# *Shadow Unlearning:* A Neuro-Semantic Approach to Fidelity-Preserving Faceless Forgetting in LLMs

**Anonymous ACL submission**

## Abstract

In a typical machine unlearning setting, both the forget set and the retain set are exposed to the unlearning methods. Existing machine unlearning approaches are largely designed under this standard assumption. However, exposing potentially sensitive data, such as medical records, makes these methods vulnerable to membership inference attacks and the misuse of Personally Identifiable Information (PII). In this paper, we introduce a stricter formulation in which the original forget set is not directly exposed to the unlearning method. Instead, only an anonymized version of the forget set is made available, ensuring that no PII is revealed. We term this novel task formulation ***Shadow Unlearning***. Since existing state-of-the-art methods are designed for the standard unlearning setting, they do not inherently satisfy the constraints of Shadow Unlearning. To address this gap, we propose a novel privacy-preserving framework, ***Neuro-Semantic Projector Unlearning (NSPU)***, specifically designed to operate under this stricter setting. To evaluate our approach, we construct the **Mu**lti-domain **F**ictitious **U**nlearning (**MuFU**) benchmark, which includes five diverse domains. We also introduce a comprehensive evaluation protocol to quantify the trade-off between knowledge retention and unlearning effectiveness. Experimental results across multiple large language models demonstrate that NSPU achieves superior unlearning performance while preserving model utility and enhancing user privacy. Furthermore, our approach is at least $10\times$ more computationally efficient than standard unlearning methods. These findings open a new direction for privacy-aware machine unlearning that effectively balances data protection and model fidelity.

## 1 Introduction

Machine unlearning selectively removes specific data points from trained models (Cao and Yang, 2015; Bourtoule et al., 2021), essential for privacy compliance (e.g., GDPR (GDPR.eu, 2018)) and model maintenance. Standard setups require a *forget set* (samples to remove) and a *retain set* (knowledge to preserve). However, this formulation introduces a privacy paradox: processing requests requires exposing the sensitive forget set to model operators (e.g., MLaaS providers), conflicting with the need to protect PII. In domains like healthcare or finance, transferring sensitive records for unlearning risks data breaches (Liu et al., 2025; Nicolazzo et al., 2025; Nguyen et al., 2025; Liu et al., 2024b; Jang et al., 2023).

Existing LLM unlearning methods require access to original forget data (and often retain sets) using techniques like gradient ascent (Thudi et al., 2022a), joint loss balancing (Liu et al., 2022a; Chundawat et al., 2023), or preference objectives (Rafailov et al., 2023; Zhang et al., 2024). However, these approaches suffer utility loss and privacy leakage under membership-inference attacks. Privacy-preserving and federated unlearning reduces data access via client-side updates or gradient sharing, but does not address anonymized forget sets in autoregressive LLMs (Wang et al., 2023a; Liu et al., 2022a). To our knowledge, no prior work has studied LLM unlearning using only anonymized forget data.

To address this critical gap, we introduce Shadow Unlearning (Figure 1), a paradigm for performing machine unlearning directly on anonymized forget sets, mitigating critical privacy risks, where data owners mask PII to facilitate secure sharing. The privacy boundary is maintained entirely at the data owner side, they perform anonymization locally and share only the anonymized forget set with the model operator. The core idea is to learn the semantic relationship between original and anonymized data in the activation space, allowing us to 'project' the influence of the anonymized data back to its original 'shadow' and unlearn it. The proposed NSPU method con-

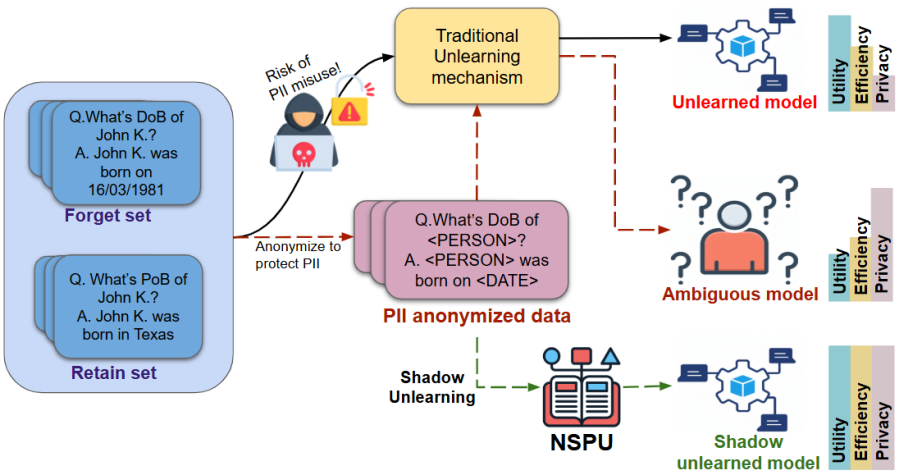


Figure 1: **The paradigm of Shadow Unlearning.**
**Motivation:** In a traditional unlearning setup, it is often required to share the retain and forget datasets for facilitating unlearning of the target model. This raises several privacy concerns related to PII. Data anonymization is a de facto way to deal with the privacy risk. This improves 'privacy', nevertheless, dents the 'utility', resulting in an 'ambiguous' model. We propose ***Shadow Unlearning*** to address this scenario by facilitating effective unlearning on anonymized data. Our approach, ***NSPU***, achieves Shadow Unlearning in a computationally 'efficient' way, preserving the 'utility' of the target model, thereby, balancing all three aspects.

sists of three stages: 1) training a lightweight projector module to learn the general transformation mapping between non-anonymized and their anonymized counterparts, not at the embedding-level, but at an activation-level, 2) identifying the principal components of the "forget space" where the dominant semantic directions must be erased, and 3) constructing an unlearning filter that modifies the model's information flow, effectively neutralizing the identified forget space concepts in real-time.

We conduct extensive experiments across five domains of the MuFU dataset. The experiments are performed on four representative LLMs of varying sizes and families to evaluate the generalizability and effectiveness of the proposed unlearning method. We compare our approach against several standard unlearning methods. While several unlearning evaluation metrics focus on assessing the model's utility and forgetting efficacy independently, we extend those metrics to assess the tradeoff between knowledge retention and forget efficacy (Zhou et al., 2024; Li et al., 2025). To this end, we employ four evaluation metrics: Harmonic Perplexity Score, Combined Efficacy Score, Harmonic ROUGE Score, and Conditional Negative Log-Likelihood Probability. Moreover, our approach is at least $10^6$ times computationally efficient than retraining from scratch and 10x faster when compared to standard unlearning methods.

The key contributions of this work include:

- We introduce a novel task in privacy-preserving unlearning – **Shadow Unlearning**, which enables selective unlearning on anonymized forget data.
- We propose **Neuro-Semantic Projector Unlearning (NSPU)**, a novel frozen-target approach for effective unlearning on anonymized forget data, achieving a strong utility-efficiency-privacy tradeoff over state-of-the-art methods.
- We compile the **MuFU** dataset across five domains: Digital informatics, Sports, Politics, Science & technology, and Finance.
- We introduce a suite of evaluation metrics designed to measure the trade-off between knowledge retention and unlearning efficacy.

## 2 Related work

**Machine unlearning for LLMs:** Unlearning approaches are broadly categorized into two categories: exact unlearning and approximate unlearning (Izzo et al., 2021). Exact unlearning mandates that the model's behavior after unlearning be indistinguishable from that of a model trained from scratch without the forget set data. To this end, early exact unlearning techniques were mainly developed on random forest (Brophy and Lowd, 2021;

Schelter et al., 2021), k-means clustering (Ginart et al., 2019). Bourtoule et al. (2021) formalized exact unlearning by introducing a general framework: sharded, isolated, sliced, aggregated (SISA). Despite the efficacy and precision of exact unlearning methods, they face significant constraints related to underlying assumptions and scalability issues (Xu et al., 2024). On the other hand, approximate unlearning (Thudi et al., 2022b) aims to forget the influence of target data points by adjusting the weights through loss-based fine-tuning strategies. These methods mainly rely on parameter optimization (Jang et al., 2023; Yao et al., 2024b; Wang et al., 2025b; Li et al., 2024; Yao et al., 2024a; Ishibashi and Shimodaira, 2023; Gu et al., 2024; Zhang et al., 2024; Lu et al., 2024; Jia et al., 2024; Tian et al., 2024; Liu et al., 2024c; Choi et al., 2024; Tang et al., 2024; Yuan et al., 2025). However, these approaches can degrade the overall performance of the models. Recent research has explored alternative approaches for LLM unlearning, including contrastive decoding (Eldan and Russinovich; Huang et al.; Wang et al., 2025a; Ji et al., 2024), task vectors (Dou et al., 2025; Liu et al., 2024c), in-context learning (Pawelczyk et al., 2024; Muresanu et al., 2025), input processing and deletion (Bhaila et al., 2025; Liu et al., 2024a) and LoRA-based unlearning (Cha et al., 2025; Liu et al.). However, most of these methods require preserving the original model's parameters, which can still raise privacy concerns. For further discussion on related works, see Appendix A.

# 3 Methodology

## 3.1 Task Formulation

Let $M_\theta$ denote a model pre-trained on a corpus $D$, composed of disjoint subsets $D = D_{\text{retain}} \cup D_{\text{forget}}$. Due to privacy constraints, the original samples are inaccessible. Instead, we assume access to a one-way anonymization function $f(\cdot)$ and the resulting anonymized forget dataset $\tilde{D}_{\text{forget}} = f(D_{\text{forget}})$. Our goal is to obtain an unlearned model $M_{\theta'}$ using only the current model state $M_\theta$ and the anonymized forget data $\tilde{D}_{\text{forget}}$. Ideally, $M_{\theta'}$ should approximate the behavior of a "gold standard" model $M^\star$ trained from scratch on the counterfactual dataset $D \setminus D_{\text{forget}}$.

## 3.2 Anonymization

To construct the anonymized forget set, a named-entity recognition (NER)-based anonymization pipeline is applied. All PII is detected using the Microsoft Presidio Anonymizer[1] and replaced with corresponding placeholder tags (e.g., <PERSON>, <LOCATION>), ensuring that direct identity information is removed while preserving the surrounding linguistic context. The same anonymization pipeline is reused in all stages of the proposed methodology to maintain consistency between pretraining-style data and the forget set.

## 3.3 Proposed Methodology

Neuro-Semantic Projection Unlearning (NSPU), comprises three key phases, as represented in Figure 2, and explained in Sections 3.4, 3.5, and 3.6.

## 3.4 Latent Representation Aligner

This preliminary, one-time phase establishes a mechanism for aligning the representations of anonymized and original data samples at the neural activation level. The aim is not to re-identify individuals, but to obtain a smooth semantic correspondence between the activations induced by original and anonymized texts. This enables the model to operate in a coherent latent geometry even when direct identification cues are removed.

To achieve this, a lightweight multi-layer perceptron (MLP) is employed as a projection function $P_\theta$. Conceptually, this MLP acts as a neuro-semantic bridge, which aligns activations of the anonymized samples with the model's original semantic structure, while keeping the information in an abstract latent form.

As a result, anonymized samples can be translated into activation vectors that are geometrically consistent with those induced by non-anonymized inputs, without exposing underlying identity attributes.

### 3.4.1 Training of the Latent Representation Aligner

Before considering the forget or retain sets, the training of the Latent representation aligner ($P_\theta$) leverages a publicly available corpus. More details of the training setup are discussed in Appendix C. From this corpus, pairs $(x_{\text{orig}}, x_{\text{anon}})$ are constructed by applying the same anonymization pipeline $f(\cdot)$ used for the forget set, where $x_{\text{anon}} = f(x_{\text{orig}})$. Both the original and anonymized texts are independently passed through the frozen target model to obtain layer-specific activations $\phi_l(x_{\text{orig}})$

[1] https://microsoft.github.io/presidio/anonymizer/

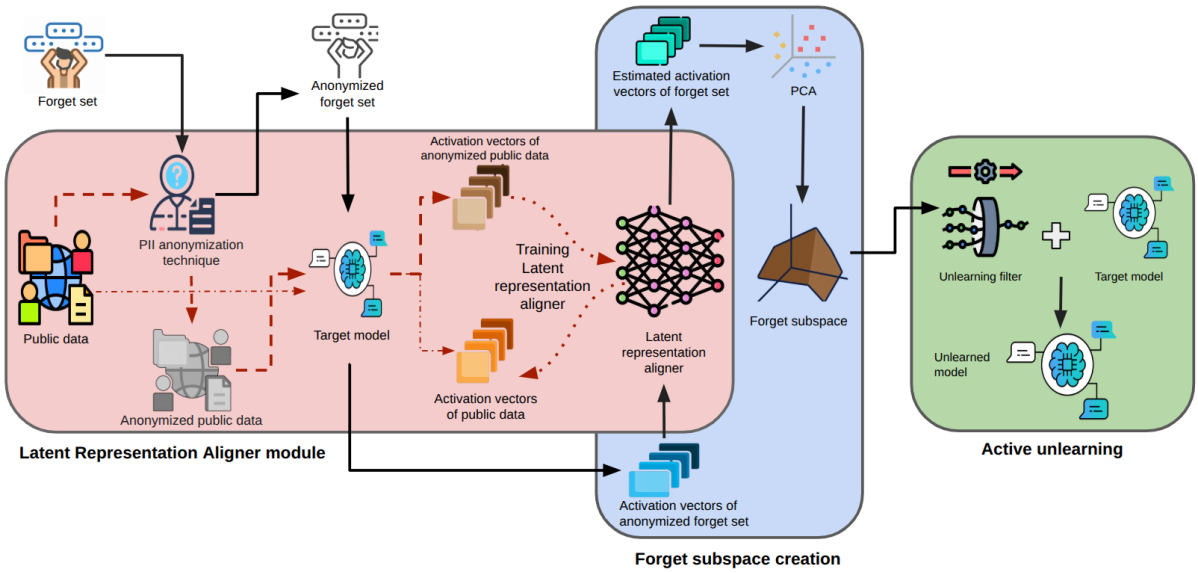


Figure 2: Neuro-Semantic Projector Unlearning (NSPU) Pipeline comprises three key phases: (i) learning a latent representation aligner, (ii) constructing a forget subspace, and (iii) integrating a linear unlearning filter.

and $\phi_l(x_{\text{anon}})$ at a chosen layer $l$. These activation pairs form the training data for the Latent representation aligner.

The Latent representation aligner $P_\theta$ is trained to minimize the semantic alignment loss:

$$\mathcal{L}_{\text{align}} = \left\| P_\theta\left(x_{\text{anon}}\right) - \phi_l(x_{\text{orig}}) \right\|_2^2$$

where $\phi_l$ denotes the layer-specific activation function of the target model. The parameters of the projector are optimized while keeping the target model frozen. The resulting trained model $P_\theta$ serves as a universal neuro-semantic bridge that estimates the neural fingerprint of the original sample from its anonymized counterpart.

**Inversion-Resistant Optimization Term.** To prevent potential inversion of the projected activations back into human-interpretable input space, we introduce an auxiliary privacy-preserving term—termed the *Inversion Optimization Score (InvOptScore)*. This score quantifies how easily an estimated activation vector could be mapped to a plausible input embedding through optimization-based inversion.

For an estimated activation, $\hat{h}_{est}$ of an anonymized sample $x_{anon}$, $\hat{h}_{\text{est}} = P_\theta(x_{\text{anon}})$ consider an inversion process that searches for a pseudo-input embedding $x' \in \mathcal{X}_{\text{emb}}$ in the model's input embedding space that reproduces $\hat{h}_{\text{est}}$. Define the inversion loss:

$$\mathcal{L}_{\text{inv-opt}}(x') = \left\| \phi_l(x') - \hat{h}_{\text{est}} \right\|_2^2$$

The optimal pseudo-input $x'^*$ is obtained via $x'^* = \arg\min_{x'} \mathcal{L}_{\text{inv-opt}}(x')$ and the corresponding inversion difficulty is summarized by

$$\text{InvOptScore}(\hat{h}_{\text{est}}) = \left\| x'^* - x_{orig} \right\|_2^2$$

The optimization over $x'$ is carried out entirely in the continuous embedding space, initialized from random embeddings and refined by gradient-based minimization of $\mathcal{L}_{\text{inv-opt}}$. No discrete token decoding is required or performed, so no explicit textual reconstruction is ever produced. A lower residual $\text{InvOptScore}(\hat{h}_{\text{est}})$ indicates that the activation is easier to approximate with some plausible input embedding and therefore carries higher inversion risk. By penalizing such low inversion losses during training, the projector is encouraged to produce activations that are semantically aligned with the model's internal geometry yet intrinsically harder to decode into human-readable content. The final training objective for the projector is

$$\mathcal{L}_{\text{proj}} = \mathcal{L}_{\text{align}} + \lambda_{\text{inv}}\big( - \text{InvOptScore}(\hat{h}_{\text{est}})\big)$$

where $\lambda_{\text{inv}} > 0$ balances semantic fidelity and inversion resistance. Larger values of $\lambda_{\text{inv}}$ prioritize privacy, while smaller values favor tighter alignment with original activations.

### 3.5 Forget Subspace

This phase constructs a forget subspace that captures the dominant directions associated with the forget set in the model's activation space. This provides a unified vector space toward which forget-related activations can be systematically diverted,

avoiding the need to reason about individual neurons or samples.

