
Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs

Xiaoyu Xu

The Hong Kong Polytechnic University
xiaoyu0910.xu@connect.polyu.hk

Xiang Yue

Carnegie Mellon University
xyue2@andrew.cmu.edu

Yang Liu

University of California, Santa Cruz
yangliu@ucsc.edu

Qingqing Ye

The Hong Kong Polytechnic University
qqing.ye@polyu.edu.hk

Haibo Hu

The Hong Kong Polytechnic University
haibo.hu@polyu.edu.hk

Minxin Du

The Hong Kong Polytechnic University
minxin.du@polyu.edu.hk

Abstract

Unlearning in large language models (LLMs) is intended to remove the influence of specific data, yet current evaluations rely heavily on token-level metrics such as accuracy and perplexity. We show that these metrics can be misleading: models often appear to forget, but their original behavior can be rapidly restored with minimal fine-tuning, revealing that unlearning may obscure information rather than erase it. To diagnose this phenomenon, we introduce a representation-level evaluation framework using PCA-based similarity and shift, centered kernel alignment, and Fisher information. Applying this toolkit across six unlearning methods, three domains (text, code, math), and two open-source LLMs, we uncover a critical distinction between *reversible* and *irreversible* forgetting. In reversible cases, models suffer token-level collapse yet retain latent features; in irreversible cases, deeper representational damage occurs. We further provide a theoretical account linking shallow weight perturbations near output layers to misleading unlearning signals, and show that reversibility is modulated by task type and hyperparameters. Our findings reveal a fundamental gap in current evaluation practices and establish a new diagnostic foundation for trustworthy unlearning in LLMs. We provide a unified toolkit for analyzing LLM representation changes under unlearning and relearning: https://github.com/XiaoyuXU1/Representational_Analysis_Tools.git.

1 Introduction

Large language models (LLMs), trained on massive corpora, have achieved remarkable success across diverse tasks, yet their capacity to memorize training snippets poses acute ethical, legal, and security risks. Memorization can unintentionally disclose sensitive, harmful, or copyrighted text [25; 13; 30], conflicting with emerging regulations, such as the EU's *Right to be Forgotten* [9].

Machine unlearning seeks to mitigate this threat by making a model act as though specified data were never seen [2]. Numerous methods have been proposed for LLMs [33; 12; 6; 26; 18; 17; 19; 20; 31], their success usually judged by token-level metrics, such as accuracy or perplexity.

However, a pivotal question remains largely unexplored: *Does LLM unlearning truly erase information, or merely suppress it, poised to “resurface” at the slightest nudge?*

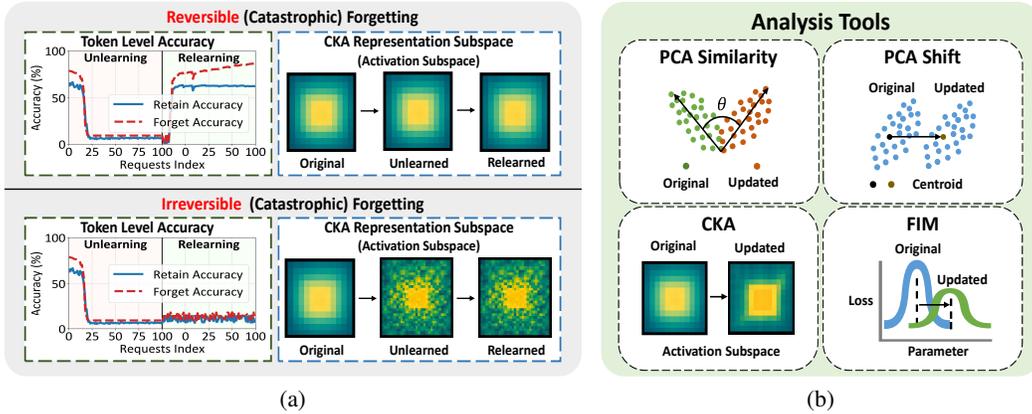


Figure 1: (a) Token-level accuracy and CKA subspaces of **reversible** (top) vs. **irreversible** (bottom) catastrophic forgetting due to *continual unlearning* then *relearning*, (b) Our four diagnostic tools

Empirically, many unlearning methods may just *appear* effective: after unlearning, a model shows near-zero accuracy or high perplexity on the *forget set*, but a brief fine-tuning step (even on unrelated data) can quickly restore its original behaviour (see Figure 1). This exposes a huge gap between surface-level metrics and the model’s internal state [21; 23], casting doubt on compliance and safety claims. If information is recoverable via simple fine-tuning (or relearning), can we truly claim that it has been “forgotten?” What looks like memory loss may, in fact, be a shallow perturbation.

In this work, we perform the first systematic analysis of **reversibility of LLM unlearning**, covering both *single-shot* and *continual* settings. The continual setting allows multiple unlearning requests over time, which we believe is a more prominent and realistic setting where the deployment of such unlearning strategies are often facing a dynamic environment. We show that standard token-level metrics prove insufficient—they can collapse even when the underlying representations remain intact. To probe deeper, we introduce a diagnostic toolkit for representational analysis, featuring PCA subspace similarity and shift [36], centered kernel alignment (CKA) [15], and Fisher information [3]. Our toolkit uncovers two distinct regimes of unlearning:

1) *Reversible (catastrophic) forgetting*: performance collapses, but feature subspaces are largely preserved, enabling rapid recovery, and 2) *Irreversible (catastrophic) forgetting*: collapse coincides with substantial representational drift, making recovery difficult/impossible. Surprisingly, both yield similar results under token-level metrics, underscoring the need for deeper representational analysis.

We further show that the transition between reversible and irreversible forgetting depends not only on the volume of unlearning requests but also on hyperparameters such as learning rate. Modest weight perturbations—especially near the output layer—can lead to token-level distortions without altering feature geometry, making “forgotten” knowledge easily recoverable.

Hence, evaluating the effectiveness of LLM unlearning must go beyond superficial token-level metrics (*e.g.*, forget accuracy declines). In safety- and privacy-critical settings, unlearning should be judged by its ability to achieve genuine erasure rather than simply representational collapse.

Contributions. We summarize our main contributions as follows:

- We present the *first* systematic study of *reversibility* in both *single* and *continual* LLM unlearning, using a feature-space toolkit, including PCA similarity, PCA shift, CKA, and Fisher information. Our analysis distinguishes **reversible** from **irreversible** (catastrophic) forgetting.
- We conduct extensive experiments with six unlearning methods (GA [33], NPO [35], and RLabel, with their variants) across three datasets (arXiv papers, GitHub code [33], and NuminaMath-1.5 [16]) on Yi-6B [34] and Qwen-2.5-7B [32]. Our results show that standard token-level metrics (*e.g.*, accuracy, perplexity, MIA susceptibility [27]) fail to capture true forgetting behavior.
- We theoretically analyze weight perturbations to explain how widespread vs. localized parameter changes relate to (ir)reversible forgetting. Small perturbations near the logits can distort token-level metrics despite intact features, hence leading to misleading assessments.

- Based on our findings and extra preliminary results, we propose several future directions, including using unlearning as a complementary form of data augmentation, and designing more robust unlearning algorithms that achieve genuine forgetting while avoiding representational collapse.

2 Preliminaries

LLM unlearning seeks to enhance privacy, improve safety, and reduce bias [33; 12; 26; 18; 17; 19]. Most work adopts the *single-unlearning* paradigm: given a training corpus \mathcal{D} and a designated *forget set* $\mathcal{D}_f \subseteq \mathcal{D}$, a model \mathcal{M} is first trained on \mathcal{D} with algorithm \mathcal{A} . An unlearning procedure \mathcal{U} then transforms \mathcal{M} into an *unlearned* model \mathcal{M}_f that should behave as if it had never encountered \mathcal{D}_f .

Ideally, \mathcal{U} should produce a model statistically indistinguishable from one retrained on the *retain set* $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$: $\mathcal{M}_f = \mathcal{U}(\mathcal{M}, \mathcal{D}_f) \approx \mathcal{M}_r = \mathcal{A}(\mathcal{M}, \mathcal{D}_r)$. While current methods strike a good forget-utility balance in controlled settings [1; 8] (e.g., a fixed \mathcal{D}_f or just a single removal request), they seldom address the practical need for *continual unlearning*, i.e., data owners can submit removal requests sequentially over time [1]. Let the successive forget sets be $\mathcal{D}_f^{(1)}, \mathcal{D}_f^{(2)}, \dots, \mathcal{D}_f^{(t)}$ (whose union is \mathcal{D}_f); the retain set after t rounds is $\mathcal{D}_r^{(t)}$. The model is then updated recursively: $\mathcal{M}_f^{(t+1)} = \mathcal{U}(\mathcal{M}_f^{(t)}, \mathcal{D}_f^{(t+1)})$, which should be similar to $\mathcal{M}_r = \mathcal{A}(\mathcal{M}, \mathcal{D}_r^{(t+1)})$ at any time t .

Retraining LLMs is prohibitively costly, so most studies rely on empirical proxies rather than formal statistically-indistinguishable guarantees [6; 24; 18; 7]. Evaluations track *forget quality* on the forget set and *utility* on the retain set, aiming to preserve both [24]. While single unlearning often yields modest declines, it is fragile: brief fine-tuning—even on benign, unrelated data—swiftly revives the “forgotten” knowledge [1; 23; 21]. The issue worsens under *continual* unlearning, where each round begins from an already degraded model, ultimately triggering catastrophic forgetting—a wholesale performance collapse [1; 28]. Prior work notes this risk but does not examine its root causes.

We hypothesize that collapse does not imply true erasure; the knowledge may remain latent in the feature space. This insight leads us to distinguish two regimes of (catastrophic) forgetting:

Definition 1 (Reversible (Catastrophic) Forgetting). Let θ_0 denote the initial model parameters, \mathcal{D}_f the forget set, and \mathcal{T} an evaluation task with metric $E(\cdot, \mathcal{T})$. Unlearning \mathcal{D}_f transforms the model to θ_u . If subsequent *relearning* on \mathcal{D}_f (or an equivalent reconstruction set) produces parameters θ_r s.t.

$$E(\theta_u, \mathcal{T}) \ll E(\theta_r, \mathcal{T}) \approx E(\theta_0, \mathcal{T}),$$

the temporary performance collapse is fully reversible; we call it *reversible catastrophic forgetting*. When the initial degradation is modest (e.g., single unlearning), we simply call it *reversible forgetting*.

Definition 2 (Irreversible (Catastrophic) Forgetting). Using the notations of Definition 1, if

$$E(\theta_u, \mathcal{T}) \approx E(\theta_r, \mathcal{T}) \ll E(\theta_0, \mathcal{T}),$$

we observe *irreversible catastrophic forgetting*: the collapse (i.e., weight perturbation) is irreversible. We refine this to *irreversible forgetting*, which further requires that the irreversible degradation be *restricted to the forget set*; performance on the retain set and on unrelated data must remain near their original levels. This condition distinguishes targeted erasure from global model failure or collapse.

To distinguish our setting from a full retrain model, where reversibility is trivially achievable, we introduce the following restriction on the relearning phase.

Relearning Restriction. After unlearning, we briefly fine-tune θ_u on a small relearning set—either the cumulative forget set $\mathcal{D}_f = \bigcup_t \mathcal{D}_f^{(t)}$, its a distribution-similar retain set $\mathcal{D}_r^{(t)}$, or an unrelated out-of-distribution corpus—to obtain θ_r without ever revisiting the full pre-training data.

3 Token-Level Evaluation

3.1 Experiment setup

Models and Datasets. We conduct experiments on two open-source models, Yi-6B [34] and Qwen-2.5-7B [32]. To ensure the generality of our findings, we use two dataset types: i) *simple tasks*—arXiv

abstracts and GitHub code from [33], and ii) a *complex* task—NuminaMath-1.5, a recent benchmark for mathematical reasoning [16]. All experiments were run on NVIDIA H100 GPUs.

Unlearning algorithms We compare six canonical methods grouped into three families.

1) Gradient-Ascent (GA) family. The unified goal is $\mathcal{L} = \mathcal{L}_{\text{forget}}(\mathcal{D}_f) + \lambda \mathcal{L}_{\text{retain}}(\mathcal{D}_r)$, where $\mathcal{L}_{\text{forget}}$ maximizes the loss on the forget set via GA, $\mathcal{L}_{\text{retain}}$ (optional) preserves utility on the retain set, and $\lambda > 0$ balances the two. Choices for $\mathcal{L}_{\text{retain}}$ give three variants: i) GA ($\mathcal{L}_{\text{retain}} = 0$), ii) GA+GD (standard cross-entropy on \mathcal{D}_r), and iii) GA+KL (KL divergence to the reference model on \mathcal{D}_r) [33].

2) Negative Preference Optimization (NPO) family. GA is replaced by an NPO loss that penalizes agreement with the forget set [35]: $\mathcal{L} = \mathcal{L}_{\text{NPO}}(\mathcal{D}_f) + \lambda \mathcal{L}_{\text{retain}}(\mathcal{D}_r)$. Variants mirror those above: NPO ($\mathcal{L}_{\text{retain}} = 0$) and NPO+KL (retain-set KL regularization).

3) Random Label (RLabel). To mimic a model that never saw \mathcal{D}_f , true labels are replaced with random ones: $\mathcal{L} = \mathcal{L}_{\text{RLabel}}(\mathcal{D}_f)$, inducing near-uniform predictions without GA/negative rewards [33].

Unlearning Scenario We consider two standard settings: i) **Single unlearning**: A trained model \mathcal{M} receives exactly one request to remove $\mathcal{D}_f \subset \mathcal{D}$, producing $\mathcal{M}_f = \mathcal{U}(\mathcal{M}, \mathcal{D}_f)$, and ii) **Continual unlearning**: The model processes a stream of requests $\mathcal{D}_f^{(1)}, \dots, \mathcal{D}_f^{(k)}$, updated iteratively by $\mathcal{M}^{(i+1)} = \mathcal{U}(\mathcal{M}^{(i)}, \mathcal{D}_f^{(i+1)})$ with $\mathcal{M}^{(0)} = \mathcal{M}$ and $\bigcup_{i=1}^n \mathcal{D}_f^{(i)} = \mathcal{D}_f$. This mirrors real-world, incremental removal demands while maintaining parity with the single-step budget.

For the *simple* tasks, we benchmark all six algorithms: GA, GA + GD, GA + KL, RLabel, NPO, and NPO+KL. For the *complex* one, lacking a clearly defined retain set, we use GA, NPO, and RLabel.

Evaluation Metrics For *single* unlearning (simple tasks only), we report: forget-set accuracy (F.Acc), retain-set accuracy (R.Acc), and privacy leakage via min- k %-prob MIA AUC [27].

