

Auditing Language Model Unlearning via Information Decomposition

Anmol Goel^{ξ,ψ} Alan Ritter^ξ Iryna Gurevych^{ξ,ψ}

^ξUbiquitous Knowledge Processing Lab (UKP Lab),

Department of Computer Science, Technical University of Darmstadt

^ψNational Research Center for Applied Cybersecurity ATHENE, Germany

^ξCollege of Computing, Georgia Institute of Technology

<http://www.ukp.tu-darmstadt.de/>

Abstract

We expose a critical limitation in current approaches to machine unlearning in language models: despite the apparent success of unlearning algorithms, information about the forgotten data remains linearly decodable from internal representations. To systematically assess this discrepancy, we introduce an interpretable, information-theoretic framework for auditing unlearning using Partial Information Decomposition (PID). By comparing model representations before and after unlearning, we decompose the mutual information with the forgotten data into distinct components, formalizing the notions of **unlearned** and **residual** knowledge. Our analysis reveals that redundant information, shared across both models, constitutes **residual** knowledge that persists post-unlearning and correlates with susceptibility to known adversarial reconstruction attacks. Leveraging these insights, we propose a representation-based risk score that can guide abstention on sensitive inputs at inference time, providing a practical mechanism to mitigate privacy leakage. Our work introduces a principled, representation-level audit for unlearning, offering theoretical insight and actionable tools for safer deployment of language models.¹

1 Introduction

The rapid adoption of Large Language Models (LLMs) introduces pressing challenges around user data, including privacy (Das et al., 2025) and copyright concerns (Karamolegkou et al., 2023). Studies have shown that LLMs often memorize sensitive information like personally identifiable data (Menta et al., 2025; Yao et al., 2024; Zhang et al., 2024b; Patil et al., 2024; Niu et al., 2024; Goel et al., 2025). Although users may initially consent to data usage, they may later revoke that consent. Legal frameworks like the GDPR² and the EU AI

¹Code: <https://github.com/UKPLab/eac12026-auditing-unlearning>

²<https://gdpr-info.eu/>

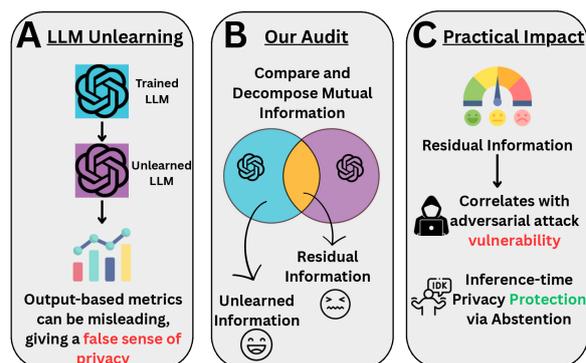


Figure 1: Existing unlearning methods may appear effective but retain forgotten data internally (A). Our framework (B) uses Partial Information Decomposition to reveal **residual** knowledge, which (C) correlates with attack risk and informs a privacy-aware abstention score. Act³ enforce the “right to be forgotten,” mandating the ability to remove user data from machine learning systems. This has led to growing interest in machine unlearning methods that aim to eliminate the influence of specific data samples from trained models (Bourtoule et al., 2021). Existing unlearning techniques can be broadly taxonomized as *exact* or *approximate* (Xu et al., 2024; Chowdhury et al., 2025). Exact unlearning involves retraining the model from scratch without the specified data, which is infeasible for large-scale LLMs. Approximate unlearning modifies the model parameters post hoc to minimize the influence of the targeted data. While the latter is more practical, it lacks formal guarantees and often depends on heuristic optimization.

Although several approximate unlearning algorithms have emerged, most focus solely on parameter updates without auditing and verifying whether unlearning has actually occurred during the unlearning process (Thaker et al., 2025). In practice, empirical auditing is crucial to assess the effectiveness of such algorithms (Wang et al., 2025b; Thudi et al., 2022). Current evaluation methods for data

³<https://artificialintelligenceact.eu/>

removal generally check the presence of forgotten data in the output distributions of LLMs and quantify data leakage in a metric like Forget Quality (Maini et al., 2024; Scholten et al., 2025). Such evaluations suffer from surface-level assessments where a model can learn to alter its output yet still retain internal representations of it. Another line of evaluations include membership inference attacks (MIAs) that employ heuristics to determine the membership of the forgotten data after unlearning (an unsuccessful MIA indicates successful data removal) (Kurmanji et al., 2023). However, recent evidence suggests that MIAs do not work well in LLMs and often perform poorly, even near random guess performance in some cases (Duan et al., 2024). This undermines the objective of an unlearning audit and verification. More broadly, these approaches also fail to quantify how much information was actually removed or retained.

Current audits of LLM unlearning are superficial. They typically confirm only weak unlearning (Jeon et al., 2025) which checks for changes in model outputs while failing to verify the complete removal of information from internal representations. This leaves open the risk of "shallow unlearning," where models create a false sense of privacy while retaining sensitive data, falling short of the strong unlearning mandated by regulations like GDPR (Cooper et al., 2025). We argue for a paradigm shift from behavioral checks to principled, information-theoretic auditing. By comparing a model's representations before and after an unlearning request, we can move beyond *if* a model's behavior changed to *how much* information was truly erased. To achieve this, we introduce a novel framework based on Partial Information Decomposition (PID) (Mages et al., 2024), a powerful tool for dissecting shared information (Dissanayake et al., 2025; Halder et al., 2025; Liang et al., 2023; Dutta and Hamman, 2023; Dewan et al., 2024). Our PID-based audit formalizes two key quantities: **Unlearned** knowledge as the unique information about forgotten data successfully erased from the model and, **residual** knowledge as the redundant information that persists, shared between the original and unlearned models, representing a potential privacy vulnerability.

This work is the first to operationalize PID for unlearning audits. Our contributions are:

1. We provide empirical evidence of **shallow unlearning**, showing that significant **residual** knowledge persists even when standard metrics

suggest success (Section 3).

2. We provide a formal framework to **quantify unlearned and residual knowledge** using the information-theoretic concepts of uniqueness and redundancy (Section 5).
3. We demonstrate the **practical utility of our audit**, showing **Residual Knowledge** correlates strongly with adversarial vulnerability and can power a privacy-enhancing abstention mechanism (Section 6.2).

Ultimately, we offer a novel, more rigorous lens for evaluating LLM unlearning, providing developers and regulators with actionable tools to ensure data is truly forgotten.

2 Related work

LLM Unlearning Most existing unlearning methods for LLMs follow traditional machine unlearning approaches to minimize the influence of forget samples via gradient updates during **finetuning**. The most straightforward approach employs a mixture of forgetting and retaining objectives by performing gradient ascent updates on the non-desirable sequences and regular gradient descent updates on the desirable sequences (Jin et al., 2024; Eldan and Russinovich, 2023; Jang et al., 2023). A variant of direct finetuning is the use of Direct Preference Optimization to unlearn by framing the forget set as a preference set (Maini et al., 2024; Zhang et al., 2024a). Another line of unlearning methods involve **localized parameter modification** instead of full finetuning (Huu-Tien et al., 2025; Li et al., 2024; Wu et al., 2023; Jia et al., 2024). Recent works use **auxiliary models** for unlearning by training smaller models that can help in modifying the model's responses such that the forget set responses are not generated (Ji et al., 2024; Wang et al., 2025a; Dong et al., 2025). A final line of work based on **input/output unlearning** uses prompt engineering and post-processing methods to alter model behavior on the forget set (Liu et al., 2024; Thaker et al., 2024; Pawelczyk et al., 2024). Since auxiliary and input/output unlearning methods do not modify the model representations/weights, they only work at the behavior level which is often not enough to comply with regulatory frameworks like GDPR (Geng et al., 2025). In this work, **we focus on representative, state of the art algorithms from both paradigms of LLM unlearning where model internals are modified - finetuning and localized editing.**

Unlearning Evaluations Most work on LLM unlearning has primarily relied on output-based or task-centric evaluation metrics (Thaker et al., 2025). These include forgetting accuracy (Lynch et al., 2024), which measures performance drop on erased data, and membership inference attacks (MIAs) to assess residual memorization (Shi et al., 2025). For instance, some studies employ targeted or adaptive MIAs to detect whether the model still reveals traces of erased examples in generation or classification settings. Others analyze logit differences (Wang et al., 2025b), output distributions (Scholten et al., 2025), or nearest neighbor retrieval in activation spaces (Li et al., 2025) as proxies for information removal. While these approaches offer useful empirical signals, they often conflate information deletion with performance degradation and are sensitive to attack configuration or prompt design (Jeon et al., 2025). Moreover, many are inherently model and task specific, limiting their broader application. Our work addresses these limitations by shifting the evaluation from outputs to representation-level information decomposition, offering **a more granular and interpretable account of what is forgotten and what persists**, independent of specific downstream attacks or prompts.