#### 3.5.1 Creation of Forget subspace

The anonymized forget set $\tilde{D}_{\text{forget}}$ is passed through the target model to obtain layer-$l$ activation vectors $\phi_l(x_{\text{anon}})$ for all $x_{\text{anon}} \in \tilde{D}_{\text{forget}}$. These anonymized activations are then passed through the trained Latent representation aligner ($P_\theta$) to produce estimated original activations $\hat{h}^{\text{orig}} = P_\theta\big(\phi_l(x_{\text{anon}})\big)$ corresponding to each sample in the $\tilde{D}_{\text{forget}}$.

Row-wise stacking of these activations yields a matrix $H \in \mathbb{R}^{n\times d}$, for $n$ forget samples and activation dimension $d$. We apply Principal Component Analysis (PCA) to $H$ to extract orthonormal directions capturing principal variance axes, which approximate dominant semantic factors across the forget set (see Appendix E for PCA rank details). The top $k$ right-singular vectors, forming an orthonormal basis for these components, comprise the columns of $U \in \mathbb{R}^{d\times k}$. This matrix $U$ defines the ***forget subspace***, the linear subspace spanned by directions most strongly associated with the forget set's internal representations.

### 3.6 Active Unlearning

Leveraging the forget subspace obtained, an unlearning filter is created as a simple adapter incorporated into the target model inference. It facilitates active unlearning, projecting away information or activations relevant to the forget set.

#### 3.6.1 Unlearning Filter

The unlearning filter $UL_{\text{filter}} \in \mathbb{R}^{d\times d}$ is defined as $UL_{\text{filter}} = I - \alpha UU^\top$, where $I$ is the $d \times d$ identity matrix and $\alpha$ is a hyperparameter governing the strength of unlearning. As depicted in Figure 3, for any activation vector $v \in \mathbb{R}^d$, the term $UU^\top v$ gives its orthogonal projection onto the forget subspace, and the filter subtracts an $\alpha$-scaled version of this projection $v_{\text{out}} = UL_{\text{filter}}\, v_{\text{in}}$. This operation shifts $v_{\text{out}}$ toward the orthogonal complement of the forget subspace, thereby attenuating information aligned with forget-related concepts while minimally disturbing components orthogonal to $U$. While retain and forget representations overlap, their dominant principal components diverge sufficiently. Projecting away from these forget axes yields an approximate yet highly effective separation. The geometric interpretation of the unlearning filter is provided in Appendix D.

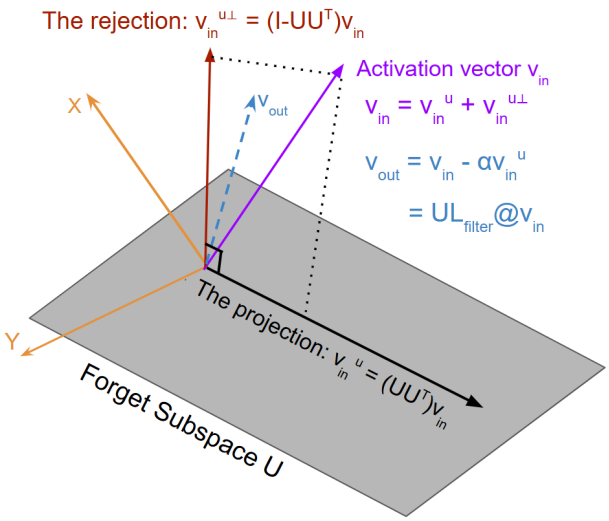


Figure 3: Functioning of Unlearning filter

The unlearning filter is implemented as a lightweight, non-trainable linear operation integrated permanently into the model architecture. Specifically, the original final layer normalization is replaced by a sequential module that first applies the unlearning transformation to the last hidden state before passing it through the standard normalization layer. This modification ensures the change is a structural part of the model, permanently steering the information flow away from the forget subspace in real-time during the forward pass.

## 4 Experiments

### 4.1 Setup

**Dataset Creation.** Existing unlearning benchmark datasets, such as TOFU (Maini et al., 2024a), often suffer from extreme homogeneity and high semantic overlap between the forget ($D_f$) and retain ($D_r$) sets. This high inter-sample correlation complicates the evaluation of unlearning efficacy. If a model retains the underlying data distribution through $D_r$, membership inference attacks (MIAs) may yield high false positives, masking the true performance of unlearning algorithms (Nguyen et al., 2025; Carlini et al., 2022).

**MuFU: Multi-Domain Fictitious Unlearning.** To address these limitations, we introduce the MuFU dataset, which evaluates unlearning across a heterogeneous environment where both forget and retain sets span multiple domains to mirror real-world complexity. Using Gemini 3 Pro, we generated 4,000 distinct samples containing personally identifiable information, distributed evenly across five domains: Digital Informatics, Finance, Sports, Sci-

ence & Technology, and Politics. Further details about the MuFU dataset creation and data variants are discussed in Appendix B.

**Models.** We conduct experiments on LLaMA-7B (Touvron et al., 2023a), Mistral-7B-instruct (Jiang et al., 2023), OLMoE-1B-7B-0924 (Muennighoff et al., 2025) and LLaMA-2-13B (Touvron et al., 2023b) variants. More details on the experimental setup are provided in Appendix E.

**Baselines.** We perform experiments by considering baselines as Gradient Ascent (GA) (Thudi et al., 2022a), Gradient Difference (GD) (Liu et al., 2022a), KL Minimization (KLM) (Chundawat et al., 2023), Direct Preference Optimization (DPO) (Rafailov et al., 2023), and Negative Preference Optimization (NPO) (Zhang et al., 2024). Detailed descriptions of the baselines are mentioned in Appendix F.

### 4.2 Evaluation

To demonstrate unlearning, evaluation is performed using the *original (non-anonymized)* questions, unlike the unlearning phase, which operates on anonymized inputs. Following prior work (Nguyen et al., 2025; Qu et al., 2024), we evaluate the trade-off between knowledge retention and forgetting using four complementary metrics. Full derivations and implementation details are provided in Appendix G.

**Harmonic Perplexity Score (HPS).** Perplexity reflects model uncertainty. Effective unlearning should increase perplexity on the forget set while keeping it low on the retain set. We define Forget Gain ($G_F$) and Retain Cost ($C_R$) as log-ratios of post- and pre-unlearning perplexity, and combine them using a harmonic formulation: $HPS = \frac{2 \cdot G_F}{G_F \cdot C_R + 1}$. This formulation penalizes methods that forget aggressively at the expense of retained knowledge.

**Combined Efficacy Score (CES).** Combined Efficacy Score (CES) utilizes the truth ratio (Maini et al., 2024a) to measure a model's preference for correct answers over perturbed ones. We sum Retain Stability (RS) and Forget Instability (FI) as: CES = RS + FI. Higher CES signifies better retention and superior suppression of forgotten content.

**Harmonic ROUGE Score (HRS).** To assess generative fidelity, we use ROUGE-L to measure semantic overlap. HRS jointly evaluates retention and forgetting by harmonically combining the Retention Ratio (RR) and Forget Ratio (FR): $HRS = \frac{2 \cdot RR}{FR \cdot RR + 1}$. This score favors high semantic preservation on the retain set and low overlap on the forget set.

**Harmonic Conditional NLL (HCNLL).** HCNLL evaluates likelihood-based behavior using conditional negative log-likelihood. Analogous to HPS, we define Forget Gain Likelihood and Retain Cost Likelihood and aggregate them harmonically: $HCNLL = \frac{2 \cdot G_{F_L}}{G_{F_L} \cdot C_{R_L} + 1}$. This metric captures probabilistic forgetting while preserving confidence in retained samples.

### 4.3 Experimental Results and Discussion

This section evaluates the model utility, unlearning efficacy, computational efficiency, and privacy via separation quality scores, tradeoff with extended analysis in Appendix Q.

### 4.4 The Utility-Efficacy Tradeoff

**GA and KLM cause catastrophic utility loss.** These methods maximize loss or divergence on forget data, leading to higher forget gain and a lower retain cost as shown in Table 1. This indicates high uncertainty in both the forget and retain sets. While this reflects strong forgetting efficacy, it comes at the cost of severe utility degradation. The negative aggregate scores for GA and KLM underscore the nature of forget-loss approaches, which lack the fundamental mechanism to preserve useful knowledge beyond the forget set.

**GD, DPO, and NPO offer limited improvements.** GD incorporates joint retention-forgetting objectives for better stability across these LLMs, while DPO and NPO suppress undesired outputs but show inconsistent behavior due to preference calibration sensitivity. Both DPO and NPO offer improved trade-offs over pure ascent techniques yet lack the robustness needed for scalable unlearning across diverse architectures.

**NSPU consistently outperforms standard unlearning methods.** NSPU achieves notably higher Aggregate scores across all tested models, while approaches such as GA and KLM consistently yield highly negative aggregate values. This indicates that NSPU effectively balances forgetting targeted data while maintaining model utility on retained data, unlike GA and KLM, which aggressively forget but cause severe degradation in overall performance.

| Model | Method | Perplexity | | | Truth Ratio | | | ROUGE-L | | | Probability | | | Aggregate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $G_F$ | $C_R$ | **HPS** (↑) | RS | FI | **CES**(↑) | RR | FR | **HRS** (↑) | $G_{F_L}$ | $C_{R_L}$ | **HCNLL** (↑) | |
| Llama2-7B | GA | 104.730 | 110.110 | 0.019 | 0.000 | -0.805 | -0.805 | 0.007 | 0.167 | 0.014 | 83.887 | 49.734 | 0.024 | -0.748 |
| | GD | 4.597 | 6.006 | 0.420 | 0.497 | 0.054 | 0.551 | 0.210 | 0.245 | 0.399 | 4.811 | 3.988 | 0.395 | 1.765 |
| | KLM | 104.730 | 110.110 | 0.019 | 0.000 | -0.808 | -0.808 | 0.007 | 0.167 | 0.014 | 83.887 | 49.734 | 0.024 | -0.751 |
| | DPO | 4.082 | 6.780 | 0.473 | 0.000 | -0.998 | -0.998 | 0.042 | 0.225 | 0.083 | 5.043 | 5.312 | 0.382 | -0.060 |
| | NPO | 39.850 | 37.772 | 0.050 | 0.000 | 0.920 | 0.920 | 0.001 | 0.031 | 0.002 | 32.027 | 17.946 | 0.062 | 1.034 |
| | **NSPU** | 1.388 | 2.055 | **1.067** | 1.021 | 0.498 | **1.519** | 0.581 | 0.705 | **0.824** | 2.590 | 2.331 | **0.662** | **4.073** |
| Mistral-7B | GA | 154.695 | 202.520 | 0.013 | 0.000 | -0.995 | -0.995 | 0.034 | 0.112 | 0.068 | 124.738 | 74.003 | 0.016 | -0.899 |
| | GD | 18.287 | 17.198 | 0.109 | 0.943 | -0.511 | 0.432 | 0.363 | 0.375 | 0.639 | 1.258 | 1.551 | 1.051 | 2.232 |
| | KLM | 193.398 | 209.753 | 0.010 | 0.000 | -1.001 | -1.001 | 0.034 | 0.112 | 0.068 | 124.738 | 74.003 | 0.016 | -0.907 |
| | DPO | 27.660 | 25.787 | 0.072 | 0.000 | -0.554 | -0.554 | 0.006 | 0.099 | 0.012 | 17.218 | 9.888 | 0.115 | -0.355 |
| | NPO | 9.549 | 7.433 | 0.207 | 0.000 | 0.996 | **0.996** | 0.001 | 0.146 | 0.001 | 9.595 | 3.561 | 0.203 | 1.406 |
| | **NSPU** | -0.084 | 2.066 | **4.995** | 0.921 | -0.536 | 0.385 | 0.876 | 0.542 | **1.188** | 0.955 | 1.650 | **1.281** | **7.849** |
| OLMOE | GA | 84.662 | 87.404 | 0.024 | 0.000 | -3.189 | -3.189 | 0.006 | 0.068 | 0.012 | 65.839 | 71.628 | 0.030 | -3.123 |
| | GD | 6.214 | 9.374 | 0.316 | 1.202 | -1.924 | -0.722 | 0.054 | 0.081 | 0.108 | 4.901 | 8.573 | 0.399 | 0.101 |
| | KLM | 84.662 | 87.404 | 0.024 | 0.000 | -3.179 | -3.179 | 0.006 | 0.077 | 0.012 | 65.839 | 71.628 | 0.030 | -3.113 |
| | DPO | 1.887 | -0.200 | -0.645 | 0.000 | -0.953 | -0.953 | 0.705 | 0.560 | 1.011 | 1.929 | 0.831 | 0.638 | 0.051 |
| | NPO | 0.095 | 0.318 | 0.617 | 1.025 | 1.156 | 2.181 | 1.007 | 0.730 | 1.160 | 0.703 | 0.636 | 0.879 | 4.837 |
| | **NSPU** | 0.631 | 2.008 | **1.771** | 2.980 | 1.548 | **4.528** | 0.857 | 0.408 | **1.270** | 1.415 | 2.483 | **1.100** | **8.669** |
| Llama-13B | GA | 62.572 | 75.388 | 0.032 | 0.000 | -0.362 | -0.362 | 0.002 | 0.164 | 0.004 | 15.695 | 27.613 | 0.127 | -0.199 |
| | GD | 0.067 | 0.077 | 0.154 | 0.598 | 0.762 | 1.360 | 0.765 | 1.454 | 0.724 | 1.329 | 0.977 | 0.850 | 3.088 |
| | KLM | 61.507 | 72.625 | 0.033 | 0.000 | -1.002 | -1.002 | 0.001 | 0.164 | 0.002 | 26.829 | 26.125 | 0.074 | -0.893 |
| | DPO | 0.065 | 0.115 | 0.228 | 0.454 | 0.710 | 1.164 | 0.651 | 0.664 | 0.909 | 1.737 | 1.367 | 0.810 | 3.111 |
| | NPO | 6.958 | 11.820 | 0.284 | 0.472 | 1.115 | 0.356 | 0.747 | 0.345 | **1.188** | 2.020 | 3.213 | 0.858 | 2.686 |
| | **NSPU** | 1.173 | 1.445 | **1.072** | 0.987 | 0.736 | **1.723** | 0.817 | 0.518 | 1.148 | 1.480 | 1.535 | **0.938** | **4.882** |

Table 1: Evaluation of NSPU against standard unlearning approaches on MuFu dataset. Aggregate score (HPS+CES+HRS+HCNLL) reflects the forgetting–utility trade-off, where higher values indicate better performance.

### 4.5 Membership Inference and Separation Quality

Membership Inference Attacks (MIAs) evaluate unlearning efficacy by testing whether an adversary can distinguish forget set samples from unseen data using the unlearned model's predictions (Shokri et al., 2017; Carlini et al., 2022; Maini et al., 2024b). Effective unlearning implies that the target model should perceive the forget set as indistinguishable from a non-member set (unseen data), while retaining performance on the retain set. To quantify this behavior, we construct the non-member set as detailed in Appendix M. To quantify MIA resistance, we compute mean Causal Language Modeling (CLM) loss across retain ($M_R$), forget ($M_F$), and non-member ($M_{NM}$) sets. The *Separation Quality Score* (SQS) measures how well the forget set loss aligns with non-members while diverging from retain samples:

$$SQS = \frac{|M_R - M_F|}{|M_R - M_F| + |M_{NM} - M_F|}$$

The numerator maximizes forget-retain divergence (successful unlearning), while the denominator minimizes forget-nonmember divergence (MIA resistance). As shown in Table 2, high SQS scores reflect the balance between effective forgetting and retention, quantifying privacy strength and unlearning fidelity.

## 5 Ablation studies

Beyond NSPU's utility, efficiency, and privacy advantages, it is critical to assess its ability to distinguish retain vs. forget information and identify domain-specific and entity-level (PII) knowledge. This motivates deeper ablations on where and how unlearning occurs.

### 5.1 Interpretable View of Knowledge Entanglement

**NSPU clearly distinguishes the forget and retain information.** Figure 4 shows layer-wise activation shifts (early, middle, last) between the original and unlearned models for both forget and retain sets. We apply the unlearning filter only at the final hidden layer, as semantic separation between forget and retain concepts is minimal in early layers and peaks in deeper layers as discussed in Appendix H. Earlier layers encode low-level features, so filtering them is unnecessary and does not expose sensitive information. Consequently, an attacker attempting to bypass the defense by probing earlier, unprotected layers would fail to extract the targeted PII, as sensitive representations have not yet formed.

Accordingly, activations remain identical up to the 31$^{st}$ layer, with unlearning effects appearing only at the final layer. The forget set shows strong separability, while the retain set exhibits minor drift with substantial overlap. This small drift is expected due to the non-orthogonality of real-world

| Method | Computational Cost (FLOPs) | | | | Separation Quality Score (SQS) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | LLaMA-7B | Mistral-7B | OLMoE-7B | LLaMA-13B | LLaMA-7B | Mistral-7B | OLMoE-7B | LLaMA-13B | Avg. (↑) |
| Retraining | $8.4 \times 10^{22}$ | $3.4 \times 10^{22}$ | $3.0 \times 10^{22}$ | $8.4 \times 10^{26}$ | – | – | – | – | – |
| GA | $1.3 \times 10^{17}$ | $1.3 \times 10^{16}$ | $1.8 \times 10^{16}$ | $1.3 \times 10^{21}$ | 0.79 | 0.78 | **0.76** | 0.403 | 0.68 |
| GD | $5.2 \times 10^{17}$ | $5.2 \times 10^{17}$ | $7.4 \times 10^{16}$ | $5.2 \times 10^{21}$ | 0.75 | 0.62 | 0.69 | 0.56 | 0.66 |
| KLM | $1.3 \times 10^{17}$ | $1.3 \times 10^{16}$ | $1.8 \times 10^{16}$ | $1.3 \times 10^{21}$ | 0.79 | 0.78 | **0.76** | 0.403 | 0.68 |
| DPO | $5.2 \times 10^{17}$ | $5.2 \times 10^{17}$ | $7.4 \times 10^{16}$ | $5.2 \times 10^{21}$ | 0.69 | 0.46 | 0.70 | 0.53 | 0.59 |
| NPO | $5.2 \times 10^{17}$ | $5.2 \times 10^{17}$ | $7.4 \times 10^{16}$ | $5.2 \times 10^{21}$ | 0.65 | 0.74 | 0.17 | 0.69 | 0.56 |
| **NSPU** | $\mathbf{7.6 \times 10^{16}}$ | $\mathbf{7.6 \times 10^{15}}$ | $\mathbf{8.7 \times 10^{15}}$ | $\mathbf{7.6 \times 10^{20}}$ | **0.83** | **0.79** | 0.74 | **0.72** | **0.77** |

Table 2: Combined comparison of computational cost (FLOPs) and separation quality score (SQS) across unlearning methods and LLMs.