Continual unlearning is evaluated on both task suites. For the simple suite, we report: F.Acc / R.Acc, F.Ppl / R.Ppl, downstream robustness on CommonsenseQA and GSM8K0-shot [29; 4], and the same MIA AUC, thus capturing utility, robustness, and privacy across the unlearning trajectory. For the complex task, we adopt MATH0-shot [10] and GSM8K_{0-shot} as primary benchmarks.

Relearning setting. To gauge how readily forgotten knowledge re-emerges, each unlearning run is followed by a controlled *relearning* phase. **Single unlearning**: We fine-tune on the whole forget set \mathcal{D}_f once, producing a single-step relearned model. **Continual unlearning**: For settings that trigger catastrophic collapse, we fine-tune with three cases: i) the cumulative forget set $\bigcup_i \mathcal{D}_f^{(i)}$, ii) the corresponding retain set $\mathcal{D}_r^{(t)}$, and iii) an unrelated auxiliary corpus. These progressively relax assumptions about access to the forgotten content, revealing the recovery potential in each case.

Hyperparameter Configuration. To comprehensively evaluate the effects of unlearning, we design multiple hyperparameter configurations that vary both the learning rate and the number of unlearning requests. For single unlearning we sweep the learning rate over $LR \in \{3, 4, 5\} \times 10^{-6}$ while fixing the request count to $N = 1$. For continual unlearning we vary both knobs: on the simple task (Yi-6B) we test $LR \in \{3, 5\} \times 10^{-6} \cup \{3 \times 10^{-5}\}$ with $N \in \{6 \rightarrow 100\}$; on the complex task (Qwen-2.5-7B) we use $LR \in \{3, 5\} \times 10^{-6}$ and 3×10^{-5} together with $N \in \{6 \rightarrow 100\}$. All runs adopt the optimizer settings of [1]: AdamW [22] ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\varepsilon = 10^{-8}$), a cosine schedule with 10% warm-up followed by decay to 10% of peak, weight decay 0.1, and gradient clipping at 1.0.

3.2 Token-Level Evaluation Results

We report quantitative results on both single unlearning and continual unlearning settings using Yi-6B and Qwen-2.5-7B across multiple configurations (Tables 1–3). For completeness, detailed experimental results and additional settings are provided in Appendix A.4.

Single Unlearning. For Yi-6B under single unlearning, Table 1 shows that all six unlearning methods reduce MIA and F.Acc, indicating some level of unlearning. The changes on retain set are modest: R.Acc only drops 2–5% for most methods, and MIA decreases by less than 30 points in most cases. Importantly, the relearned models often recover original performance—both GA+KL and RLabel restore R.Acc to near 65.0%, and F.Acc rebounds above 77%. These results suggest that under single unlearning, most methods achieve seemingly successful forgetting at the token level, but as we show

Table 1: Yi-6B: MIA / F.Acc / R.Acc (%) simple task using three LRs under single unlearning

Phase	Method	LR=3×10 ⁻⁶			LR=4×10 ⁻⁶			LR=5×10 ⁻⁶		
		MIA	F.Acc	R.Acc	MIA	F.Acc	R.Acc	MIA	F.Acc	R.Acc
Original	—	70.9	78.9	65.5	70.9	78.9	65.5	70.9	78.9	65.5
Unlearn	GA	45.5	65.4	54.0	43.8	62.4	52.3	41.2	60.3	50.9
	GA+GD	65.4	75.1	64.6	58.2	73.8	65.8	55.3	68.5	63.5
	GA+KL	48.9	71.0	58.5	47.6	70.6	58.1	44.8	68.4	55.4
	NPO	67.2	76.2	64.7	65.2	75.8	62.8	62.2	75.2	62.7
	NPO+KL	66.5	76.3	64.8	67.2	76.4	63.2	64.5	75.6	61.2
	RLabel	69.6	77.7	64.7	69.2	76.5	64.5	68.7	75.4	63.3
Relearn	GA	67.2	76.6	65.2	68.6	77.6	62.8	67.6	76.9	65.5
	GA+GD	68.6	77.0	65.3	68.8	76.9	65.3	68.8	77.2	65.3
	GA+KL	67.9	77.6	65.3	68.3	75.5	65.2	67.7	77.2	65.2
	NPO	68.2	77.1	65.3	68.2	77.2	65.2	68.3	77.0	65.1
	NPO+KL	68.9	77.1	65.3	67.9	76.3	63.0	68.6	76.9	65.2
	RLabel	68.3	78.8	65.6	68.9	76.4	65.3	68.8	78.9	65.2

in Section 4.2 to Section 4.2, the underlying representation changes are minimal—indicating the phenomenon of *reversible forgetting*.

Continual Unlearning. By examining post-relearning recoverability in Table 2 and Table 3, we identify two distinct forms of catastrophic forgetting. When the model regains both utility (e.g., F.Acc, R.Acc) and privacy (e.g., MIA AUC) to levels near or exceeding the original after relearning, we classify the behavior as *reversible catastrophic forgetting*. This suggests that the underlying representational structure remains intact, enabling efficient recovery via lightweight retraining. Such reversibility is consistently observed in methods like NPO and NPO+KL, particularly under low learning rates or small removal batches.

Conversely, when relearning fails to restore utility—reflected in persistently low F.Acc and R.Acc despite partial MIA recovery—we categorize it as *irreversible catastrophic forgetting*. This scenario frequently arises with methods like GA and RLabel under aggressive hyperparameters (e.g., LR = 3 × 10⁻⁵), where damage accumulates across layers and results in irreversible representational collapse. Importantly, MIA AUC alone can be misleading. Models may exhibit near-complete privacy recovery while remaining functionally impaired.

Table 2: Yi-6B: MIA / F.Acc / R.Acc (%) for simple task under four unlearning settings

Phase	Method	LR=3×10 ⁻⁵ , N=100			LR=5×10 ⁻⁶ , N=100			LR=3×10 ⁻⁶ , N=100			LR=3×10 ⁻⁵ , N=6		
		MIA	F.Acc	R.Acc	MIA	F.Acc	R.Acc	MIA	F.Acc	R.Acc	MIA	F.Acc	R.Acc
Original	—	70.8	78.9	65.5	70.8	78.9	65.5	70.8	78.9	65.5	70.8	78.9	65.5
Unlearn	GA	26.1	0.0	0.0	23.2	9.1	6.2	25.2	16.8	14.4	29.6	36.3	36.1
	GA+GD	16.8	9.7	2.3	28.7	3.6	3.1	69.4	78.8	65.5	66.9	77.0	64.0
	GA+KL	17.8	9.0	6.2	27.3	9.1	6.2	18.9	3.8	3.2	29.5	52.9	41.5
	NPO	60.1	37.8	37.9	50.6	51.0	52.3	68.4	78.3	64.1	68.7	71.6	59.4
	NPO+KL	59.0	64.3	55.9	65.4	77.6	64.3	66.7	78.8	65.5	67.9	67.6	56.1
	RLabel	65.1	0.0	0.0	63.6	0.1	0.4	61.4	0.4	0.7	62.7	72.7	61.1
Relearn	GA	74.5	2.1	1.8	68.0	80.0	65.0	68.6	80.8	65.2	68.2	70.5	58.7
	GA+GD	68.1	2.2	2.6	69.8	81.2	65.1	70.0	81.8	65.5	67.0	61.6	54.4
	GA+KL	70.7	1.7	1.6	68.3	81.1	64.8	70.7	81.0	63.2	65.0	66.6	56.2
	NPO	70.0	57.0	45.6	68.0	82.7	65.5	69.9	81.2	65.4	68.4	71.2	59.4
	NPO+KL	67.7	60.7	54.2	69.5	83.8	65.6	69.9	83.8	65.4	69.0	67.6	56.1
	RLabel	69.5	4.3	2.8	70.4	80.8	65.3	70.0	80.5	65.3	65.2	72.7	61.1

4 A Unified Representational Analysis

4.1 Representational Analysis Tools

We monitor representational drift with four layer-wise diagnostics: *PCA Similarity*, *PCA Shift*, *CKA*, and the diagonal *Fisher Information Matrix* (FIM), summarized in Figure 1(b); implementation details (including their precise definitions) are deferred to Appendix A.3.

Table 3: Qwen-2.5-7B: MIA / MATH / GSM8K Accuracy (%) for complex task under four settings

Phase	Method	LR= 3×10^{-5} , N=6			LR= 3×10^{-6} , N=6			LR= 5×10^{-6} , N=6			LR= 5×10^{-6} , N=100		
		MIA	MATH	GSM8K	MIA	MATH	GSM8K	MIA	MATH	GSM8K	MIA	MATH	GSM8K
Original	—	99.3	9.0	80.1	99.3	9.0	80.1	99.3	9.0	80.1	99.3	9.0	80.1
Unlearn	GA	5.9	0.0	0.0	0.9	0.0	0.0	3.8	0.0	0.0	5.5	0.0	0.0
	NPO	95.9	0.0	0.2	97.4	21.5	74.1	67.4	24.1	71.8	94.7	0.0	0.4
	RLabel	35.5	0.0	0.0	69.6	0.0	1.5	11.2	0.0	0.0	2.9	0.0	0.0
Relearn	GA	97.6	0.0	1.1	99.3	5.1	83.2	99.4	9.3	77.8	99.2	0.0	0.0
	NPO	95.8	0.0	0.0	99.4	4.7	82.6	99.4	16.5	75.7	99.2	0.0	0.0
	RLabel	99.5	0.0	0.0	99.3	5.3	83.3	99.3	10.0	77.2	99.6	0.0	0.0

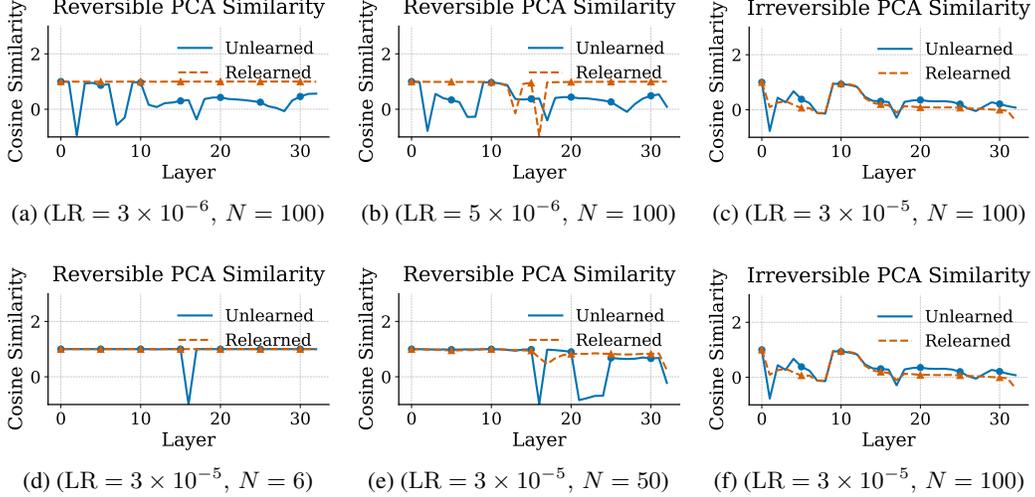


Figure 2: Layer-wise PCA Similarity for GA on Yi-6B (simple task). (a–c) vary LR $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ at $N = 100$; (d–f) vary $N \in \{6, 50, 100\}$ at LR $= 3 \times 10^{-5}$. Similarity near 1 signals *reversible (catastrophic) forgetting*; sustained low similarity signals *irreversible (catastrophic) forgetting*.

PCA Similarity & Shift. For each layer i , we collect activations $\mathbf{H}_i^{\text{orig}}$, $\mathbf{H}_i^{\text{unl}}$, and $\mathbf{H}_i^{\text{rel}}$ on a probe set \mathcal{X} . Let $\mathbf{c}_{i,1}^{(*)}$ and $p_{i,1}^{(*)}$ be the first principal direction and its mean projection for state $(*) \in \{\text{orig}, \text{unl}, \text{rel}\}$. The cosine between $\mathbf{c}_{i,1}^{\text{orig}}$ and $\mathbf{c}_{i,1}^{(*)}$ yields PCA similarity; the signed difference $p_{i,1}^{(*)} - p_{i,1}^{\text{orig}}$ gives PCA shift. Small angles and shifts indicate stable features; otherwise, catastrophic forgetting [36].

Centered Kernel Alignment (CKA). With centered activation matrices X_i^{orig} and $X_i^{(*)}$, we compute $\text{CKA}(X_i^{\text{orig}}, X_i^{(*)}) \in [0, 1]$; values ≈ 1 mean nearly identical subspaces, those ≈ 0 are orthogonal.

Fisher information. We estimate the diagonal empirical FIM by averaging squared gradients over \mathcal{X} . Comparing FIM^{orig} , FIM^{unl} , and FIM^{rel} reveals how unlearning flattens the loss landscape and whether relearning restores parameter importance [14; 11].

We compute all diagnostics not only on the forget set but also on the retain set and on unrelated data, allowing us to distinguish targeted forgetting from broader representational disruption.

4.2 Representational Results

Principal Component Analysis: Similarity and Shift. Figures 2 and 3 show that, in continual unlearning, larger learning rates or more removal requests drive PCA Similarity sharply downward and leave large, unrecovered PCA Shifts—clear evidence of *irreversible catastrophic forgetting*. Under gentler hyper-parameters, similarity stays high and shifts remain bounded; relearning then realigns both metrics, signalling *reversible catastrophic forgetting*. Remarkably, the choice of relearning data and analyzed data—forget, retain, or even unrelated—makes little difference: all three restore the feature geometry, implying the knowledge was suppressed rather than erased (Figure 10 and 14).

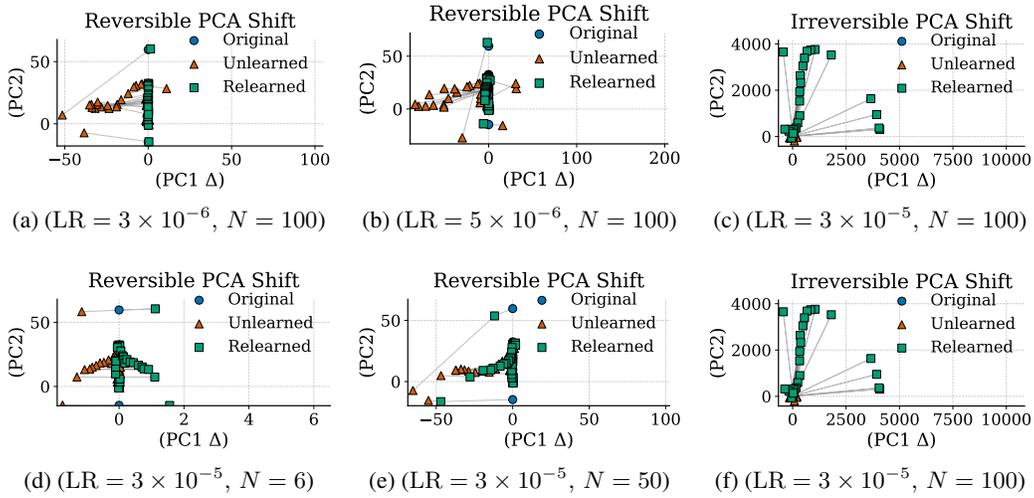


Figure 3: PCA Shift for GA on Yi-6B, simple task. (a–c) LR $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ with $N = 100$; (d–f) LR = 3×10^{-5} with $N \in \{6, 50, 100\}$. Shift magnitude reflects feature displacement: large, unrecovered shifts indicate severe, *irreversible (catastrophic)* forgetting, while small or fully recovered shifts indicate mild, *reversible (catastrophic)* forgetting.

