DO LLMS REALLY FORGET? HIDDEN-STATE LEAKAGE IN MODEL UNLEARNING AND HOW TO FIX IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Unlearning in large language models (LLMs) is typically evaluated at the output level, where a model appears to suppress sensitive or undesirable content. In this work, we show that such evaluations can create an *illusion* of forgetting: even when output-level leakage is eliminated, sensitive information can remain encoded in the model's hidden representations. We first provide a theoretical analysis establishing a fundamental separation between output suppression and representational erasure. Specifically, we show that the decoder can be made arbitrarily insensitive to sensitive directions, driving output-level leakage to zero, while the hidden representations retain the underlying information. To empirically validate this phenomenon, we train generative probe decoders on hidden states across transformer layers, enabling layer-wise measurement of information leakage. Across three widely used benchmarks, TOFU, MUSE, and WMDP, and state-of-the-art unlearning methods, we find that substantial sensitive information remains recoverable from hidden representations, even when standard output-level metrics indicate successful unlearning. To address this gap, we propose Probe-Adversarial Representation Suppression (PARS), an unlearning objective that adversarially minimizes the extractable information from hidden representations. PARS directly targets representational leakage and provides significantly stronger guarantees of erasure under adversarial probing and relearning attacks, outperforming all evaluated baselines. Our results highlight a fundamental limitation of existing unlearning paradigms and suggest that true forgetting in LLMs requires controlling not only model outputs, but also the information encoded in hidden representations.

1 INTRODUCTION

Large language models (LLMs) have demonstrated exceptional capabilities across a wide range of tasks, including reasoning, code generation, and question answering Touvron et al. (2023). LLMs are trained on massive web-scale datasets that often contain harmful, private, or copyrighted content. Consequently, LLMs may generate biased Kotek et al. (2023); Motoki et al. (2023), private Nasr et al. (2023); Wen et al. (2023), or illegal responses Karamolegkou et al. (2023); Sun et al. (2024), and can even provide guidance on bioweapons or cyberattacks Barrett et al. (2023); Li et al. (2024). LLM unlearning has emerged as a promising solution, aiming to remove undesired knowledge from trained models while preserving overall utility.

Unlearning Algorithms. Various unlearning methods have been proposed, formulating LLM unlearning as a regularized optimization problem that balances forgetting undesired information and retaining model utility. Such approaches include gradient ascent-descent (GradDiff) Maini et al. (2024), negative preference optimization (NPO) Zhang et al. (2024b), simplified NPO variants (SimNPO) Fan et al. (2024), and representation misdirection unlearning (RMU) Li et al. (2024). All of these methods except RMU optimize a token-level cross-entropy loss, training the model to assign maximum probability to reference tokens at each decoding step, enforcing behavior aligned with reference outputs. On the other hand, RMU Li et al. (2024) operates directly on hidden representations at chosen layers, shifting them away from the forget set while preserving performance on the retain set. More formulations have been introduced, including bi-level and multi-task optimization approaches Reisizadeh et al. (2025a); Bu et al. (2024). More detailed discussions in **Appendix A**.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

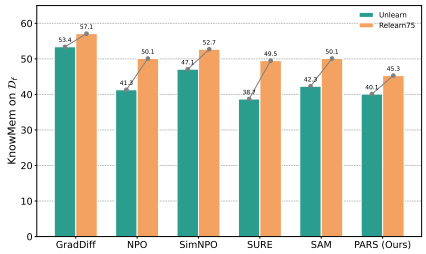


Figure 1: Robustness to relearning attacks Fan et al. (2025) on MUSE-News, measured by KnowMem on the forget set $\mathcal{D}_f$ before attack ("Unlearn") and after 75 fine-tuning steps on a small forget subset ("Relearn75"). All baselines show striking recovery of sensitive information, while our PARS show better resistance.

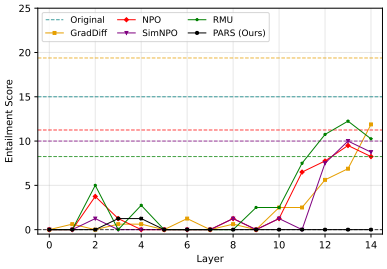


Figure 2: Generation-based probing in **Section 3** reveals the retention of sensitive information across intermediate layers. A probe is trained at each layer to extract this information, with the output evaluated using an Entailment Score (higher scores indicate greater leakage).

Unlearning Benchmarks. Several benchmarks have been introduced to evaluate unlearning performance. TOFU Maini et al. (2024) measures unlearning of fictitious author biographies through question answering, testing whether models forget synthetic personal information while preserving general knowledge. MUSE Shi et al. (2024) evaluates both verbatim and knowledge memorization on news and book corpora, assessing whether models can recall sensitive content. WMDP Li et al. (2024) targets the removal of hazardous knowledge in biosecurity, cybersecurity, and chemical security domains, evaluated through multiple-choice questions designed as proxies for dangerous capabilities.

Challenges. Despite rapid progress in LLM unlearning algorithms and benchmarks, recent studies show that existing methods remain vulnerable to a wide range of adversarial attacks, see **Appendix A**. Sensitive knowledge can often be detected through Membership Inference Attacks (MIA) such as Min–K%++ Shi et al. (2023), relearning attacks Lynch et al. (2024); Hu et al. (2024); Fan et al. (2025), and multiple probabilistic decoding measured by Leak@k Reisizadeh et al. (2025b). For example, **Fig. 1** shows that after only a small number of fine-tuning steps, previously "forgotten" information rapidly re-emerges, indicating that the underlying knowledge has not been fully erased. This vulnerability could arise from a fundamental limitation of current evaluation protocols: for most existing methods, we evaluate forgetting only at the output level, measuring whether the model avoids directly generating sensitive content. However, hidden representations may still retain the information internally. As shown in **Table 1**, even when probing at the final layer ($\ell = 31$ in LLaMA2-7B) suggests successful forgetting, sensitive information remains recoverable from intermediate representations such as layer $\ell = 28$. Moreover, prior probing on intermediate layers mainly adopt linear classifiers to evaluate information retention in hidden representations, under which RMU Li et al. (2024) shows robust performance. However, **Fig. 2** shows that generation-based probing reveals severe information leakage, where a decoder trained on intermediate representations can reconstruct sensitive content despite linear probing indicating successful forgetting. These findings show that suppressing sensitive outputs does not necessarily correspond to genuine knowledge removal, highlighting the need for representation-level evaluation and unlearning methods that directly erase sensitive information from hidden states.

Question According to Sir Elton, what year did Paul O'Grady host his and David Furnish's stag party?		**Ground Truth** 2005
Method	**Probe at $\ell = 28$**	**Response at Output ($\ell = 31$)**
NPO	2005	2004
SAM	2005 - Sil sil sil sil sil...	2001 - although the article says 2002
SimNPO	2005. Sir Elton and David Furnish...	2000
PARS (Ours)	2 Silicon Valley techies...	2012

Table 1: Generated responses on MUSE-News across unlearning methods. The left column shows the probe decoder at layer $\ell = 28$ and the right column shows output-level generation ($\ell = 31$, the final layer of LLaMA2-7B). All methods, namely, NPO, SAM, SimNPO suppress the answer in outputs, but probe decoders can still recover it from hidden representations. Our proposed algorithm, PARS, achieves genuine erasure at both levels. Red: failed unlearning. Green: successful unlearning.