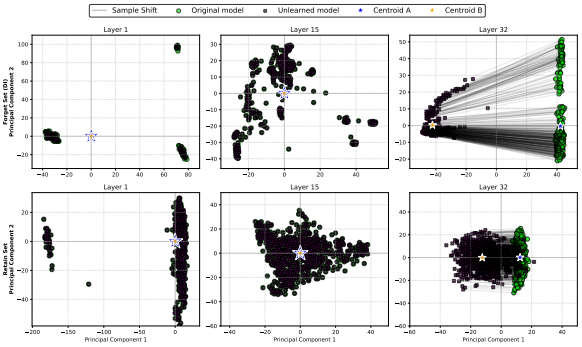


Figure 4: Impact of unlearning on forget (top-row) and retain datasets (bottom-row) before and after applying the unlearning filter (Mistral-7B).

concepts, however, it is negligible compared to the large separation in the forget set. This demonstrates that NSPU effectively removes target knowledge while preserving utility. Similar behavior is observed in Llama-13B, as shown in Appendix N.

### 5.2 Identification of desired information to unlearn

To validate that NSPU targets intended information rather than spurious knowledge, we conduct an ablation to examine the forgotten content. Since Shadow Unlearning operates on anonymized data, verifying that the model erases the specific entity per request is essential. The Latent Representation Aligner bridges anonymized and original representations while maintaining inversion resistance to prevent leakage. While Appendix C Table 8 details overall performance, Table 3 assesses how closely reconstructed activation vectors approximate original activations across domains and models. We use an *InvOptScore* regularization term to penalize excessive similarity, balancing semantic alignment with privacy. The results confirm meaningful semantic correspondence in projected vectors, validating the module's ability to localize correct knowledge under privacy constraints. Similarly, entity-level analysis (Table 4) shows NSPU consistently erases targeted representations 83% of the time on average while retaining remaining domain knowledge.

| Domain | Llama-7B | Mistral-7B | OLMoE-1B-7B | Llama-13B |
|---|---|---|---|---|
| DI | 0.64 | 0.49 | 0.63 | 0.82 |
| Politics | 0.80 | 0.70 | 0.76 | 0.87 |
| Finance | 0.79 | 0.70 | 0.74 | 0.87 |
| Sports | 0.76 | 0.60 | 0.70 | 0.84 |
| Science | 0.58 | 0.47 | 0.57 | 0.79 |
| **Average** | **0.71** | **0.59** | **0.68** | **0.84** |

Table 3: Domain-wise similarity between original and estimated activation vectors.

| PERSON | LOCATION | PHONE_NUMBER | DATE_TIME | Misc |
|---|---|---|---|---|
| 0.85 | 0.84 | 0.84 | 0.80 | 0.83 |

Table 4: PII attribute-wise similarity between original and estimated activation vectors for Llama-13B.

### 5.3 Additional results

We provide more results and ablations, including: MuFU dataset creation (Appendix B), training of Latent representation aligner (Appendix C), human evaluation (Appendix I.2), computational efficiency analysis (Appendix L) based on FLOPS and VRAM, domain-specific performance evaluation (Appendix J), NSPU performance on downstream tasks (Appendix K), error analysis (Appendix O) and robustness evaluation of NSPU (Appendix P).

## 6 Conclusion

We introduce Shadow Unlearning, a paradigm for privacy-preserving unlearning on anonymized forget sets, and propose Neuro-Semantic Projector Unlearning (NSPU) to realize it. NSPU outperforms baselines with superior knowledge retention and forgetting efficacy, demonstrated across comprehensive metrics. It offers computational efficiency, confirmed by FLOPs analysis, and ensures privacy protection, validated through membership-inference attacks and separation-quality assessments. Overall, our work establishes a new direction for privacy-aware unlearning, balancing data protection, utility, and efficiency.

## 7 Limitations

NSPU relies on PCA to construct the forget subspace, which captures directions of maximum variance rather than exact causal influence. Nevertheless, these directions effectively approximate dominant semantic factors in the forget set, which is sufficient for approximate unlearning. The pipeline is primarily optimized for batch unlearning; its behavior under extremely low-resource settings (e.g., a single forget sample) remains unexplored. Additionally, handling sequential and continuous unlearning requests without cumulative utility degradation is an open challenge. NSPU applies the unlearning filter only at the final hidden layer to preserve earlier representations, leaving intermediate activations structurally unchanged. Although interpretability analysis suggests a low likelihood of extracting semantic PII from early layers, we do not provide formal guarantees against advanced white-box inversion attacks targeting these layers.

Theoretical guarantees of NSPU ensure geometric projection away from the forget subspace but do not imply equivalence to a model retrained from scratch without the forget data. Furthermore, the method depends on access to internal activations, restricting applicability to open-source models. Finally, our NER-based anonymization pipeline introduces a potential edge case where distinct samples map to identical anonymized representations. In such collisions, selectively unlearning only a subset may confuse the Latent Representation Aligner (LRA), leading to unintended retention or forgetting. Future work can address this by designing more fine-grained, entity-level unlearning filters to better distinguish structurally identical samples.

## 8 Reproducibility Statement

To ensure reproducibility of our results, we have thoroughly documented all experimental details and evaluation procedures with proper grounding. All experiments were conducted using fixed random seeds to guarantee deterministic behavior across runs. We plan to release our implementation code, detailed experimental scripts, and the dataset to facilitate the exact reproduction of our reported results.

## 9 Ethics Statement

This work advances responsible and privacy-aware deployment of large language models through Shadow Unlearning. The proposed paradigm enables effective machine unlearning using anonymized forget data, reducing exposure of personally identifiable information during the unlearning process. By decoupling unlearning from direct access to sensitive data, our approach supports compliance with privacy regulations such as the GDPR's Right to be Forgotten. It also resolves a critical privacy paradox present in existing unlearning methods.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Karuna Bhaila, Minh-Hao Van, and Xintao Wu. 2025. Soft prompting for unlearning in large language models. 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 4046–4056.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE.

Jonathan Brophy and Daniel Lowd. 2021. Machine unlearning for random forests. In *International Conference on Machine Learning*, pages 1092–1104. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. 2025. Towards robust and parameter-efficient knowledge unlearning for llms. In *The Thirteenth International Conference on Learning Representations*.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. 2021.

When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 896–911.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2024. Snap: Unlearning selective knowledge in large language models with negative instructions. *arXiv e-prints*, pages arXiv–2406.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. 2023. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354.

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*.

Josep Domingo-Ferrer, Najeeb Jebreel, and David Sánchez. 2025. Efficient unlearning with privacy guarantees. *arXiv preprint arXiv:2507.04771*.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. 2025. Avoiding copyright infringement via large language model unlearning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5176–5200.

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

Ji Gao, Sanjam Garg, Mohammad Mahmoody, and Prashant Nalini Vasudevan. 2022. Deletion inference, reconstruction, and compliance in machine (un) learning. *Proceedings on Privacy Enhancing Technologies*, 2022(3).

Xiangshan Gao, Xingjun Ma, Jingyi Wang, Youcheng Sun, Bo Li, Shouling Ji, Peng Cheng, and Jiming Chen. 2024. Verifi: Towards verifiable federated unlearning. *IEEE Transactions on Dependable and Secure Computing*, 21(6):5720–5736.

GDPR.eu. 2018. Article 17: Right to be forgotten. [Accessed: 2025-11-14].

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Kang Gu, Md Rafi Ur Rashid, Najrin Sultana, and Shagufta Mehnaz. 2024. Second-order information matters: Revisiting machine unlearning for large language models. *arXiv preprint arXiv:2403.10557*.

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

Hongsheng Hu, Shuo Wang, Tian Dong, and Minhui Xue. 2024. Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 3257–3275. IEEE.

James Y Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. Offset unlearning for large language models. *Transactions on Machine Learning Research*.

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

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. 2021. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, pages 2008–2016. PMLR.

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.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana R Kompella, Sijia Liu, and Shiyu Chang. 2024. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *Advances in Neural Information Processing Systems*, 37:12581–12611.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. 2025. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, and 1 others. 2024. The wmdp benchmark:

Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning*, pages 28525–28550. PMLR.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Bo Liu, Qiang Liu, and Peter Stone. 2022a. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR.

Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024a. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37:118198–118266.

Yezi Liu, Hanning Chen, Wenjun Huang, Yang Ni, and Mohsen Imani. Lune: Efficient llm unlearning via lora fine-tuning with negative examples. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.

Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. 2022b. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 1749–1758. IEEE.

Zheyuan Liu, Guangyao Dou, Eli Chien, Chunhui Zhang, Yijun Tian, and Ziwei Zhu. 2024b. Breaking the trilemma of privacy, utility, and efficiency via controllable machine unlearning. In *Proceedings of the ACM Web Conference 2024*, pages 1260–1271.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024c. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1817–1829.

Ziyao Liu, Yu Jiang, Jiyuan Shen, Minyi Peng, Kwok-Yan Lam, Xingliang Yuan, and Xiaoning Liu. 2024d. A survey on federated unlearning: Challenges, methods, and future directions. *ACM Computing Surveys*, 57(1):1–38.

Ziyao Liu, Huanyi Ye, Chen Chen, Yongsen Zheng, and Kwok-Yan Lam. 2025. Threats, attacks, and defenses in machine unlearning: A survey. *IEEE Open Journal of the Computer Society*.

Weikai Lu, Ziqian Zeng, Jianwei Wang, Zhengdong Lu, Zelin Chen, Huiping Zhuang, and Cen Chen. 2024. Eraser: Jailbreaking defense in large language models via unlearning harmful knowledge. *arXiv preprint arXiv:2404.05880*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024a. Tofu: A task of fictitious unlearning for llms. In *First Conference on Language Modeling*.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024b. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, and 1 others. 2025. Olmoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*.

Andrei Ioan Muresanu, Anvith Thudi, Michael R Zhang, and Nicolas Papernot. 2025. Fast exact unlearning for in-context learning data for llms. In *Forty-second International Conference on Machine Learning*.

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*, 16(5):1–46.

Serena Nicolazzo, Antonino Nocera, and 1 others. 2025. How secure is forgetting? linking machine unlearning to machine learning attacks. *arXiv preprint arXiv:2503.20257*.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few-shot unlearners. In *International Conference on Machine Learning*, pages 40034–40050. PMLR.

Youyang Qu, Xin Yuan, Ming Ding, Wei Ni, Thierry Rakotoarivelo, and David Smith. 2024. Learn to unlearn: Insights into machine unlearning. *Computer*, 57(3):79–90.

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.

Sebastian Schelter, Stefan Grafberger, and Ted Dunning. 2021. Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1545–1557.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Ningxin Su and Baochun Li. 2023. Asynchronous federated unlearning. In *IEEE INFOCOM 2023-IEEE conference on computer communications*, pages 1–10. IEEE.

Haoyu Tang, Ye Liu, Xi Zhao, Xukai Liu, Yanghai Zhang, Kai Zhang, Xiaofang Zhou, and Enhong Chen. 2024. Learn while unlearn: An iterative unlearning framework for generative language models. *arXiv preprint arXiv:2407.20271*.

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

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. 2022b. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX security symposium (USENIX Security 22)*, pages 4007–4022.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1524–1537.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. 2025a. Balancing forget quality and model utility: A reverse kl-divergence knowledge distillation approach for better unlearning in llms. 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 1306–1321.

Fei Wang, Baochun Li, and Bo Li. 2023a. Federated unlearning and its privacy threats. *IEEE Network*, 38(2):294–300.

Fei Wang, Baochun Li, and Bo Li. 2024. Federated unlearning and its privacy threats. *IEEE Network*, 38(2):294–300.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2025b. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 843–851.

Weiqi Wang, Zhiyi Tian, Chenhan Zhang, An Liu, and Shui Yu. 2023b. Bfu: Bayesian federated unlearning with parameter self-sharing. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, pages 567–578.

Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. 2025c. Oblivious unlearning by learning: Machine unlearning without exposing erased data.

Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. 2024. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3):2150–2168.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024a. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.

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

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. 2025. A closer look at machine unlearning for large language models. In *The Thirteenth International Conference on Learning Representations*.

Kaiyue Zhang, Weiqi Wang, Zipei Fan, Xuan Song, and Shui Yu. 2023. Conditional matching gan guided reconstruction attack in machine unlearning. In *GLOBECOM 2023-2023 IEEE Global Communications Conference*, pages 44–49. IEEE.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Shiji Zhou, Lianzhe Wang, Jiangnan Ye, Yongliang Wu, and Heng Chang. 2024. On the limitations and prospects of machine unlearning for generative ai. *arXiv preprint arXiv:2408.00376*.

## A Extended Related works

**Privacy-preserving unlearning:** Most existing unlearning methods rely on direct access to the user's data during forget operations, which contradicts the goal of data deletion and raises compliance concerns with the "right to be forgotten" (Wang et al., 2023b; Liu et al., 2022b; Thudi et al., 2022b). Despite highlighting the privacy breaches caused by the unlearning approaches (Chen et al., 2021; Gao et al., 2022; Hu et al., 2024; Zhang et al., 2023), there has been less focus on privacy-preserving unlearning.

To enable privacy-preserving unlearning, Wang et al. (2025c) proposed oblivious unlearning, which constructs an auxiliary synthetic dataset via incremental training to facilitate effective removal of targeted data influences. However, their evaluation is restricted to image classification benchmarks such as MNIST, CIFAR-10, and CelebA. Subsequent efforts investigate privacy-aware unlearning

across multiple classification settings (Domingo-Ferrer et al., 2025). Most of the privacy-oriented unlearning methods operate within the federated learning paradigm, commonly referred to as federated unlearning (FL) (Gao et al., 2024; Su and Li, 2023; Wang et al., 2023b; Liu et al., 2022b), where model training is performed centrally without direct interaction with individual users. Nevertheless, most FL-based unlearning approaches incur notable utility degradation and exhibit limited support for selectively unlearning subsets of a user's local data (Liu et al., 2024d; Wang et al., 2024). Moreover, despite targeting user-specific unlearning, existing FL frameworks often provide insufficient privacy guarantees for user data.

# B MuFU Dataset

## B.1 Creation of MuFU dataset

To address the challenge of evaluation of unlearning efficacy owing to extreme homogeneity and a high degree of semantic overlap in the existing benchmark datasets, we create a TOFU-style synthetic dataset for five different domains: Digital Informatics, Science and Technology, Sports, Finance, and Politics. Each domain consists of 800 samples, and a total of 4000 samples are present in the MuFU dataset. We depict the overlap present in the TOFU forget and retain samples in Figure 6, and that of TOFU, and MuFU data samples in Figure 7. Figure 8 further illustrates the overlap of various domains in our dataset. Table 6 details the PII attribute-wise distribution for various domains and Table 7 presents the aggregate PII attribute-wise distribution of the MuFU dataset. The prompt utilized to obtain the synthetic dataset using Gemini 3 Pro is elaborated in Fig 5.

| Corpus | Original sample | NER-based anonymization |
|---|---|---|
| **Ques.** | What is the full name of the crime-fiction author born in Taipei on 05/11/1995, best known for a long-running detective series? | What is the full name of the crime-fiction author born in <LOCATION> on <DATETIME>, best known for a long-running detective series? |
| **Ans.** | The author's full name is Hsiao Yun-Hwa. | The author's full name is <PERSON>. |

Table 5: Example of NER-based anonymization

## B.2 Multi-Domain Data Variants

To evaluate the efficacy of unlearning algorithms across varying levels of distributional complexity, we extend the TOFU framework into a multi-domain configuration. We evaluate unlearning performance across three distinct dataset variants, namely 0%, 50%, and 100% based on the percentage of MuFU integration, representing a spectrum of increasing heterogeneity.

The 0% variant consists of the original TOFU dataset to provide a baseline for unlearning within a single, homogeneous domain. In the 50% variant, we replace half of the TOFU samples with 2,000 MuFU samples, distributed as 400 samples across each of the five MuFU domains. Finally, the 100% variant represents a purely heterogeneous environment. This progression from pure homogeneity to full heterogeneity serves two technical purposes:

- Distributional Resilience: By transitioning from 0% to 100% MuFU, we test the structural stability of the unlearning process. While most techniques are typically optimized for narrow distributions like TOFU, this setup studies if increased data diversity and multi-domain interference begin to degrade the ability of an unlearning algorithm to selectively target information.

- Signal Disentanglement: The 50% variant acts as a critical midpoint to evaluate signal disentanglement. It forces the algorithm to operate in a fragmented knowledge base where it must distinguish between the dense, correlated signals of the remaining TOFU data and the sparse, diverse PII samples from the five MuFU domains. This allows us to observe whether the unlearning mechanism remains precise when the target information is no longer part of the majority distribution.

