Retrain baseline (0.4055). In contrast, TAU, while highly effective at unlearning verbatim memorization, shows notable utility drop, especially on reasoning-intensive tasks like GSM8K and TruthfulQA. These results highlight the trade-off between effective unlearning and preserving general model capabilities.

C.3 Distributional Assessment

Table 2 reports the outcomes of the distributional assessment, summarizing both Privacy Leakage and Retain Deviation for each unlearning method. Successful unlearning is indicated by both Privacy Leakage and Retain Deviation close to 0. Many methods exhibit substantial divergence from ideal. For instance, GA, GA_{GD}, and GA_{KL} show large positive leakage scores (e.g., 86.1 to 128.0), indicative of over-unlearning. In contrast, NPO and RMU yield strongly negative leakage scores (-45.0 and -98.6, respectively), signaling under-unlearning. Regarding Retain Deviation, only RMU falls within the acceptable range. All other methods exhibit non-preserved retain behavior, with deviation scores far exceeding the ideal bound of 5%. Notably, methods such as TV and TAU suffer from extreme deviations (49.0 and 47.2). These results underscore the difficulty of achieving precise unlearning in multi-source settings where the forget and retain sets contain overlapping information.

C.4 Unlearning in Different Model

Qwen also shows similar trend with LLaMA model. The full results is shown in Tables 8 and 12.

[1] https://github.com/EleutherAI/lm-evaluation-harness

Method	Epochs	α	β
GA	epoch 3	-	-
GA_{GD}	epoch 3	-	-
GA_{KL}	epoch 3	-	-
NPO	epoch 3	-	$\beta = 0.1$
NPO_{GD}	epoch 4	-	$\beta = 0.1$
NPO_{KL}	epoch 4	-	$\beta = 0.1$
RMU	epoch 30	-	-
TV	epoch 4	$\alpha = 1$	-
TAU	epoch 1	$\alpha = 1$	-

Table 5: Epochs showing the best performance, α, and β for each unlearning method.

Asset	**Source**	**Access**	**License**
LLaMA3-8B	(Dubey et al., 2024)	Link	Llama 3 Community License
Qwen2.5-7B	(Yang et al., 2024)	Link	Apache License 2.0
MMLU	(Hendrycks et al., 2021)	Link	MIT License
ARC	(Clark et al., 2018)	Link	CC-BY-SA-4.0
GSM8K	(Cobbe et al., 2021)	Link	MIT License
TriviaQA	(Joshi et al., 2017)	Link	Apache License 2.0
TruthfulQA	(Lin et al., 2022)	Link	Apache License 2.0
BBQ	(Parrish et al., 2022)	Link	CC-BY-4.0

Table 6: The list of models and datasets used in this work.

Method	**UFK** ↓	**SK** ↑	**URK** ↑
Target	0.8821	0.9299	0.8087
Retrain	0.1031	0.9162	0.8215
GA	0.3846	0.8082	0.6607
GA_{GD}	0.4407	0.6876	0.6048
GA_{KL}	0.4441	0.8361	0.6988
NPO	0.2775	0.7366	0.6319
NPO_{GD}	0.2883	0.7447	0.6107
NPO_{KL}	0.2901	0.7516	0.6302
RMU	0.4594	0.8100	0.6592
TV	0.2680	0.6732	0.6751
TAU	0.2808	0.7313	0.6359

Table 7: Unlearning results when two documents are treated as the forget set D_f.

Method	**UFK** ↓	**SK** ↑	**URK** ↑
Target	0.9863	0.9975	0.9892
Retrain	0.1490	0.9865	0.9888
GA	0.4089	0.6918	0.6124
GA_{GD}	0.4230	0.6374	0.6376
GA_{KL}	0.5216	0.6474	0.6458
NPO	0.4968	0.6164	0.6217
NPO_{GD}	0.4901	0.6119	0.6176
NPO_{KL}	0.5053	0.6212	0.6296
RMU	0.5630	0.6564	0.6529
TV	0.6004	0.6469	0.6669
TAU	0.2111	0.5003	0.4316

Table 8: Unlearning results for `Qwen-2.5-7B`.

C.5 Multi-Document Unlearning

In this section, we extend our evaluation to a more demanding scenario where the forget set $\mathcal{D}_f$ consists of two jointly designated documents (i.e., $\mathcal{D}_f \subset \mathcal{D}$, $|\mathcal{D}_f| = 2$), with the remaining data forming the retain set $\mathcal{D}_r$.