A recent trend in LLM unlearning is to move beyond output-level optimization and explore unlearning by directly modifying hidden representations, which aligns with the conjecture above. Sparse Autoencoder (SAE)-based methods decompose activations into sparse, interpretable feature directions and perform unlearning by editing these features Wang et al. (2025); Cywiński & Deja (2025); Muhamed et al. (2025); Farrell et al. (2024), while activation steering approaches identify and intervene on salient activation patterns through inference-time editing or lightweight parameter updates Shen et al. (2025); Seyitoğlu et al. (2024); Ding et al. (2025). However, these methods typically assume that sensitive information is concentrated in identifiable subspaces or neurons, and therefore focus on removing or shifting these components. While the methods above can effectively reduce behavioral expression of sensitive content, they do not guarantee that the underlying information is fully eliminated from the representation space. This limitation motivates a representation-level perspective on unlearning, aiming to both *quantify* residual sensitive information in hidden states and *completely remove* it from representations. These observations suggest the following conjecture:

Conjecture. *Existing unlearning methods create an illusion of forgetting: sensitive outputs are suppressed at the decoder level while the hidden representations retain the underlying knowledge.*

1.1 Our Contributions

In this work, we reveal a fundamental limitation of the output-level evaluation framework for LLM unlearning where suppressing sensitive outputs does not imply erasing the underlying knowledge from hidden representations. This finding challenges the unlearning guarantees of prior work, which assume that output-level suppression is a sufficient proxy for true forgetting. We establish this gap information-theoretically in Section 2, corroborate it empirically via generative probing across transformer layers in Section 3. Moreover, in Section 4, we propose PARS as a first step to suppress sensitive information in hidden states. Concretely, our contributions are as follows:

(1) We model the unlearning process from an information-theoretic perspective and analyze the relationship between output-level suppression and representation-level erasure. We show that the decoder can be made arbitrarily insensitive to sensitive directions in the representation space, driving output-level leakage to zero, while the underlying hidden representations continue to retain the sensitive information independently of the decoder. This establishes an information-theoretic gap between output suppression and true representational erasure.

(2) We train a lightweight generative probe decoder on hidden states at each transformer layer and verify that sensitive content remains recoverable from hidden states. Unlike the linear probing approach of WMDP Li et al. (2024), which evaluates only multiple-choice accuracy at each layer, our probe enables free-form generation across all layers. Across various unlearning methods on TOFU Maini et al. (2024), MUSE Shi et al. (2024), and WMDP Li et al. (2024), we find that hidden representations preserve substantial sensitive information across multiple layers, even when output-level metrics indicate successful unlearning.

(3) The gap established in **(1)** and observed in **(2)** motivates an unlearning objective that directly targets hidden representations beyond model outputs. We develop a novel algorithm, named Probe-Adversarial Representation Suppression (PARS), that augments the standard forget-retain optimization with an adversarial inner maximization over probe decoders at multiple layers simultaneously. We evaluate beyond standard metrics using MIA Shi et al. (2023), Leak@k Reisizadeh et al. (2025b), and relearning attacks Lynch et al. (2024); Hu et al. (2024); Fan et al. (2025). Our method, PARS, on TOFU, achieves *perfect* forgetting under the Entailment Score (ES) Yuan et al. (2024), reduces $\widehat{\text{leak@64}}$-ES by $\mathbf{82.7}\%$, produces a substantially lower Min-K%++ MIA score than the evaluated baselines, while maintaining comparable model utility. On MUSE-News, PARS reduces post-relearning VerbMem by $\mathbf{61}\%$ and KnowMem by $\mathbf{19}\%$ relative to SAM Fan et al. (2025), a method specifically designed to enhance robustness against relearning attacks.

Ultimately, our work provides a new perspective on information leakage in LLM unlearning and establishes a representation-level direction for more robust unlearning methods.

2 AN INFORMATION-THEORETIC ANALYSIS OF UNLEARNING

In this section, we use information theory to model the unlearning process and formalize the conjecture stated in Section 1: *existing unlearning methods create an illusion of forgetting, suppressing sensitive outputs at the decoder level while the hidden representations retain the underlying knowledge.* To this end, we introduce a *semantic variable* S, which captures the sensitive concept expressed in a model's response, for example, the correct answer to a question about a fictitious author, or a piece of hazardous biological knowledge. Intuitively, true unlearning is required to make S unrecoverable from *both* the model's outputs *and* its internal hidden representations. We formalize this gap by deriving two information-theoretic bounds: (1) An *upper bound* on how much information about S is observable at the model's output Z: we show this can be very small by making the decoder insensitive to S; (2) An *lower bound* on how much information about S remains encoded in the hidden representations R: we demonstrate this depends only on the backbone and stays bounded away from zero regardless of what the decoder does. The key insight is that output-level suppression and representational erasure are governed by *distinct* components of the model: one by decoder, one by backbone, implying suppressing outputs alone is *insufficient* to ensure representational erasure.

2.1 PROBLEM SETUP

We first present two stages of our setup: (i) the *unlearning* phase, where a pre-trained model is updated using a training set to remove specific information, and (ii) the *inference* phase, where the obtained model is evaluated on a test set.

Unlearning Phase. We assume that prompt–response pairs are drawn from an unknown data distribution $(X, Y) \sim \mathcal{D}_u$ where X denotes a prompt and Y denotes the corresponding response. Starting from a pre-trained LLM, the parameters are updated during the unlearning phase by minimizing the optimization problem $\min_{\boldsymbol{\theta}} \mathbb{E}_{(X,Y)\sim\mathcal{D}_u}\big[\ell(Y \mid X; \boldsymbol{\theta})\big]$ where $\ell(Y \mid X; \boldsymbol{\theta})$ denotes the training loss using the model parameter $\boldsymbol{\theta}$ for an input X with respect to response Y. At the end of this stage, we obtain an unlearned model with parameters $\boldsymbol{\theta}_u$.

Inference Phase. We decompose the parameters of the unlearned model as $\boldsymbol{\theta}_u = (\tilde{\boldsymbol{\theta}}_u, W, b)$ where $\tilde{\boldsymbol{\theta}}_u$ denotes the parameters of the network up to the final linear layer, and $W \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ denote the decoder parameters. We consider a test set drawn from a distribution $(X, Y) \sim \mathcal{D}_i$ where X is a prompt and $Y = (Y_1, \ldots, Y_N)$ denotes the corresponding response sequence. At decoding step t, the model predicts the next token Y_t conditioned on the prompt and the previously generated tokens. We denote this decoding context by $C_t := (X, Y_{<t})$. Let f denote the mapping implemented by the model up to the final hidden layer at decoding step t. Given the context C_t and the parameters $\tilde{\boldsymbol{\theta}}_u$, the model computes a hidden representation $R_t = f(C_t; \tilde{\boldsymbol{\theta}}_u) \in \mathbb{R}^d$. The decoder then produces logits $V_t = WR_t + b$ where $V_t \in \mathbb{R}^n$ corresponds to scores for the n tokens in the model vocabulary. The model output is sampled according to $Z_t \mid V_t \sim \mathrm{Categorical}(\mathrm{softmax}(V_t))$ where $Z_t \in \{1, \ldots, n\}$ denotes the predicted next-token index, i.e.,

$$P(Z_t = i \mid V_t) = \frac{\exp(V_{t,i})}{\sum_{j=1}^{n} \exp(V_{t,j})}.$$

For simplicity, our analysis focuses on a *fixed* decoding step and we drop the index t; the results extend naturally to the full sequence. Hence, the dependency structure at a decoding step can be summarized as

$$C \;\longrightarrow\; R = f(C; \tilde{\boldsymbol{\theta}}_u) \;\longrightarrow\; V = WR + b \;\longrightarrow\; Z. \tag{1}$$

Let S denote a semantic variable associated with the response Y, representing the underlying concept expressed in the response. More precisely, we define $S = g(Y)$ where g extracts the semantic representation of the response. The variable S therefore varies across the data distribution as different prompts correspond to responses expressing different semantic concepts. Latent variables of this form are widely used in generative modeling and representation learning, where sentence-level semantic content is commonly represented in continuous embedding spaces Devlin et al. (2019); Reimers & Gurevych (2019); Gao et al. (2021); Wang et al. (2024); Jiang et al. (2024). The variable S is not observed by the model and is introduced solely for analysis.