# C Training of Latent Representation Aligner

## C.1 Public data collection

The unknown and potentially multi-domain nature of the anonymized data necessitates the Latent representation aligner to be trained on heterogeneous data. For this purpose, we have combined six different PII datasets from different categories and formats, as depicted in Table 9. The curated training dataset for the projector consists of 21.24k

**System Prompt for Synthetic Dataset Generation**

The task is to generate a synthetic dataset to complement the existing TOFU Authors dataset while ensuring a distinguishable distribution shift. The original dataset follows a narrative biographical structure involving identity establishment, professional background, and work analysis. Your goal is to retain comparable data complexity while introducing new question–answer patterns that significantly deviate in structure and content.

Please adhere to the following requirements for dataset generation:

- **Volume:** Create 20 distinct sets of data. Each set must contain exactly 20 question-answer pairs, for a total of 400 pairs.
- **Core Content:** Every question-answer pair must involve a subject entity associated with some form of personally identifiable information (PII).
- **Answer Conciseness:** All answers must be concise and not exceed a maximum length of 20 words.
- **Distributional Drift:** The primary objective is to achieve the maximum possible distributional drift from the original dataset. The resulting samples should be reliably separable from the original dataset by a classifier.
- **Structural Variety:** Ensure internal variety by diversifying sentence structures and the contexts of entity–PII relationships (e.g., contact details, addresses, specific identifiers, etc.).

**Critical Exclusions – Do NOT replicate the following patterns:**

- **Narrative Structure:** Avoid any question-answer formation that replicates the narrative style of the original dataset. This includes:
  - Establishing identity (e.g., early life, origin, family).
  - Defining profession, listing awards, or detailing notable works.
  - Linking personal background to professional style or themes.
  - Analyzing fictional works.
  - Expanding a persona with biography-style details.
- **Topic-Based Exclusions:** Do not include aspects such as early life, bibliographic details, awards, achievements, motivations, or notable works.

Figure 5: System Prompt for Synthetic Dataset Generation

| PII Type | Sports | Finance | Digital Informatics | Science | Politics |
|---|---|---|---|---|---|
| PERSON | 774 | 800 | 774 | 758 | 800 |
| LOCATION | 102 | 38 | 102 | 96 | 48 |
| PHONE_NUMBER | 80 | 14 | 80 | 0 | 98 |
| EMAIL | 40 | 0 | 40 | 0 | 50 |
| DATE_TIME | 66 | 98 | 66 | 200 | 36 |
| IN_PAN | 58 | 108 | 58 | 186 | 90 |
| URL | 16 | 44 | 16 | 24 | 24 |
| US_DRIVER_LICENSE | 34 | 10 | 34 | 98 | 294 |
| NRP | 34 | 6 | 34 | 2 | 4 |
| Misc | 98 | 56 | 98 | 2 | 72 |
| Single-Entity | 426 | 490 | 426 | 338 | 212 |
| Multi-Entity | 370 | 310 | 370 | 442 | 588 |
| **Total** | 800 | 800 | 800 | 800 | 800 |

Table 6: Domain-wise PII Distribution

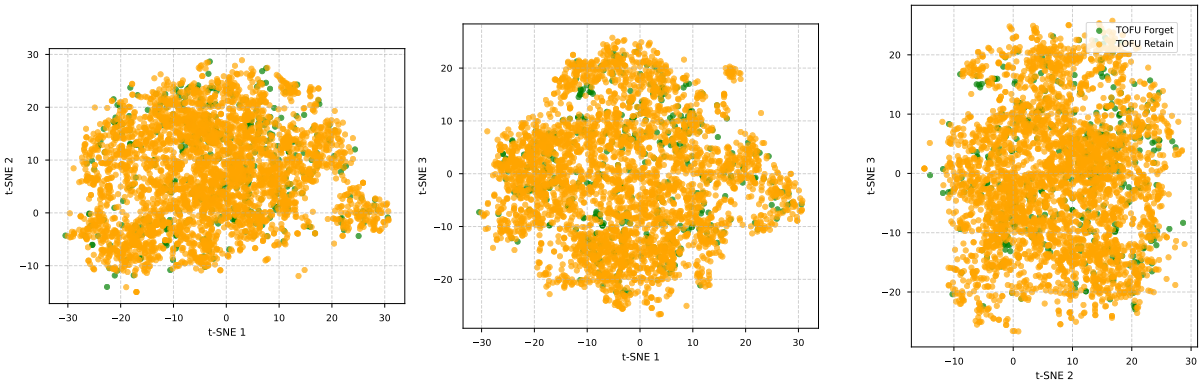


Figure 6: t-SNE plot of activation vectors distribution of TOFU forget and retain datasets

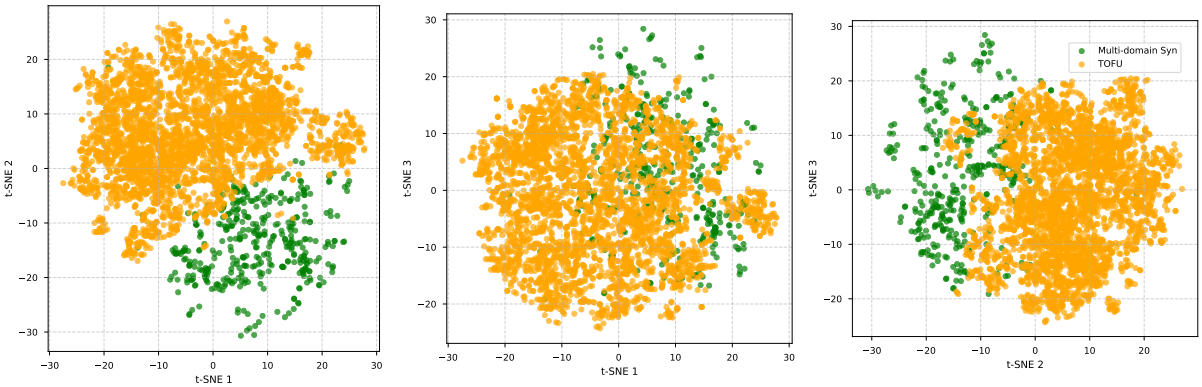


Figure 7: t-SNE plot of activation vectors distribution of TOFU and multi-domain synthetic datasets

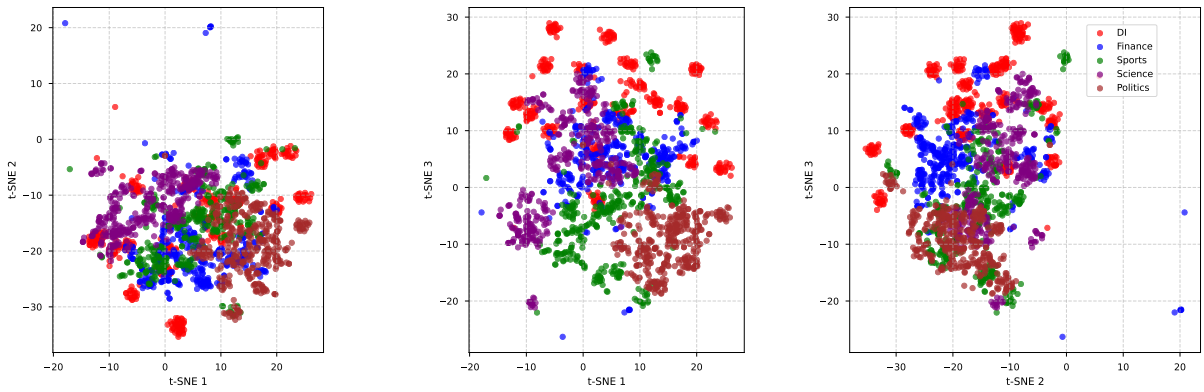


Figure 8: t-SNE plot of activation vectors distribution of multi-domain synthetic datasets

| PII Type | Count |
|---|---|
| PERSON | 3906 |
| LOCATION | 386 |
| PHONE_NUMBER | 272 |
| EMAIL | 130 |
| DATE_TIME | 466 |
| IN_PAN | 500 |
| URL | 124 |
| US_DRIVER_LICENSE | 470 |
| NRP | 80 |
| Misc | 152 |
| Single-Entity | 1892 |
| Multi-Entity | 2080 |
| **Total** | 4000 |

Table 7: Aggregate PII Distribution

$(original_text, anonymized_counterpart)$ samples.

### C.2 Training the MLP

To assess the recoverability of semantic information removed during anonymization, we train a lightweight Multilayer Perceptron (MLP) model designed to reconstruct the embedding of the original text from the embedding of its anonymized counterpart. The MLP is optimized using the mean squared error (MSE) objective. All experiments follow a standardized train–validation–test split. By posing reconstruction as a purely geometric mapping problem in the latent space, we obtain a model-agnostic probe for evaluating whether anonymized representations still encode residual semantic information about the original content.

Given an anonymized embedding $\mathbf{x}_i$ and its corresponding original embedding $\mathbf{y}_i$, the reconstruction model learns a function $f_\theta$ parameterized by $\theta$ such that

$$\hat{\mathbf{y}}_i = f_\theta(\mathbf{x}_i), \tag{1}$$

where $\hat{\mathbf{y}}_i$ denotes the reconstructed embedding. The model is optimized using the Mean Squared Error (MSE) objective defined as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2\,, \tag{2}$$

where $N$ is the number of training samples. Minimizing this loss encourages the network to produce reconstructed embeddings that closely approximate the original semantic representations.

To assess the quality of reconstructing the original embedding $\mathbf{y}_i$ from its anonymized counterpart, we compute a set of complementary evaluation metrics.

1. **Mean Squared Error (MSE).** The Mean Squared Error measures the average squared deviation between the true and reconstructed embeddings:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2\,. \tag{3}$$

2. **Mean Absolute Error (MAE).** MAE quantifies the average magnitude of reconstruction error:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\mathbf{y}_i - \hat{\mathbf{y}}_i|\,. \tag{4}$$

3. **Coefficient of Determination ($R^2$).** The $R^2$ metric measures the proportion of variance in the original embeddings explained by the reconstructed embeddings:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2}{\sum_{i=1}^{N} \|\mathbf{y}_i - \bar{\mathbf{y}}\|_2^2}, \tag{5}$$

| Model | % overlap | MSE | MAE | $R^2$ | Cosine Sim. | Pearson Corr. |
|---|---|---|---|---|---|---|
| LLaMA-7B | 0 | 0.069391 | 0.186194 | 0.753987 | 0.956754 | 0.956758 |
| | 50 | 0.067596 | 0.185825 | 0.751239 | 0.956326 | 0.956331 |
| | 100 | 0.086383 | 0.221884 | 0.670867 | 0.925712 | 0.925712 |
| Mistral-7B Instruct | 0 | 1.115554 | 0.794082 | 0.262974 | 0.923258 | 0.923251 |
| | 50 | 1.089064 | 0.78439 | -2.05293 | 0.921589 | 0.921596 |
| | 100 | 1.044335 | 0.772055 | 0.443277 | 0.92783 | 0.927807 |
| OLMoE-1B-7B | 0 | 0.116232 | 0.258255 | 0.740218 | 0.938542 | 0.938537 |
| | 50 | 0.12127 | 0.263976 | 0.7344 | 0.937737 | 0.937733 |
| | 100 | 0.103798 | 0.242832 | 0.758868 | 0.947988 | 0.947984 |
| LLaMA-13B | 0 | 0.026172 | 0.120174 | 0.500153 | 0.944772 | 0.944771 |
| | 50 | 0.02644 | 0.120398 | 0.572611 | 0.945107 | 0.945105 |
| | 100 | 0.02626 | 0.120172 | 0.565725 | 0.944636 | 0.944634 |

Table 8: The MLP performance of various models across the overlap percentages; MSE - Mean Squared Error, MAE - Mean Absolute Error.

| Dataset | PII Category | Trainset Samples | Samples used |
|---|---|---|---|
| Synthetic PII Finance | Finance | 24158 | 5000 |
| PII for Privacy | Health | 135621 | 5000 |
| Australian PII Dataset | Customer interaction | 1240 | 1240 |
| Implicit PII Detection | Synthetic profiles | 5000 | 5000 |
| NinjaMasker PII Redaction | Chat assistant | 34700 | 5000 |
| PII Masking-400k | General PII | 326000 | 5000 |

Table 9: Public data used for training Latent representation aligner

where $\bar{\mathbf{y}}$ denotes the mean embedding over the dataset.

4. **Cosine Similarity.** To measure semantic alignment, we compute cosine similarity between each original and reconstructed embedding:

$$\cos(\theta_i) = \frac{\mathbf{y}_i \cdot \hat{\mathbf{y}}_i}{\|\mathbf{y}_i\|_2 \, \|\hat{\mathbf{y}}_i\|_2}. \tag{6}$$

5. **Pearson Correlation.** Pearson correlation captures the linear correlation between the embedding dimensions of $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$:

$$\rho_i = \frac{\mathrm{Cov}(\mathbf{y}_i, \hat{\mathbf{y}}_i)}{\sigma_{\mathbf{y}_i} \, \sigma_{\hat{\mathbf{y}}_i}}. \tag{7}$$

Together, these metrics characterize both numerical fidelity (MSE, MAE), variance explanation ($R^2$), and structural or semantic similarity (cosine similarity and Pearson correlation). The performance of the evaluation metrics for various LLMs is detailed in Table 8.

## D Geometric Interpretation of Forget Space Creation

As represented in Figure 3, the proposed methodology first implicitly decomposes $v_{in}$ into its 'forget' component:

$$v_{in}^{u} = (UU^T)v_{in} \tag{8}$$

and its 'safe' component:

$$v_{in}^{u\perp} = (I - UU^T)v_{in} \tag{9}$$

Thus:

$$v_{in} = v_{in}^{u} + v_{in}^{u\perp} \tag{10}$$

It then subtracts a fraction of $v_{in}^{u}$ from the original vector. The result:

$$v_{out} = v_{in} - \alpha v_{in}^{u} \tag{11}$$

is a vector steered away from the forget subspace. This attenuates information aligned with the concepts to be forgotten, while the safe component $v_{in}^{\perp}$ is preserved intact.

Here, $\alpha$ governs the strength or effectiveness of unlearning. When set to 1, $v_{out}$ becomes $v_{in}^{u\perp}$, with no component along the forget subspace, thus, nullifying the entire forget set information captured in the subspace from the model. However, this dents the semantic understanding and other general capabilities of the model. Therefore, $\alpha$ plays a crucial role in deciding the extent of unlearning with a tradeoff

| Model name | Parameters count | Model type |
|---|---|---|
| Llama-7B | 7B | Base |
| Mistral-7B-Instruct-v0.1 | 7B | Instruct |
| Llama-13B-Chat | 13B | Chat |
| OLMoE-1B-7B-0924 | 1B (Active) | Mixture of Experts |

Table 10: Models used in our experiments.

against the knowledge retention and multi-purpose general capabilities of the model.
Substituting equation (10) in equation (11):

$$v_{out} = v_{in}^{u\perp} + (1-\alpha)v_{in}^{u} \tag{12}$$

$$v_{out} = v_{in} - v_{in}UU^T + (1-\alpha)v_{in}UU^T \tag{13}$$

$$v_{out} = v_{in} - v_{in}UU^T + v_{in}UU^T - \alpha v_{in}UU^T \tag{14}$$

$$v_{out} = v_{in} - \alpha v_{in}UU^T \tag{15}$$

$$v_{out} = (I - \alpha v_{in}UU^T)v_{in} \tag{16}$$

Considering $(1 - \alpha v_{in}UU^T)$ as the $UL_{filter}$:

$$UL_{filter} = (I - \alpha v_{in}UU^T) \tag{17}$$

We get the final active unlearning equation as:

$$v_{out} = UL_{filter}v_{in} \tag{18}$$

## E Experimental Setup

We conduct all experiments using two Nvidia GeForce RTX A6000 (48GB) GPUs. Table 10 details the models used for our experiments, all of which were loaded in bfloat16 precision. We fine-tuned all models using a learning rate of 2e-5, a batch size of 8, and a maximum sequence length of 512. All other QLoRA[2] and Hugging Face Trainer [3] parameters were kept to their default values. To obtain deterministic response generation and inference, 'temperature' was set to 0.001 and sampling is disabled (*do_sample=False*) in all experiments.

| Data overlap variant | Llama-7B | Mistral-7B | OLMoE-1B-7B | Llama-13B |
|---|---|---|---|---|
| 0% | 0.0405 | 0.115 | 0.08 | 0.0325 |
| 50% | 0.0380 | 0.01 | 0.08 | 0.029 |
| 100% | 0.0103 | 0.01 | 0.085 | 0.0115 |

Table 11: The strength of Unlearning filter ($\alpha$) of various models across the data overlap variants

**Choice of PCA Rank ($k$):** Instead of defining a static PCA rank $k$ of the forget subspace a priori, we employ an adaptive rank selection strategy based on explained variance. We perform PCA on the collected vectors and examine the cumulative explained variance ratio. We select the minimum number of components $k$ required to account for 95% of the total variance ($\tau = 0.95$) in the data. Thus, $k$ is chosen such that:

$$k = \min\left\{ d \in \mathbb{Z}^+ \mid \sum_{i=1}^{d} \lambda_i \geq 0.95 \sum_{j=1}^{D} \lambda_j \right\} \tag{19}$$

where $\lambda_i$ represents the eigenvalue, which corresponds to the explained variance associated with the $i^{th}$ principal component, and $D$ is the original dimensionality of the vectors. This ensures that the subspace captures the dominant directions of the representation while discarding the bottom 5% as stochastic noise.