Due to space limitations, complete figures, single-unlearning results, and additional methods appear in Appendix A.5 and Appendix A.6.

Centered Kernel Alignment Analysis. Figure 4 tracks layer-wise CKA. With gentle unlearning, CKA stays near 1 and relearning restores it completely—typical of *reversible catastrophic forgetting*. Stronger updates or many requests push deep-layer CKA sharply lower, and alignment cannot be fully recovered, signalling *irreversible catastrophic forgetting*. The choice of relearning or probe data—forget, retain, or unrelated—barely affects this outcome: once relearning begins, latent structure resurfaces regardless of input (Figure 18). Due to space limitations, complete figures and additional settings appear in Appendix A.5 and Appendix A.6.

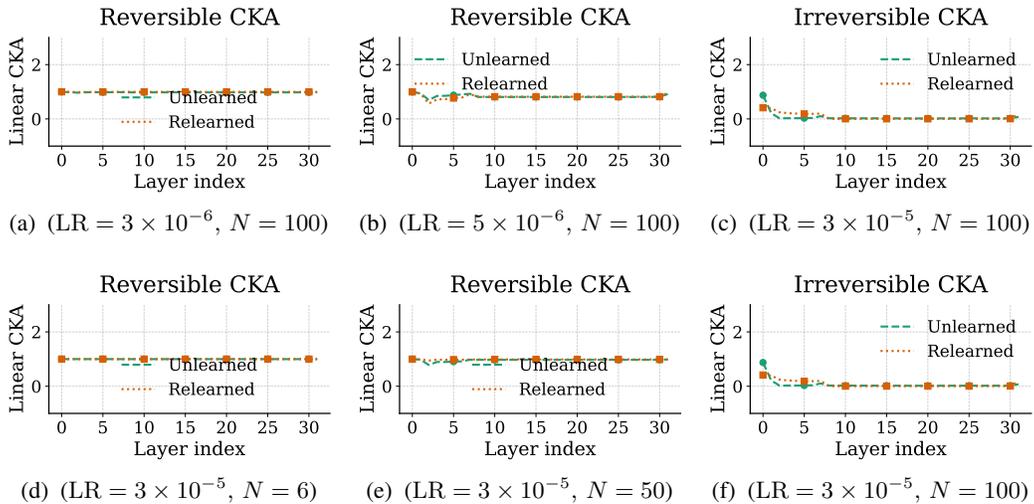


Figure 4: CKA for GA on Yi-6B, simple task. (a–c) LR $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ with $N = 100$; (d–f) LR = 3×10^{-5} with $N \in \{6, 50, 100\}$. High CKA (near 1) indicates strong subspace alignment and *reversible (catastrophic)* forgetting, whereas low CKA (near 0) reflects severe representational drift and *irreversible (catastrophic)* forgetting.

Fisher Information Analysis. Continual unlearning progressively flattens the loss landscape (Figures 5). Higher learning rates or larger request counts push the FIM spectra sharply left, and—at extreme settings—the shift persists after relearning, signalling *irreversible catastrophic forgetting*. Under gentler hyper-parameters the spectra recentre, indicating *reversible catastrophic forgetting*. Relearning with forget, retain, or unrelated data realigns the FIM almost equally well, confirming that the lost sensitivity is suppressed rather than erased (Figure 34-38). Due to space limitations, complete figures and additional settings appear in Appendix A.5 and Appendix A.6.

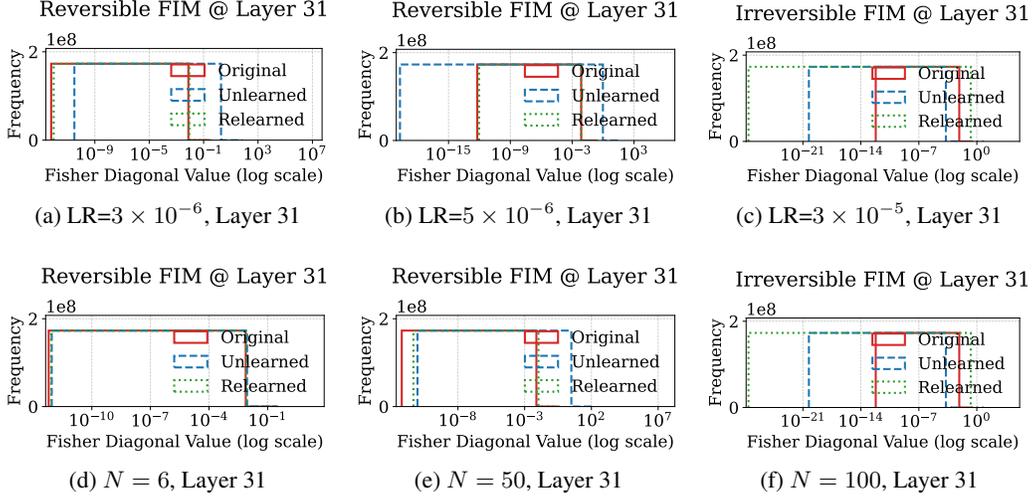


Figure 5: FIM (layer 31) for GA on Yi-6B, simple task. (a–c) $\text{LR} \{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ with $N = 100$; (d–f) $\text{LR} = 3 \times 10^{-5}$ with $N \in \{6, 50, 100\}$. Larger leftward shifts indicate a greater flattening of the loss landscape and *irreversible* (catastrophic) forgetting; spectra that recentre on the original peak denote *reversible* (catastrophic) forgetting.

4.3 Representational Theoretical Analysis

To explain the empirical distinction between *reversible* and *irreversible* (catastrophic) forgetting, we present a perturbation model linking unlearning updates to structural collapse across layers. Consider an L -layer feedforward neural network $f(x) = \sigma(W_L \sigma(\dots \sigma(W_1 x) \dots))$, with activations σ and weights $W_{i=1}^L$. Unlearning is modeled as layer-wise perturbations $\tilde{W}_i = W_i + E_i$ with $|E_i| = O(\text{LR}, N)$, where LR is the learning rate and N the number of unlearning steps. A Neumann-series expansion yields $\tilde{f}(x) - f(x) = \sum_{\emptyset \neq S \subseteq \{1, \dots, L\}} (W_L \circ \dots \circ E_{i_k} \circ \dots \circ W_1)(x)$.

When small perturbations are confined to a few layers, first-order terms dominate, and the effect is *reversible* (catastrophic) forgetting. In contrast, comparable perturbations spread across many layers, higher-order terms accumulate, producing *irreversible* (catastrophic) forgetting.

PCA Similarity. Let X_i and $Y_i = X_i + E'_i$ be the centered activations at layer i before and after unlearning. By Davis–Kahan theorem [5], $\cos \angle(\mathbf{c}_i^{\text{orig}}, \mathbf{c}_i^{\text{upd}}) \approx 1 - O(\|E'_i\| / (\lambda_{1,i} - \lambda_{2,i}))$, with top two eigenvalues $\lambda_{1,i}, \lambda_{2,i}$. The layer-averaged PCA similarity is $\bar{S}_{\text{PCA}} \approx 1 - O((1/L) \sum_i \|E'_i\|)$.

PCA Shift. Along the first principal component, the activation-centroid shift is $\Delta p_i = \mu_{i,1}^{\text{upd}} - \mu_{i,1}^{\text{orig}} = O(\|E'_i\|)$. Large perturbations $\|E'_i\|$ spanning many layers cause irreversible representational drift; otherwise, the shifts remain localized and reversible.

CKA. Let $\tilde{K}_{Y_i} = \tilde{K}_{X_i} + \Delta K_i$ be the perturbed Gram matrix. Then, CKA_i is computed as $1 - O(\|\Delta K_i\|_* / \|\tilde{K}_{X_i}\|_*)$, which implies that $\bar{C} \approx 1 - O((1/L) \sum_i \|\Delta K_i\|_*)$.

Fisher Information. Given update $\delta w_i = O(\|E_i\|)$, the Fisher diagonal behaves as $F_{ii}(w + \delta w) = F_{ii}(w) + O(\|\delta w_i\|)$, so the average Fisher becomes $\bar{F} = (1/P) \sum_i F_{ii} = F_0 - O((1/P) \sum_i \|E_i\|)$.

Token-level metrics, such as accuracy, MIA, and AUC, may report total collapse even when the model’s internal geometry is largely preserved. In *reversible* forgetting, a few parameter changes (*e.g.*, in output heads or layer norms) can drastically perturb token probabilities while leaving deeper representations intact. For the soft-max output $\log p(y|x; \theta + \delta\theta) \approx \log p(y|x; \theta) + \nabla_{\theta} \log p(y|x; \theta)^{\top} \delta\theta + O(\|\delta\theta\|^2)$, a small $\delta\theta$ in high-sensitivity regions (near the logits) can dominate the first-order term, producing large drops in accuracy or anomalous AUC scores despite minimal representational drift.

Unlearning on the forget set applies a weight update $E_i = LR \times \nabla_{W_i} L(D_f)$, which both removes over-fitting to \mathcal{D}_f and accentuates its principal feature subspace. After relearning, parameters may return close to their originals, yet amplified patterns persist, sometimes yielding *better* performance on an augmented \mathcal{D}_f than the baseline. Therefore, unlearning can inadvertently act as a contrastive regularizer, further illustrating the mismatch between surface metrics and feature subspaces.

Summary. *Reversible (catastrophic) forgetting* occurs when perturbations touch only a few layers; PCA similarity/shift, CKA, and FIM remain near baseline. In contrast, *irreversible (catastrophic) forgetting* emerges from large, distributed updates that collapse the model’s representational structure.

Unlearning acts as a contrastive perturbation: it removes memorized content yet “reinforces” salient features of the forget set, so subsequent relearning can even outperform the original model on related inputs, revealing its “dual role” in both removal and refinement.

Token-level metrics (*e.g.*, accuracy, MIA) are overly sensitive to small shifts in high-impact parameters and can misclassify the regime. Structural diagnostics, complemented by augmented evaluation, can provide a more reliable assessment of whether forgetting is truly irreversible.

5 Discussion and Takeaways

Beyond theoretical justifications in Section 4, we summarize main empirical and analytical insights.

(1) Single vs. continual unlearning, and the role of GA/RLabel. Single unlearning rarely produces *irreversible* collapse: performance is recoverable and representational drift is slight. Continual unlearning, especially with large learning rates, often drives the model into permanent failure: ~ 100 sequential requests can push both forget- and retain-set accuracy to near zero. GA and RLabel already over-forget in the single scenario and magnify this damage when applied continually. Adding retain-set terms, as in GA+KL or NPO(+KL) [33; 35; 31], markedly improves stability.

(2) Collapse stems from structural drift, not true erasure. PCA-Similarity/Shift, CKA, and the FIM consistently expose this breakdown: irreversible collapse coincides with large rotations of principal directions, centroid shifts, and vanishing Fisher mass across many layers. When perturbations remain local (small LR, few unlearning requests), these diagnostics stay near baseline—*reversible* forgetting. Token-level metrics alone are unreliable: small updates in high-sensitivity parameters (*e.g.*, logits or layer norms) can tank accuracy or inflate MIA AUC while internal geometry is intact.

(3) Unlearning can act as implicit augmentation. In several continual runs, subsequent relearning on the forget set can often yield *higher* accuracy than that of the original model. This surprising outcome suggests that unlearning is not merely a memory deletion mechanism but may also serve as a form of implicit contrastive regularization. As detailed in Section 4, unlearning amplifies the feature subspace associated with the forget set, and relearning on augmented inputs can reinforce semantic structure while promoting robustness. This process reorganizes internal representations to better capture generalizable patterns, acting as a form of curriculum learning.

(4) Diagnostics guide irreversible (benign) forgetting. Tools such as PCA Similarity and Shift, CKA, and Fisher Information reveal not just the presence of structural drift, but also where and how it arises across layers. This enables targeted control: effective unlearning can be guided toward perturbing only those parameters responsible for the forget set, while preserving the representation structure on the retain set and unrelated data. This opens a path to *targeted, irreversible forgetting*, a permanent and isolated removal of information without collateral collapse, offering actionable insights for building safer unlearning algorithms.

6 Conclusion

We revisit machine unlearning for LLMs through a systematic study of *reversibility*. Token-level metrics alone can mislead: a model may appear collapsed yet remain fully recoverable.

To diagnose this gap, we introduce a feature-space toolkit—PCA Similarity, PCA Shift, CKA, and FIM—that cleanly separates *reversible* from *irreversible* catastrophic forgetting. Our empirical and theoretical results show that true forgetting arises only when many layers undergo coordinated, large-magnitude perturbations; by contrast, minor updates in high-sensitivity regions (*e.g.*, output logits) can slash accuracy or inflate perplexity while leaving internal representations intact.

These findings call for evaluation protocols that go beyond token-level scores and for algorithms that actively control representational drift. We further observe that unlearning, followed by proper relearning, can refine representations and even boost downstream performance. Together, these insights chart a path toward safer, more interpretable unlearning in LLMs.

References

- [1] Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O’Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernández-Orallo, Mor Geva, and Yarin Gal. Open problems in machine unlearning for AI safety. *arXiv:2501.04952*, 2025.
- [2] Lucas Bourtole, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *S&P*, pages 141–159, 2021.
- [3] Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. Towards robust and cost-efficient knowledge unlearning for large language models. *arXiv:2408.06621*, 2024.
- [4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.
- [5] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.
- [6] Ronen Eldan and Mark Russinovich. Who’s harry potter? approximate unlearning in llms. *arXiv:2310.02238*, 2023.
- [7] Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. *arXiv:2410.02760*, 2024.
- [8] Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning. In *ICLR*, 2025.
- [9] Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *NeurIPS*, pages 3513–3526, 2019.
- [10] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021.
- [11] Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *ICLR*, 2022.
- [12] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *ACL*, pages 14389–14408, 2023.
- [13] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In *EMNLP*, pages 7403–7412, 2023.