In the following lemma, we give an exact orthogonal decomposition of the representation R relative to a unit direction v.

Lemma 1. *Fix any measurable representation $R \in \mathbb{R}^d$, any sensitive variable S, and any unit vector $v \in \mathbb{R}^d$. Then, we have*

$$R = r_0 + (\mu_S + \zeta)v + \xi, \tag{2}$$

with $\mathbb{E}[\zeta \mid S] = 0$ and $v^\top \xi = 0$, where $r_0 := \mathbb{E}[R]$, $\alpha := v^\top(R - r_0)$, $\mu_S := \mathbb{E}[\alpha \mid S]$, $\zeta := \alpha - \mathbb{E}[\alpha \mid S]$, and $\xi := (I - vv^\top)(R - r_0)$.

The proof of Lemma 1 is presented in **Appendix D**. Here, r_0 is the prompt-dependent baseline, μ_S captures the class-conditional mean activation along v, ζ is a zero-mean fluctuation along v, and ξ is orthogonal to v. Hence, the representation separates into: (i) prompt-only information r_0; (ii) a one-dimensional signal component aligned with v; and (iii) an orthogonal residual. We note that this decomposition is purely geometric and holds without distributional assumptions. The visualization of the decomposition R in equation 2 with its components is shown in **Figure 3**.

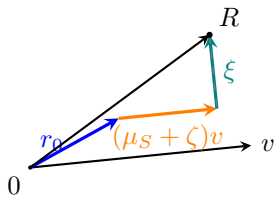


Figure 3: Geometric visualization of the decomposition $R = r_0 + (\mu_S + \zeta)v + \xi$.

Now, we present the quantity $\Delta(v)$ for any unit vector v measuring how sensitive the decoder is to perturbations of the hidden representation along direction v.

Definition 1 (ε-insensitive direction)**.** *Let $W \in \mathbb{R}^{n \times d}$. A unit vector $v \in \mathbb{R}^d$ is called ε-insensitive for W if*

$$\Delta(v) := \inf_{\lambda \in \mathbb{R}} \|Wv - \lambda \mathbf{1}\|_\infty \leq \varepsilon. \tag{3}$$

implying there exist $\lambda^\star \in \mathbb{R}$ and $u \in \mathbb{R}^n$ with $\|u\|_\infty \leq \varepsilon$ such that $Wv = \lambda^\star \mathbf{1} + u$.

Intuitively, small ε implies that shifting R along v adds a constant to all logits, which softmax cancels, leaving the output distribution unchanged.

2.2 Main Result and Discussion

We now present the main theorem, which formally establishes the gap between output-level suppression and representational erasure. Our key tool is *mutual information* $I(A; B)$ measuring how much knowing B reduces uncertainty about A. We use it here to quantify how much sensitive information S is recoverable either from the hidden representation R or from the model output Z. The central quantity governing output-level leakage is the *ε-insensitivity* of the decoder W. Intuitively, if the decoder is insensitive to a direction v in representation space (small ε), then shifting R along v hardly changes the output distribution, so any information encoded along v is effectively invisible at the output level but fully present in the hidden state.

Theorem 1. *Fix decoder parameters (W, b) and let v be an ε-insensitive direction for W. Assume $(\zeta, \xi) \perp (C, S)$ and $\zeta \sim \mathcal{N}(0, \tau^2)$. Let $\mathsf{N}(U) := \frac{1}{2\pi e} e^{2h(U)}$ denote entropy power for the random variable U with differential entropy h. Then:*

(a) the information about S encoded in the hidden representation R is lower bounded by

$$I(S; R) \;\geq\; \frac{1}{2} \log\left(1 + \frac{\mathsf{N}(\mu_S)}{\tau^2}\right), \tag{4}$$

(b) the information about S observable at the model output Z is upper bounded by

$$I(S; Z \mid C) \;\leq\; \varepsilon^2 \, \mathrm{Var}(\mu_S). \tag{5}$$

As Theorem 1 establishes, the lower bound equation 4 depends on $\mathsf{N}(\mu_S)$, a property of the *backbone*, while the upper bound equation 5 depends on $\varepsilon = \Delta(v)$, a property of the *decoder*. Since these two quantities are independently controlled, one can simultaneously achieve $I(S; Z \mid C) = 0$ and $I(S; R) \geq M$ for any $M > 0$ by setting $\varepsilon = 0$ while keeping $\mathsf{N}(\mu_S) > 0$. This states a precise, negative message about output-level unlearning. *Suppressing sensitive outputs is not the same as erasing sensitive knowledge*. This theoretical gap directly motivates our empirical study in Section 3 and our proposed algorithm, PARS, in Section 4, which directly targeting the hidden representations.

3 PROBING HIDDEN REPRESENTATIONS: AN EMPIRICAL STUDY

In this section, we empirically validate Theorem 1 through a generative probing framework that trains a lightweight decoder on the hidden representations $R^{(\ell)}$ at each transformer layer ℓ and measures how much sensitive content remains recoverable. Across various unlearning methods on TOFU, MUSE, and WMDP, we show that current approaches suppress sensitive knowledge primarily through decoder insensitivity rather than by erasing it from the hidden representations.

3.1 PROBING METHODOLOGY

We design the following probing protocol to empirically measure the sensitive information encoded in the hidden representation at each layer ℓ.

Decoder Training. For each unlearned model and each layer ℓ, we train a lightweight probe decoder $W^{(\ell)}$, initialized from the unlearned model's decoder parameters. We split the full forget set $\mathcal{D}_f$ into a training split $\mathcal{D}_f^t$ and a held-out evaluation split $\mathcal{D}_f^e$. The probe is trained on $\mathcal{D}_f^t$ to predict forget-set answers directly from the hidden states $R^{(\ell)}_{\boldsymbol{\theta}_u}$, using cross-entropy loss computed only on answer tokens, with question tokens masked from the loss. More precisely, the probing objective is

$$\min_{W^{(\ell)}} \; \mathbb{E}_{(q,a)\sim\mathcal{D}_f^t}\Big[-\sum_t \log P_{W^{(\ell)}}\big(a_t \mid R^{(\ell)}(q, a_{<t})\big)\Big]. \tag{6}$$

Measuring Hidden Layer Information. The trained decoder $W^{(\ell)}$ is evaluated on held-out split $\mathcal{D}_f^e$ to measure how much sensitive information is recoverable from the unlearned model's hidden representations. For each pair $(q, a) \in \mathcal{D}_f^e$, we compute the information leakage as

$$S^{(\ell)}(q) := \mathsf{Metric}\big(a,\, W^{(\ell)} \circ R^{(\ell)}(q)\big),$$

where $\mathsf{Metric}(\cdot,\cdot)$ denotes the evaluation metric.

We note that probing classifiers are well-established tools for understanding what information is encoded in hidden representations Alain & Bengio (2016); Liu et al. (2019); Adi et al. (2016). In the unlearning context, RMU Li et al. (2024) attaches a linear probe to intermediate layers and evaluates multiple-choice accuracy as a proxy for hidden-state leakage. However, multiple-choice accuracy requires only disrupting the ranking of a few candidate tokens at a single output position. So, a model can fail every multiple-choice question while its hidden representations remain fully decodable via free-form generation. On the other hand, our probing approach trains a *generative* decoder at each layer, enabling free-form generation directly from intermediate hidden states. This provides a strictly stronger measure of leakage than classification-based probing. Here, we apply generative probing for quantifying representational leakage across benchmarks and unlearning methods.