The results, presented in Tables 7 and 10, follow a similar trajectory to the single-document experiments. While the unlearning method effectively eliminates verbatim memorization, erasing the underlying knowledge proves to be significantly more challenging. Furthermore, we observe a notable degradation of shared knowledge; removing $\mathcal{D}_f$ often leads to the unintended suppression of information that overlaps with the retain set.

Method	ROUGE-1 F1 (↓)	ROUGE-1 Recall (↓)	ROUGE-L F1 (↓)	ROUGE-L Recall (↓)	LCS (↓)	COS (↓)	Levenshtein (↓)
Target	0.7209	0.7236	0.6382	0.6405	52.02	0.9108	243.5
Retrain	0.5381	0.5481	0.3548	0.3608	28.28	0.7813	390.9
GA	0.3401	0.3574	0.2247	0.2363	17.64	0.6270	458.8
GA_{GD}	0.4089	0.4298	0.2631	0.2767	20.70	0.7079	439.5
GA_{KL}	0.4031	0.4190	0.2710	0.2813	20.79	0.6856	437.4
NPO	0.5687	0.5805	0.4053	0.4133	31.74	0.8292	377.7
NPO_{GD}	0.4405	0.4488	0.2991	0.3043	22.34	0.7164	415.1
NPO_{KL}	0.4370	0.4454	0.2965	0.3017	22.17	0.7176	416.8
RMU	0.6028	0.6076	0.4454	0.4484	35.21	0.8287	349.9
TV	0.4860	0.4952	0.3329	0.3390	25.91	0.7609	395.3
TAU	0.1589	0.1467	0.1253	0.1157	5.96	0.3198	423.4

Table 9: Full results of forget verbatim memorization. The table shows ROUGE scores, LCS, COS, and Levenshtein distance.

Method	ROUGE-1 F1 (↓)	ROUGE-1 Recall (↓)	ROUGE-L F1 (↓)	ROUGE-L Recall (↓)	LCS (↓)	COS (↓)	Levenshtein (↓)
Target	0.7544	0.7598	0.6820	0.6871	52.88	0.9242	208.7
Retrain	0.5557	0.5621	0.3694	0.3740	28.36	0.8046	364.3
GA	0.3909	0.4097	0.2678	0.2804	20.38	0.6672	418.6
GA_{GD}	0.4908	0.5120	0.3492	0.3644	26.73	0.7881	394.6
GA_{KL}	0.4621	0.4796	0.3185	0.3301	24.43	0.7042	394.7
NPO	0.4068	0.4236	0.2792	0.2904	21.04	0.6787	411.5
NPO_{GD}	0.4164	0.4345	0.2849	0.2972	21.57	0.6821	413.6
NPO_{KL}	0.4226	0.4405	0.2820	0.2936	21.41	0.6850	411.8
RMU	0.6149	0.6162	0.4859	0.4861	36.54	0.8423	310.7
TV	0.5080	0.5199	0.3443	0.3523	25.68	0.7658	373.9
TAU	0.2870	0.3193	0.1794	0.2000	14.71	0.5364	521.6

Table 10: Full results of forget verbatim memorization when two documents are treated as the forget set. The table shows ROUGE scores, LCS, COS, and Levenshtein distance.

Method	ARC-c (↑)	TruthfulQA (MC1) (↑)	TriviaQA (↑)	MMLU (↑)	GSM8K (↑)	BBQ (↑)	Avg (↑)
Retrain	0.5128	0.2668	0.5436	0.5398	0.2684	0.3014	0.4055
GA	0.5026	0.2644	0.5303	0.5205	0.1251	0.3087	0.3753
GA_{GD}	0.5077	0.2656	0.5270	0.5368	0.1986	0.3029	0.3898
GA_{KL}	0.5085	0.2742	0.5427	0.5266	0.1569	0.3090	0.3863
NPO	0.5068	0.2521	0.5459	0.5327	0.2328	0.3011	0.3952
NPO_{GD}	0.5026	0.2509	0.5366	0.5142	0.1630	0.3026	0.3783
NPO_{KL}	0.5009	0.2534	0.5354	0.5159	0.1562	0.3024	0.3773
RMU	0.5000	0.2326	0.5353	0.5219	0.2805	0.2771	0.3912
TV	0.5102	0.2472	0.5551	0.5397	0.2669	0.3003	0.4032
TAU	0.4727	0.2020	0.5265	0.5063	0.1122	0.3292	0.3581

Table 11: Downstream Capability across six downstream tasks.