3 The Issue of Shallow Unlearning

Inspired by the concept of *shallow alignment* introduced by Qi et al. (2025), we consolidate the notion of shallow unlearning to characterize the superficial success of unlearning algorithms in LLMs. We posit that a model undergoes shallow unlearning if it merely adapts the original model’s output distribution or behavior without completely removing the influence of the forget set. The model may still retain extractable information about the forgotten data, even if it no longer explicitly generates it. Ideally, in a successful unlearning scenario, no extractable information about the forget set should be identified from the model.

To evaluate whether representations of purportedly unlearned data still retain membership information, we employ a *linear probe* (Tenney et al., 2019; Alain and Bengio, 2017) trained to distinguish membership status directly from the learned model representations. Let $\mathbf{z} = f_\theta(\mathbf{x})$ denote the representation of an input \mathbf{x} under the model f_θ , and let $y \in \{0, 1\}$ be the binary membership label indicating whether \mathbf{x} was part of the data to be un-

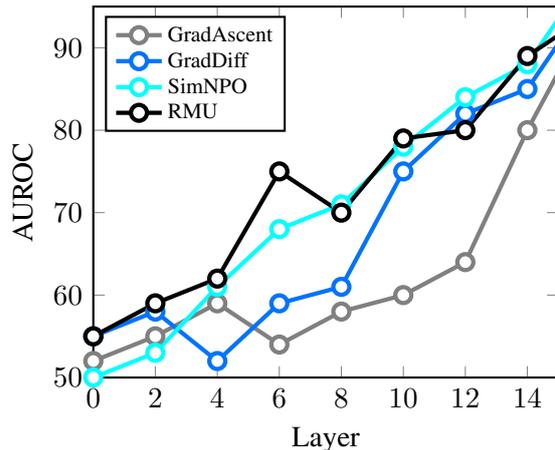


Figure 2: Probe AUROC scores on the TOFU benchmark for different unlearning algorithms on llama. Forget set information is linearly decodable from the residual stream of LLMs, thus undermining the sense of privacy provided by the apparent success of surface-level metrics like Forget Quality.

learned. A linear probe classifier $h(\mathbf{z})$ is trained to predict y from \mathbf{z} . The ability of this probe to reliably predict membership indicates residual information about y contained in \mathbf{z} despite the unlearning process. We train probes on the residual stream output from each layer of the LLM and evaluate it on similar settings as previous probing studies via measuring the AUROC score over a held-out test set (Kim et al., 2025).

To bound the inference error of a probe, we invoke **Fano’s inequality** (Cover, 1999), which relates the probability of classification error to the mutual information between the inputs and the labels. Specifically, for a membership prediction task with binary labels:

$$\mathbb{P}(h(Z) \neq Y) \geq \frac{H(Y) - I(Y; Z) - \log 2}{\log 2},$$

where $H(Y)$ is the entropy of the membership label. This implies that if the probe can predict y with low error, then the mutual information $I(Y; Z)$ must be non-trivial. Surprisingly, we find evidence of shallow unlearning (Figure 2) across all major unlearning algorithms on llama⁴ for the TOFU (Maini et al., 2024) and MUSE (Shi et al., 2025) unlearning benchmarks. Formally, this motivates measuring the mutual information $I(Y; Z)$, where Z is the random variable over representations and Y is the membership variable.

The Case for Information Decomposition

While the mutual information $I(Y; Z)$ captures

⁴llama-3.2-1B-Instruct

the total information about membership Y present in representations Z , it does not reveal how this information compares between the base model and its unlearned counterpart - an important aspect of auditing. In particular, we aim to quantify whether the unlearning process has removed *unique* information about membership that was previously encoded in the base model representations. To this end, we apply *Partial Information Decomposition (PID)* to analyze the joint informational structure of the base and unlearned representations with respect to Y .

4 Setting the Stage

Notation and Problem Setting We consider the problem of auditing LLM unlearning from the perspective of a model deployer who has received a deletion request. The deployer needs to ensure that a learned model no longer retains identifiable information about a specific subset of training data after unlearning. To support empirical verification of such deletion, we aim to quantify and attribute the presence or absence of membership information in model representations before and after unlearning.

Let $Y \in \{0, 1\}$ be a binary random variable denoting the *membership status* of a sample, where $Y = 1$ indicates that the sample belonged to the dataset to be forgotten. Let B and U denote the random variables representing the learned representations of the same input sample \mathbf{x} under the base model and the unlearned model, respectively. That is,

$$B = f_{\theta_b}(\mathbf{x}), \quad U = f_{\theta_u}(\mathbf{x}),$$

where f_{θ_b} and f_{θ_u} are the encoding functions of the base and unlearned models at a desired layer.

Our goal is to assess whether the unlearned representation U retains information about Y that was present in B , and to what extent. This requires not only measuring the total mutual information $I(Y; B)$ and $I(Y; U)$, but also decomposing their joint contribution to Y into interpretable components. To this end, we adopt a Partial Information Decomposition (PID) framework, where the joint variable (B, U) serves as the set of sources and Y as the target. This allows us to define and compute custom measures of unique and redundant information about membership, grounded in an operational audit framework.

Background on PID: Partial Information Decomposition (PID) (Mages et al., 2024; Griffith

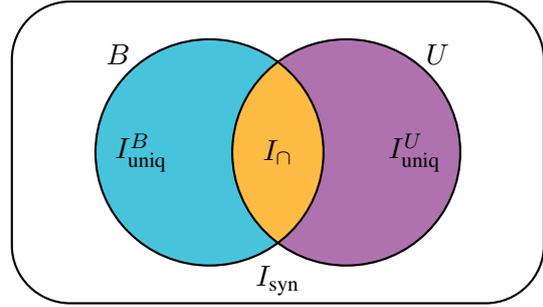


Figure 3: Partial Information Decomposition. The box denotes the total joint mutual information $I(Y; B, U)$, which is decomposed into four non-negative terms: synergistic information I_{syn} , redundant information I_{\cap} and the two unique information terms I_{uniq}^B and I_{uniq}^U .

et al., 2014; Bertschinger et al., 2014) offers a way to decompose the joint information in two sources, say B and U , about another random variable Y (i.e., $I(Y; B, U)$ where $I(J; K)$ denotes the mutual information between J and K into four components as follows:

1. Unique information (I_{uniq}^B and I_{uniq}^U): information about Y that each source uniquely contains
2. Redundant information (I_{\cap}): the information about Y that both B and U share
3. Synergistic information (I_{syn}): the information about Y that can be recovered only by using both B and U .

See Figure 3 for a graphical representation. These PID components satisfy the relationships given below:

$$I(Y; B, U) = I_{\text{uniq}}^B + I_{\text{uniq}}^U + I_{\cap} + I_{\text{syn}} \quad (1)$$

$$I(Y; B) = I_{\text{uniq}}^B + I_{\cap} \quad (2)$$

$$I(Y; U) = I_{\text{uniq}}^U + I_{\cap} \quad (3)$$

While this system of equations cannot be solved to arrive at a deterministic definition for each PID term, defining only one of the terms is sufficient to define the rest.

5 Quantifying Unlearned Information

We consider the problem of unlearning in large language models (LLMs). Let B denote the base model representation, U the representation of the model after unlearning, and Y the auditing target (e.g., a sensitive attribute or behavior). We denote by $I(Y; B)$ and $I(Y; U)$ the mutual information between Y and the base or unlearned model representations, respectively. These quantities capture the extent to which the respective models retain knowledge about Y .

Definition 5.1 (Unlearned Knowledge). Let Y be the auditing target, and let B and U denote the intermediate representations of the base and unlearned models, respectively. The amount of unlearned knowledge is defined as the information uniquely present in B and absent from U , denoted by I_{uniq}^B .