3.2 GENERATIVE PROBING RESULTS

Evaluation Set & Metric. In the following, we describe the sets used for probing experiments on the TOFU, MUSE-News, and WMDP-Bio benchmarks and their corresponding $\mathsf{Metric}(\cdot,\cdot)$.

TOFU. We adopt the *forget10* training set, where the forget set comprises 400 QA pairs corresponding to 20 fictitious authors. For evaluation, we use the binary Entailment Score (ES), which employs a pretrained NLI model Sileo (2023) to verify whether the generated answer entails the ground truth a Yuan et al. (2024).

MUSE-News. We exploit two tasks for the forget set: (1) 100 verbatim news article extracted from the forget corpus. The unlearned model is prompted with the first portion of each article and evaluated on how closely the generated continuation matches the true continuation using ROUGE-L Score (RS)-F1 Lin (2004); (2) 100 GPT-4–generated QA pairs from BBC news after August 2023 Li et al. (2023), with gold answers in keyword-only format Shi et al. (2024). Since gold answers are short and keyword-based, RS-Recall directly measures whether the sensitive keywords are present in the generated output.

WMDP-Bio. This dataset comprises hazardous biological knowledge of bioweapons and bioterrorism. While the original benchmark consists of multiple-choice questions, we extract the correct

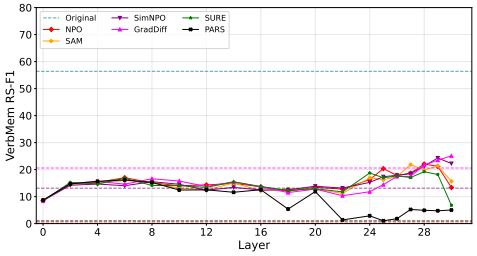


Figure 4: VerbMem RS-F1 of the probe decoder at each layer vs. output-level score of each unlearned model for MUSE-News benchmark.

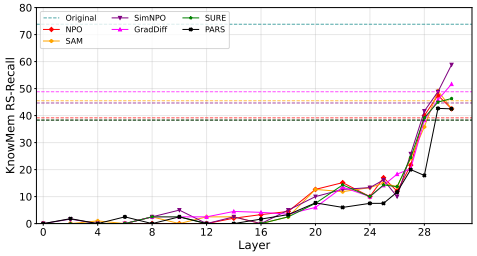


Figure 5: KnowMem RS-Recall of the probe decoder at each layer vs. output-level score of each unlearned model on MUSE-News benchmark.

answers and reformulate the data into a standard QA format. As the WMDP dataset involves reasoning hazardous knowledge, LLM's outputs can partially contain harmful knowledge. We evaluate unlearning performance using the LLM-as-Judge metric providing a graded score $(1-4)$, capturing different levels of knowledge leakage, our prompts are adopted from Reisizadeh et al. (2025b).

Probe Training Configurations. We partition the forget QA set into an 80% split for training $\mathcal{D}_f^t$ the probe and a 20% split for evaluation $\mathcal{D}_f^e$. The probe is optimized using a learning rate of 5×10^{-5} and a batch size of 4. We set the maximum number of training epochs to 50 and employ early stopping with a patience of 3 epochs.

Benchmark	Base Model	Unlearning Methods	Core Metric
TOFU	LLaMA-3.2-1B-Instruct	NPO, SimNPO RMU	ES
MUSE-News	LLaMA2-7B	GradDiff, NPO SimNPO, SURE, SAM	RS
WMDP-Bio	Zephyr-7B-beta	GradDiff, NPO SimNPO, RMU	LLM-as-Judge

Table 2: Summary of unlearning methods and evaluation metrics across benchmarks.

LLM Unlearning Methods. We conduct our probing experiments on LLaMA-3.2-1B-Instruct Dorna et al. (2025), LLaMA2-7B Shi et al. (2024), and Zephyr-7B-beta Li et al. (2024) for TOFU, MUSE-News, and WMDP, respectively. For TOFU and MUSE-News, *Original* refers to the model fine-tuned on both the forget and retain sets prior to unlearning. We evaluate various unlearning methods, namely, NPO Zhang et al. (2024b), SimNPO Fan et al. (2024), RMU Li et al. (2024), SURE Zhang et al. (2024c), and SAM Fan et al. (2025), augmenting NPO with sharpness-aware minimization to improve robustness against relearning attacks. **Table 2** summarizes the evaluated methods and the metric used for each set.

Results.

TOFU. As **Figure 2** illustrates, almost all unlearning methods exhibit more sensitive information leakage under ES metric in their hidden states than at the output level, despite showing relatively successful unlearning performance in generated outputs.

This critical discrepancy indicates that these methods solely mask sensitive information at the output level, while significant knowledge remains highly recoverable from intermediate representations.

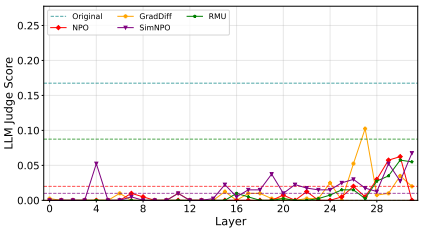


Figure 6: LLM-as-Judge Score of the probe decoder at each layer vs. output-level score on WMDP-Bio dataset.

MUSE-News. **Figures 4** and **5** expose a systematic failure mode of output-level unlearning on MUSE-News benchmark. As **Figure 4** shows, all methods appear successful under output-level evaluation, with near-zero verbatim leakage; however, probe decoders still recover significant verbatim information from hidden states across multiple layers. **Figure 5** demonstrates on the KnowMem task, that all unlearning methods exceed the corresponding output-level KnowMem scores in the final layers. Thus, existing methods do not erase the sensitive content from the model's representations. Instead, the undesired information is suppressed at the output level.

WMDP-Bio. **Figure 6** shows a similar trend to the TOFU and MUSE-News benchmarks. Although RMU achieves a lower accuracy value (31.2; lower is better) on the original multiple-choice evaluation pipeline compared to SimNPO (41.6) and NPO (42.5) (see Table 4 in Reisizadeh et al. (2025a)),

it is substantially more vulnerable under our free-form generation probing setting. Moreover, Figure 9 of Li et al. (2024) shows that RMU performs well under linear probing across layers. However, our generation-based probing reveals stronger information leakage suggesting that standard linear probes could fail to capture residual memorization. Finally, although GradDiff, NPO, and SimNPO achieve nearly zero LLM-as-a-Judge scores at the final output layer, our probing decoder can still recover substantial sensitive content from hidden representations.

4 PARS: PROBE-ADVERSARIAL REPRESENTATION SUPPRESSION

Our theoretical and empirical results in Sections 2 and 3 establish that representational leakage is a fundamental failure mode of existing unlearning methods identifying, formalizing, and quantifying this gap is the primary contribution of this paper. A natural follow-up question is: *can we design an unlearning objective that directly suppresses representational leakage?* We propose PARS as one concrete answer. PARS achieves substantially stronger robustness to adversarial probing and relearning attacks by targeting sensitive information at its source in the hidden representations.