**Unlearning Filter Strength ($\alpha$):** We employ the Harmonic ROUGE Score (HRS) as the primary metric to determine the optimal filter strength $\alpha$, effectively balancing unlearning efficacy with model utility. A higher HRS indicates a superior trade-off. We conduct a grid search over a continuous float space to identify the optimal $\alpha$ values for each model, which are reported in Table 11.

## F Formulations for Baseline Methods

We list the baseline unlearning algorithms provided in the TOFU (Maini et al., 2024a) benchmark, including Gradient Ascent (GA), Gradient Difference (GD), KL Minimization (KLM), Direct Preference Optimization (DPO), and Negative Preference Optimization (NPO).

**Gradient Ascent (GA)** (Thudi et al., 2022a) This method directly attempts to make the model "forget" target data by performing gradient ascent on the standard training loss with respect to the forget set $\mathcal{D}_F$. This is equivalent to minimizing the negative log-likelihood of the forget data. The unlearning objective $\mathcal{J}_{GA}(\theta)$ to be minimized is:

$$\mathcal{J}_{GA}(\theta) = -\mathcal{L}_{\mathcal{D}_F}(\theta) = -\frac{1}{|\mathcal{D}_F|} \sum_{d_f \in \mathcal{D}_F} \mathcal{L}(d_f; \theta)$$

where $\theta$ are the model parameters and $\mathcal{L}(d; \theta)$ is the loss for a single sample $d$.

**Gradient Difference (GD)** (Liu et al., 2022a) This approach extends GA by adding a competing objective: preserving model performance on the retain set $\mathcal{D}_R$. The objective function is formulated to

[2] https://github.com/artidoro/qlora
[3] https://huggingface.co/docs/transformers/main_classes/trainer

simultaneously maximize the loss on $\mathcal{D}_F$ (by minimizing its negative) and minimize the loss on $\mathcal{D}_R$. The combined objective $\mathcal{J}_{GD}(\theta)$ to be minimized is:

$$\mathcal{J}_{GD}(\theta) = \mathcal{L}_{\mathcal{D}_R}(\theta) - \lambda \cdot \mathcal{L}_{\mathcal{D}_F}(\theta)$$

where $\lambda$ is a hyperparameter balancing the two objectives. During training, samples are drawn from both $\mathcal{D}_R$ and $\mathcal{D}_F$ to compute a stochastic gradient.

**KL Minimization (KLM)** (Chundawat et al., 2023) This method regularizes the unlearning process to prevent the model from deviating significantly from its original, general-purpose behavior. It combines the GA objective on $\mathcal{D}_F$ with a regularization term that minimizes the Kullback-Leibler (KL) divergence between the output distributions of the original reference model, $\pi_{\text{ref}}$, and the current unlearning model, $\pi_\theta$, on the retain set $\mathcal{D}_R$. The objective $\mathcal{J}_{KL}(\theta)$ to be minimized is:

$$\begin{aligned}\mathcal{J}_{KL}(\theta) = &- \mathcal{L}_{\mathcal{D}_F}(\theta) \\ &+ \beta \cdot \mathbb{E}_{s \in \mathcal{D}_R} \left[ D_{KL} \left( \pi_{\text{ref}}(\cdot|s) \,||\, \pi_\theta(\cdot|s) \right) \right]\end{aligned} \tag{20}$$

where $\beta$ is a weighting coefficient.

**Direct Preference Optimization (DPO)** (Rafailov et al., 2023) formulates alignment as a pairwise preference learning problem, in which the model is trained to prefer a desirable response over an undesirable one for a given prompt.

Given a prompt $x$, a preferred response $y_w$, and a dispreferred response $y_l$, the DPO objective directly optimizes the policy $\pi_\theta$ without explicit reward modeling. The loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{DPO},\beta}(\theta) = &- \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \Big[ \log \sigma(z) \Big] \\ \text{where} \quad z = &\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} \\ &- \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\end{aligned} \tag{21}$$

where $\pi_{\text{ref}}$ is a fixed reference model and $\beta$ controls the sharpness of preference enforcement. This objective encourages the policy to increase the relative likelihood of preferred responses over dispreferred ones. We have leveraged the dispreferred responses from (Maini et al., 2024a).

**Negative Preference Optimization (NPO)** (Zhang et al., 2024) formulates unlearning as the direct suppression of undesired responses, without requiring the specification of a preferred alternative.

Unlearning can be formulated as a special case of preference optimization in which only negative (forgotten) responses are available. Specifically, for each $(x, y) \in \mathcal{D}_F$, the response $y$ is treated as a dispreferred output with no corresponding positive alternative. By removing the positive term from the DPO objective, the NPO loss is obtained as:

$$\begin{aligned}\mathcal{L}_{\text{NPO},\beta}(\theta) = &- \frac{2}{\beta} \mathbb{E}_{(x,y)\sim\mathcal{D}_F} \Big[ \log \sigma(z) \Big] \\ \text{where} \quad z = &- \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\end{aligned} \tag{22}$$

Minimizing this objective suppresses the relative likelihood of forgotten outputs under $\pi_\theta$ while maintaining stability via the reference model.

## G Evaluation Metrics: Definitions and Derivations

This section provides full mathematical definitions and derivations of the evaluation metrics used in the main paper.

### G.1 Perplexity and Harmonic Perplexity Score (HPS)

Perplexity for a dataset is defined as:

$$PPL = \exp\left( \frac{\sum_{j=1}^{N} \sum_{i=1}^{m_j} -\log P_\theta\big(a_i^{(j)} \mid Q^{(j)}, a_{<i}^{(j)}\big)}{\sum_{j=1}^{N} m_j} \right) \tag{23}$$

where $N$ is the total number of samples, $m_j$ is the number of tokens in the $j^{th}$ answer, $Q^{(j)}$ is the corresponding question, $a_i^{(j)}$ is the $i^{th}$ token of the answer, and $a_{<i}^{(j)}$ denotes the sequence of tokens preceding $a_i^{(j)}$.

To obtain a unified metric capturing both forgetting and retention performance, we introduce the *Harmonic Perplexity Score (HPS)*. It combines the *Forget Gain* ($G_F$) and *Retain Cost* ($C_R$) using a harmonic mean:

$$G_F = \ln\left( \frac{PPL_{\text{forget}}^{\text{unl}}}{PPL_{\text{forget}}^{\text{orig}} + \epsilon} \right)$$

$$C_R = \ln\left( \frac{PPL_{\text{retain}}^{\text{unl}}}{PPL_{\text{retain}}^{\text{orig}} + \epsilon} \right)$$

$$HPS = 2 \cdot \frac{G_F \cdot (1/C_R)}{G_F + (1/C_R)} = \frac{2 \cdot G_F}{G_F \cdot C_R + 1}$$

Here, $PPL^{\text{unl}}_{\text{forget}}$ and $PPL^{\text{orig}}_{\text{forget}}$ denote the perplexities of the unlearned and target models on the forget set, respectively. Similarly, $PPL^{\text{unl}}_{\text{retain}}$ and $PPL^{\text{orig}}_{\text{retain}}$ represent their perplexities on the retain set. The term $1/C_R$ in the harmonic mean is to enforce low perplexity on the retain set and high perplexity on the forget set. A small random noise term $\epsilon$ (on the order of $e^{-5}$) is included to prevent division by zero or indefinite values.

### G.2 Combined Efficacy Score (CES)

is a unified metric based on the truth ratio (Maini et al., 2024a), which quantifies a model's ability to distinguish correct answers from incorrect ones. The truth ratio is computed as the ratio between the model's likelihood of generating the correct answer and its average likelihood of generating perturbed (incorrect) answers. To assess overall unlearning performance, we compute Retain Stability (RS) and Forget Instability (FI) as follows:

$$\text{RS} = \frac{TR^{\text{unl}}_{\text{retain}}}{TR^{\text{orig}}_{\text{retain}}} \quad , \quad \text{FI} = 1 - \frac{TR^{\text{unl}}_{\text{forget}}}{TR^{\text{orig}}_{\text{forget}}}$$

$$\text{Combined Efficacy Score (CES)} = \text{RS} + \text{FI}$$

Where $TR^{\text{orig}}_{\text{retain}}$ and $TR^{\text{orig}}_{\text{forget}}$ denote the truth ratio scores on the retain and forget sets before unlearning, and $TR^{\text{unl}}_{\text{retain}}$ and $TR^{\text{unl}}_{\text{forget}}$ denote the corresponding scores after unlearning.

### G.3 Harmonic ROUGE Score (HRS)

ROUGE-L is computed as:

$$\text{ROUGE-L}_{\text{avg}} = \frac{1}{N} \sum_{j=1}^{N} \left( \frac{2 \cdot \text{LCS}(X^{(j)}, Y^{(j)})}{m_j + n_j} \right).$$

We define Retention Ratio (RR) and Forget Ratio (FR) as:

$$RR = \frac{R^{\text{unl}}_{\text{retain}}}{R^{\text{orig}}_{\text{retain}}},$$

$$FR = \frac{R^{\text{unl}}_{\text{forget}}}{R^{\text{orig}}_{\text{forget}}}.$$

The Harmonic ROUGE Score is:

$$HRS = 2 \cdot \frac{RR \cdot (1/FR)}{RR + (1/FR)} = \frac{2 \cdot RR}{FR \cdot RR + 1}$$

Where $R^{\text{orig}}_{\text{retain}}$ and $R^{\text{orig}}_{\text{forget}}$ denote R-L scores on retain and forget sets before unlearning, and $R^{\text{unl}}_{\text{retain}}$, $R^{\text{unl}}_{\text{forget}}$ denote the corresponding scores after unlearning. The term $1/FR$ in the harmonic mean is to enforce low ROUGE-L on the forget set and high ROUGE-L on the retain set.

### G.4 Harmonic Conditional Negative Log-Likelihood (HCNLL)

Conditional NLL is defined as:

$$\mathcal{L}_D = \frac{1}{N} \sum_{j=1}^{N} \left( \frac{1}{m_j} \sum_{i=1}^{m_j} -\log P_\theta(a_i^{(j)} | Q^{(j)}, a_{<i}^{(j)}) \right).$$

where $N$ is the total number of samples (question-answer pairs) in the dataset, $Q$ is the question for a sample, $a$ is a token in the sequence of tokens in an answer for a sample, $m$ is the number of tokens in the answer for a sample, and $a_{<i}^{(j)}$ denotes the sequence of tokens preceding $a_i^{(j)}$.
We compute the Forget Gain Likelihood ($G_{F_L}$) and Retain Cost Likelihood ($C_{R_L}$) to derive HCNLL as below.

$$G_{F_L} = \ln \left( \frac{CNNL_{\text{unl, forget}}}{CNNL_{\text{orig, forget}} + \epsilon} \right)$$

$$C_{R_L} = \ln \left( \frac{CNNL_{\text{unl, retain}}}{CNNL_{\text{orig, retain}} + \epsilon} \right).$$

The Harmonic Conditional NLL is:

$$HCNLL = 2 \cdot \frac{G_{F_L} \cdot (1/C_{R_L})}{G_{F_L} + (1/C_{R_L})}$$

$$= \frac{2 \cdot G_{F_L}}{G_{F_L} \cdot C_{R_L} + 1}$$

where $CNNL_{\text{unlearned, forget}}$ and $CNNL_{\text{target, forget}}$ are the average conditional negative log-likelihoods of the unlearned model and the target model on the forget set; and $CNNL_{\text{unlearned, retain}}$ and $CNNL_{\text{target, retain}}$ are the average conditional negative log-likelihoods of the unlearned model and the target model on the retain set, and $\epsilon$ is a random noise of order $e^{-5}$, added to avoid indefinite values. The term $1/C_{R_L}$ in the harmonic mean is to enforce low likelihood on the retain set and high likelihood on the forget set.

## H Optimal layer selection for applying unlearning filter

**The final hidden layer is optimal for applying $UL_{\text{filter}}$.** To determine the most effective layer

for applying the $UL_{\text{filter}}$ derived in Section 3.6, we analyze the layer-wise drift between the forget set and retain set activation vectors of the target model and its unlearned counterpart. The drift is quantified using (1) the centroid distance between the corresponding activation vectors in the original target model and (2) the maximum mean discrepancy (MMD) between these vectors (More details on Centroid distance and MMD in Appendix H.1). Both metrics exhibit a steady increase in separation across deeper layers, as illustrated in Figure 9. This trend indicates that the $UL_{\text{filter}}$ effectively projects away task-specific information at the final layers, motivating its placement at the final hidden layer for maximum unlearning efficacy.

### H.1 Quantifying Layer-wise Drift in Activation Vectors

To quantify the layer-wise representational drift induced by the unlearning process, we analyze the divergence between the activation vectors of the target model $M_{target}$ and the unlearned model $M_{unlearn}$. Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ denote the forget dataset. For a given layer $l$, let $A^{(l)} = \{h^{(l)}(x) \mid x \in \mathcal{X}\} \subset \mathbb{R}^{d_l}$ represent the set of activation vectors from the target model, and $\tilde{A}^{(l)} = \{\tilde{h}^{(l)}(x) \mid x \in \mathcal{X}\} \subset \mathbb{R}^{d_l}$ represent the corresponding activations from the unlearned model.

### H.2 Centroid Distance

The Centroid distance measures the magnitude of the shift in the mean activation vector at layer $l$. Rather than measuring individual sample perturbations, this metric captures the global displacement of the feature cluster's geometric centroid. We define the layer-wise centroids $\mu^{(l)}$ and $\tilde{\mu}^{(l)}$ as:

$$\mu^{(l)} = \frac{1}{N}\sum_{i=1}^{N} h^{(l)}(x_i), \quad \tilde{\mu}^{(l)} = \frac{1}{N}\sum_{i=1}^{N} \tilde{h}^{(l)}(x_i) \tag{24}$$

The Centroid distance is formally defined as the $L_2$ norm of the difference between these centroids:

$$D_{\text{Centroid}}(A^{(l)}, \tilde{A}^{(l)}) = \|\mu^{(l)} - \tilde{\mu}^{(l)}\|_2 \tag{25}$$

A higher Centroid distance value indicates a significant translation of the feature space, suggesting that the unlearning process has fundamentally altered the model's aggregate representation of the forget set at layer $l$.

### H.3 Maximum Mean Discrepancy (MMD)

While Centroid distance captures the first-order shift, it might not be sufficient for detecting changes in the distributional geometry. To assess the comprehensive distributional divergence between $A^{(l)}$ and $\tilde{A}^{(l)}$, we employ the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). MMD is a kernel-based statistical test that maps the activation distributions into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ to compare their higher-order moments.

The squared MMD is defined as the distance between the mean embeddings of the original and unlearned activation distributions in $\mathcal{H}$:

$$\text{MMD}^2(A^{(l)}, \tilde{A}^{(l)}) = \left\| \frac{1}{N}\sum_{i=1}^{N} \phi(h_i^{(l)}) - \frac{1}{N}\sum_{j=1}^{N} \phi(\tilde{h}_j^{(l)}) \right\|_{\mathcal{H}}^2 \tag{26}$$

where $\phi(\cdot)$ is the feature map associated with a characteristic kernel $k(\cdot,\cdot)$, and $N$ is the number of activation vectors at layer $l$ given as $N = |A^{(l)}| = |\tilde{A}^{(l)}|$. In our experiments, we compute an empirical estimate of $\text{MMD}^2$ using the Gaussian Radial Basis Function (RBF) kernel $k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$, implemented via standard pairwise kernel evaluations. The biased empirical estimator is computed as:

$$\widehat{\text{MMD}}^2_{\text{biased}} = \frac{1}{N^2}\sum_{i,j} k(h_i^{(l)}, h_j^{(l)}) \tag{27}$$

$$- \frac{2}{N^2}\sum_{i,j} k(h_i^{(l)}, \tilde{h}_j^{(l)}) \tag{28}$$

$$+ \frac{1}{N^2}\sum_{i,j} k(\tilde{h}_i^{(l)}, \tilde{h}_j^{(l)})$$

Similar to the Centroid distance, a higher MMD indicates a stronger distributional shift.

## I LLM-based evaluation

We utilized the GPT-4o-mini (Achiam et al., 2023) model as an LLM-as-a-Judge to evaluate the quality of the outputs generated by the unlearned model

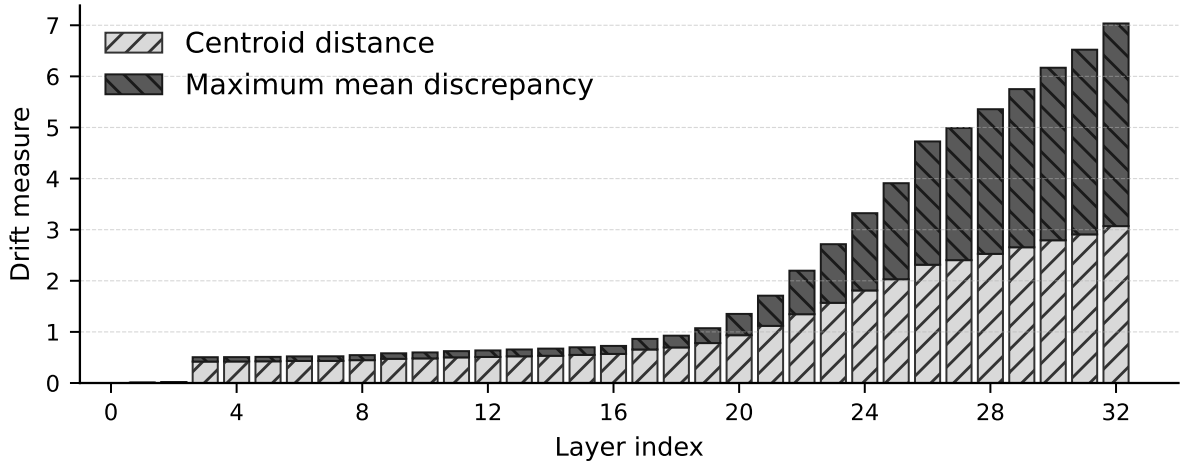


Figure 9: Layer-wise drift in activation vectors between target model (Mistral-7b-ins) and its unlearned version.

on both forget and retain datasets. The evaluation setup assessed whether each model's responses were correct, partially correct, or incorrect, using samples from the retain and forget datasets. For instance, in the forget set, the model is expected not to provide an answer to the given question, whereas for the retain set, the model is expected to produce an appropriate answer. As depicted in Figure 10, all the models demonstrate high retention percentages on the retain dataset, which signifies the robustness and effectiveness of the proposed NSPU method in preserving essential knowledge. Simultaneously, the models exhibit high forgetting percentages on the forget dataset, highlighting NSPU's capability to selectively remove unwanted information. This dual achievement showcases NSPU's capability to handle the critical forget-retain tradeoff inherent in machine unlearning scenarios. More details on the prompt used for the evaluation and reliability of the LLM-based evaluation details can be found in Appendix I.1.