- [14] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *arXiv:1612.00796*, 2016.
- [15] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In *ICML*, pages 3519–3529, 2019.
- [16] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. [<https://huggingface.co/AI-MO/NuminaMath-1.5>] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.
- [17] Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. In *NeurIPS*, 2024.
- [18] 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 B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhruhu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Kiran Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *ICML*, 2024.
- [19] Zitong Li, Qingqing Ye, and Haibo Hu. Funu: Boosting machine unlearning efficiency by filtering unnecessary unlearning. *arXiv:2501.16614*, 2025.
- [20] Junxu Liu, Mingsheng Xue, Jian Lou, Xiaoyu Zhang, Li Xiong, and Zhan Qin. Muter: Machine unlearning on adversarially trained models. In *ICCV*, pages 4869–4879, 2023.
- [21] Michelle Lo, Fazl Barez, and Shay B. Cohen. Large language models relearn removed concepts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *ACL*, pages 8306–8323, 2024.
- [22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
- [23] Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv:2402.16835*, 2024.
- [24] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for llms. *arXiv:2401.06121*, 2024.
- [25] 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:2311.17035*, 2023.
- [26] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *ICML*, 2024.
- [27] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *ICLR*, 2024.
- [28] 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:2407.06460*, 2024.

- [29] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL*, pages 4149–4158, 2019.
- [30] Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. Unveiling the implicit toxicity in large language models. In *EMNLP*, pages 1322–1338, 2023.
- [31] Xiaoyu Xu, Minxin Du, Qingqing Ye, and Haibo Hu. Obliviate: Robust and practical machine unlearning for large language models. arXiv:2505.04416, 2025.
- [32] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv:2412.15115, 2024.
- [33] Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In *ACL*, pages 8403–8419, 2024.
- [34] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai. arXiv:2403.04652, 2024.
- [35] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. arXiv:2404.05868, 2024.
- [36] Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. arXiv:2501.13453, 2025.

A Appendix

A.1 Limitations

Our experiments target two LLMs and a handful of tasks and unlearning methods; although our diagnostic framework is model-agnostic and designed to scale, empirical validation on much larger models and production-scale pipelines remains to be done. The constrained relearning protocol and selected metrics provide clear insights into representational drift but are not exhaustive and do not offer formal privacy guarantees. Extending these analyses to diverse architectures, richer unlearning strategies, and rigorous theoretical bounds is an important direction for future work.

A.2 Broader impacts

This work provides new tools and theoretical insights for evaluating and guiding machine unlearning in LLMs, with the potential to improve privacy guarantees and regulatory compliance (e.g., GDPR “right to be forgotten”). By distinguishing reversible from irreversible forgetting, our diagnostics can help developers ensure that sensitive or copyrighted material is truly expunged rather than merely hidden. At the same time, adversaries could exploit reversible forgetting pathways to mask malicious or biased content and then restore it later, highlighting the need for robust defenses. The computational overhead of continual unlearning and repeated diagnostics may increase energy consumption, underscoring the importance of efficient implementations. Finally, by exposing the fragility of token-level metrics, our work advocates for more trustworthy evaluation standards that balance privacy, utility, and environmental considerations in real-world deployments.

A.3 Detailed Analysis Tools

PCA Similarity and PCA Shift. For each Transformer layer, we perform PCA on the hidden activations of the *original* and *updated* models. Let $\mathbf{c}_{i,1}^{\text{orig}}$ and $\mathbf{c}_{i,1}^{\text{upd}}$ denote the first principal component (PC1) directions of layer i . The *PCA Similarity* is defined as

$$\text{PCA-Sim}(i) = \cos(\mathbf{c}_{i,1}^{\text{orig}}, \mathbf{c}_{i,1}^{\text{upd}}) = \frac{(\mathbf{c}_{i,1}^{\text{orig}})^\top \mathbf{c}_{i,1}^{\text{upd}}}{\|\mathbf{c}_{i,1}^{\text{orig}}\| \|\mathbf{c}_{i,1}^{\text{upd}}\|} \in [-1, 1],$$

where values near 1 indicate stable directional alignment, and values near -1 suggest a near-orthogonal shift in dominant directions.

To capture translational drift, we also compute the mean projection of activations along PC1 and PC2:

$$\text{PCA-Shift}(i) = p_{1,\text{upd}} - p_{1,\text{orig}}, \quad \text{Principle}(i) = p_{2,\text{upd}},$$

where PCA-Shift quantifies displacement along PC1 and Principle captures orthogonal deviation along PC2. These metrics reflect how the representation center drifts within the top subspace.

Centered Kernel Alignment (CKA). To assess subspace alignment, we use linear Centered Kernel Alignment (CKA) [15], which compares activation matrices $X, Y \in \mathbb{R}^{N \times D}$ from before and after unlearning. First, we compute the centered Gram matrices:

$$\tilde{K}_X = HXX^\top H, \quad \tilde{K}_Y = HYY^\top H, \quad H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top.$$

The CKA score is then given by:

$$\text{CKA}(X, Y) = \frac{\text{Tr}(\tilde{K}_X \tilde{K}_Y)}{\sqrt{\text{Tr}(\tilde{K}_X^2)} \sqrt{\text{Tr}(\tilde{K}_Y^2)}} \in [0, 1],$$

where values near 1 indicate highly overlapping subspaces, and values near 0 signal near-orthogonality.

Fisher Information. To measure parameter-level importance, we compute the diagonal of the empirical Fisher Information Matrix (FIM). For each parameter w_i and input distribution \mathcal{D}_{dis} , the diagonal entry is approximated as:

$$\text{FIM}_{ii} \approx \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_{\text{dis}}} \left[(\partial_{w_i} \log p(y | \mathbf{x}; \mathbf{w}))^2 \right].$$

Larger values indicate that w_i has a stronger influence on the model’s predictions. A substantial leftward shift in the Fisher spectrum after unlearning implies a flattened loss landscape and diminished parameter sensitivity.

Together, these tools form a feature-space diagnostic suite: FIM captures global sensitivity, CKA measures subspace preservation, and PCA-based metrics expose fine-grained geometric drift across layers—enabling a robust assessment of representational degradation during unlearning.

A.4 Complete Experiment

Table 4 summarizes token-level performance under four unlearning configurations on Yi-6B, varying both learning rates and request counts. Most metrics—including forget/retain accuracy, perplexity, CSQA, and GSM8K—follow consistent trends: milder methods (e.g., NPO, NPO+KL) yield moderate degradation and support effective recovery via relearning, while aggressive methods (e.g., GA, RL) under high learning rates lead to severe, often irrecoverable performance drops. These results align closely with our theoretical distinction: *reversible (catastrophic) forgetting* corresponds to localized perturbations and structurally recoverable models, while *irreversible (catastrophic) forgetting* reflects widespread, irreversible damage.

MIA, however, behaves differently. Even when downstream performance collapses, MIA scores often remain high and recover rapidly. This suggests MIA is especially sensitive to shallow changes (e.g., output logits or normalization), rather than deeper representational shifts. As a result, MIA often correlates with reversible forgetting—capturing surface-level instability without revealing structural collapse. These observations reinforce that while token-level metrics offer partial signals, only internal diagnostics can reliably differentiate reversible from irreversible forgetting.

A.5 Single Unlearning

Figure 6 illustrates feature-level changes under single unlearning using PCA Similarity, PCA Shift, CKA, and Fisher Information. In (a), PCA Similarity remains consistently high across layers, with cosine scores near 1, indicating that dominant activation directions are well preserved. Slight dips in shallow and final layers are quickly restored after relearning, suggesting minimal and reversible geometric drift. Subfigure (b) confirms that PC1 shifts and orthogonal deviations are small, with relearned centers closely matching the original. In (c), CKA shows near-perfect alignment between the original, unlearned, and relearned representations, reinforcing the conclusion that subspace structure remains intact. Fisher Information spectra in (d–f) reveal only mild leftward shifts, indicating slight loss flattening and reduced parameter sensitivity, which are fully recovered after relearning. Overall, these results confirm that single unlearning causes only minor, reversible structural perturbations—highlighting the fragility of token-level evaluations in capturing irreversible forgetting.

A.6 Detailed Analysis Results

A.6.1 Principal Component Analysis: Similarity and Shift

Across the same hyper-parameter grid, Figures 7–13 plot the *PCA-Shift* trajectories—layer-wise displacements of activation centroids along the first two principal directions. For GA-based objectives the pattern mirrors their Similarity curves: as LR rises the orange triangles (unlearned) shoot far from the blue circles (original), especially in deeper layers, and the green squares (relearned) fail to return, producing long grey rays that diagnose *irreversible* drift. With GA+GD the spread is smaller but still widens sharply at 3×10^{-5} , confirming that doubling the loss term does not prevent global collapse.

NPO and NPO+KL behave differently. Even at aggressive LR the shifts remain tightly clustered—most layers move < 100 units on PC1 and almost none on PC2—and green squares consistently fall back a little onto the original line segment. RLabel shows an intermediate picture: early layers barely move, while late layers fan out as LR or N grow; nonetheless the rays shorten markedly after relearning, indicating that most distortion is still recoverable.

Task complexity amplifies divergence. On Qwen-2.5-7B the GA rays explode more quickly (Figure 13c,f,i), spanning thousands of PC1 units and driving PC2 to extreme negative values—the hallmark of a high-order, multi-layer perturbation predicted by our theory in Section 4. Conversely,

Table 4: **Yi-6B simple-task metrics under four** (LR, N) **settings**. For each block: forget/retain perplexity (F.Ppl / R.Ppl), forget/retain accuracy (F.Acc / R.Acc), CommonsenseQA (CSQA), GSM8K, and membership-inference AUC (MIA).

Phase	Method	F.Ppl	R.Ppl	F.Acc	R.Acc	CSQA	GSM8K	MIA
LR= 3×10^{-5} , $N = 100$								
Original	—	3.8	7.8	78.9	65.5	73.1	39.6	70.9
Unlearn	GA	∞	∞	0.0	0.0	19.3	0.0	26.1
	GA+GD	∞	∞	9.7	2.3	19.7	0.0	16.8
	GA+KL	∞	∞	9.0	6.2	19.6	0.0	17.8
	NPO	31296.5	597.9	37.8	37.9	62.2	1.0	60.1
	NPO+KL	348080.2	4482.0	64.3	55.9	64.9	1.4	59.0
	Rlable	63791.7	65903.4	0.0	0.0	20.9	0.0	65.1
Relearn	GA	137094.5	758443.5	2.1	1.8	19.7	0.0	74.5
	GA+GD	5274.5	9568.6	2.2	2.6	19.6	0.0	68.1
	GA+KL	5037.1	15019.9	1.7	1.6	20.6	0.0	70.7
	NPO	16.6	41.7	57.0	45.6	51.8	0.6	70.0
	NPO+KL	21.8	16.2	60.7	54.3	48.0	0.9	67.7
	Rlable	4056.1	15048.6	4.3	2.8	19.7	0.0	69.5
LR= 5×10^{-6} , $N = 100$								
Unlearn	GA	∞	∞	9.1	6.2	19.6	0.0	23.2
	GA+GD	∞	∞	3.6	3.1	24.5	0.0	28.7
	GA+KL	∞	∞	9.1	6.2	19.6	0.0	27.3
	NPO	3017.7	1110.6	50.1	52.3	72.9	37.5	50.6
	NPO+KL	38.5	232.4	77.6	64.3	73.1	37.6	65.4
	Rlable	57035.4	53377.1	0.1	0.4	19.1	0.0	63.6
Relearn	GA	3.7	7.8	80.0	64.9	70.2	39.9	68.0
	GA+GD	3.6	7.6	81.2	65.1	72.1	39.0	69.8
	GA+KL	3.6	8.4	81.1	64.8	71.6	40.7	68.3
	NPO	3.5	7.6	82.7	65.5	74.0	39.7	68.0
	NPO+KL	3.5	7.8	83.8	65.6	74.1	39.7	69.5
	Rlable	3.6	7.7	80.8	65.3	71.8	39.2	70.3
LR= 3×10^{-6} , $N = 100$								
Unlearn	GA	∞	∞	16.8	14.4	69.5	12.3	25.2
	GA+GD	3.3	7.6	78.8	65.5	77.0	37.5	69.4
	GA+KL	∞	∞	35.4	40.6	63.2	18.3	18.9
	NPO	3.7	7.9	78.3	65.0	73.3	38.7	68.4
	NPO+KL	3.8	8.1	78.4	65.1	73.6	38.6	66.7
	Rlable	36794.7	32562.0	3.8	3.2	19.3	2.2	61.4
Relearn	GA	3.7	7.6	80.8	65.2	73.4	39.9	68.6
	GA+GD	3.6	7.4	81.8	65.5	72.1	39.0	70.0
	GA+KL	3.6	10.3	81.0	63.3	67.2	40.7	70.7
	NPO	3.5	7.5	81.2	65.4	72.9	39.7	69.9
	NPO+KL	3.5	7.5	83.8	65.5	73.0	39.7	69.9
	Rlable	3.6	7.6	80.5	65.3	72.2	39.2	70.0
LR= 3×10^{-5} , $N = 6$								
Unlearn	GA	inf	inf	36.3	36.1	69.1	5.8	29.6
	GA+GD	209.3	20.6	77.0	64.0	70.0	37.8	66.9
	GA+KL	inf	inf	53.0	41.5	68.3	2.0	29.5
	NPO	12.3	10.7	71.6	59.4	71.7	24.7	68.7
	NPO+KL	8.9	10.7	74.7	62.1	72.8	32.2	67.9
	Rlable	51589.2	40622.9	0.4	0.7	19.8	0.0	62.6
Relearn	GA	6.8	11.4	70.5	58.7	64.5	18.4	68.2
	GA+GD	12.3	11.5	61.6	54.4	61.3	7.3	67.1
	GA+KL	17.1	11.6	66.6	56.2	60.6	3.0	65.0
	NPO	6.0	11.6	71.2	59.4	59.4	2.0	68.4
	NPO+KL	7.3	11.6	67.6	56.1	42.9	1.6	69.0
	Rlable	6.4	11.4	72.7	61.1	67.5	28.9	65.2

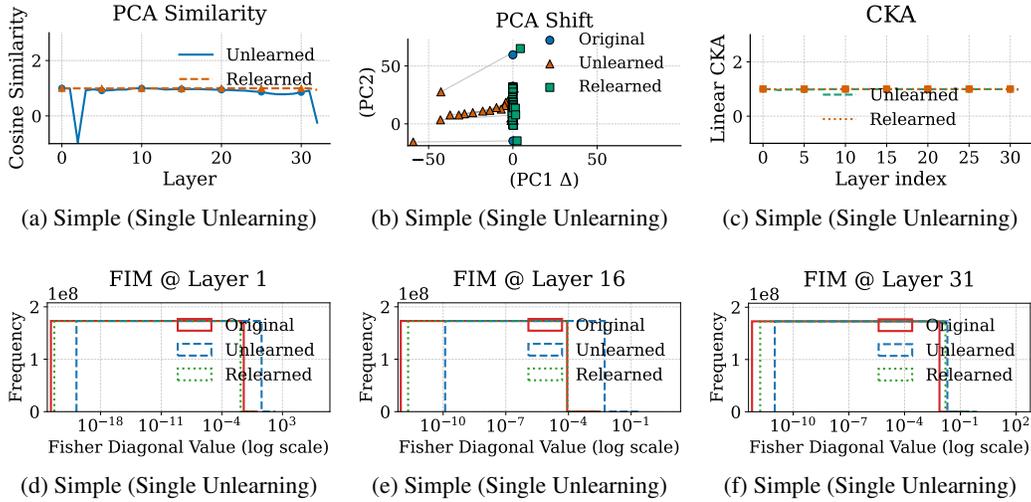


Figure 6: Single unlearning analysis on Yi-6B for GA under a simple task. PCA Similarity, PCA Shift, CKA, and Fisher information (FIM) distributions are reported across key layers to assess representation and parameter changes before and after unlearning and relearning.