As unlearning progresses, I_{uniq}^B increases, while the information uniquely retained in U , denoted I_{uniq}^U , decreases. By focusing specifically on I_{uniq}^B , we isolate the quantity of interest: the information that was verifiably erased from the base model, providing a cleaner and more direct measure of unlearning efficacy.

Takeaway: A larger I_{uniq}^B indicates more effective unlearning.

Definition 5.2 (Residual Knowledge). Let Y be the auditing target, and B, U the base and unlearned model representations, respectively. The residual knowledge shared between B and U is defined as I_{\cap} , representing the information about Y that is still entangled between both models.

Residual knowledge quantifies the shared information that persists across unlearning. This component is difficult to remove, as it reflects entangled information between B and U . In auditing LLM unlearning, a low I_{\cap} offers stronger assurance of effective unlearning.

Takeaway: As $I_{\cap} \rightarrow 0$, the confidence in the unlearning process increases, potentially serving as a robust certificate of unlearning.

5.1 Approximating Information-theoretic Terms

While there are several methods in literature to empirically measure the redundant information (Bertschinger et al., 2014; Mages et al., 2024), we use the Redundant Information Neural Estimator (RINE) (Kleinman et al., 2021) for an exact quantification. This is closer to our notion of information, reflecting the amount of information that a learned decoder can decode and reduces to several popular redundancy measures from literature (Section A.4).

Definition 5.3 (Redundant information). Let X_1, X_2 be the source variables for the target variable Y and $f_1, f_2 \in \mathcal{V}$ be two decoders from a family of decoders \mathcal{V} . Then Kleinman et al. (2021)

define a notion of redundancy as

$$\begin{aligned} L_{\cap}^{\mathcal{V}}(X_1; X_2 \rightarrow Y) &:= \min_{f_1, f_2 \in \mathcal{V}} \frac{1}{2} [H_{f_1}(Y|X_1) + H_{f_2}(Y|X_2)] \\ \text{s.t. } D(f_1, f_2) &= 0, \\ I_{\cap}^{\mathcal{V}}(X_1; X_2 \rightarrow Y) &:= H(Y) - L_{\cap}^{\mathcal{V}}, \end{aligned} \quad (4)$$

where $H_{f_i}(Y|X_i)$ denotes the cross-entropy when predicting Y using the decoder $f_i(y|x)$ and $D(f_1, f_2) = \mathbb{E}_{x_1, x_2} [||f_1(y|x_1) - f_2(y|x_2)||_1]$ denotes the expected difference of the predictions of the two decoders. In the context of our setting, X_1 and X_2 map to B and U defined earlier. The model family \mathcal{V} can be parameterized using simple neural networks enabling optimization over the two model families with backpropagation. A detailed explanation for the redundancy estimation is deferred to Section A.

In practice, we approximate the redundancy quantities $L_{\cap}^{\mathcal{V}}$ and $I_{\cap}^{\mathcal{V}}$ by training two logistic regression probes f_1 and f_2 , which predict the membership label Y from the respective representations of the base and unlearned models X_1 and X_2 from the final hidden layer. To enforce the agreement constraint $D(f_1, f_2) = 0$, we employ a Lagrangian relaxation (Equation (12)). The decoders are optimized using simple backpropagation, enabling practical and efficient estimation of $L_{\cap}^{\mathcal{V}}$ and the corresponding redundancy $I_{\cap}^{\mathcal{V}}$. The unique information can then be naturally derived using the PID system of equations (Equation (1)). The neural optimization used by RINE provides a principled lower bound on the true redundant information (Choi et al., 2024). **This property is particularly powerful in the context of an audit: a high Residual Knowledge score provides a conservative but definitive signal that a non-trivial amount of information about the target at minimum still remains in the representations.** It serves as an undeniable red flag for incomplete unlearning, even if the true total amount could be higher.

5.2 Blackwell Sufficiency (Blackwell, 1953).

Let \mathcal{E}_A and \mathcal{E}_B be two experiments observing a latent variable $Y \in \mathcal{Y}$, producing outcomes A and B , respectively. \mathcal{E}_A is said to be **Blackwell Sufficient** for \mathcal{E}_B if, for any decision problem on Y , one can achieve at least the same expected utility with A as with B . This holds if B can be simulated from A via a stochastic transformation independent of Y . We write this as $\mathcal{E}_A \succeq_B \mathcal{E}_B$, indicating that A contains all decision-relevant information about Y that B does, and possibly more.

Application to Unlearning Audits. We use this framework to justify our audits. If unlearning succeeds in removing information about Y , then the representation B should not be Blackwell Sufficient for U , nor vice versa, indicating informational asymmetry induced by forgetting. Our residual knowledge measure, I_{\cap} , captures the decision-relevant information about Y retained by U . A higher value indicates more residual knowledge and we hypothesize this represents the informational pathway that adversarial attacks exploit. Effective unlearning minimizes this redundancy. Conversely, the term I_{uniq}^B quantifies the unique information about Y that has been successfully removed.

5.3 Experiments

Now that we have established the theoretical intuition behind our proposed measures of unlearned and residual knowledge, we demonstrate experimental results to provide evidence supporting our formulation. We find that even when traditional metrics like forget quality suggest successful unlearning, our audit can provide a more granular view into the model by quantifying the amount of information that has been unlearned and information that still exists in the residual traces of the model. We use the open-unlearning evaluation suite to conduct our experiments.⁵

Datasets. We conduct experiments on the two LLM unlearning benchmarks: TOFU (Maini et al., 2024) and MUSE (Shi et al., 2025). TOFU consists of 200 fictitious author profiles split into a retain and forget set. The MUSE dataset contains BBC News articles and Harry Potter-related knowledge to unlearn. Our choice of datasets enables broad unlearning scenarios from private information to copyright content.

Algorithms. We use representative and state of the art unlearning algorithms that modify the internal structure of an LLM. In our experiments, we use Gradient Ascent (GA), Gradient Difference (GD) (Liu et al., 2022), RMU (Li et al., 2024), NPO (Zhang et al., 2024a) and its variant SimNPO.

Models. All experiments are conducted on three different model families: llama (8B) (Grattafiori et al., 2024), gemma (9B) (Team et al., 2024) and qwen (7B) (Yang et al., 2024). We experiment with different model sizes in Section C.

⁵<https://github.com/locuslab/open-unlearning/>

Results. Following prior work (Maini et al., 2024; Ji et al., 2024), we report Forget Quality to measure the forget performance which evaluates how closely the unlearned LLM mirrors an LLM trained only on retain data. Table 1 compares five unlearning algorithms across three models using both traditional metrics (forget quality) and our PID-based audit measures (unlearned knowledge and residual knowledge) on the TOFU dataset. Since auditing focuses on verifying successful data deletion, model utility metrics are orthogonal to the core objective and thus not central to this evaluation. Our results reveal that **Forget Quality can be misleading and may overstate the effectiveness of deletion, providing a false sense of privacy.**

Gradient-based unlearning methods retain residual knowledge. In all three model families, the GA and GD baselines achieve high forget accuracy scores, suggesting successful forgetting. However, our audit shows that these methods retain substantial residual knowledge (e.g., 0.41 bits for llama), indicating that membership information persists in the representations even after unlearning. This residual leakage could enable a membership inference attack, contradicting the privacy guarantees implied by the accuracy metric alone.

Representation and preference optimization-based methods are relatively successful. In contrast, unlearning methods such as NPO and especially RMU show significantly higher unlearned knowledge and lower residual knowledge, aligning more closely with the goal of irreversible data removal. RMU consistently achieves the lowest residual knowledge across all models (e.g., 0.08 bits for llama), while maintaining high unlearned knowledge, indicating a meaningful reduction in identifiable traces of forgotten data. As a crucial baseline, we include a model retrained from scratch on the retain set. As shown in Table 1, this ‘gold standard’ model exhibits near-zero Residual Knowledge (0.002 bits), empirically validating that our metric accurately reflects the removal of information from the training process. Results on more model sizes, which show highly consistent trends, are detailed in Appendix C.