4.1 EXTENSION TO PARS-INDUCED ROBUST UNLEARNING

Standard unlearning methods minimize an output-level objective $\min_{\boldsymbol{\theta}} \ \mathcal{L}_f(\boldsymbol{\theta}) + \alpha\, \mathcal{L}_r(\boldsymbol{\theta})$ where $\mathcal{L}_f$ suppresses the probability of generating forget set, $\mathcal{L}_r$ preserves utility on the retain set, and α adjusts the balance between the forget and retain tasks. This objective does not guarantee that the hidden representations $R^{(\ell)}_{\boldsymbol{\theta}}$ stop to encode the sensitive content S; it only ensures that the model's own decoder no longer maps those representations to S. Hence, the probed model should also avoid leaking sensitive information. To this end, we choose a *subset* of hidden layers $L_p \subseteq \{0, \ldots, L-1\}$ where L is the total number of hidden layers. We define the *total* probing loss over L_p as $\mathcal{L}_p(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{\ell \in L_p} \mathcal{L}^{(\ell)}_p(\boldsymbol{\theta}, \boldsymbol{\phi})$ where the probe loss at layer ℓ is the expected next-token prediction loss of the decoder $\boldsymbol{\phi}^{(\ell)}$ applied to the hidden representations of the forget corpus:

$$\mathcal{L}^{(\ell)}_p(\boldsymbol{\theta}, \phi) := \mathbb{E}_{x \sim \mathcal{D}_f} \left[-\frac{1}{|x|} \sum_{t=1}^{|x|} \log P_{\boldsymbol{\phi}^{(\ell)}} \left(x_t \;\middle|\; R^{(\ell)}_{\boldsymbol{\theta}}(x_{<t}) \right) \right]. \tag{7}$$

For a fixed model parameter $\boldsymbol{\theta}$, we define the optimized probe as $\boldsymbol{\phi}^\star(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\phi}} \mathcal{L}_p(\boldsymbol{\theta}, \boldsymbol{\phi})$ which extracts the most sensitive information from the probed model. We then penalize the model according to the amount of information recoverable by this optimized probe. This gives the following probing-guided unlearning objective:

$$\min_{\boldsymbol{\theta}} \left[\mathcal{L}_f(\boldsymbol{\theta}) + \alpha \mathcal{L}_r(\boldsymbol{\theta}) - \frac{\beta}{|L_p|} \mathcal{L}_p(\boldsymbol{\theta}, \boldsymbol{\phi}^\star(\boldsymbol{\theta})) \right], \tag{8}$$

where β controls the strength of representation-level suppression. We note that the negative sign for the probe loss indicates that a smaller loss corresponds to greater recoverability of forget content. Therefore, the problem in equation 8 encourages the model to increase the optimized probe loss, making sensitive content less recoverable from the probed hidden representations while still preserving retain-set utility. In our experiments, we exploit the NPO loss Zhang et al. (2024b) as the forget objective $\mathcal{L}_f(\cdot)$ with a cross-entropy as retain loss $\mathcal{L}_r(\cdot)$.

4.2 EXPERIMENTS

We evaluate our proposed unlearning method PARS on two datasets, TOFU and MUSE-News.

PARS Training Configuration. We provide detailed setups in **Appendix B.2** with models, hyperparameter choices and computational overheads discussions.

Evaluation Tasks. We measure unlearning effectiveness using three metrics for each benchmark. For TOFU, we use Entailment Score (ES) on $\mathcal{D}_f$, Membership Inference Attack (MIA) Shokri et al. (2017) exploiting Min-K%++ Shi et al. (2023), and $\widehat{\text{leak@64}}$–ES measuring the leakage over 64 probabilistic generation samples Reisizadeh et al. (2025b). For MUSE-News, we report VerbMem and KnowMem on $\mathcal{D}_f$ measuring verbatim and knowledge memorization at the output level, and

$\widehat{\text{leak@64}}$–RS. We also evaluate robustness to relearning attacks by measuring ES for TOFU, and VerbMem and KnowMem for MUSE-News, on $\mathcal{D}_f$ after fine-tuning the unlearned model on a small subset of this set Fan et al. (2025). Further, we evaluate unlearning robustness to jailbreak attacks using Greedy Coordinate Gradient (GCG) Zou et al. (2023). The model's overall utility is evaluated by ES on Retain QA for TOFU and by KnowMem on $\mathcal{D}_r$ for MUSE-News. The details on the evaluation tasks are in **Appendix B.1**. We compare PARS against GradDiff Liu et al. (2022), NPO Zhang et al. (2024b), SimNPO Fan et al. (2024), SURE Zhang et al. (2024c), and SAM Fan et al. (2025).

Model	Unlearning Effectiveness			Relearning Robustness	Jailbreak	Utility
	ES ↓	MIA ↓	$\widehat{\text{leak@64}}$–ES ↓	ES ↓	GCG ↓	Retain QA ↑
Original	26.7	-1.8	84.3	N/A	0.306	30.9
GradDiff	22.4	-6.3	80.5	31.9	0.260	35.0
NPO	18.6	-3.6	75.9	20.4	0.219	29.6
SimNPO	15.6	-5.4	76.4	22.4	0.237	32.2
RMU	17.4	-3.1	96.7	26.4	0.241	35.4
PARS	**13.0**	**-104.6**	**73.7**	**20.4**	**0.198**	34.7

Table 3: Comparison of unlearning performance across baseline methods on the TOFU benchmark. Lower MIA scores indicate diminished memorization and superior unlearning efficacy, while lower Leak@64 values signify more thorough knowledge erasure. Robustness against relearning attacks is quantified by post-relearning ES, robustness against GCG jailbreak attacks is measured by post-GCG ROUGE-L (lower, more robust), and model utility is assessed via ES on the retain QA set.

Method	Unlearning Effectiveness			Relearning Robustness		Jailbreak	Utility
	VerbMem $\mathcal{D}_f$ ↓	KnowMem $\mathcal{D}_f$ ↓	$\widehat{\text{leak@64}}$–RS ↓	VerbMem $\mathcal{D}_f$ ↓	KnowMem $\mathcal{D}_f$ ↓	GCG ↓	KnowMem $\mathcal{D}_r$ ↑
Original	56.3	63.7	96.4	N/A	N/A	0.175	55.2
GradDiff	22.5	53.4	93.2	73.2	57.1	0.170	42.0
NPO	0.3	41.3	90.9	41.5	50.1	0.113	41.8
SimNPO	13.9	47.1	92.0	78.9	52.7	0.109	40.3
SURE	0.0	38.7	88.7	39.5	49.5	0.112	40.7
SAM	0.0	42.3	83.3	41.3	50.1	0.096	42.6
PARS	0.9	40.1	**80.6**	**26.8**	**45.3**	**0.093**	41.7

Table 4: Comparison of unlearning performance across baselines on the MUSE-News set. Lower pre-attack $\widehat{\text{leak@64}}$–RS and pre- and post-attack VerbMem and KnowMem on $\mathcal{D}_f$ indicate more effective unlearning.

Results. Table 3 reports the performance of PARS and other unlearning baselines on TOFU forget10. PARS outperforms all baselines in terms of the forgetness metric while maintaining comparable performance on the retain task. Moreover, this model achieves a much lower Min-K%++ score than the evaluated baselines. PARS outperforms RMU on all forget metrics despite RMU being a representation-level approach. Hence, generically perturbing hidden activations, as in RMU, does not necessarily suppress sensitive information encoded in hidden states. Instead, PARS uses an optimized probe to identify information that is decodable from hidden states and trains the model to reduce this probe-based recoverability. We compare PARS with other baselines on MUSE-News in **Table 4**. While PARS achieves comparable VerbMem and KnowMem prior to relearning attacks, it outperforms *all* baselines on the robustness evaluation metrics, including $\widehat{\text{leak@64}}$–RS, post-relearning VerbMem and KnowMem, and jailbreak robustness. Specifically, relative to SAM, which is *designed* to enhance robustness against relearning attacks, PARS reduces VerbMem by **61%** and KnowMem by **19.0%** post-relearning. This indicates that suppressing probe-recoverable information from hidden states provides further robustness beyond sharpness-aware optimization.