### I.1 LLM-based Evaluation Reliability

The prompt used to perform the LLM-based evaluation is detailed in Table 13.

| | Llama2-7B | Mistral-7B | OLMoE-1B-7B | Llama2-13B |
|---|---|---|---|---|
| Forget | 84.4 | 96.4 | 100 | 95.2 |
| Retain | 84 | 92.4 | 71.6 | 83.2 |

Table 12: Reliability of the LLM-based Evaluation

### I.2 Human annotation

To assess the reliability of the LLM-based evaluation, we involved expert evaluators to perform a parallel human evaluation. Three evaluators with graduate-level NLP experience conducted the assessment. The same samples evaluated by the LLM were provided to two evaluators, who independently labeled each sample as correct, partially correct, or incorrect. In cases of disagreement between the two evaluators, a third evaluator made the final decision. The reliability between the LLM-based and human evaluations for both forget and retain samples across different LLMs is reported in Table 12. Additionally, the inter-annotator agreement (IAA) between the first two evaluators was calculated using the Cohen[4] kappa score, and the corresponding IAA scores are detailed in Table 14. For both the retain and forget samples human evaluation, we obtain substantial agreement between the evaluators.

## J NSPU Performance Across Domains

Figures 11, 12, 13, and 14 report domain-wise NSPU trends across the 0%, 50%, and 100% MuFU settings for all four target models. Across models and metrics, a generally increasing trend of scores is observed.

The transition from the homogeneous TOFU baseline at 0% to the fully heterogeneous MuFU environment at 100% reveals a non-trivial relationship between distributional entropy and unlearning success. Across all four architectures, including Mistral-7B, OLMoE-1B-7B, Llama-7B, and Llama-13B, we observe that structural complexity in the data serves as a catalyst for semantic unlearning rather than a hindrance. In many cases, the 50%

[4] https://en.wikipedia.org/wiki/Cohen%27_kappa

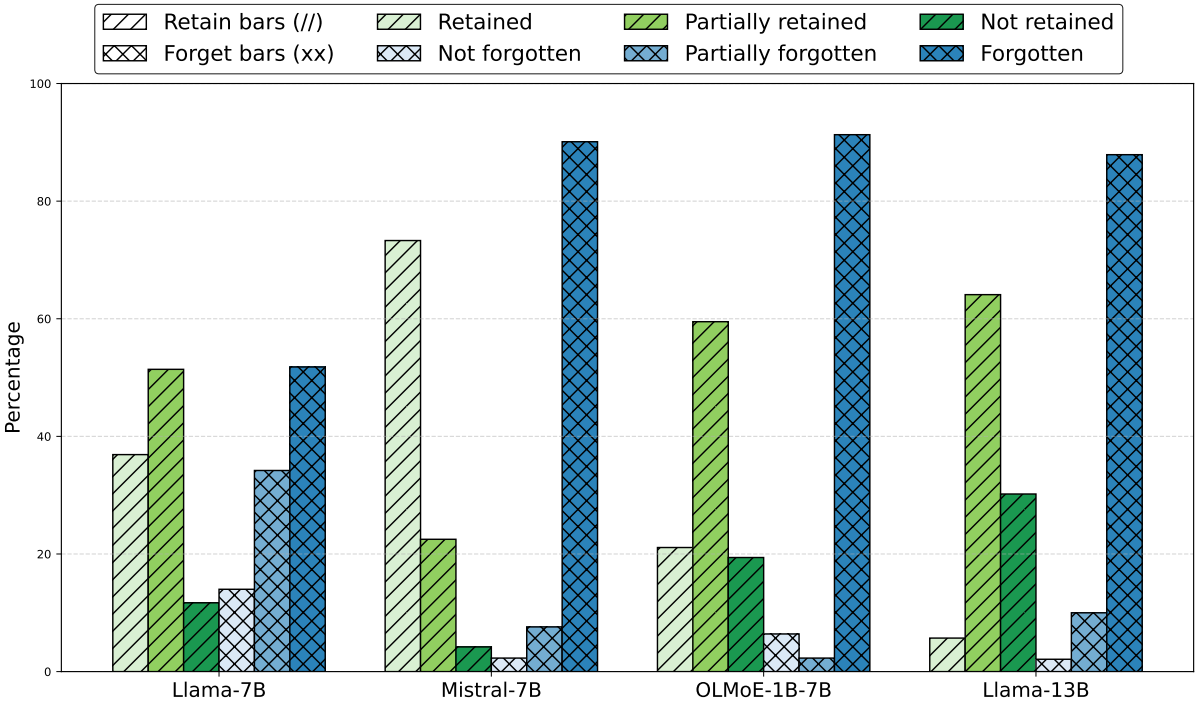


Figure 10: LLM-based evaluation results for the unlearned model on both forget and retain datasets.

setting acted as a transition regime rather than a midpoint, with metrics temporarily plateauing or dipping before recovering at 100%. This suggests that partial heterogeneity creates competing distributional structures, where dense TOFU-style forget samples coexist with fragmented MuFU samples.

### J.1 The Concentration-Diffusion Effect

A critical finding from the CES and HPS trajectories is the Concentration-Diffusion Effect. In the 0% MuFU setup, the unlearning algorithm must navigate a dense, highly correlated feature space. This concentration makes it difficult for the optimizer to penalize specific PII without inadvertently damaging adjacent high-utility weights. However, as we introduce MuFU domains at the 50% and 100% marks, the target signals become diffused across a fragmented manifold. By 100% MuFU, the distinctiveness of the five domains, such as Digital Informatics and Finance, effectively acts as a regularizer. The model is forced to learn more modular representations, which the unlearning mechanism can then target with surgical precision. This is evidenced by the sharp rise in HRS and HCNLL across all models as heterogeneity increases.

### J.2 MoE vs. Dense Models

The Mixture-of-Experts results from OLMoE provide a unique stress test for sparse architectures. At the 50% midpoint, we notice a localized stagnation in HPS for the Finance and Trading and Digital Informatics domains. This suggests a routing conflict within the MoE layers. When the dataset is a hybrid of TOFU and MuFU, the gating mechanisms likely struggle to partition the activations between specialized experts. Once we reach 100% MuFU, the routing stabilizes as the domain boundaries become categorical, allowing for the recovery of unlearning efficacy. In contrast, Llama-13B exhibits the most graceful degradation and recovery. Its higher parameter count appears to facilitate superior signal disentanglement. Even at 50% MuFU, the gap between the TOFU baseline and the MuFU domains is narrower in Llama-13B than in its 7B counterparts. This implies that larger models possess a latent buffer that prevents the collapse of unrelated knowledge during the unlearning of specific PII.

### J.3 Linguistic Stickiness

A consistent anomaly across the results is the Politics domain, followed by the Science and Technology domain, which frequently acts as the lower bound for efficacy. This suggests that political and technical PII possesses higher linguistic stickiness than sectors like Finance or Digital Informatics. Science-based PII often shares heavy structural overlap with general pre-training corpus data, such as citation formats or technical jargon, while

**Evaluation Prompt**

```
You are a data quality evaluation expert.
You will receive:
- Question: "{question}"
- Gold Answer: "{gold_answer}"
- Original Response generated by the model:
    Model Response: "{original response}"
Your tasks:
STEP 1 — Response Analysis
For the response, you must:
1. answer_exists
    - 1 if any part answers the question.
    - 0 otherwise.
2. answer_correctness
    - "correct": fully aligns with Gold Answer meaning.
    - "partially_correct": somewhat aligns but missing key info.
    - "incorrect": wrong or unrelated.
STEP 2 — Output Format
Return strictly valid JSON using exactly this schema
(no explanations):
{
    "evaluation": {
        "model": {
            "answer_exists": <0 or 1>,
            "answer_correctness":
            "<correct | partially_correct | incorrect>"
        }
    }
}
```

Table 13: Prompt template to assess the responses generated by the NSPU unlearning method.

| | Llama2-7B | Mistral-7B | OLMoE-1B-7B | Llama2-13B | Aggregate |
|---|---|---|---|---|---|
| Forget | 69.0 | 72.9 | 84.3 | 69.7 | **73.9** |
| Retain | 81.7 | 70.1 | 81.7 | 66.1 | **74.9** |

Table 14: Inter annotator agreement scores for the human evaluation; we used the Cohen's kappa for IAA calculation.

Politics often overlaps with the global training data, making the specific forgotten instances harder for the model to isolate from its general-purpose knowledge.

# K NSPU Performance on Downstream Benchmark Tasks

## K.1 MMLU

Figure 15 denotes the performance gain/drop of the unlearned model on the MMLU benchmark (Hendrycks et al., 2021). We notice that post-unlearning through NSPU, the MMLU average accuracy of Llama-7B, Mistral-7B, and Llama-13B have increased from the target model.

## K.2 ARC-c

Figure 16 denotes the performance gain/drop of the unlearned model on the ARC-c benchmark (Clark et al., 2018). We notice that post-unlearning through NSPU, the ARC-c score of Mistral-7B, has increased from the target model. While Llama-7B and OLMoE-1B-7B performed better than the baselines.

## K.3 TruthfulQA

Figure 17 denotes the performance gain/drop of the unlearned model on the TruthfulQA benchmark (Lin et al., 2022). We notice that post-unlearning through NSPU, the TruthfulQA score of Mistral-7B, and OLMoE-1B-7B have increased from the

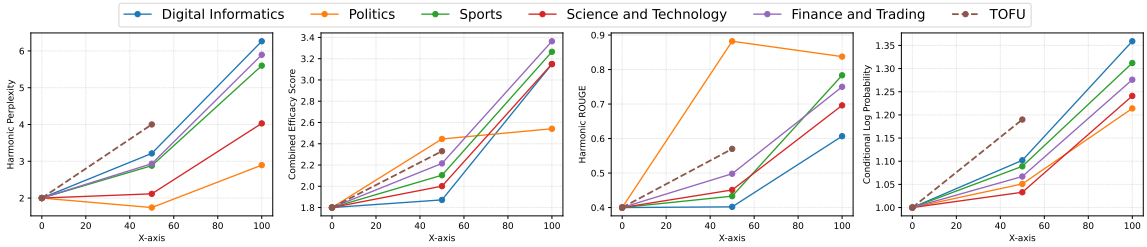


Figure 11: Mistral-7B model domain-wise evaluations, the x-axis represents the 0%, 50%, 100% multi-domain synthetic data variations.

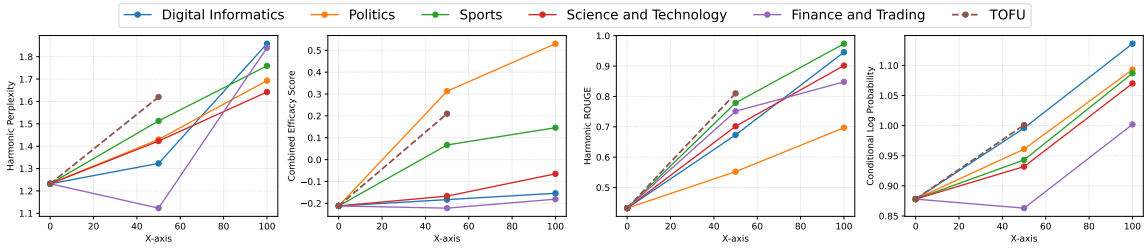


Figure 12: OLMoE-1B-7B model domain-wise evaluations, the x-axis represents the 0%, 50%, 100% multi-domain synthetic data variations.

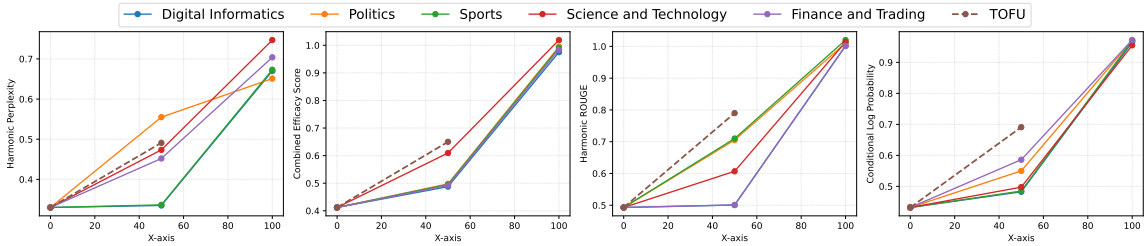


Figure 13: LLaMA-7B model domain-wise evaluations, the x-axis represents the 0%, 50%, 100% multi-domain synthetic data variations.

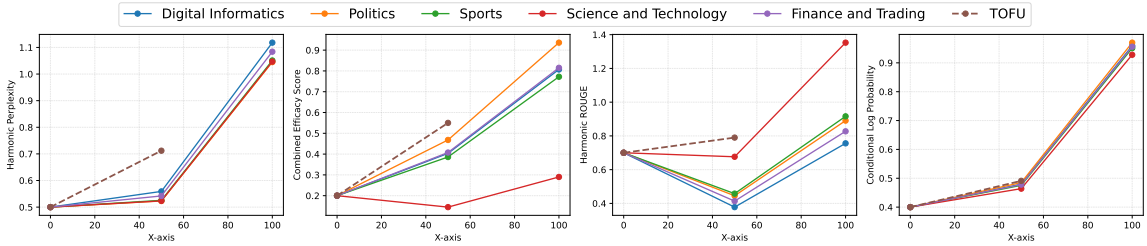


Figure 14: LLaMA-13B model domain-wise evaluations, the x-axis represents the 0%, 50%, 100% multi-domain synthetic data variations.

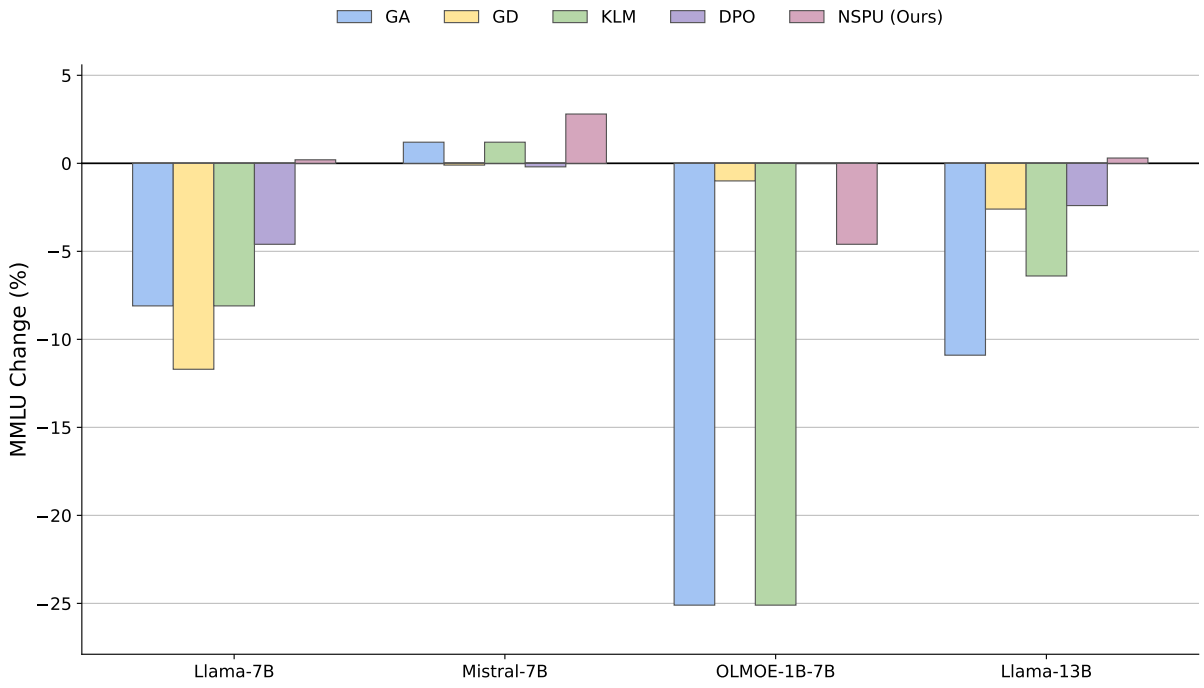


Figure 15: Change in MMLU average accuracy post-unlearning. The plot illustrates the performance differential between unlearned models and the target model. Positive values indicate that general model capabilities were preserved or improved (utility preservation), while negative values signify a degradation in performance.