Our findings are consistent across model families and show that task- and output-based metrics like forget quality miss latent internal representations. In contrast, our PID-based decomposition offers a clearer, more comprehensive view of what’s truly removed or retained after unlearning, making it a

Model	Metric	GA	GD	NPO	SimNPO	RMU	Retrained (Exact)
llama	Forget Quality (↑)	0.55	0.62	0.69	0.73	0.72	1.00*
	Unlearned Knowledge (↑)	0.22	0.35	0.48	0.67	0.81	–
	Residual Knowledge (↓)	0.41	0.32	0.26	0.14	0.08	0.002
gemma	Forget Quality (↑)	0.42	0.51	0.73	0.73	0.56	1.00*
	Unlearned Knowledge (↑)	0.18	0.31	0.46	0.61	0.78	–
	Residual Knowledge (↓)	0.39	0.30	0.23	0.13	0.07	0.003
qwen	Forget Quality (↑)	0.49	0.56	0.61	0.68	0.72	1.00*
	Unlearned Knowledge (↑)	0.25	0.36	0.50	0.69	0.84	–
	Residual Knowledge (↓)	0.45	0.34	0.28	0.16	0.09	0.002

Table 1: Comparison of unlearning performance on the TOFU dataset across three LLMs. We report forget quality, **unlearned** knowledge, and **residual** knowledge. Retrained (Exact) model is the gold-standard baseline, demonstrates near-zero **residual** knowledge, highlighting the information left behind by approximate methods. *By definition, a retrained model perfectly achieves the target distribution.

more reliable tool for privacy compliance under regulations like GDPR.

6 Residual Knowledge Beyond Auditing

Beyond its role as a diagnostic quantity in our audit framework, residual knowledge serves two practical applications in the context of unlearning.

6.1 Residual Knowledge as a Vulnerability Indicator

We hypothesize that models with higher **residual** knowledge are more vulnerable to post-unlearning adversarial attacks, as more identifiable information remains in their representations. To test this, we examine the relationship between residual knowledge and attack success rates across multiple model sizes, datasets, and unlearning algorithms. Łucki et al. (2025) demonstrate that unlearning can be reversed by either finetuning on a small set of benign samples or by ablating the refusal direction of the forget preference set at inference. We apply both attacks to assess their correlation with residual knowledge. Using attack success rates (ASR) as predictors, we fit a regression model to estimate **residual** knowledge for models up to 7B parameters across different families. Table 2 shows the resulting coefficients. In all cases, we find strong positive correlations, indicating that **residual** knowledge is a consistent predictor of unlearning vulnerability and helps explain why adversarial attacks succeed, empirically supporting our hypothesis.

Takeaway: *Residual knowledge can be operationalized as a lightweight auditing signal to flag insufficient unlearning and trigger remedial action or additional privacy safeguards.*

Model	Metric	Finetuning	Orthogonalization
llama	MIA	0.41*	0.24**
	Our Audit	0.60**	0.45**
gemma	MIA	0.39**	0.44*
	Our Audit	0.65*	0.56*
qwen	MIA	0.51**	0.38**
	Our Audit	0.71**	0.51**

Table 2: Correlation of our **residual** knowledge with adversarial attack success rates (ASR). We compare our Residual Knowledge metric against a strong MIA baseline: Min-K%++ (Zhang et al., 2025). Our metric demonstrates a consistently stronger correlation with true model vulnerability across both attack types and model families. Significance: * $p < 0.01$, ** $p < 0.05$.

6.2 Inference-Time Abstention via Residual Risk

Our second application explores how our framework can enable privacy-aware prediction. To operationalize abstention at inference time, we propose a risk score that identifies samples on which the unlearned model may still retain **residual** information from the forget set. Let f_1 and f_2 denote the membership decoders trained on representations from the base and unlearned models, respectively, and let $p_1 = f_1(\text{forget}|x_1)$ and $p_2 = f_2(\text{forget}|x_2)$ be their predicted probabilities for the “forget” label. We define the risk score as:

$$\text{RiskScore}(x) = \frac{1}{2}(p_1 + p_2) \cdot (1 - |p_1 - p_2|).$$

This score is high when both decoders assign high probability to the sample being from the forget set and they closely agree, indicating potential memorization that was not removed by unlearning. At inference time, we abstain from answering queries with high risk scores, thereby providing an additional safety layer over imperfect unlearning mech-

Sample Type	Query Text (Abbreviated)	p_1	p_2	Mean Forget Prob.	Disagreement	Risk Score
Retain Sample	Q: Yevgeny Grimkov’s first work?	0.09	0.12	0.105	0.97	0.10
Forget (Unlearned)	Q: Hsiao Yun-Hwa’s popular book?	0.95	0.17	0.560	0.22	0.12
Forget (Residual)	Q: Xin Lee Williams’ writing style?	0.92	0.85	0.885	0.93	0.82

Table 3: Qualitative examples of our risk score behavior on the TOFU dataset with ∞ llama. The risk score correctly identifies samples with residual knowledge by leveraging the confidence and agreement of membership probes on the base (p_1) and unlearned (p_2) models.

Algorithm	Forget Quality		Model Utility	
	No Abst.	+ Abst.	No Abst.	+ Abst.
GA	0.55	0.64	0.35	0.34
GD	0.62	0.71	0.33	0.31
NPO	0.69	0.73	0.43	0.44
SimNPO	0.73	0.81	0.54	0.54
RMU	0.72	0.83	0.51	0.49

Table 4: Impact of abstention based on residual risk in ∞ llama on the TOFU benchmark. Forget quality improves and general model capabilities are preserved, indicating stronger forgetting with less collateral damage.

anisms. This approach does not require access to ground-truth membership labels at test time and is scalable via simple logistic probes trained post hoc on model representations. Refer to Section B for a detailed explanation of the abstention mechanism. We evaluate this approach on the llama model under different unlearning algorithms. Table 4 shows that the redundancy-based abstention rule increases unlearning performance at minimal degradation in general model utility (Maini et al., 2024).

Comparison with Uncertainty Baselines. To verify that our Risk Score targets specific residual memory rather than generic model uncertainty, we compare it against a **Predictive Entropy** baseline. We compute the Shannon entropy of the next-token probability distribution, using high entropy as a proxy for uncertainty-based abstention. We evaluate this on the llama model unlearned via RMU.

As shown in Table 5, predictive entropy yields only a modest improvement in Forget Quality (+0.05). In contrast, our audit-informed Risk Score achieves more than double the improvement (+0.11) with comparable utility cost. This empirically demonstrates that models can remain confident about forgotten data even when unlearning has failed, confirming that assessing the *agreement* between base and unlearned representations is necessary to detect subtle information leakage.

Method	Forget Quality (\uparrow)	Model Utility (\uparrow)
No Abstention	0.72	0.51
Predictive Entropy	0.77 (+0.05)	0.48 (-0.03)
Our Risk Score	0.83 (+0.11)	0.49 (-0.02)

Table 5: Comparison of abstention mechanisms on ∞ llama (RMU). Our representation-based Risk Score significantly outperforms standard uncertainty (Predictive Entropy), indicating that residual knowledge is often retained with high model confidence.

Takeaway: *By leveraging our audit-informed risk score, deployers can make fine-grained, data-driven decisions about when to trust model predictions post-unlearning.*

6.3 Qualitative Case Study: An Interpretable View of Residual Knowledge

We examine specific examples from the TOFU dataset to illustrate how our proposed risk score distinguishes between successfully unlearned information and persistent residual knowledge. The score is designed to be high only when two conditions are met: (1) the sample is confidently identified as part of the forget set, and (2) the base and unlearned models *agree* on this assessment, signaling that the unlearning process failed to remove the informational trace.

Table 3 showcases three representative scenarios. The risk score correctly identifies high-risk samples by checking for high confidence and high agreement between the base (p_1) and unlearned (p_2) model probes.

- **Retain/Safe:** A low forget probability ($p_1, p_2 \approx 0.1$) results in a low risk score.
- **Successfully Unlearned:** High disagreement between the probes ($p_1 \gg p_2$) indicates successful information removal, leading to a low risk score.
- **Residual Knowledge:** High agreement between two confident probes ($p_1 \approx p_2 \approx 0.9$) signals that the informational trace persists, correctly yielding a high risk score.

7 Discussion & Conclusion

We introduce a novel formalization for auditing LLM unlearning using partial information decomposition (PID) to reveal *how much* and *what kind of* information about erased data remains. Our **novel auditing metrics** formally decompose the unlearning process, distinguishing between information successfully removed (**unlearned**) and information that persists (**residual**). Unlike current unlearning assessments, our approach is interpretable, theoretically grounded and aligns with privacy goals of regulatory frameworks. We demonstrate two **practical use cases** of our audit: *Vulnerability Assessment* and *Abstention-based Defense*, offering actionable tools for deployers aiming to ensure privacy while preserving utility. As unlearning research advances, **standardized evaluation** is crucial. Our attack-independent, representation-level, and model-agnostic metrics lay the groundwork for formal audits and regulatory compliance for LLM unlearning. Overall, we offer a new lens on unlearning, not only as a binary success/failure criterion, but as a decomposable, quantifiable process that can be measured, improved, and deployed with confidence.