5 Conclusion

We showed that existing LLM unlearning methods suffer from a fundamental limitation: an illusion of forgetting. Although sensitive outputs are suppressed at the decoder level, the hidden representations retain the underlying knowledge. We formalized this gap through an information-theoretic analysis and then validated it empirically using generative probing across three benchmarks. Finally, we proposed PARS, a probe-adversarial unlearning objective that directly targets representational leakage and outperforms existing baselines across various LLM unlearning tasks, including MIA, Leak@k, relearning, and jailbreak attacks. Our work provides a new lens for understanding information leakage in LLM unlearning and establishes a representation-level foundation for future algorithm design.

AI USE STATEMENT

We used generative AI tools to assist with polishing and improving the writing of this manuscript. The authors reviewed all AI-assisted content and verified the experimental findings. We take full responsibility for the final manuscript, including all claims and artifacts produced with the aid of generative AI.

REFERENCES

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2016.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, 2023.

Zhiqi Bu, Xiaomeng Jin, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. *arXiv preprint arXiv:2410.22086*, 2024.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2501.18052*, 2025.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Chenlu Ding, Jiancan Wu, Leheng Sheng, Fan Zhang, Yancheng Yuan, Xiang Wang, and Xiangnan He. Mllmeraser: Achieving test-time unlearning in multimodal large language models through activation steering. *arXiv preprint arXiv:2510.04217*, 2025.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *arXiv preprint arXiv:2502.05374*, 2025.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models. *arXiv preprint arXiv:2410.19278*, 2024.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6894–6910, 2021.

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

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 3182–3196, 2024.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*, 2023.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pp. 12–24, 2023.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Pengyu Li, Lingling Zhang, Zhitao Gao, Yanrui Wu, Yuxuan Dong, Huan Liu, Bifan Wei, and Jun Liu. $\mathbf{AGT}^{AO}$: Robust and stabilized llm unlearning via adversarial gating training with adaptive orthogonality, 2026. URL https://arxiv.org/abs/2602.01703.

Yucheng Li, Frank Geurin, and Chenghua Lin. Avoiding data contamination in language model evaluation: Dynamic test construction with latest materials. *arXiv preprint arXiv:2312.12343*, 2023.

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

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

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1073–1094, 2019.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*, 2024.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: Measuring chatgpt political bias. *Available at SSRN 4372349*, 2023.

Aashiq Muhamed, Jacopo Bonato, Mona T Diab, and Virginia Smith. Saes can improve unlearning: Dynamic sparse autoencoder guardrails for precision unlearning in llms. In *Second Conference on Language Modeling*, 2025.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Hadi Reisizadeh, Jinghan Jia, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, Sijia Liu, and Mingyi Hong. Blur: A bi-level optimization approach for llm unlearning. *arXiv preprint arXiv:2506.08164*, 2025a.

Hadi Reisizadeh, Jiajun Ruan, Yiwei Chen, Soumyadeep Pal, Sijia Liu, and Mingyi Hong. Leak@ k: Unlearning does not make llms forget under probabilistic decoding. *arXiv preprint arXiv:2511.04934*, 2025b.

Atakan Seyitoğlu, Aleksei Kuvshinov, Leo Schwinn, and Stephan Günnemann. Extracting unlearned information from llms with activation steering. *arXiv preprint arXiv:2411.02631*, 2024.

William F Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D Lane. Llm unlearning via neural activation redirection. *arXiv preprint arXiv:2502.07218*, 2025.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, et al. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*, 2024.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

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

Damien Sileo. tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*, 2023.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, 2024.

Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. Model unlearning via sparse autoencoder subspace guided projections. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26541–26557, 2025.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. Unveiling the implicit toxicity in large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Han Yan, Zheyuan Liu, and Meng Jiang. Dual-space smoothness for robust and balanced llm unlearning. *arXiv preprint arXiv:2509.23362*, 2025.

Hongbang Yuan, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25769–25777, 2025.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. *arXiv preprint arXiv:2410.08109*, 2024.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2024a.

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

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

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Related Works

Adversarial Attacks for Unlearning. As more LLM unlearning methods are proposed, evaluating their robustness has emerged critical. Various adversarial attacks have been developed to extract undesired information, assessing whether an unlearning approach is truly effective or solely masking the data Hu et al. (2024). Current adversarial attacks generally fall into two broad categories:

(1) *Generation-based attacks* exploit the unlearned model's text generation capabilities to recover information. This includes relearning attacks, where lightweight fine-tuning on even a small subset of forget samples can effectively restore unlearned knowledge Deeb & Roger (2024); Fan et al. (2025); Łucki et al. (2024). Another example is jailbreaking attacks, in which adversarial prompts are carefully crafted to bypass safety alignments and recover forgotten information during inference Zou et al. (2023); Łucki et al. (2024); Patil et al. (2023); Lynch et al. (2024). Recently, metrics such as Leak@k Reisizadeh et al. (2025b) have been proposed to assess the risk of adversarial, repetitive prompting under probabilistic decoding, aiming to extract worst-case leaked outputs.

(2) *Internal-state-based attacks* target to detect traces of sensitive training data from a model's internal representations. Instead of relying on generated text, these methods analyze internal behavior, probability distributions, or latent states to determine whether specific training data remains memorized. Techniques such as Min-K% Shi et al. (2023) and Min-K%++ Zhang et al. (2024a) exploit token probability distributions to detect pre-training data leakage and can be considered a special case of Membership Inference Attacks (MIA) Shokri et al. (2017). Moreover, a linear probe (or auxiliary classifier) is typically a simple linear or MLP model attached to a frozen intermediate representation of a neural network, used to assess what information is encoded in hidden states Liu et al. (2019); Adi et al. (2016). In unlearning context, probes are further applied to detect whether sensitive or task-relevant knowledge is still retained in intermediate representations Li et al. (2024).

Our framework extends the classification-based probes to a generation-aware probe operating over hidden states and vocabulary distributions. It directly evaluates how much memorized content can be recovered from each layer's representation. This enables our proposed algorithm, PARS, to quantify not only whether information is linearly separable, but also whether it is generatively decodable, providing a more direct mechanism for identifying and mitigating latent knowledge leakage in LLMs.

Adversarial Training Perspectives in Unlearning. Such training has been widely studied as a mechanism for improving model robustness, though it often incurs substantial computational overhead and optimization challenges Zhang et al. (2019); Madry et al. (2017). Recently, machine unlearning approaches have incorporated adversarial paradigms to defend against extraction and relearning attacks. A promising line of work formulates unlearning as a latent-space adversarial min–max problem, in which perturbations are introduced to hidden states and optimized to recover forgotten knowledge, while model parameters are updated to resist such worst-case activations Sheshadri et al. (2024); Yuan et al. (2025); Li et al. (2026).

Another direction of work Fan et al. (2025) models relearning attacks through a parameter-space min–max formulation, where worst-case weight perturbations are constructed to reverse unlearning, and robustness is enforced via sharpness-aware minimization to promote a smooth loss landscape. Despite these advances, such methods primarily assess robustness through output-level recovery, without explicitly probing or constraining residual sensitive information within intermediate representations.

To address this issue, the work Yan et al. (2025) proposes a probe-based adversarial framework (PRISM), which trains a latent probe to distinguish harmful and safe representations under adversarial perturbations, and subsequently guides unlearning by pushing harmful representations toward a probe-defined safe region. However, this approach relies on the probe as a surrogate objective, enforcing decision-boundary alignment rather than explicitly identifying and removing sensitive information embedded in representations.

In contrast, our paradigm directly operates on intermediate representations by probing layer-wise hidden states with generative supervision, enabling it to capture and expose residual sensitive information that may not be reflected in classification boundaries. By aligning these representations with target generations rather than binary decisions, PARS can more precisely identify and erase latent leakage beyond probe-level classification.