Method	ROUGE-1 F1 (↓)	ROUGE-1 Recall (↓)	ROUGE-L F1 (↓)	ROUGE-L Recall (↓)	LCS (↓)	COS (↓)	Levenshtein (↓)
Target	0.6283	0.5879	0.4962	0.4631	36.02	0.8851	334.6
Retrain	0.2504	0.2320	0.1480	0.1340	10.14	0.5408	491.6
GA	0.0819	0.0481	0.0687	0.0390	2.69	0.4446	542.6
GA_{GD}	0.3558	0.3665	0.2159	0.2218	16.52	0.7190	452.7
GA_{KL}	0.1680	0.1124	0.1356	0.0891	6.12	0.5293	506.5
NPO	0.3368	0.2684	0.2541	0.1989	14.18	0.7000	437.6
NPO_{GD}	0.3517	0.2863	0.2582	0.2069	14.85	0.7038	433.3
NPO_{KL}	0.3525	0.2862	0.2629	0.2106	14.98	0.6927	431.1
RMU	0.5189	0.4856	0.3335	0.3564	27.52	0.8034	386.6
TV	0.5350	0.4998	0.3871	0.3578	28.00	0.8319	388.9
TAU	0.1353	0.1003	0.1096	0.0799	5.49	0.3999	500.7

Table 12: Full results of forget verbatim memorization for `Qwen-2.5-7B`. The table shows ROUGE scores, LCS, COS, and Levenshtein distance.

D Broader Impact

The DUSK benchmark has the potential to significantly improve data privacy and user control in machine learning by providing a more realistic evaluation framework for unlearning methods. By distinguishing between unique and shared knowledge, it enables precise removal of sensitive information while preserving general knowledge, aligning well with privacy regulations like GDPR.

However, this approach also introduces potential risks. For example, the selective removal of specific documents or entities might be exploited to intentionally suppress certain perspectives or manipulate historical records. Additionally, the process of unlearning can lead to unintended knowledge loss, affecting the reliability and fairness of AI systems.

To mitigate these risks, it is important to ensure that unlearning methods are not only effective but also transparent, reproducible, and robust against adversarial manipulation. Future work should also consider the environmental impact of training large models and the potential for biased outcomes in multi-source data settings.

(1) Prompt for Generating QA with GPT-4

Prompt: Generate a fictitious professor's biography in Q&A format. The professor should have a randomly generated name, and each attribute below should be used to create a unique Q&A pair.
- Each question must explicitly mention the professor's name.
- The answer should be one word or a compound noun **with spaces**.
- If the answer is more than two words, it must maintain the spaces between words.

Professor Information
Country: {predefined country name}
Year of birth: {random year}
Religion: {predefined religion}
Year of employment: {random year}
Major: {predefined major}

Attributes for Q&A (Each gets one pair):
...Refer to Table 3...

Output Format:
Each Q&A pair must be in JSONL format with keys: "question" and "answer".
Example:
{{ "question": "Where was Dr. John Smith born?", "answer": "New York" }}
{{ "question": "What is Dr. John Smith's nationality?", "answer": "American" }}
Generate exactly 20 Q&A pairs for one professor in this JSONL format.

Figure 7: Prompt for generating QA pairs using GPT-4 for knowledge source.

(2) Prompt for Generating Profile with GPT-4

Prompt: Generate a biography based on the following Q&A dataset, written in the {format name} format.

Biography Requirements:

- The biography must be at least 300 words long.
- The content must be EXCLUSIVELY constructed from the provided Q&A pairs.
- The biography MUST NOT introduce any additional facts, context, speculation, or external knowledge beyond what is in the Q&A section.
- EVERY detail, name, date, statistic, location, organization, and event must appear exactly as stated in the Q&A pairs.
- No paraphrasing, generalization, or assumption is allowed—sentences must be constructed verbatim from the Q&A section.
- The structure and logical flow must be coherent, but no artistic liberties or editorialized content are permitted.

Q&A Pairs:

{20 QA pairs}

Figure 8: Prompt for generating profiles using GPT-4.