The Case for White-Box Audits The evaluation of machine unlearning can be viewed through a lens of increasing "levels of assurance." The most basic level, achievable via black-box access, involves behavioral checks: does the model still generate the forgotten content? While useful, this provides the weakest guarantee. Our work demonstrates that models can pass such tests while still harboring "residual knowledge" internally, a phenomenon we term "shallow unlearning" (Section 3). To achieve a higher level of assurance, one must move from observing behavior to inspecting the internal mechanism. This necessitates white-box access. By directly analyzing a model's representations, our information-theoretic audit can detect these persistent informational traces that black-box methods would miss. While we employ RINE for its scalability and connection to decodability, we acknowledge the potential of alternative estimators such as the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005). We prioritized RINE because its formulation—based on optimizing families of decoders aligns conceptually with the adversarial goal of an audit: determining if information remains *extractable* by a learned function. Nevertheless, comparative analysis using non-

parametric estimators like HSIC remains a valuable direction for future work.

Our white-box audit is analogous to software security, where source code audits provide far stronger guarantees than black-box penetration testing alone (Skandylas and Asplund, 2025). Therefore, our PID-based framework is designed for the specific, high-stakes scenario of a developer, deployer, or regulator who requires verifiable proof of data removal. *In this context, white-box access is not a limitation but a prerequisite for a meaningful and trustworthy audit.*

Limitations

We investigate the residual information with respect to data membership, future works should look into specific sensitive attributes relevant for deployers. **Second**, the limitations of current LLM unlearning benchmarks carry over to our work as well: we only investigate English-language unlearning and our setting is restricted to unlearning finetuned LLMs. However, our conceptual framework should still provide valid audits for future benchmarks since our approach is data-independent and only relies on residual information encoded in the LLM representations. Additionally, the strength of our framework lies in the generalizable nature of our proposed information-theoretic measures. We only require access to two versions of the LLM. In the current study, we use the base and fine tuned model, but the core principles remain valid for other combinations an auditor might be interested in, like unlearning from instruction-tuned models, or unlearning from pre training data. **Third**, unlike exact unlearning, we are unable to provide certified guarantees for the unlearning process. However, we argue that our proposed measures are a practical and interpretable way to investigate unlearning in a more principled manner. **Finally**, our framework currently audits LLM unlearning at the representation level. Our audit can be naturally extended to weights and parameters of the model which we leave as part of future work. For a black-box setting in frontier LLMs, our audit can naturally be extended assuming access only to model logits and log probabilities.

Our current work focuses on establishing a reliable framework to quantify if and how much information persists and a localisation study would be out of scope of our current work. We believe our framework lays the essential groundwork for future

work to perform such fine-grained localization by applying our PID audit layer-by-layer or head-by-head. Our framework also serves as a fundamental diagnostic tool that can guide better algorithms, for example, by minimising the redundant information during unlearning.

On the Use of Information-Theoretic Estimators. Our framework relies on a neural estimator (RINE) to quantify information-theoretic measures from high-dimensional LLM representations. It is important to contextualize this choice. Accurately estimating mutual information in continuous, high-dimensional spaces is a notoriously challenging open problem. Variational estimators like RINE represent the state-of-the-art approach, but they provide a lower bound on the true information quantities.

From an auditing perspective, we argue this lower bound is not a critical weakness but a practical strength. An audit’s primary goal is often risk detection. If our audit reveals a high ‘Residual Knowledge’ score, it provides sufficient evidence to fail the unlearning process, as it proves a substantial amount of information persists. The utility of this approach is further validated by our empirical findings: ‘Residual Knowledge’, even as a bound, correlates far more strongly with real-world adversarial vulnerability than standard MIA baselines (Table 2). While determining the precise tightness of these bounds for LLM representations remains a valuable direction for future theoretical work, our results show that the current estimators are already effective tools for practical privacy auditing.

Ethical Considerations

This work directly engages with ethical and legal challenges surrounding user privacy in LLMs. In particular, we focus on compliance with data protection regulations such as the GDPR and the EU AI Act, which mandate the right to data deletion. Our proposed auditing framework aims to provide verifiable, interpretable evidence of data removal, enabling greater accountability in machine unlearning systems. We emphasize that our methods are intended for responsible use by model developers, auditors, and policymakers to assess and improve privacy guarantees. However, the auditing techniques could potentially be misused to probe for residual information in models not explicitly designed for unlearning. All experiments in this work are conducted on publicly available benchmarks

and synthetic data, without processing personally identifiable or sensitive real-world information. We advocate for future research to uphold similar ethical standards and to include auditing as a core component of privacy-preserving ML system design.

Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

- Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, Toulon, France.
- Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. 2014. Quantifying unique information. *Entropy*, 16(4):2161–2183.
- David Blackwell. 1953. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272.
- Lucas Bourtole, 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, Virtual. IEEE.
- Kwanghee Choi, Jee-weon Jung, and Shinji Watanabe. 2024. Understanding probe behaviors through variational bounds of mutual information. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5655–5659, Seoul, Korea. IEEE.
- Somnath Basu Roy Chowdhury, Krzysztof Choromanski, Arijit Sehanobish, Avinava Dubey, and Snigdha Chaturvedi. 2025. Towards scalable exact machine unlearning using parameter-efficient fine-tuning. In *International Conference on Learning Representations*, Singapore.
- A Feder Cooper, Christopher A Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, et al. 2025. Machine unlearning doesn’t do what you think: Lessons for generative ai policy, research, and practice. In *Advances in neural information processing systems*, San Diego, USA.
- Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

- Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39.
- Shaurya Dewan, Rushikesh Zawar, Prakanshul Saxena, Yingshan Chang, Andrew Luo, and Yonatan Bisk. 2024. Diffusion pid: Interpreting diffusion via partial information decomposition. In *Advances in Neural Information Processing Systems*, volume 37, pages 2045–2079, Vancouver, Canada.
- Pasan Dissanayake, Faisal Hamman, Barproda Halder, Iliia Sucholutsky, Qiuyi Zhang, and Sanghamitra Dutta. 2025. Quantifying knowledge distillation using partial information decomposition. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4474–4482, Mai Khao, Thailand. PMLR.
- Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2025. [UNDIAL: Self-distillation with adjusted logits for robust 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 8827–8840, Albuquerque, New Mexico. Association for Computational Linguistics.
- Michael Duan, Anshuman Suri, Niloofar Miresghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, Philadelphia, PA, USA.
- Sanghamitra Dutta and Faisal Hamman. 2023. A review of partial information decomposition in algorithmic fairness and explainability. *Entropy*, 25(5):795.
- Ronen Eldan and Mark Russinovich. 2023. [Who’s harry potter? approximate unlearning in llms](#). *CoRR*, abs/2310.02238.
- Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. [A comprehensive survey of machine unlearning techniques for large language models](#). *CoRR*, abs/2503.01854.
- Anmol Goel, Yaxi Hu, Iryna Gurevych, and Amartya Sanyal. 2025. Differentially private steering for large language model alignment. In *International Conference on Learning Representations*, Singapore.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.
- Virgil Griffith, Edwin KP Chong, Ryan G James, Christopher J Ellison, and James P Crutchfield. 2014. Intersection information based on common randomness. *Entropy*, 16(4):1985–2000.
- Virgil Griffith and Tracey Ho. 2015. Quantifying redundant information in predicting a target random variable. *Entropy*, 17(7):4644–4653.
- Aaron J Gutknecht, Michael Wibrals, and Abdullah Makkeh. 2021. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A*, 477(2251):20210110.
- Barproda Halder, Faisal Hamman, Pasan Dissanayake, Qiuyi Zhang, Iliia Sucholutsky, and Sanghamitra Dutta. 2025. Towards formalizing spuriousness of biased datasets using partial information decomposition. *Transactions on machine learning research*.
- Dang Huu-Tien, Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. 2025. [On effects of steering latent representation for large language model unlearning](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’25/IAAI’25/EAAI’25*, Philadelphia, Pennsylvania, USA. AAAI Press.
- 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, Toronto, Canada. Association for Computational Linguistics.
- Dongjae Jeon, Wonje Jeung, Taeheon Kim, Albert No, and Jonghyun Choi. 2025. An information theoretic evaluation metric for strong unlearning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40.
- 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. In *Advances in Neural Information Processing Systems*, volume 37, pages 12581–12611, Vancouver, Canada.
- Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. 2024. Wagle: Strategic weight attribution for effective and modular unlearning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 55620–55646, Vancouver, Canada.
- Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu,