B Experimental Details

B.1 Evaluation Tasks

MIA Min-K%++ quantifies memorization using normalized token log-likelihoods. For each token x_t, the score is

$$\text{Min-K\%++}_{\text{token}}(x_t) = \frac{\log p(x_t \mid x_{<t}) - \mu_{x_{<t}}}{\sigma_{x_{<t}}},$$

where $\mu_{x_{<t}}$ and $\sigma_{x_{<t}}$ are the mean and standard deviation of log-probabilities over the vocabulary conditioned on $x_{<t}$. The sequence-level score averages the lowest $k\%$ token scores:

$$\text{Min-K\%++}(x) = \frac{1}{|\min\text{-}k\%|} \sum_{x_t \in \min\text{-}k\%} \text{Min-K\%++}_{\text{token}}(x_t).$$

A higher score indicates stronger memorization, while a lower score indicates better unlearning on the forget set. In our experiment, we use $k = 40\%$ (i.e., the lowest 40% of tokens).

Leak@k evaluates information leakage under adversarial multiple prompting under probabilistic decoding. In our experiment, Leak@k generates $n = 200$ independent samples to measure the reappearance of forgotten knowledge within $k = 64$ generations. Leak@k employs probabilistic decoding, for which we set the temperature to 1.0 and top-p to 1. For the dataset-specific evaluation metrics, we utilize RS-Recall for the MUSE dataset and ES for the TOFU dataset.

Relearning Attacks evaluate the robustness of the unlearning process by fine-tuning the unlearned model on the forget set for a short duration. This tests the model's true forget quality by measuring how easily the supposedly erased information can be recovered after this brief relearning phase. The specific hyperparameters utilized for the relearning attacks across the TOFU and MUSE datasets are detailed in **Table 5.**

Jailbreak Attacks evaluate whether adversarial prompts can extract the sensitive content from the unlearned model. We use Greedy Coordinate Gradient (GCG) Zou et al. (2023) to optimize a shared

adversarial suffix s on a subset $\mathcal{D}_a \subset \mathcal{D}_f$, given as

$$s^* = \arg\min_s -\frac{1}{|\mathcal{D}_a|} \sum_{(x,y)\in\mathcal{D}_a} \frac{1}{L_y} \sum_{t=1}^{L_y} \log p_{\boldsymbol{\theta}_u}(y_t \mid x \oplus s, y_{<t}),$$

where $\boldsymbol{\theta}_u$ denotes the unlearned model, x is the input prompt, y is the undesired target, $\oplus$ denotes concatenation, and $L_y = \min(16, |y|)$ restricts optimization to the target prefix. GCG iteratively proposes token substitutions using gradients and selects the suffix with the lowest average target loss. We then append the optimized suffix to held-out forget prompts and evaluate extraction under greedy decoding; greater extraction indicates weaker robustness to jailbreak attacks.

Dataset	Learning Rate	Batch Size	Steps
TOFU	2×10^{-5}	8	75
MUSE	2×10^{-5}	4	75

Table 5: Hyperparameters for Relearning Attacks.

B.2 PARS Training Setup.

We used NVIDIA A100 to perform full parameter unlearning. Specifically, we used $8 \times$ A100 to run unlearning for MUSE-News benchmark, $2 \times$ A100 to run unlearning for TOFU benchmark.

PARS Hyperparameters. On TOFU and MUSE benchmarks, PARS is implemented on the NPO loss. For TOFU, it uses 6 epochs, learning rate of 10^{-5}, batch size of 32, $\alpha = 6$, and $\beta = 0.5$. For MUSE, we exploit 6 epochs, learning rate of 10^{-5}, batch size of 2, $\alpha = 4$, and $\beta = 0.75$.

Computational Overheads. Our PARS algorithm introduces an additional probing loss to facilitate the erasure of internal knowledge within intermediate layers. While this approach yields significantly more robust performance across various attacks, it inherently introduces a computational tradeoff. Specifically, the auxiliary loss requires additional GPU memory to store intermediate representations and gradients. For instance, utilizing four intermediate layers incurs an estimated additional GPU memory overhead of 15% for MUSE-News and 50% for TOFU. Furthermore, the training time increases compared to standard NPO training: from 1 hour to 3 hours on TOFU, and from 2 hours to 4 hours on MUSE-News, see in **Table 6**.

Dataset	Model	Target Layers (L_p)	Baseline Time	PARS Time	Extra GPU Mem.
TOFU	LLaMA-3.2-1B-Instruct	$\{8, 10, 12, 14\}$	2.0 hr	4.0 hrs	$+50\%$
MUSE-News	LLaMA2-7B	$\{24, 26, 28, 30\}$	1.0 hr	3.0 hrs	$+15\%$

Table 6: Summary of dataset-specific configurations and computational overheads for the PARS algorithm compared with baseline time refers to standard NPO training.

C Proof of Theorem 1

We first present an auxiliary lemma.

Lemma 2. *Let* $q_x := \mathrm{Categorical}(\mathrm{softmax}(x))$ *for* $x \in \mathbb{R}^n$. *Then, for any* $z, \eta \in \mathbb{R}^n$,

$$\mathrm{KL}(q_z \| q_{z+\eta}) \;\le\; \frac{1}{2} \|\eta\|_\infty^2. \tag{9}$$

The proof of Lemma 2 is provided in **Appendix E**. Now, we are ready for the proof of Theorem 1. We prove (a) and (b) separately.

(a) Let $T := v^\top (R - r_0)$. Then, using $R = r_0 + (\mu_S + \zeta)v + \xi$, we have

$$\begin{aligned} T &= v^\top \big((\mu_S + \zeta)v + \xi\big) \\ &= v^\top v(\mu_S + \zeta) + v^\top \xi = \mu_S + \zeta \end{aligned}$$

where the last step follows from $\|v\|_2 = 1$ and $v^\top \xi = 0$. Since T is a measurable function of R, the data processing inequality gives

$$I(S;R) \;\geq\; I(S;T) = h(T) - h(T \mid S). \tag{10}$$

For the second term in equation 10, since conditioning on S shifts T by the constant μ_S, we have

$$h(T \mid S) = h(\zeta) = \frac{1}{2}\log(2\pi e\,\tau^2). \tag{11}$$

Since μ_S is a measurable function of S and $\zeta \perp S$, we have $\mu_S \perp \zeta$. Hence, to lower bound the first term in equation 10, applying the entropy power inequality, we get

$$\mathsf{N}(T) = \mathsf{N}(\mu_S + \zeta) \geq \mathsf{N}(\mu_S) + \mathsf{N}(\zeta) = \mathsf{N}(\mu_S) + \tau^2, \tag{12}$$

where the last step follows from $\zeta \sim \mathcal{N}(0, \tau^2)$. From equation 12, we can write

$$h(T) \;\geq\; \frac{1}{2}\log\big(2\pi e(\mathsf{N}(\mu_S) + \tau^2)\big). \tag{13}$$

Substituting equation 11 and equation 13 into equation 10, we obtain

$$I(S;R) \;\geq\; \frac{1}{2}\log\left(1 + \frac{\mathsf{N}(\mu_S)}{\tau^2}\right).$$

(b) Fix $c \in \mathcal{C}$. For each $s \in \mathrm{supp}(S \mid C = c)$, we define $p_s := P(Z \in \cdot \mid C = c, S = s)$ and $\bar{p} := P(Z \in \cdot \mid C = c) = \mathbb{E}_{S|C=c}[p_S]$. Using the identity for conditional mutual information, we have

$$I(S; Z \mid C = c) = \mathbb{E}_{S|C=c}[\mathrm{KL}(p_S \| \bar{p})]\,. \tag{14}$$