NPO’s clusters expand only modestly, even under the same LR, and contract once relearning is applied.

Taken together, PCA-Shift complements PCA-Similarity: Similarity captures angular mis-alignment, while Shift quantifies translational drift. Their joint reading confirms that GA (with or without GD/KL) is prone to large, irreversible representational displacements, whereas NPO variants and, to a lesser extent, RLabel confine shifts to a regime that remains correctable—consistent with the reversible versus irreversible forgetting boundary observed in our utility experiments.

A.6.2 Centered Kernel Alignment Analysis

Figures 15– 17 report layer-wise linear CKA between the original model and its unlearned / relearned counterparts. Across both Yi-6B and Qwen-2.5-7B, GA again stands out: as LR or N grows its CKA curve close to zero in the final third of the network and never returns, revealing a deep sub-space fracture that matches the irreversible PCA trends. GA+GD and GA+KL modestly attenuate this dip, but still fail to restore full alignment after relearning.

By contrast, NPO and NPO+KL keep CKA higher GA through almost all layers—even under 3×10^{-5} or $N=100$ —and relearning lifts the few layers back to baseline, confirming their perturbations are lower than GA series. RLabel occupies an intermediate position: as LR or N increases, its CKA curve drops rapidly—mirroring GA’s behavior—and ultimately exhibits irreversible forgetting.

Task complexity does not change the ordering but amplifies the gaps: on the math-heavy Qwen benchmark GA’s tail layers fall to almost zero at high LR, whereas NPO keeps higher than GA. Taken together with the PCA-Shift results, CKA shows that only GA-style objectives consistently destroy the encoder–decoder sub-space, while NPO families maintain higher stability than GA series and RLabel induces a moderate, recoverable tilt.

A.6.3 Fisher Information Analysis

Figures 19–33 trace the empirical Fisher spectra layer-by-layer. Across both Yi-6B (simple) and Qwen2.5-7B (complex), GA and GA,variants exhibit a pronounced *leftward* translation of the diagonal histogram as LR or N increases—the peak moves several orders of magnitude in deep and mid layers, signalling a flattened loss surface and vanishing parameter salience. Crucially, these shifts persist after relearning, marking the transition to *irreversible* forgetting. By contrast, NPO, NPO+KL, and RL exhibit smaller leftward shifts under moderate LR or N , with their Fisher spectra recentring after relearning—signalling primarily reversible drift. However, when pushed to extreme regimes

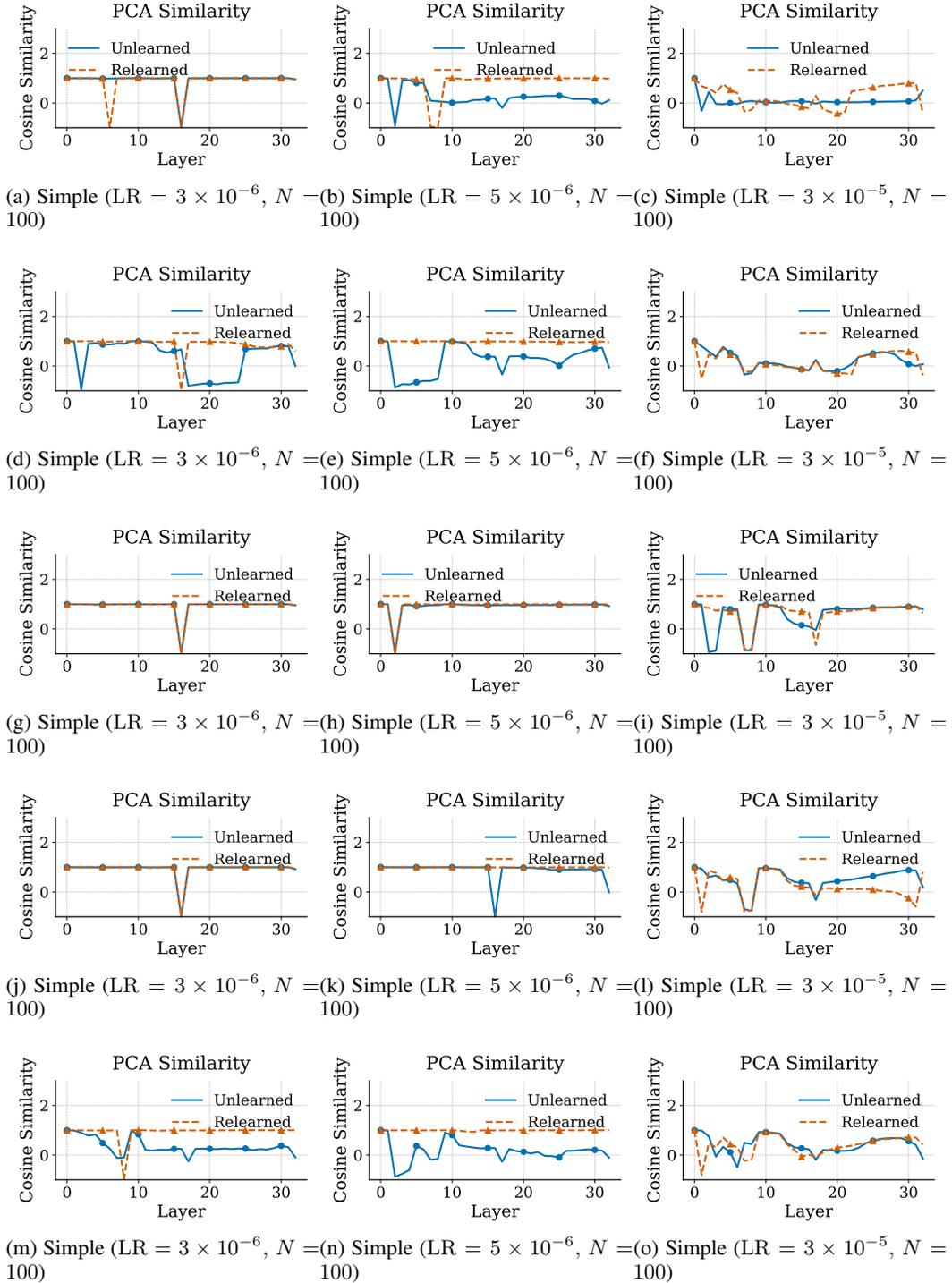


Figure 7: PCA Similarity Across Layers. Each row shows results under different unlearning methods: GA+GD (a-c), GA+KL (d-f), NPO (g-i), NPO+KL (j-l), and Rlable (m-o). All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

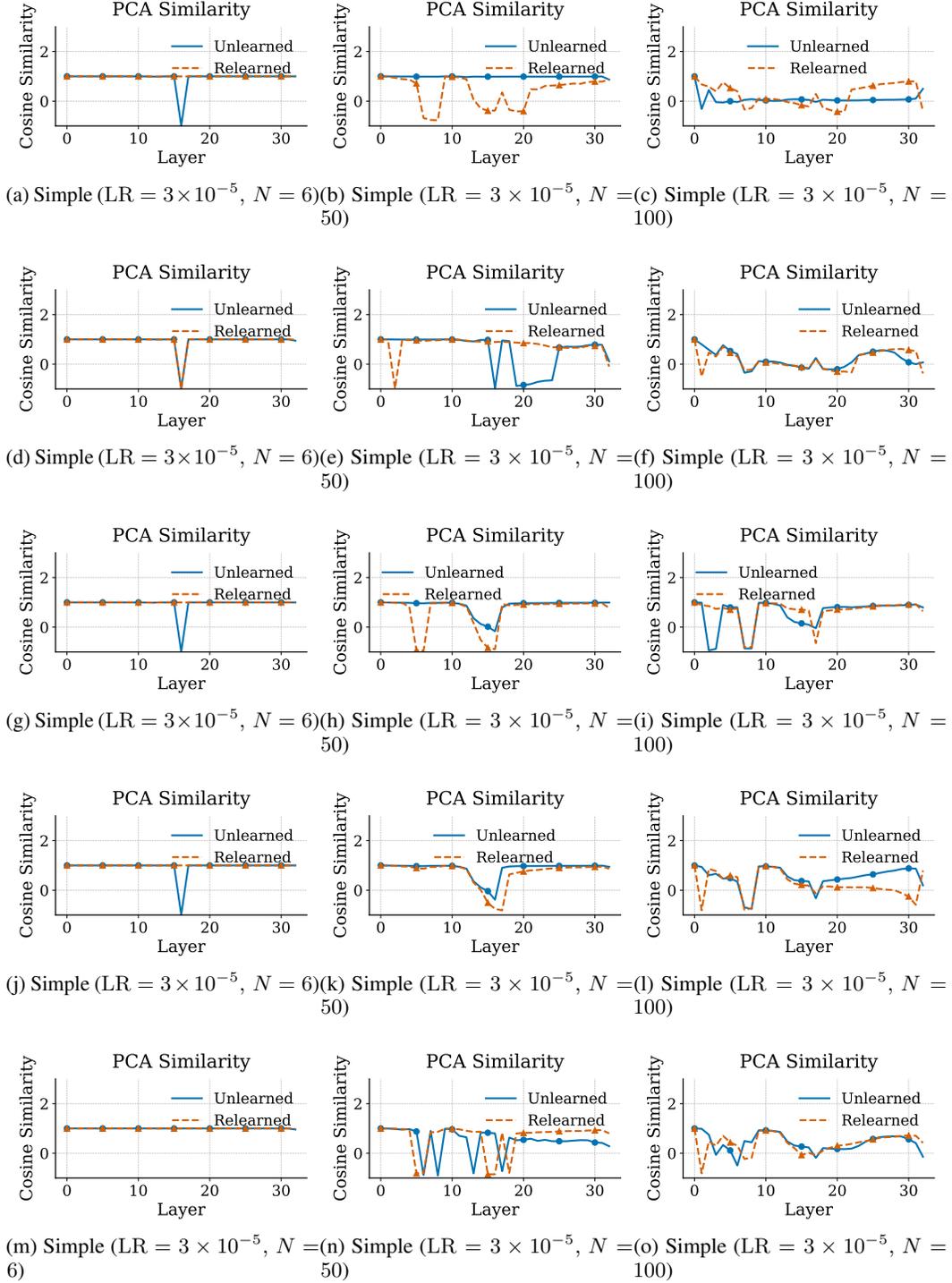


Figure 8: PCA Similarity Across Layers. Each row shows results under different unlearning methods: GA+GD (a–c), GA+KL (d–f), NPO (g–i), NPO+KL (j–l), and Rlable (m–o). Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

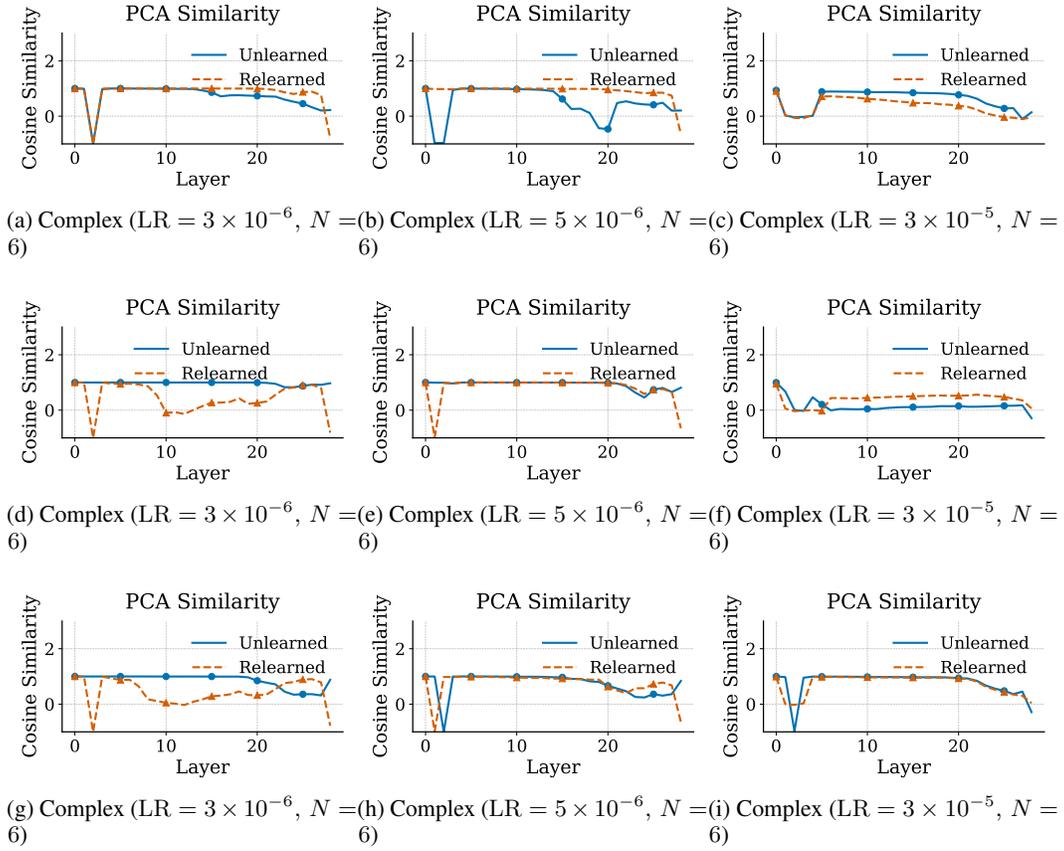


Figure 9: PCA Similarity Across Layers. Each row shows results under different unlearning methods: GA (a-c) NPO (d-f), Rlable (g-j). All plots are for the complex task on Qwen2.5-7B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 6$.

(e.g. LR = 3×10^{-5} or $N = 100$), these methods also show persistent leftward displacement in some layers, indicating milder yet still irreversible forgetting. Varying N (Figures 19–30) reinforces this: at $N = 6$ all methods stay near the original spectrum; at $N = 50$ – 100 , GA series objectives flatten most layers, while the NPO family and RLabel flatten more narrowly and recover more fully, but not perfectly. The complex task echoes the simple-task trends (Figures 31, 32, 33): GA again drives layer-24/28 peaks leftward by $\sim 10^4$, whereas NPO variants shift by less than one decade and rebound. Taken together, Fisher statistics confirm our geometric findings: irrecoverable forgetting is characterised by a global, unrecoverable loss-of-curvature, while reversible forgetting leaves the curvature profile largely intact and easily restorable.