- and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 98213–98263, Vancouver, Canada.
- Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. [Copyright violations and large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics.
- Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. In *International Conference on Learning Representations*, Singapore.
- Michael Kleinman, Alessandro Achille, Stefano Soatto, and Jonathan C. Kao. 2021. [Redundant information neural estimation](#). *Entropy*, 23(7).
- Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2023. Towards unbounded machine unlearning. In *Advances in neural information processing systems*, volume 36, pages 1957–1987, New Orleans, Louisiana, USA.
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhurugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28525–28550. PMLR.
- Yongce Li, Chung-En Sun, and Tsui-Wei Weng. 2025. [Effective skill unlearning through intervention and abstention](#). 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 6358–6371, Albuquerque, New Mexico. Association for Computational Linguistics.
- Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. 2023. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pages 27351–27393, New Orleans, Louisiana, USA.
- Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254, Montreal, Canada. PMLR.
- Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024. Large language model unlearning via embedding-corrupted prompts. In *Advances in Neural Information Processing Systems*, volume 37, pages 118198–118266, Vancouver, Canada.
- Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. 2025. An adversarial perspective on machine unlearning for ai safety. *Transactions on Machine Learning Research*.
- Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. [Eight methods to evaluate robust unlearning in llms](#). *CoRR*, abs/2402.16835.
- Tobias Mages, Elli Anastasiadi, and Christian Rohner. 2024. [Non-negative decomposition of multivariate information: From minimum to blackwell-specific information](#). *Entropy*, 26(5).
- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. In *Conference on Language Modeling*, Philadelphia, PA.
- Tarun Ram Menta, Susmit Agrawal, and Chirag Agarwal. 2025. [Analyzing memorization in large language models through the lens of model attribution](#). 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 10661–10689, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? In *International Conference on Learning Representations*, Vienna, Austria.
- Vaidehi Patil, Peter Hase, and Mohit Bansal. 2024. [Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks](#). In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.
- Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. Safety alignment should be

- made more than just a few tokens deep. In *International Conference on Learning Representations*, Singapore.
- Yan Scholten, Stephan Günnemann, and Leo Schwinn. 2025. [A probabilistic perspective on unlearning and alignment for large language models](#). In *The Thirteenth International Conference on Learning Representations*, Singapore.
- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Maladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2025. Muse: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, Singapore.
- Charilaos Skandylas and Mikael Asplund. 2025. Automated penetration testing: Formalization and realization. *Computers & Security*, 155:104454.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. [BERT rediscovers the classical NLP pipeline](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. 2025. [Position: Llm unlearning benchmarks are weak measures of progress](#). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 520–533, Copenhagen, Denmark.
- Pratiksha Thaker, Yash Maurya, and Virginia Smith. 2024. I’m not familiar with the name harry potter: Prompting baselines for unlearning in llms. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- Anvith Thudi, Hengrui Jia, Iliia Shumailov, and Nicolas Papernot. 2022. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX security symposium (USENIX Security 22)*, pages 4007–4022, Boston, Massachusetts, USA.
- Praveen Venkatesh, Corbett Bennett, Sam Gale, Tamina Ramirez, Gregory Heller, Severine Durand, Shawn Olsen, and Stefan Mihalas. 2023. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. In *Advances in Neural Information Processing Systems*, volume 36, pages 74602–74635, New Orleans, Louisiana, USA.
- 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, Albuquerque, New Mexico. Association for Computational Linguistics.
- WeiQi Wang, Zhiyi Tian, An Liu, and Shui Yu. 2025b. [Tape: Tailored posterior difference for auditing of machine unlearning](#). In *Proceedings of the ACM on Web Conference 2025, WWW ’25*, page 3061–3072, New York, NY, USA. Association for Computing Machinery.
- Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. [DEPN: Detecting and editing privacy neurons in pre-trained language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore. Association for Computational Linguistics.
- 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.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. [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.
- Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. [Min-k%++: Improved baseline for detecting pre-training data from large language models](#). In *International Conference on Learning Representations*, Singapore.
- Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024a. [Negative preference optimization: From catastrophic collapse to effective unlearning](#). In *Conference on Language Modeling*, Philadelphia, PA.
- Zhexin Zhang, Junxiao Yang, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. 2024b. [Safe unlearning: A surprisingly effective and generalizable solution to defend against jailbreak attacks](#). *CoRR*, abs/2407.02855.

A Partial Information Decomposition and Redundancy Estimation

Information theory provides powerful tools like mutual information $I(X; Y)$ to quantify the dependency between random variables X and Y (Cover, 1999). However, standard mutual information does not naturally describe how information about a target variable Y is distributed among multiple source variables X_1, \dots, X_n . For instance, given two sources X_1 and X_2 , the mutual information $I(X_1, X_2; Y)$ tells us the total information about Y contained in (X_1, X_2) , but it doesn't reveal how much information is unique to X_1 , unique to X_2 , common to both (redundant), or only available when considering them jointly (synergistic).

Understanding this distribution of information is crucial in various domains, such as designing multi-sensor systems or analyzing neural activity recorded from different brain areas. To address this limitation, Mages et al. (2024) proposed the Partial Information Decomposition (PID), a principled framework for decomposing the total information $I(X_1, \dots, X_n; Y)$ into components reflecting unique, redundant, and synergistic contributions from the sources.

For the case of two sources X_1 and X_2 and a target Y , the total mutual information $I(X_1, X_2; Y)$ is decomposed as:

$$I(X_1, X_2; Y) = I_{\text{uniq}}^{X_1} + I_{\text{uniq}}^{X_2} + I_{\text{syn}}(X_1, X_2; Y) + I_{\cap}(X_1, X_2; Y) \quad (5)$$

where $I_{\text{uniq}}^{X_i}$ represents the unique information that source X_i provides about Y that is not available from the other sources, $I_{\text{syn}}(X_1, X_2; Y)$ is the synergistic information about Y that is only present when X_1 and X_2 are considered jointly, and $I_{\cap}(X_1, X_2; Y)$ is the redundant information about Y that is common to both X_1 and X_2 . These components relate to standard information-theoretic quantities as follows:

$$I(X_1; Y) = I_{\text{uniq}}^{X_1} + I_{\cap}(X_1, X_2; Y) \quad (6)$$

$$I(X_2; Y|X_1) = I_{\text{uniq}}^{X_2} + I_{\text{syn}}(X_1, X_2; Y) \quad (7)$$

(Note that $I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1)$, which can be seen by summing equations 6 and 7).

While the PID framework is appealing, defining and computing its constituents, particularly the redundant information I_{\cap} , has been challenging. Ex-

isting definitions often involve difficult optimization problems that are only feasible for discrete variables over small alphabets (Griffith et al., 2014; Bertschinger et al., 2014; Kleinman et al., 2021; Venkatesh et al., 2023; Gutknecht et al., 2021).

A.1 Existing Notions of Redundant Information

The definition of redundant information $I_{\cap}(X_1, \dots, X_n; Y)$ has been a central focus of research within the PID framework. Despite some disagreement on specific properties, several desirable characteristics are widely accepted for a notion of redundancy:

- **Symmetry:** $I_{\cap}(X_1; \dots; X_n \rightarrow Y)$ is invariant to the permutation of X_1, \dots, X_n .
- **Self-redundancy:** $I_{\cap}(X_i \rightarrow Y) = I(X_i; Y)$ for any source X_i . The redundancy of a single source with itself is simply its mutual information with the target.
- **Monotonicity:** $I_{\cap}(X_1; \dots; X_n \rightarrow Y) \leq I_{\cap}(X_1; \dots; X_{n-1} \rightarrow Y)$. Redundancy cannot increase by adding more sources.