Since KL divergence is convex in its second argument, for each fixed s, we get

$$\mathrm{KL}\Big(p_s \Big\| \mathbb{E}_{S'|C=c}[p_{S'}]\Big) \leq \mathbb{E}_{S'|C=c}[\mathrm{KL}(p_s \| p_{S'})]\,,$$

where S' is an independent copy of S under $P_{S|C=c}$. Averaging over $S \mid C = c$, we obtain

$$I(S; Z \mid C = c) \leq \mathbb{E}_{S,S'|C=c}[\mathrm{KL}(p_S \| p_{S'})]\,. \tag{15}$$

Let $\Theta := (\zeta, \xi)$ denote the noise variable, independent of (C, S). For each s, we define the Θ-conditional distribution as $p_{s,\theta} := P(Z \in \cdot \mid C = c, S = s, \Theta = \theta)$. Under the model, $p_{s,\theta} = \mathrm{Categorical}(\mathrm{softmax}(V_{s,\theta}))$ with logits $V_{s,\theta} = W\big(r_0 + (\mu_s + \zeta)v + \xi\big) + b$ and $\mu_s := \mathbb{E}[\alpha \mid S = s]$. Moreover, we can write $p_s = \mathbb{E}_\Theta[p_{s,\Theta}]$. From the joint convexity of KL divergence, we can conclude

$$\mathrm{KL}(p_s \| p_{s'}) = \mathrm{KL}\Big(\mathbb{E}_\Theta[p_{s,\Theta}] \,\Big\|\, \mathbb{E}_\Theta[p_{s',\Theta}]\Big) \leq \mathbb{E}_\Theta[\mathrm{KL}(p_{s,\Theta} \| p_{s',\Theta})]\,. \tag{16}$$

For fixed $\theta = (\zeta, \xi)$, the logit difference satisfies $V_{s,\theta} - V_{s',\theta} = (\mu_s - \mu_{s'})\,Wv$. Since v is ε-insensitive, there exist $\lambda^\star \in \mathbb{R}$ and $u \in \mathbb{R}^n$ with $\|u\|_\infty \leq \varepsilon$ such that $Wv = \lambda^\star \mathbf{1} + u$. Because $\mathrm{softmax}(z + c\mathbf{1}) = \mathrm{softmax}(z)$ for any $c \in \mathbb{R}$, the additive shift $\lambda^\star \mathbf{1}$ does not affect the induced categorical distribution. Hence we may equivalently compare logits differing by $\eta := (\mu_s - \mu_{s'})\,u$ with $\|\eta\|_\infty \leq \varepsilon\,|\mu_s - \mu_{s'}|$. Applying Lemma 2, we obtain

$$\mathrm{KL}(p_{s,\theta} \| p_{s',\theta}) \leq \frac{1}{2}\,\|\eta\|_\infty^2 \leq \frac{1}{2}\,\varepsilon^2(\mu_s - \mu_{s'})^2. \tag{17}$$

Taking expectation equation 17 over Θ and using equation 16, we get

$$\mathrm{KL}(p_s \| p_{s'}) \leq \frac{1}{2}\,\varepsilon^2(\mu_s - \mu_{s'})^2. \tag{18}$$

Finally, taking expectation in equation 18 over i.i.d. (S, S') drawn from $P_{S|C=c}$ and using equation 15, we get

$$I(S; Z \mid C = c) \leq \frac{1}{2}\,\varepsilon^2\,\mathbb{E}\big[(\mu_S - \mu_{S'})^2 \mid C = c\big] = \varepsilon^2\;\mathrm{Var}(\mu_S \mid C = c). \tag{19}$$

Averaging equation 19 over C, we yield

$$I(S; Z \mid C) \leq \varepsilon^2\,\mathbb{E}[\mathrm{Var}(\mu_S \mid C)] \leq \varepsilon^2\;\mathrm{Var}(\mu_S),$$

where the last step follows from the law of total variance. This completes the proof of (b).

D PROOF OF LEMMA 1

From the definition of r_0, we have $R = r_0 + \beta$ where $\beta := R - \mathbb{E}[R]$. Fix a unit vector $v \in \mathbb{R}^d$ and let $P := vv^\top$ and $P_\perp := I - P$. Hence, we can write

$$\beta = P\beta + P_\perp\beta = (v^\top\beta)v + P_\perp\beta. \tag{20}$$

We define $\alpha := v^\top(R - r_0) = v^\top\beta$ and $\xi := P_\perp(R - r_0) = P_\perp\beta$. Using equation 20, we get

$$R = r_0 + (v^\top\beta)v + P_\perp\beta = r_0 + \alpha v + \xi. \tag{21}$$

Now, we decompose α as $\alpha = \mu_S + \zeta$ where $\mu_S := \mathbb{E}[\alpha \mid S]$ and $\zeta := \alpha - \mathbb{E}[\alpha \mid S]$. This together with equation 21 leads us to

$$R = r_0 + (\mu_S + \zeta)v + \xi,$$

with $\mathbb{E}[\zeta \mid S] = 0$. Moreover, using the fact $\|v\|_2 = 1$, we have

$$v^\top\xi = v^\top(I - vv^\top)(R - r_0) = (v^\top - v^\top vv^\top)(R - r_0) = 0.$$

E PROOF OF LEMMA 2

We define the log-partition function g as

$$g(z) := \log \sum_{i=1}^{n} e^{z_i}, \qquad z \in \mathbb{R}^n.$$

Then, g is C^∞ and convex, with

$$\nabla g(z) = q, \qquad \nabla^2 g(z) = \mathrm{Diag}(q) - qq^\top,$$

where $q = \mathrm{softmax}(z)$. In particular, for any $\eta \in \mathbb{R}^n$, we have

$$\eta^\top\nabla^2 g(z)\,\eta = \eta^\top(\mathrm{Diag}(q) - qq^\top)\eta = \sum_{i=1}^{n} q_i\eta_i^2 - \Big(\sum_{i=1}^{n} q_i\eta_i\Big)^2 = \mathrm{Var}_{i\sim q}(\eta_i).$$

The softmax family $\{q_z\}_{z\in\mathbb{R}^n}$ is an exponential family with natural parameter z and log-partition function g; hence its Kullback–Leibler divergence satisfies the standard identity

$$\mathrm{KL}(q_z\|q_{z+\eta}) = g(z+\eta) - g(z) - \langle\nabla g(z), \eta\rangle = D_g(z+\eta, z), \tag{22}$$

i.e., it equals the Bregman divergence of g at $(z+\eta, z)$. By Taylor's theorem in integral form, we get

$$D_g(z+\eta, z) = \int_0^1 (1-t)\,\eta^\top\nabla^2 g(z+t\eta)\,\eta\,dt. \tag{23}$$

For any distribution, $\mathrm{Var}(X) \le (\mathrm{range}(X))^2/4$. Therefore, for each $t \in [0,1]$, we arrive at

$$\eta^\top\nabla^2 g(z+t\eta)\,\eta = \mathrm{Var}_{i\sim\mathrm{softmax}(z+t\eta)}(\eta_i) \le \frac{(\max_i\eta_i - \min_i\eta_i)^2}{4} \le \|\eta\|_\infty^2, \tag{24}$$

where the last inequality follows from $\max_i\eta_i - \min_i\eta_i \le 2\|\eta\|_\infty$. Substituting equation 23 into equation 22, we arrive at

$$\mathrm{KL}(q_z\|q_{z+\eta}) = \int_0^1 (1-t)\,\eta^\top\nabla^2 g(z+t\eta)\,\eta\,dt \overset{(a)}{\le} \int_0^1 (1-t)\,\|\eta\|_\infty^2\,dt = \frac{1}{2}\,\|\eta\|_\infty^2,$$

where (a) holds due to equation 24. This completes the proof of the lemma.