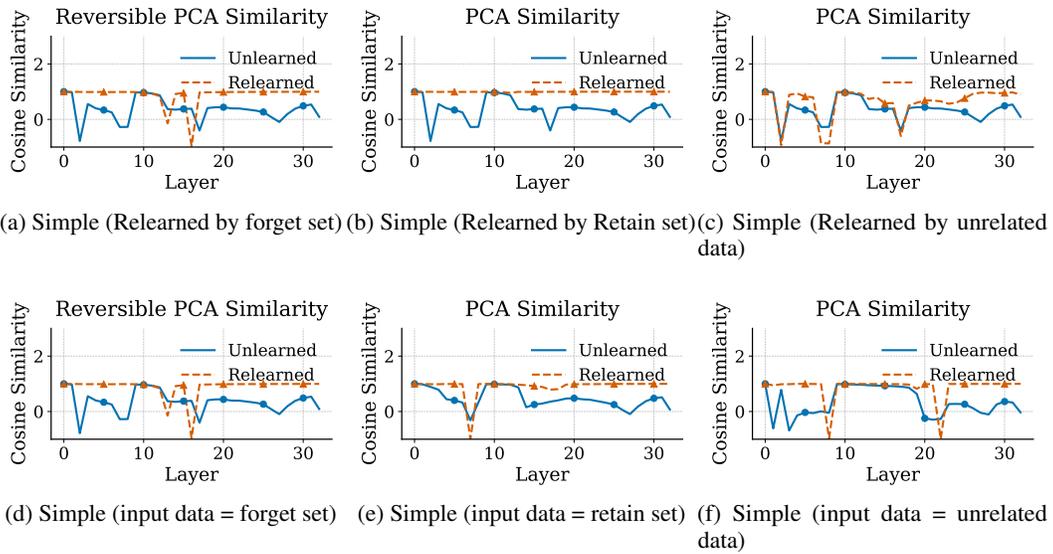


Figure 10: PCA Similarity Analysis for GA under Varied Relearning and Evaluation Inputs on Yi-6B (Simple Task). (a–c): Relearning is performed using the forget set, retain set, or unrelated data respectively. (d–f): PCA similarity is measured using the forget set, retain set, or unrelated data as evaluation input.

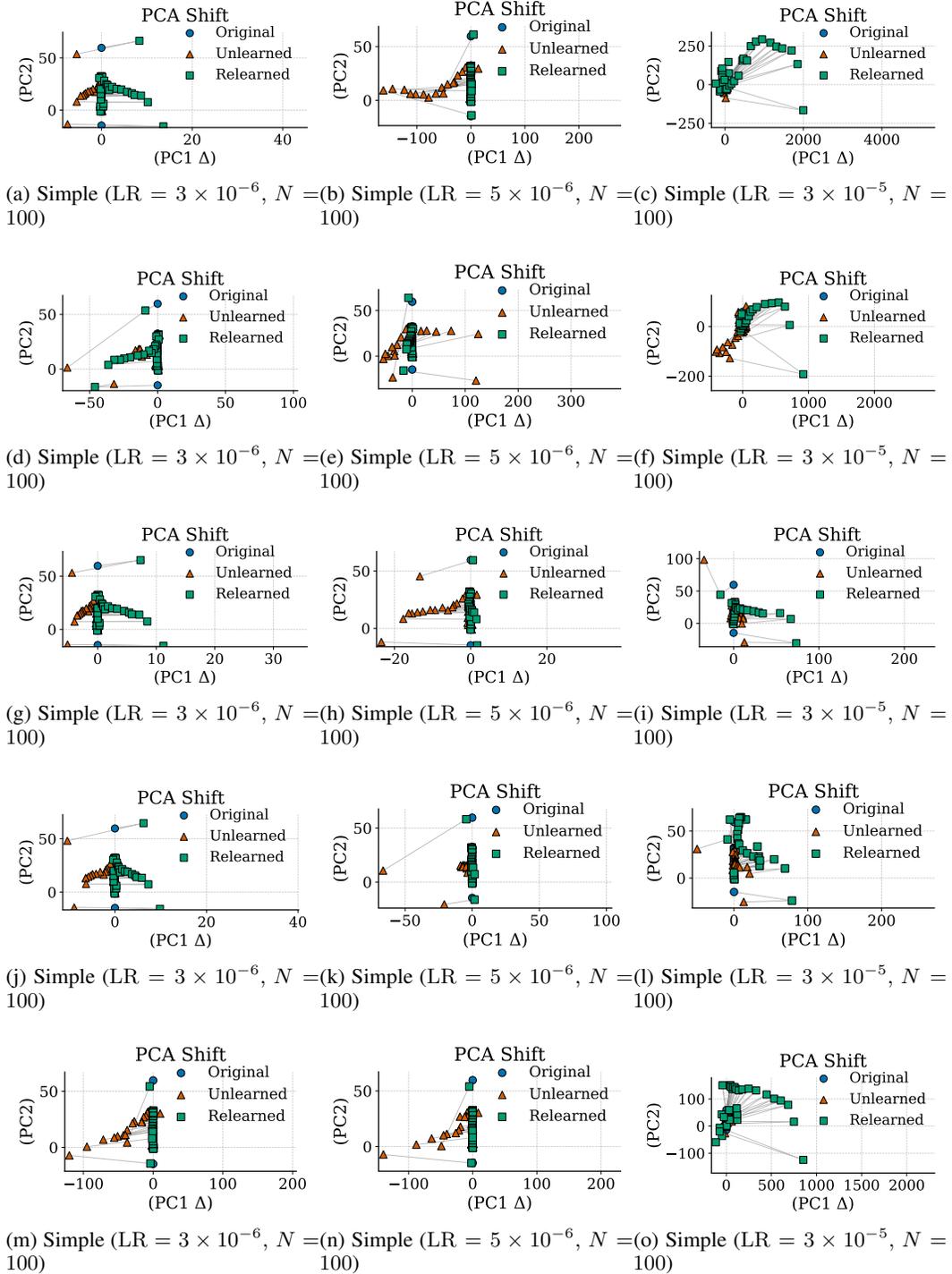


Figure 11: PCA Shift Across Layers. Each row shows results under different unlearning methods: GA+GD (a–c), GA+KL (d–f), NPO (g–i), NPO+KL (j–l), and Rlable (m–o). All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

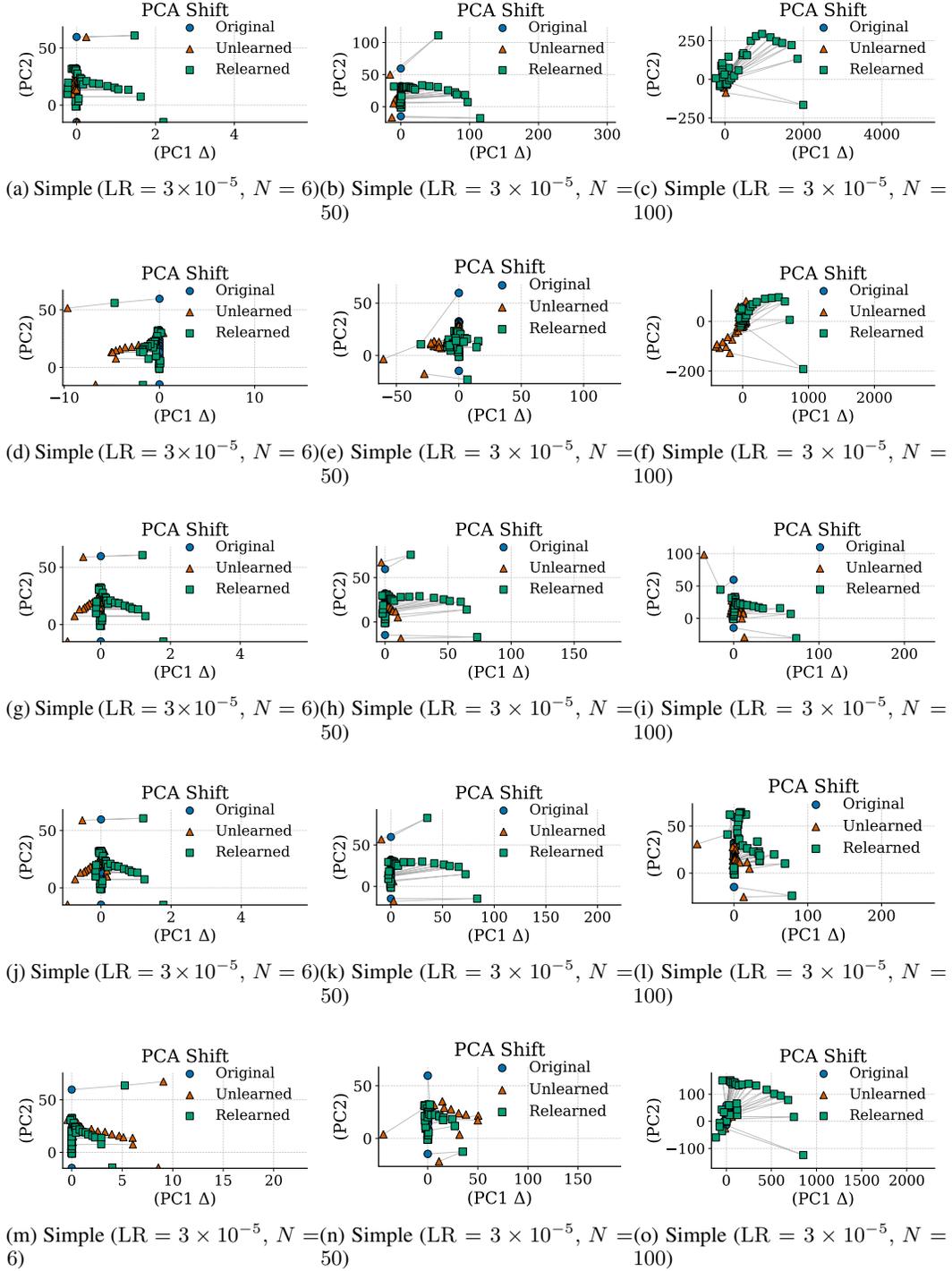


Figure 12: PCA Shift Across Layers. Each row shows results under different unlearning methods: GA+GD (a-c), GA+KL (d-f), NPO (g-i), NPO+KL (j-l), and Rlable (m-o). Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

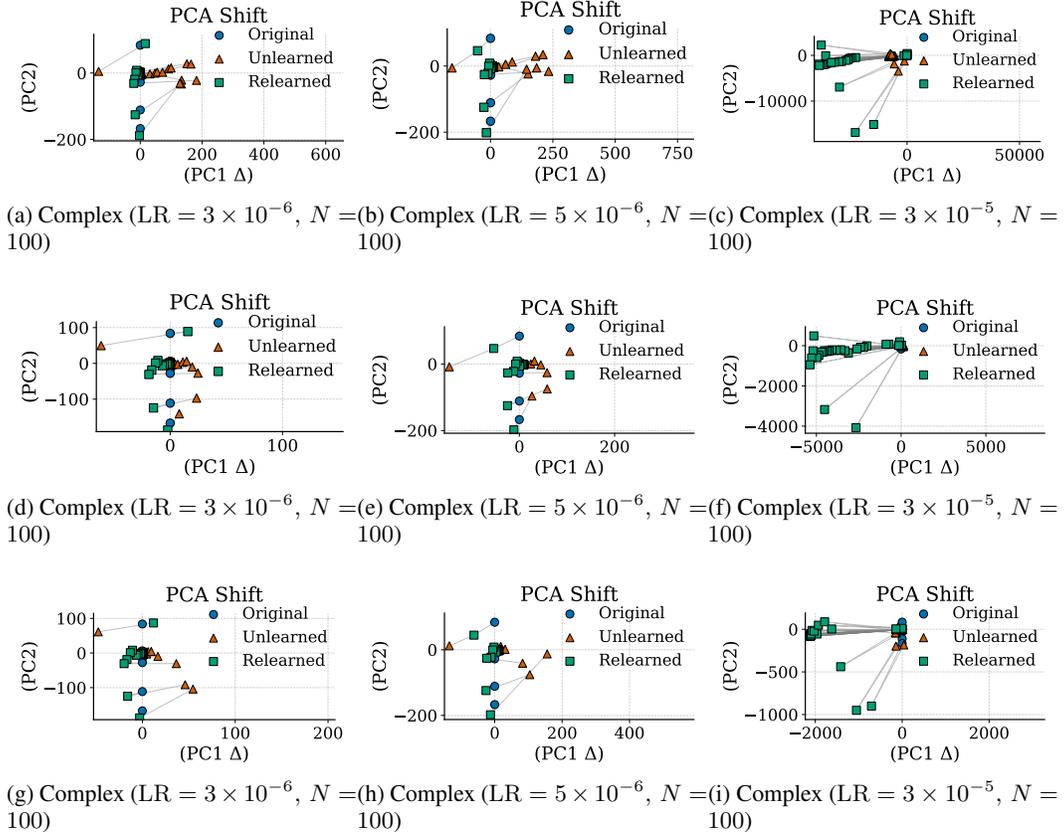


Figure 13: PCA Shift Across Layers. Each row shows results under different unlearning methods: GA (a-c) NPO (d-f), Rlable (g-j). All plots are for the complex task on Qwen2.5-7B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 6$.

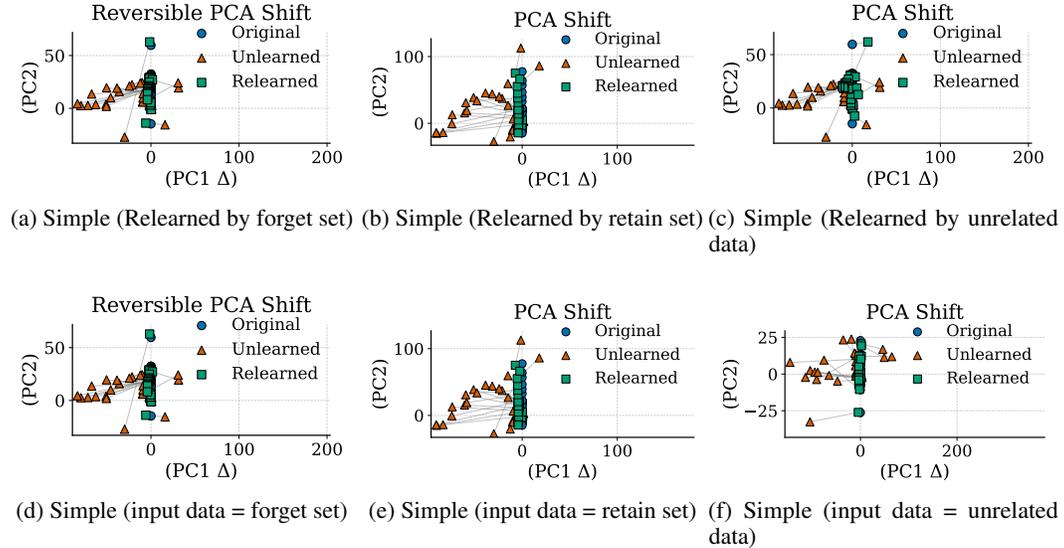


Figure 14: PCA Shift Analysis under Varied Relearning and Evaluation Inputs on Yi-6B (Simple Task). (a-c): Relearning is performed using the forget set, retain set, or unrelated data respectively. (d-f): PCA shift is measured using the forget set, retain set, or unrelated data as evaluation input.