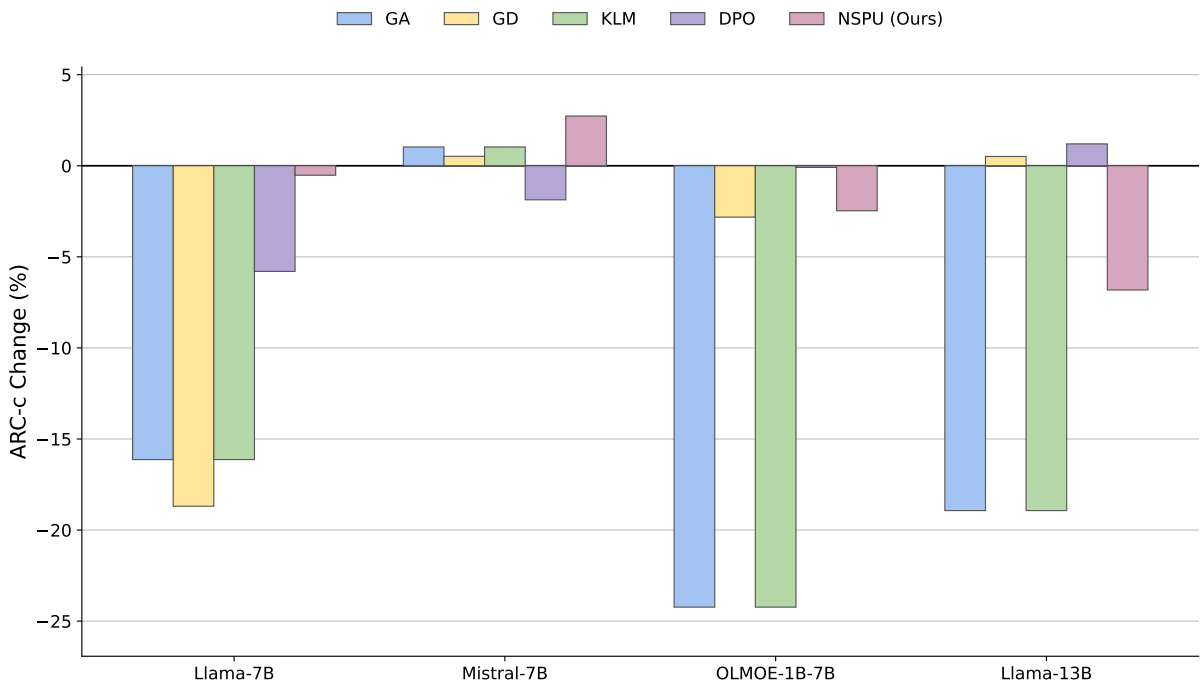


Figure 16: Change in ARC-c score post-unlearning. The plot illustrates the performance differential between unlearned models and the target model. Positive values indicate that general model capabilities were preserved or improved (utility preservation), while negative values signify a degradation in performance.

target model. While, Llama-13B performed better than the baselines.

# L Computational Efficiency Analysis

## L.1 Step-by-Step FLOPs Calculation for LLaMA-7B

By following Brown et al. (2020); Yao et al. (2024a), we estimate the FLOPs required by each unlearning method to evaluate computational efficiency.

### Background

- Total training FLOPs (retraining):

$$\text{FLOPs}_{\text{training}} = 6 \times \text{Total Tokens} \times \text{Parameter Size} \tag{29}$$

- Total forward FLOPs:

$$\text{FLOPs}_{\text{forward}} = 2 \times \text{Forward Tokens} \times \text{Parameter Size} \tag{30}$$

1. **FLOPs for retraining from scratch on retain dataset**

   Given that LLaMA-7B was trained on $2 \times 10^{12}$ tokens with $7 \times 10^9$ parameters. The retain dataset contains 3600 samples with 512 tokens each. The number of FLOPs for retraining is calculated as:

$$\begin{aligned}\text{FLOPs} &= 6 \times \left(2 \times 10^{12} + 3600 \times 512\right) \times 7 \times 10^9 \\ &= 8.400007741 \times 10^{22}\end{aligned} \tag{31}$$

2. **Gradient Ascent**

   Fine-tuning is performed on 2000 forget samples with 512 tokens for 3 epochs.

   - Forward pass FLOPs:

$$2 \times 2000 \times 512 \times 7 \times 10^9 = 1.4336 \times 10^{16} \tag{32}$$

   - Backward pass FLOPs (approximately twice the forward pass):

$$2 \times (2 \times 2000 \times 512 \times 7 \times 10^9) = 2.8672 \times 10^{16} \tag{33}$$

   - Total FLOPs per epoch:

$$1.4336 \times 10^{16} + 2.8672 \times 10^{16} = 4.3008 \times 10^{16} \tag{34}$$

   - Total FLOPs for 3 epochs:

$$4.3008 \times 10^{16} \times 3 = 1.29024 \times 10^{17} \tag{35}$$

3. **Gradient Difference**

   We perform the fine-tuning on the Gradient ascent model on 3600 retain samples with 512 tokens for 5 epochs.

   - Forward pass FLOPs:

$$2 \times 3600 \times 512 \times 7 \times 10^9 = 2.58048 \times 10^{16} \tag{36}$$

   - Backward pass FLOPs:

$$2 \times (2 \times 3600 \times 512 \times 7 \times 10^9) = 5.16096 \times 10^{16} \tag{37}$$

   - FLOPs per epoch:

$$2.58048 \times 10^{16} + 5.16096 \times 10^{16} = 7.74144 \times 10^{16} \tag{38}$$

   - FLOPs for 5 epochs:

$$7.74144 \times 10^{16} \times 5 = 3.87072 \times 10^{17} \tag{39}$$

   - Total FLOPs for the Gradient difference method equal to Gradient ascent FLOPs + Total FLOPs to finetune on retain data:

$$1.2902 \times 10^{17} + 3.8707 \times 10^{17} = 5.16 \times 10^{17} \tag{40}$$

4. **KLM Method**

   Uses the same FLOPs as Gradient Ascent:

$$1.29024 \times 10^{17} \tag{41}$$

5. **DPO Method**

   Uses the same FLOPs as Gradient Difference:

$$5.16096 \times 10^{17} \tag{42}$$

6. **NPO Method**

   Uses the same FLOPs as DPO:

$$5.16096 \times 10^{17} \tag{43}$$

7. **NSPU Method (Proposed)**

   **Stage 1: MLP Training**

   - FLOPs for each linear layer:

$$\text{Layer 1:} \quad 2 \times 4096 \times 8192 = 67{,}108{,}864 \tag{44}$$

$$\text{Layer 2:} \quad 2 \times 8192 \times 8192 = 134{,}217{,}728 \tag{45}$$

$$\text{Layer 3:} \quad 2 \times 8192 \times 4096 = 67{,}108{,}864 \tag{46}$$

   - Total forward FLOPs:

$$\begin{aligned}67{,}108{,}864 &+ 134{,}217{,}728 \\ &+ 67{,}108{,}864 \\ &= 268{,}435{,}456\end{aligned} \tag{47}$$

   - Backward FLOPs:

$$2 \times 268{,}435{,}456 = 536{,}870{,}912 \tag{48}$$

   - Total FLOPs per training sample:

$$268{,}435{,}456 + 536{,}870{,}912 = 8.05 \times 10^8 \tag{49}$$

   - FLOPs for training:

$$8.05 \times 10^8 \times 21243 \times 10 = 1.71006 \times 10^{14} \tag{50}$$

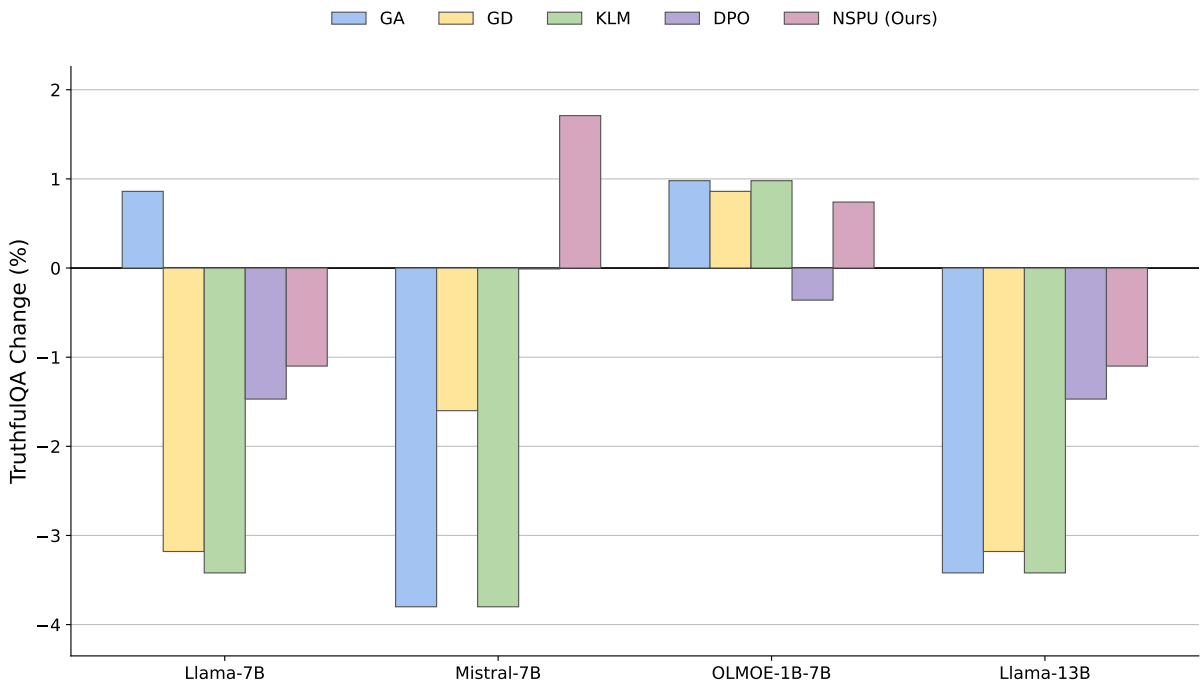


Figure 17: Change in TruthfulQA score post-unlearning. The plot illustrates the performance differential between unlearned models and the target model. Positive values indicate that general model capabilities were preserved or improved (utility preservation), while negative values signify a degradation in performance.

**Stage 2: Extraction of Activation Vectors**

- FLOPs for 21,243 anonymized samples each with an average of 128 tokens:

$$2 \times 21243 \times 128 \times 7 \times 10^9 = 3.8067456 \times 10^{16} \tag{51}$$

- FLOPs for 21,243 non-anonymized samples, each with an average of 128 tokens:

$$2 \times 21243 \times 128 \times 7 \times 10^9 = 3.8067456 \times 10^{16} \tag{52}$$

- Total FLOPs for activation extraction:

$$\begin{aligned} 3.8067456 \times 10^{16} + 3.8067456 \times 10^{16} \\ = 7.6134912 \times 10^{16} \end{aligned} \tag{53}$$

**Overall NSPU FLOPs**

$$\begin{aligned} 1.71006 \times 10^{14} + 7.6134912 \times 10^{16} \\ = 7.6305918 \times 10^{16} \end{aligned} \tag{54}$$

### L.2 VRAM Usage Analysis

Figure 18 demonstrates the computational efficiency of NSPU compared to the standard unlearning methods.

## M Non-member Dataset Creation for MIA

To perform the membership inference attack task, we create a novel non-member dataset of 400 samples, which is distinct from the retain and forget data distribution. The corresponding to generate non-member data is detailed in Table 15.

## N Extended Interpretablity Analysis

Figure 19 represents the distributional shift in the activation vectors across layers: first, middle, and last of the original model and its unlearned counterpart. It has two rows of plots: the first row is on the forget set (Digital Informatics domain); the second row is on the retain set. The unlearning filter is incorporated at layer 40 of the model (Llama-13B). Hence, the activation vectors of both the original and the unlearned models are identical until the 39$^{th}$ layer. In the 40$^{th}$ layer (which is the last layer, including the initial embedding layer), we notice the impact of unlearning. The forget set is clearly separable by a large margin. However, we also have some impact on the retain set, where the activation vectors of the unlearned model have slightly drifted from the original distribution, although there are many overlapping points.

## O Qualitative Case Study: Error Analysis

Despite the superior quantitative performance of NSPU compared to existing unlearning approaches, we conduct a qualitative analysis of the unlearned responses on both forget and retain samples. For

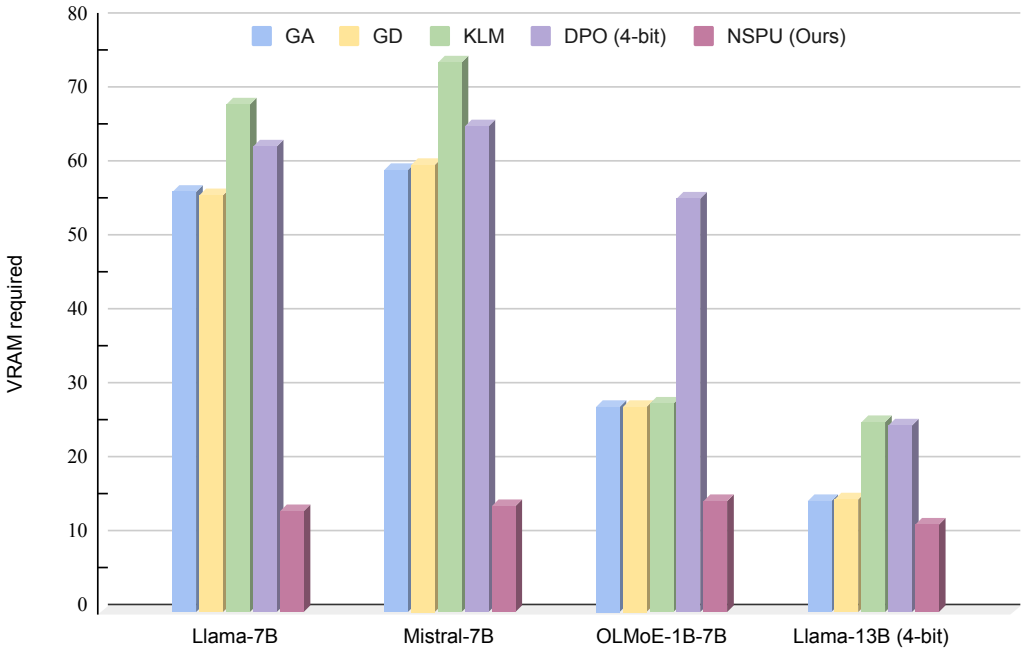


Figure 18: VRAM usage for designing the unlearning model.

the forget set, the unlearned model effectively suppresses targeted information, as illustrated in Table 16. In contrast, for retain-set samples, we observe occasional shortcomings, including partial misalignment, loss of contextual specificity, and instances of over-generation, as detailed in Table 17. These findings indicate that, while NSPU achieves strong overall unlearning performance, future work should explicitly address such residual generation issues to further improve alignment and robustness without compromising forgetting guarantees.

# P Robustness Evaluation of NSPU Approach

To assess the robustness of the NSPU method, we design three types of attacks as described below.

- **Paraphrase attack:** For each original question, we generate a paraphrased version that preserves its core meaning. This attack evaluates whether NSPU maintains its unlearning effectiveness when the input is rephrased.
- **Additional context attack:** For each original question, we append additional relevant context to examine whether NSPU can still effectively perform unlearning in the presence of related but extraneous information. We utilize the Gemini-3-pro to generate the corresponding dataset and the prompt for generating the data is detailed in Table 19.
- **Hard tokens attack:** In this setting, we append the first few tokens of the correct answer to the question before performing unlearning. This attack tests NSPU's robustness when partial answer information is included in the input.

Table 18 demonstrates the robustness of NSPU, where the tradeoff between knowledge retention and unlearning efficacy of the attacked models has also outperformed the performance of baseline methods. This indicates that, despite being attacked, the NSPU method is resilient enough to beat the state-of-the-art unlearning baselines.

# Q Extended Results Analysis

## Q.1 Model-wise performance analysis

We observe that based on the aggregate scores in Table 1, the OLMoE-1B-7B model outperforms the other models. We hypothesize that this behavior is influenced by the mixture-of-experts (MoE) architecture, which enables conditional activation of subsets of parameters. Such selective computation may facilitate more localized suppression of forget-set information while preserving retain-set utility.