Several notions satisfying these properties have been proposed. Two notable examples, formulated as optimization problems over a random variable Q related to the sources and target, are:

Intersection Information I^{\wedge} Proposed by Griffith et al. (2014), I^{\wedge} is defined as the maximum information that a random variable Q can have about Y , subject to Q being a deterministic function of each source X_i :

$$I^{\wedge}(X_1; \dots; X_n \rightarrow Y) := \max_Q I(Y; Q) \quad (8)$$

s.t. $\forall i, \exists f_i$ (deterministic): $Q = f_i(X_i)$.

This definition captures information about Y that is common to all sources, as it must be extractable from each source independently via a deterministic function.

Common Information I^{GH} A more general notion was proposed by Griffith and Ho (2015), defined as:

$$I^{GH}(X_1; \dots; X_n \rightarrow Y) := \max_Q I(Y; Q) \quad (9)$$

s.t. $\forall i, I(Y; Q|X_i) = 0$.

The constraint $I(Y; Q|X_i) = 0$ means that $Y \leftrightarrow X_i \leftrightarrow Q$ forms a Markov chain for each i . This

is a weaker constraint than $Q = f_i(X_i)$, allowing Q to be a stochastic function of X_i . I^{GH} thus reflects the maximum information between Y and a variable Q that is dependent on each X_i in a way that renders Y conditionally independent of Q given X_i .

A key limitation of these existing definitions, when computed directly via optimization over Q or the joint distribution, is that they are generally only feasible for discrete variables with very small alphabets, making them impractical for high-dimensional continuous data typical in machine learning applications.

A.2 Conceptual Grounding: Blackwell Sufficiency and Residual Knowledge

To justify our proposed measure of *residual knowledge* in an unlearned model regarding a forgotten dataset, we draw on the concept of Blackwell Sufficiency (Blackwell, 1953), a foundational tool for comparing the informational content of statistical experiments (in our case, model representations).

A.3 Variational Redundant Information Neural Estimator (RINE)

To overcome the computational challenges and enable the estimation of redundant information in high-dimensional settings, Kleinman et al. (2021) reformulate the problem as a variational optimization over a restricted family of functions.

For two sources X_1, X_2 and a target Y , RINE is defined by an optimization problem over a chosen model family \mathcal{V} of functions (e.g., neural networks), representing potential decoders f_1, f_2 :

The core of the optimization is to find decoders $f_1, f_2 \in \mathcal{V}$ that are highly similar in their predictions about Y from X_1 and X_2 respectively, while minimizing their prediction error on Y . Let $f_i(y|x_i)$ denote the conditional probability distribution over Y given $X_i = x_i$ as modeled by decoder f_i . RINE defines a loss function $\mathcal{L}_{\mathcal{V}}(X_1; X_2 \rightarrow Y)$ that seeks functions $f_1, f_2 \in \mathcal{V}$ minimizing the average prediction loss (cross-entropy) while being constrained to produce similar predictions:

$$\begin{aligned} \mathcal{L}_{\mathcal{V}}(X_1; X_2 \rightarrow Y) &:= \min_{f_1, f_2 \in \mathcal{V}} \left[\frac{1}{2} H_{f_1}(Y|X_1) \right. \\ &\quad \left. + \frac{1}{2} H_{f_2}(Y|X_2) \right] \\ \text{s.t. } D(f_1, f_2) &= 0 \end{aligned} \quad (10)$$

where $H_f(Y|X) = \mathbb{E}_{(x,y) \sim p(x,y)}[-\log f(y|x)]$ is the expected cross-entropy when predicting Y using decoder f , and $D(f_1, f_2) =$

$\mathbb{E}_{x_1, x_2 \sim p(x_1, x_2)}[\|f_1(Y|x_1) - f_2(Y|x_2)\|_1]$ is the expected difference between the output distributions of the two decoders (e.g., measured by L1 norm, though other divergence measures could be used). The $1/2$ terms are included for symmetry. The constraint $D(f_1, f_2) = 0$ implies that $f_1(Y|X_1)$ and $f_2(Y|X_2)$ are equal in expectation over the data distribution. If f_1 and f_2 are deterministic functions, this constraint means $f_1(X_1) = f_2(X_2)$ almost surely. If they are stochastic functions, it means their output distributions are the same.

The \mathcal{V} -redundant information, denoted $I_{\mathcal{V}}(X_1, X_2; Y)$, is then defined based on this minimized loss:

$$I_{\mathcal{V}}(X_1, X_2; Y) := H(Y) - \mathcal{L}_{\mathcal{V}}(X_1; X_2 \rightarrow Y) \quad (11)$$

where $H(Y)$ is the entropy of the target variable. This definition is analogous to the definition of standard mutual information $I(X; Y) = H(Y) - H(Y|X)$, where $H(Y|X)$ is the minimum possible cross-entropy when predicting Y from X using **any** function. Here, we minimize cross-entropy using functions from \mathcal{V} subject to the similarity constraint, effectively isolating the information about Y that is extractable by similar functions from both sources.

To solve the constrained minimization problem subject to (10) in practice, we can minimize the corresponding Lagrangian:

$$\begin{aligned} \mathcal{L}_{\mathcal{V}}(X_1; X_2 \rightarrow Y, \beta) &:= \min_{f_1, f_2 \in \mathcal{V}} \left[\frac{1}{2} H_{f_1}(Y|X_1) \right. \\ &\quad \left. + \frac{1}{2} H_{f_2}(Y|X_2) \right. \\ &\quad \left. + \beta D(f_1, f_2) \right] \end{aligned} \quad (12)$$

where $\beta \geq 0$ is a Lagrange multiplier. As $\beta \rightarrow \infty$, the solution to the Lagrangian approaches the solution to the constrained problem, forcing $D(f_1, f_2) \rightarrow 0$.

A key insight is that by choosing the function family \mathcal{V} appropriately, RINE generalizes the existing notions of redundancy I^{\wedge} and I^{GH} :

Proposition 1 Let \mathcal{V} be the family of deterministic functions. Then, $I_{\mathcal{V}}(X_1, X_2; Y) = I^{\wedge}(X_1, X_2; Y)$. If, instead, \mathcal{V} is the family of stochastic functions, then $I_{\mathcal{V}}(X_1, X_2; Y) = I^{GH}(X_1, X_2; Y)$.

This proposition is formally proven below. This connection shows that optimizing over a family of functions, as done in RINE, provides a unified

framework that recovers established definitions of redundancy when the function family is sufficiently rich. By using restricted families of functions like neural networks, we can approximate this redundancy in high-dimensional settings where exact computation is intractable. The optimization is performed over the parameters of the functions $f_1, f_2 \in \mathcal{V}$ (e.g., neural network weights) using gradient descent.

A.4 Proof of Proposition 1

We provide an intuitive proof of Proposition 1, showing how the RINE objective recovers known definitions of redundant information depending on the type of functions allowed. We borrow the proofs directly from Kleinman et al. (2021) but provide an intuitive explanation below.

Recall the RINE loss:

$$\mathcal{L}_{\mathcal{V}}(X_1; X_2 \rightarrow Y) := \min_{f_1, f_2 \in \mathcal{V}} \left[\begin{array}{l} 1/2 H_{f_1}(Y|X_1) \\ + 1/2 H_{f_2}(Y|X_2) \end{array} \right] \quad (13)$$

subject to the constraint that the outputs of f_1 and f_2 are identical:

$$D(f_1, f_2) := \mathbb{E}_{x_1, x_2} [\|f_1(Y|x_1) - f_2(Y|x_2)\|_1] = 0 \quad (14)$$

This means both functions must output the same predictive distribution over Y , even though they receive different inputs (X_1 and X_2 respectively). Let this shared output be a variable Q .

Then the loss becomes:

$$\mathcal{L}_{\mathcal{V}}(X_1; X_2 \rightarrow Y) = H(Y|Q) \quad (15)$$

and the estimated redundancy is:

$$I_{\mathcal{V}}(X_1, X_2; Y) = H(Y) - H(Y|Q) = I(Y; Q) \quad (16)$$

So the amount of redundant information depends on how much Q tells us about Y , where Q must be derivable from both X_1 and X_2 through functions in \mathcal{V} .

Case 1: Deterministic Functions. Suppose \mathcal{V} is the set of all deterministic functions. Then Q must be a deterministic function of each input:

$$Q = f_1(X_1) = f_2(X_2)$$

This matches the definition of I^{\wedge} :

$$I^{\wedge}(X_1, X_2 \rightarrow Y) = \max_{Q: Q=f_i(X_i)} I(Y; Q) \quad (17)$$

Therefore, under deterministic \mathcal{V} , the RINE estimate recovers I^{\wedge} .