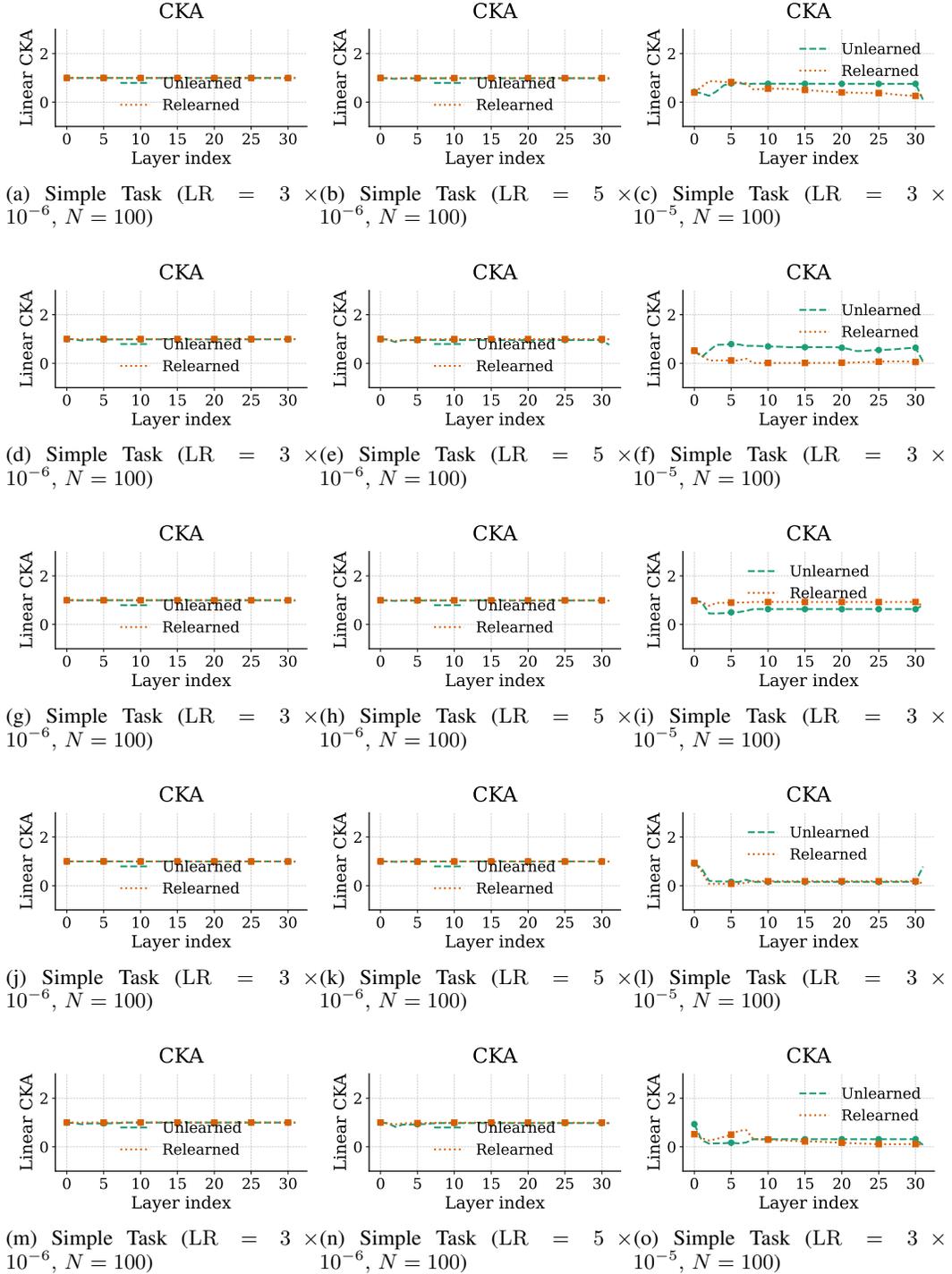


Figure 15: CKA Across Layers. Each row shows results under different unlearning methods: GA+GD (a–c), GA+KL (d–f), NPO (g–i), NPO+KL (j–l), and Rlable (m–o). All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

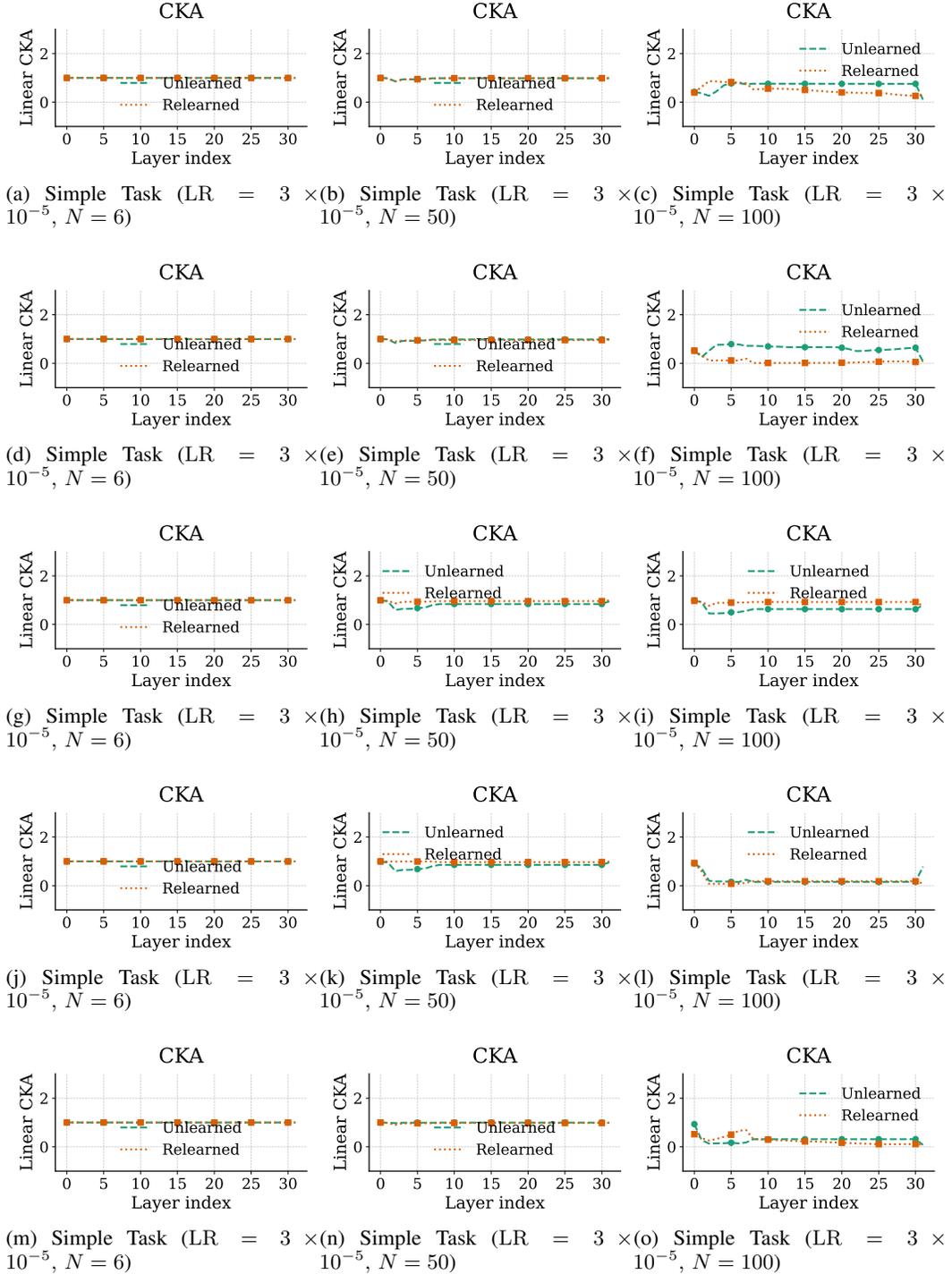


Figure 16: CKA Across Layers. Each row shows results under different unlearning methods: GA+GD (a–c), GA+KL (d–f), NPO (g–i), NPO+KL (j–l), and Rlable (m–o). Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

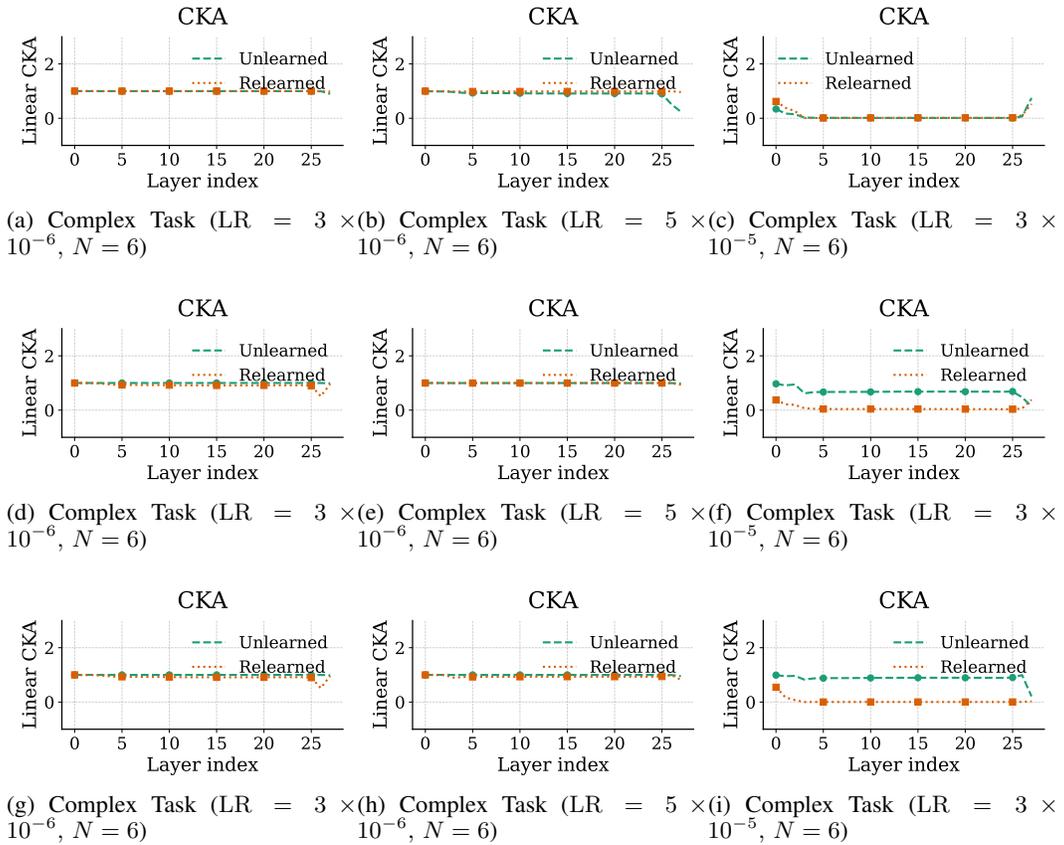


Figure 17: CKA Across Layers. Each row shows results under different unlearning methods: GA (a-c) NPO (d-f), Rlable (g-j). All plots are for the complex task on Qwen2.5-7B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 6$.

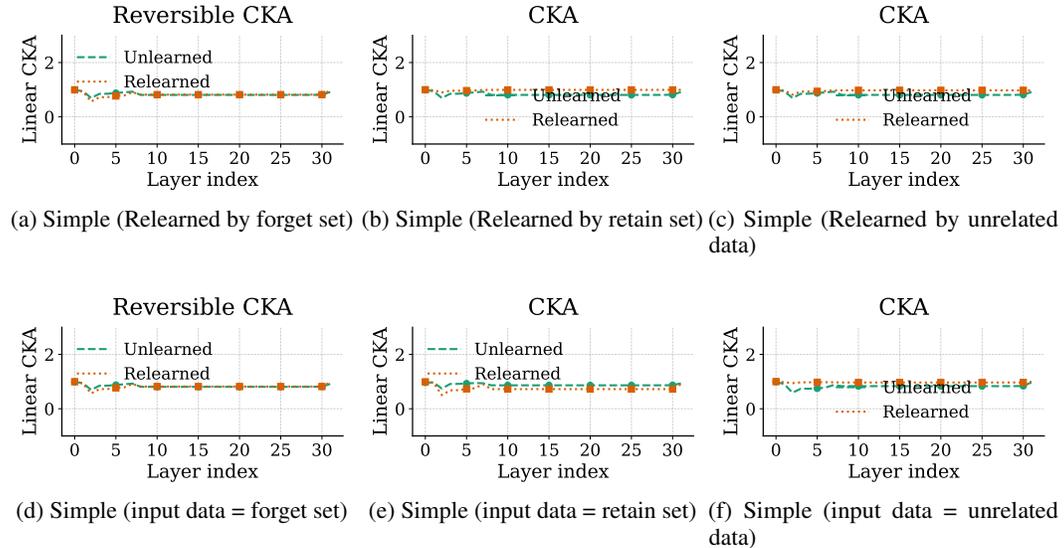


Figure 18: CKA Analysis under Varied Relearning and Evaluation Inputs on Yi-6B (Simple Task). (a-c): Relearning is performed using the forget set, retain set, or unrelated data respectively. (d-f): CKA is measured using the forget set, retain set, or unrelated data as evaluation input.