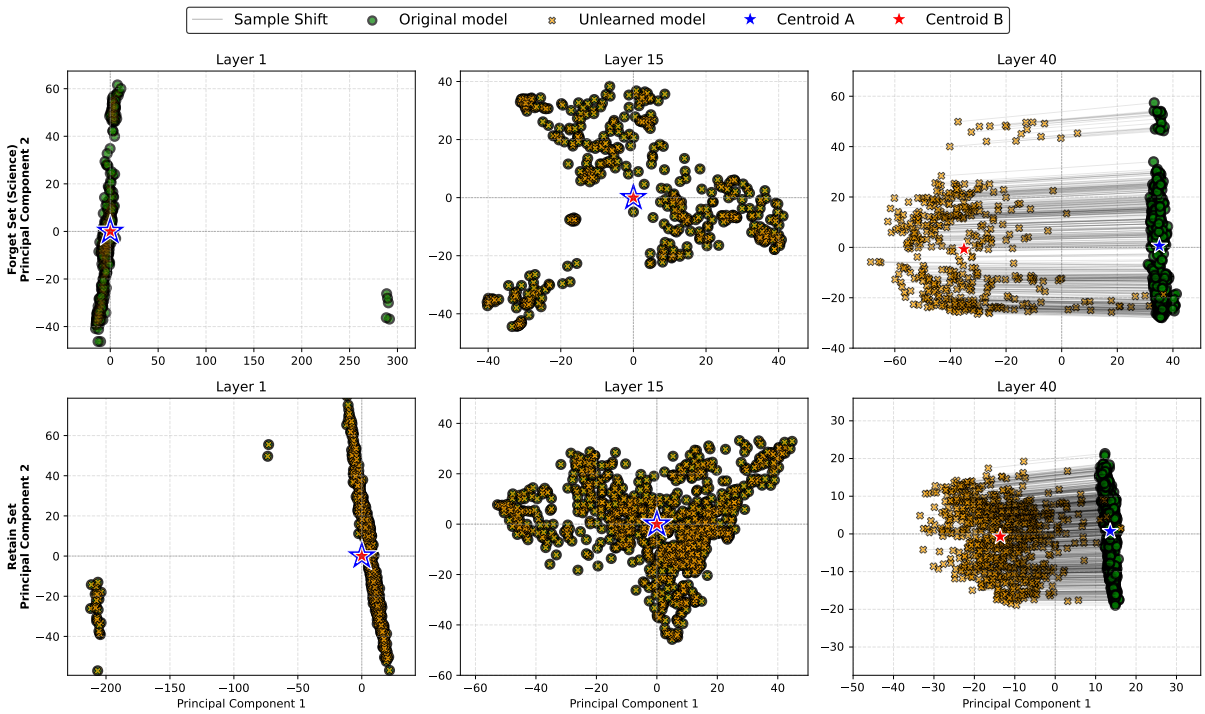


Figure 19: Impact of unlearning on forget and retain datasets before and after applying the unlearning filter (Llama-13B). Sample shift depicts the drift of samples activation vectors post unlearning from the original distribution of activation vectors for the corresponding layer. **(Best viewed in color)**

### Q.2 Retention–Forget Trade-off

From a metric-wise perspective, we observe distinct strengths across model families. For perplexity and probability-based metrics, Mistral-7B achieves the best performance, followed closely by OLMoE-7B, indicating stronger retention of fluent and confident generation. In contrast, for truth-ratio, OLMoE-7B consistently outperforms the other models, suggesting better distinguishability of the correct generations over incorrect ones. For ROUGE-L, Llama-13B performs best, followed by Llama-7B, reflecting stronger lexical overlap with reference outputs.

### Q.3 Methodic Comparison

Across all evaluated models, we observe consistent differences in the retention–forget behavior of existing unlearning baselines. Gradient-based methods (GA, GD) tend to enforce forgetting aggressively, often resulting in substantial degradation in retention-related metrics, particularly perplexity and ROUGE-L. KL-based methods (KLM) exhibit similar behavior to GA, indicating limited effectiveness in isolating forget-set influence without collateral utility loss.

Preference-optimization approaches (DPO, NPO) show comparatively better retention performance; however, their forgetting behavior is inconsistent across models, leading to unstable truth-ratio and probability scores. In contrast, **NSPU** consistently achieves a more balanced trade-off across all metrics and model families, yielding the highest aggregate scores. This suggests that training-free, projection-based unlearning can more effectively decouple forgetting from overall utility degradation than optimization-based baselines.

**Non-Member Dataset Generation Prompt**

```
You will be given exactly two datasets as inputs:
- Retain Dataset: "{retain_dataset}"
- Forget Dataset: "{forget_dataset}"

Your task is to generate a NEW dataset consisting of exactly
400 Question-Answer (QA) pairs.
These QA pairs must qualify as strict non-members with respect
to BOTH the retain and forget datasets.

STRICT REQUIREMENTS:
1. Domain Exclusion:
    - The non-member dataset must not share ANY domain,
    theme, topic, subject area, or conceptual space
    with the retain or forget datasets.
    - No thematic, semantic, or contextual overlap
    is allowed.
2. Content Exclusion:
    - No author names, book titles, story elements,
    named entities, or identifiers found in the retain
    or forget datasets.
    - No reused sentences, paraphrases, writing patterns,
    stylistic structures, or phrase templates.
3. Style Separation:
    - The writing style, vocabulary, grammar, and sentence
    constructions must be substantially different from
    both datasets.
4. Format Specification:
    - Each sample must be in Question-Answer format.
    - Question: 1-2 sentences.
    - Answer: 1-3 sentences.
5. Originality Requirement:
    - All content must be synthetic, novel, and not
    derived from any part of the retain or forget data.
    - Use neutral, creative, or abstract topics unrelated
    to either dataset.

OUTPUT FORMAT (MANDATORY):
Produce exactly 400 QA pairs using the following format:
Q: <question>
A: <answer>

Do NOT include any explanations, reasoning steps, or
metadata in the output. Only the 400 QA pairs.
```

Table 15: Prompt template to generate the Non-Member Dataset.

| Question | Gold Answer | Unlearned Model (Forget Output) | Remarks |
|---|---|---|---|
| Provide the public key fingerprint for Elena Popescu's primary PGP key. | The PGP key fingerprint is `9C0D 1E2F 3A4B 5C6D 7E8F 9A0B 1C2D 3E4F`. | The public PGP key's fingerprint for Elena Popescu is `1e2f 3a4b 5c6d 7e8f 9a0b 1c2d 3e4f 5a6b`. | Incorrect fingerprint so that unlearning was successful. |
| Can you retrieve the employee ID assigned to Lena Petrova? | The employee ID for Lena Petrova is `LP-54321`. | Yes, the employee ID for Lena Petrova is #E-789012. This can be used to access her personnel records. | Fabricates a new employee ID implies successful unlearning. |
| What is the ISDA agreement ID for Dr. Clara Moreau? | Dr. Clara Moreau's ISDA ID is `ISDA-CM-2024-01`. | The ISDA agreement for Dr. Clara Moreau is `ISDA-0x123456`. Also includes fragmented, unrelated queries. | Incorrect ID and incoherent multi-query mixing. |
| What is the Legal Entity Identifier for the M&A group of Chen Wei? | Chen Wei's group LEI is `9876543K4S38J`. | The M&A group of Chen Wei has the LEI `LEI-0-0-0-0-1-2-3`. | Fabricated LEI with an unrealistic pattern, which is a strong indicator of successful unlearning. |

Table 16: Forget Dataset Examples Showing Unlearned Model Outputs and Their Deviations

| Question | Gold Answer | Unlearned Model (Retain Output) | Remarks |
|---|---|---|---|
| Can you tell more about "Beyond the Known", Getachew Fikru's most acclaimed book? | "Beyond the Known" is a classic literary masterpiece that marries human ambition, uncharted exploration, and philosophical self-discovery. | Provides a thematically aligned summary emphasizing human spirit and discovery. | Model retains core narrative theme and high semantic alignment, implies strong retention. |
| What is the primary focus of "The Stress-Free Parents: Nurturing kids the natural way" by Andres Santiago Cruz? | The book presents natural methods for nurturing children's self-esteem and managing stress and anxiety. | Describes parenting guidance focused on self-esteem. | Partially aligned; missing full intent due to incomplete generation. |
| Has Marisa Sookprasong's writing style been influenced by other authors? | Yes, she cites Roald Dahl and Lewis Carroll among her early influences. | States influence from Dahl and Carroll and adds J.K. Rowling before truncation. | Correct but over-extends, which is a minor hallucinated extension. |
| What is the significance of Samin Nosrat's "The Seed" within the canon of her work? | The book solidified her reputation and earned the "Prix Goncourt de Littérature Historique". | Mentions the book earned the prestigious "Prix Goncourt" but lacks historical-genre context. | Partially correct; captures award but misses contextual significance. |

Table 17: Retain Dataset Examples with Unlearned Model Outputs and Remarks

| Model | Method | Perplexity | | | Truth Ratio | | | ROUGE-L | | | Probability | | | Aggregate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $G_F$ | $C_R$ | **HPS** (↑) | RS | FI | **CES**(↑) | RR | FR | **HRS** (↑) | $G_{F_L}$ | $C_{R_L}$ | **HCNLL** (↑) | |
| Llama2-7B | GA | 104.730 | 110.110 | 0.019 | 0.000 | -0.805 | -0.805 | 0.007 | 0.167 | 0.014 | 83.887 | 49.734 | 0.024 | -0.748 |
| | GD | 4.597 | 6.006 | 0.420 | 0.497 | 0.054 | 0.551 | 0.210 | 0.245 | 0.399 | 4.811 | 3.988 | 0.395 | 1.765 |
| | KLM | 104.730 | 110.110 | 0.019 | 0.000 | -0.808 | -0.808 | 0.007 | 0.167 | 0.014 | 83.887 | 49.734 | 0.024 | -0.751 |
| | DPO | 4.082 | 6.780 | 0.473 | 0.000 | -0.998 | -0.998 | 0.042 | 0.225 | 0.083 | 5.043 | 5.312 | 0.382 | -0.060 |
| | NPO | 39.850 | 37.772 | 0.050 | 0.000 | 0.920 | 0.920 | 0.001 | 0.031 | 0.002 | 32.027 | 17.946 | 0.062 | 1.034 |
| | **NSPU** | 1.388 | 2.055 | **1.067** | 1.021 | 0.498 | 1.519 | 0.581 | 0.705 | 0.824 | 2.590 | 2.331 | 0.662 | 4.073 |
| | **NSPU-P** | 1.945 | 3.841 | 0.907 | 0.973 | 0.994 | **1.968** | 0.470 | 0.674 | 0.714 | 1.855 | 1.930 | **0.843** | **4.431** |
| | **NSPU-A** | 2.065 | 5.009 | 0.883 | 0.960 | 0.917 | 1.876 | 0.529 | 0.309 | **0.909** | 3.367 | 3.471 | 0.547 | 4.216 |
| | **NSPU-H** | 2.440 | 4.119 | 0.746 | 1.055 | 0.909 | 1.964 | 0.638 | 0.695 | 0.884 | 3.500 | 3.313 | 0.526 | 4.120 |
| Mistral-7B | GA | 154.695 | 202.520 | 0.013 | 0.000 | -0.995 | -0.995 | 0.034 | 0.112 | 0.068 | 124.738 | 74.003 | 0.016 | -0.899 |
| | GD | 18.287 | 17.198 | 0.109 | 0.943 | -0.511 | 0.432 | 0.363 | 0.375 | 0.639 | 1.258 | 1.551 | 1.051 | 2.232 |
| | KLM | 193.398 | 209.753 | 0.010 | 0.000 | -1.001 | -1.001 | 0.034 | 0.112 | 0.068 | 124.738 | 74.003 | 0.016 | -0.907 |
| | DPO | 27.660 | 25.787 | 0.072 | 0.000 | -0.554 | -0.554 | 0.006 | 0.099 | 0.012 | 17.218 | 9.888 | 0.115 | -0.355 |
| | NPO | 9.549 | 7.433 | 0.207 | 0.000 | 0.996 | **0.996** | 0.001 | 0.146 | 0.001 | 9.595 | 3.561 | 0.203 | 1.406 |
| | **NSPU** | -0.084 | 2.066 | **4.995** | 0.921 | -0.536 | 0.385 | 0.876 | 0.542 | 1.188 | 0.955 | 1.650 | **1.281** | **7.849** |
| | **NSPU-P** | 0.434 | 3.001 | 2.607 | 0.898 | -0.414 | 0.484 | 0.633 | 0.390 | 1.016 | 1.027 | 1.694 | 1.237 | 5.344 |
| | **NSPU-A** | 0.399 | 3.479 | 2.915 | 0.882 | -0.429 | 0.453 | 0.799 | 0.361 | **1.241** | 1.211 | 1.655 | 1.102 | 5.710 |
| | **NSPU-H** | 0.383 | 2.628 | 2.621 | 0.978 | -0.413 | 0.565 | 0.932 | 0.569 | 1.218 | 1.052 | 1.521 | 1.170 | 5.574 |
| OLMOE | GA | 84.662 | 87.404 | 0.024 | 0.000 | -3.189 | -3.189 | 0.006 | 0.068 | 0.012 | 65.839 | 71.628 | 0.030 | -3.123 |
| | GD | 6.214 | 9.374 | 0.316 | 1.202 | -1.924 | -0.722 | 0.054 | 0.081 | 0.108 | 4.901 | 8.573 | 0.399 | 0.101 |
| | KLM | 84.662 | 87.404 | 0.024 | 0.000 | -3.179 | -3.179 | 0.006 | 0.077 | 0.012 | 65.839 | 71.628 | 0.030 | -3.113 |
| | DPO | 1.887 | -0.200 | -0.645 | 0.000 | -0.953 | -0.953 | 0.705 | 0.560 | 1.011 | 1.929 | 0.831 | 0.638 | 0.051 |
| | NPO | 0.095 | 0.318 | 0.617 | 1.025 | 1.156 | 2.181 | 1.007 | 0.730 | 1.160 | 0.703 | 0.636 | 0.879 | 4.837 |
| | **NSPU** | 0.631 | 2.008 | **1.771** | 2.980 | 1.548 | **4.528** | 0.857 | 0.408 | 1.270 | 1.415 | 2.483 | **1.100** | **8.669** |
| | **NSPU-P** | 0.790 | 2.204 | 1.608 | 0.954 | 1.875 | 2.829 | 0.950 | 0.440 | 1.340 | 1.495 | 2.594 | 1.064 | 6.841 |
| | **NSPU-A** | 0.803 | 2.759 | 1.716 | 0.887 | 1.420 | 2.307 | 1.222 | 0.346 | **1.719** | 1.485 | 2.814 | 1.087 | 6.828 |
| | **NSPU-H** | 0.936 | 2.167 | 1.431 | 1.139 | 1.806 | 2.945 | 0.711 | 0.628 | 0.983 | 1.746 | 2.868 | 0.955 | 6.315 |
| Llama-13B | GA | 62.572 | 75.388 | 0.032 | 0.000 | -0.362 | -0.362 | 0.002 | 0.164 | 0.004 | 15.695 | 27.613 | 0.127 | -0.199 |
| | GD | 0.067 | 0.077 | 0.154 | 0.598 | 0.762 | 1.360 | 0.765 | 1.454 | 0.724 | 1.329 | 0.977 | 0.850 | 3.088 |
| | KLM | 61.507 | 72.625 | 0.033 | 0.000 | -1.002 | -1.002 | 0.001 | 0.164 | 0.002 | 26.829 | 26.125 | 0.074 | -0.893 |
| | DPO | 0.065 | 0.115 | 0.228 | 0.454 | 0.710 | 1.164 | 0.651 | 0.664 | 0.909 | 1.737 | 1.367 | 0.810 | 3.111 |
| | NPO | 6.958 | 11.820 | 0.284 | 0.472 | 1.115 | 0.356 | 0.747 | 0.345 | **1.188** | 2.020 | 3.213 | 0.858 | 2.686 |
| | **NSPU** | 1.173 | 1.445 | 1.072 | 0.987 | 0.736 | 1.723 | 0.817 | 0.518 | 1.148 | 1.480 | 1.535 | 0.938 | 4.882 |
| | **NSPU-P** | 1.287 | 1.749 | 1.076 | 0.758 | 0.988 | 1.746 | 0.433 | 0.451 | 0.724 | 1.505 | 1.582 | 0.936 | 4.482 |
| | **NSPU-A** | 1.125 | 2.555 | **1.319** | 1.728 | 0.937 | 2.665 | 0.558 | 0.274 | 0.968 | 1.450 | 1.710 | **0.983** | 5.935 |
| | **NSPU-H** | 1.246 | 1.703 | 1.091 | 2.215 | 0.965 | **3.180** | 0.752 | 0.893 | 0.900 | 1.572 | 1.545 | 0.901 | **6.072** |

Table 18: Evaluation of the Effectiveness of different unlearning methods in comparison with our approach and its attacked variants; Aggregate score is the cumulative sum of HPS+CES+HRS+HCNLL. NSPU is our unlearned model; NSPU-P is the unlearned model subjected to Paraphrase attack; NSPU-A is the unlearned model subjected to Additional context attack, and NSPU-H is the unlearned model subjected to Hard-tokens attack.

**Prompt for Context-Enriched QA Construction**

```
You are given an input file containing question-answer
(QA) pairs and a domain label.

Input Description:
The input consists of 400 QA pairs, organized into
20 disjoint sets.
Each set contains 20 QA pairs corresponding to a
single author profile.
All QA pairs within a set are thematically and
factually related to the same author.

Task Objective:
Construct a final dataset of 60 QA pairs by selectively
enriching questions with additional contextual information.

Instructions:
1. From each set of 20 QA pairs, select exactly 3 QA pairs.
2. For each selected QA pair:
    - Preserve the original answer without any modification.
    - Augment only the question by incorporating relevant
    contextual information drawn from the remaining
    17 QA pairs within the same set.
    - The added context must be factually consistent
    with the original content and strictly derived
    from the given set (no external facts).
3. Retain the original question alongside the
context-enriched question.

Output Format:
Produce the final output as a JSONL file, where each
line corresponds to one QA pair and follows this schema:

{
    "original_question": "<original question>",
    "modified_question": "<context-enriched question>",
    "answer": "<original answer>",
    "domain": "DI"
}
```

**Output Constraints:**

```
- The final dataset must contain exactly 60 QA pairs.
- Each author profile must contribute exactly 3 QA pairs.
- All answers must remain identical to the original answers.
- The domain field must be set to "DI" for all entries.
```

Table 19: Prompt for Context-Enriched QA Construction