Case 2: Stochastic Functions. Now let \mathcal{V} be the set of stochastic functions. Then Q can be a stochastic function of X_1 and X_2 , and the constraint $D(f_1, f_2) = 0$ still requires that both X_1 and X_2 can generate the same distribution over Q .

This corresponds to the constraint:

$$I(Y; Q|X_1) = 0 \quad \text{and} \quad I(Y; Q|X_2) = 0$$

which means $Y \leftrightarrow X_i \leftrightarrow Q$ forms a Markov chain for each i . That is the definition of I^{GH} :

$$I^{GH}(X_1, X_2 \rightarrow Y) = \max_{Q: I(Y; Q|X_i)=0} I(Y; Q) \quad (18)$$

So when \mathcal{V} allows stochastic mappings, RINE recovers I^{GH} .

Conclusion. RINE unifies both I^{\wedge} and I^{GH} under a single optimization framework. By choosing different function families \mathcal{V} (deterministic or stochastic), the method recovers different classical definitions of redundant information.

B Inference-Time Abstention via Risk Scoring

In this section, we provide a detailed explanation of our inference-time abstention mechanism designed to mitigate residual privacy risks in unlearned models. Even after training-time unlearning procedures are applied, models may retain partial information from forgotten samples. Our approach provides an additional safety barrier by abstaining from answering queries that are likely to belong to the forget set and are insufficiently unlearned.

B.1 Motivation

Two natural candidates for an inference-time abstention score are the *redundant information* or *disagreement* between the base and unlearned models to identify forgotten examples. However, these methods have key limitations:

- **Redundancy-based scoring:** Redundancy metrics, such as $I_{\cap}^{\mathcal{V}}(X_1; X_2 \rightarrow Y)$, require access to ground-truth membership labels Y for every test sample, which are unavailable in real-world inference settings.
- **Disagreement-based scoring:** Simple disagreement metrics like $|f_1(y|x_1) - f_2(y|x_2)|$ may misclassify successfully forgotten samples as risky (due to decoder divergence) and

benign retained samples as safe (due to decoder agreement), thus leading to high false positives and negatives.

To address these issues, we propose a more targeted risk scoring mechanism that accounts for both the confidence and agreement of decoders specifically on the “forget” class.

B.2 Proposed Risk Score and Abstention Algorithm

Let f_1 and f_2 be binary membership decoders trained on representations x_1 and x_2 from the base and unlearned models, respectively. Let:

$$p_1 = f_1(\text{forget}|x_1), \quad p_2 = f_2(\text{forget}|x_2).$$

We define the risk score as:

$$\text{RiskScore}(x) = \underbrace{\frac{1}{2}(p_1 + p_2)}_{\text{Mean Forget Probability}} \cdot \underbrace{(1 - |p_1 - p_2|)}_{\text{Agreement Term}}.$$

This risk score is high when both decoders assign high probability to the sample being from the forget set *and* they closely agree on this assessment, suggesting residual memorization. ⁶

Algorithm 1 Inference-Time Abstention via Risk Scoring

- 1: Trained decoders f_1, f_2 , representations x_1, x_2 of input x , abstention threshold τ
 - 2: $p_1 \leftarrow f_1(\text{forget}|x_1)$
 - 3: $p_2 \leftarrow f_2(\text{forget}|x_2)$
 - 4: $\text{mean_forget} \leftarrow 0.5 \cdot (p_1 + p_2)$
 - 5: $\text{agreement} \leftarrow 1 - |p_1 - p_2|$
 - 6: $\text{risk_score} \leftarrow \text{mean_forget} \cdot \text{agreement}$
 - 7: **if** $\text{risk_score} > \tau$ **then**
 - 8: **Abstain**
 - 9: **else**
 - 10: **Answer**
-

B.3 Discussion

The final abstention score balances two intuitions:

- **High average forget probability** suggests that the sample looks like a forgotten point.
- **Low disagreement** indicates the unlearned model still agrees with the base model, implying forgetting may have failed.

⁶We evaluate the performance on different thresholds and observe best performance at 0.48. This is a parameter deployers can optimize depending on their privacy-compliance.

Table 6 provides a conceptual example to clarify the behavior of the proposed risk score across different possible types of samples. Only when both of these hold do we trigger abstention, which makes the mechanism robust against false positives on benign queries and false negatives on unsuccessfully forgotten ones. This strategy also avoids requiring access to ground-truth labels or expensive PID optimization at inference time.

C Scalability Experiments with Smaller Models

To verify that our auditing framework is effective across different model scales, we conducted additional experiments on smaller language models: llama-3-1B-Instruct and gemma-2B. The primary goal was to assess whether the trends observed in the 7B-9B parameter models hold true for smaller, more accessible model sizes.

As shown in Table 7, the results are highly consistent with our main findings. The relative performance ranking of the unlearning algorithms remains largely stable. For instance, RMU continues to be one of the most effective methods, achieving the lowest ‘Residual Knowledge’ and highest ‘Unlearned Knowledge’. Conversely, simpler methods like Gradient Ascent (GA) still leave behind significant residual information, even in these smaller models.

This consistency across scales (from 1B to 9B parameters) suggests that the phenomena of “shallow unlearning” and the utility of our PID-based audit are fundamental properties of the unlearning process in current LLMs, rather than artifacts of a specific model size.

D Experimental Details

In all experiments, we use the open-unlearning library for access to the unlearning algorithms and benchmarks. We train simple logistic regression models for probing experiments. For redundancy computation, we train linear probes on base and unlearned model representations from the final layer, using Adam optimization in PyTorch. All experiments are conducted on NVIDIA A100 80GB GPU. For the adversarial attacks, we reuse the codebase from Łucki et al. (2025) and compute ASRs on each model family (all models upto 7B parameters) and dataset combination, then run the regression reported in Section 6.1. Our framework can be applied to any unlearning algorithm and model as

Table 6: Example of risk score behavior for different sample types.

Sample Type	p_1	p_2	Mean Forget	Disagreement	Risk Score	Action
Retain Sample	0.1	0.1	0.1	0.0	0.10	Answer
Forget (Unlearned)	0.9	0.1	0.5	0.8	0.10	Answer
Forget (Residual)	0.9	0.8	0.85	0.1	0.77	Abstain

Table 7: Unlearning performance audit on smaller models (llama-3-1B and gemma-2B) using the TOFU dataset. Results show consistent trends with the larger models presented in the main paper (Table 1), demonstrating the scalability of our auditing framework.

Model	Metric	GA	GD	NPO	SimNPO	RMU
Llama-3 1B	Forget Quality (\uparrow)	0.48	0.55	0.61	0.65	0.68
	Unlearned Knowledge (\uparrow)	0.31	0.36	0.51	0.62	0.80
	Residual Knowledge (\downarrow)	0.29	0.24	0.19	0.18	0.08
Gemma 2B	Forget Quality (\uparrow)	0.45	0.53	0.68	0.70	0.59
	Unlearned Knowledge (\uparrow)	0.20	0.29	0.48	0.64	0.80
	Residual Knowledge (\downarrow)	0.33	0.29	0.18	0.15	0.09

long as we have white-box access.

The threshold for abstention-based experiments was determined by performing a sweep on a held-out validation set to balance forget quality improvement and utility preservation. This can be adapted by deployers based on their specific privacy-utility trade-offs.

Model	Metric	GA	GD	NPO	SimNPO	RMU
llama	Forget Quality (\uparrow)	0.53	0.60	0.70	0.74	0.71
	Unlearned Knowledge (\uparrow)	0.24	0.34	0.47	0.65	0.80
	Residual Knowledge (\downarrow)	0.42	0.33	0.27	0.15	0.09
gemma	Forget Quality (\uparrow)	0.44	0.50	0.71	0.74	0.57
	Unlearned Knowledge (\uparrow)	0.19	0.29	0.45	0.60	0.76
	Residual Knowledge (\downarrow)	0.40	0.29	0.24	0.12	0.06
qwen	Forget Quality (\uparrow)	0.48	0.57	0.63	0.66	0.73
	Unlearned Knowledge (\uparrow)	0.26	0.38	0.52	0.68	0.83
	Residual Knowledge (\downarrow)	0.44	0.35	0.29	0.17	0.08

Table 8: Audit results on the MUSE dataset.