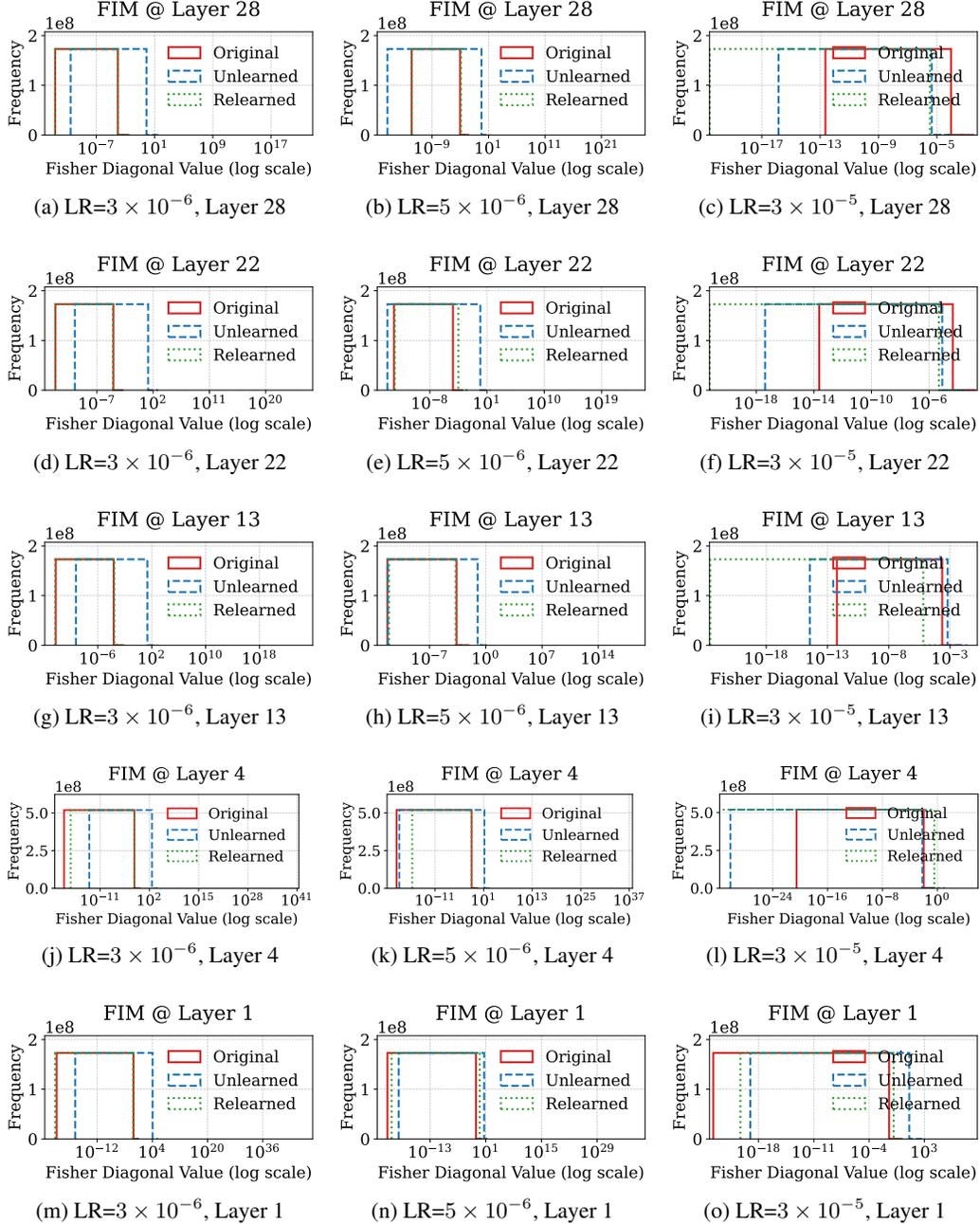


Figure 19: FIM for GA Across Layers. All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

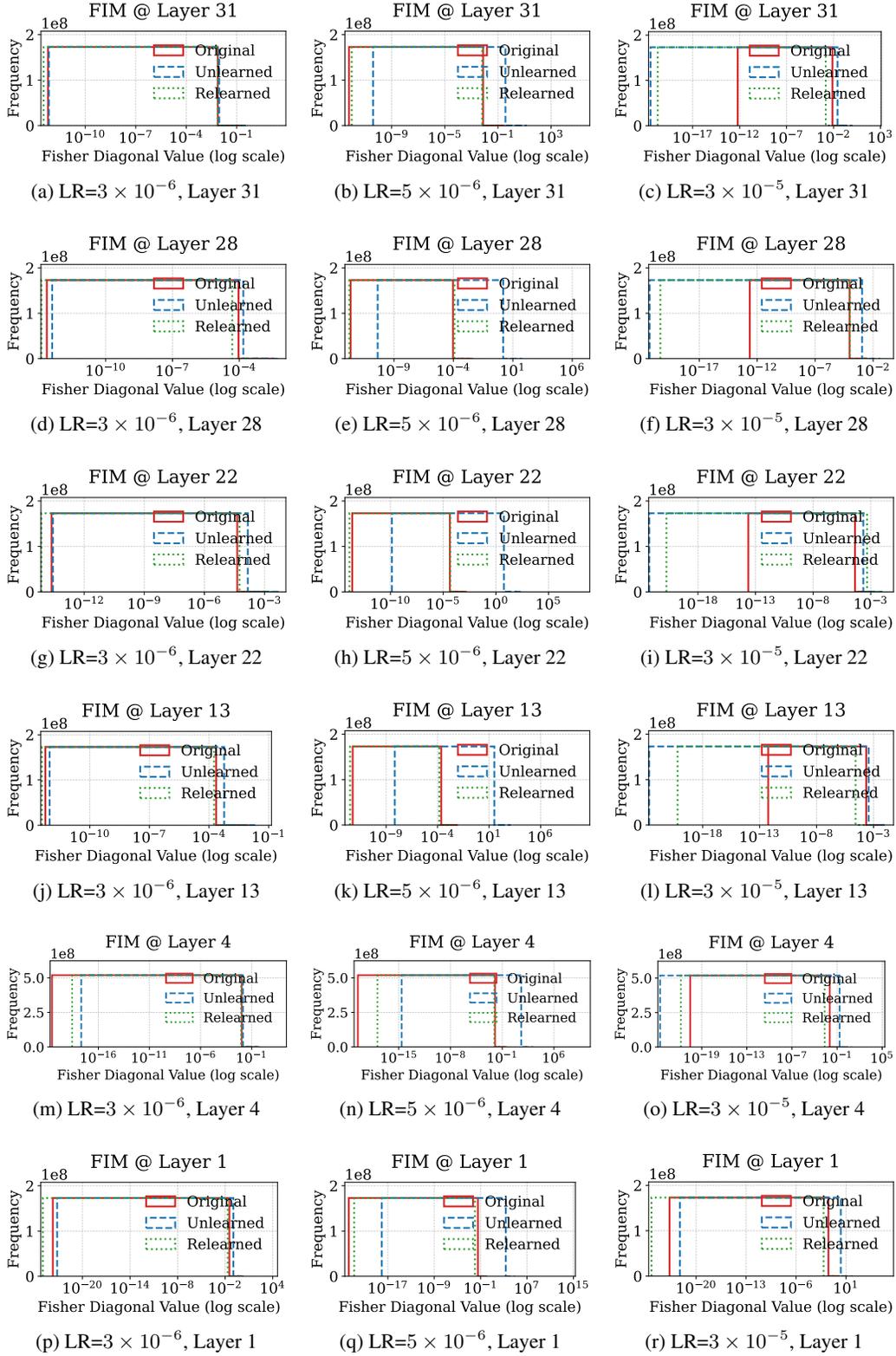


Figure 20: FIM for GA+GD Across Layers. All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

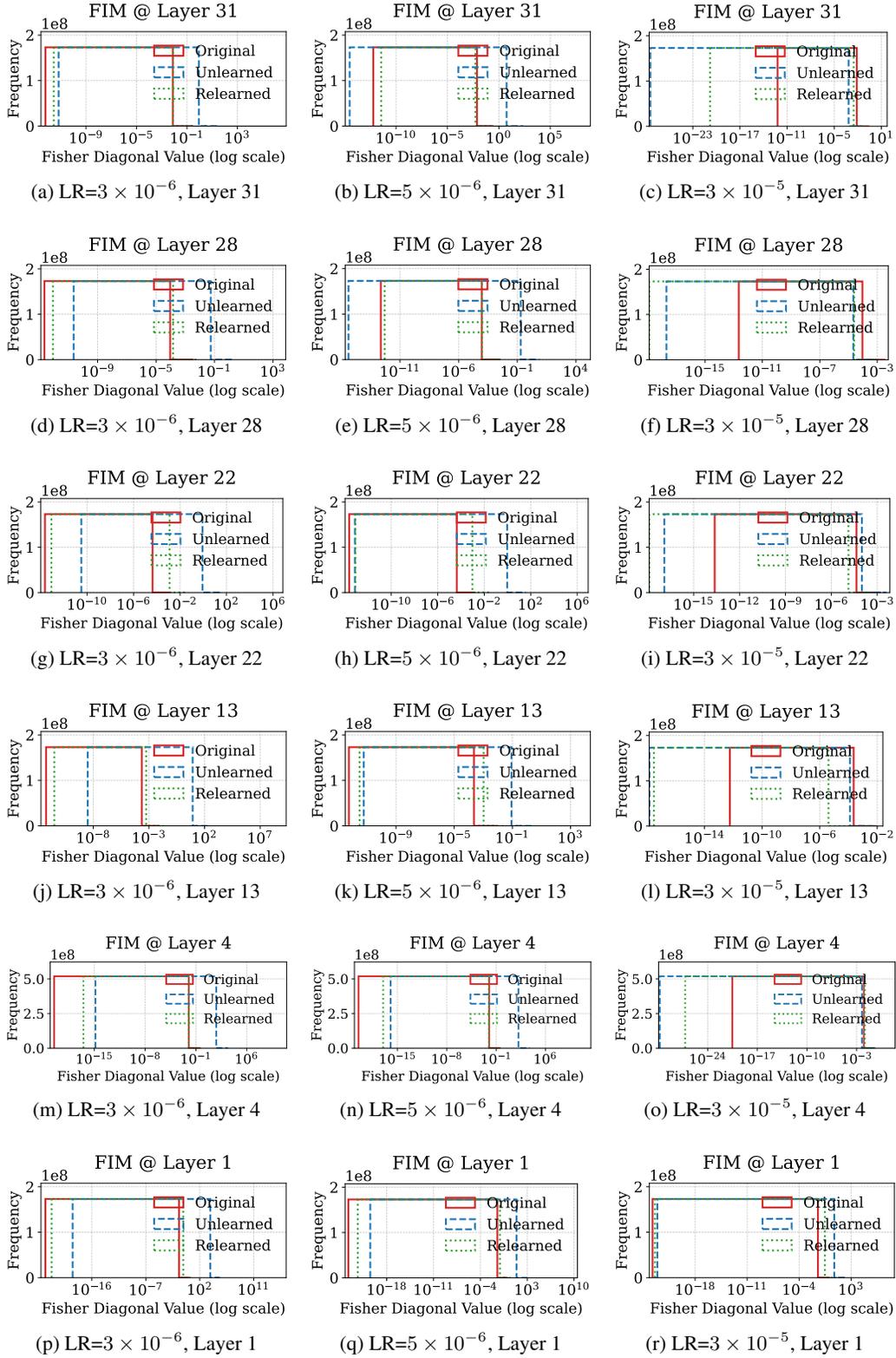


Figure 21: FIM for GA+KL Across Layers. All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

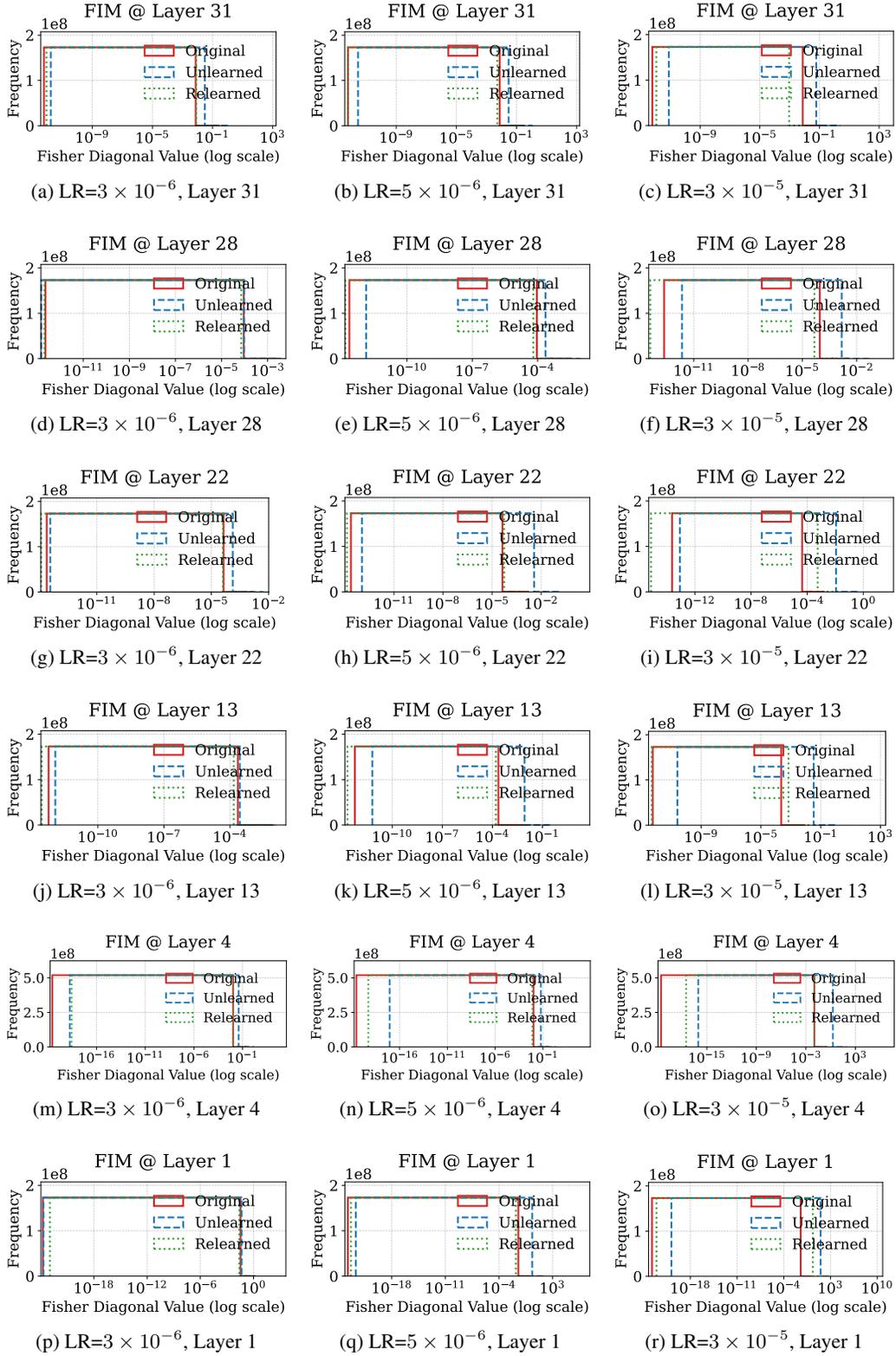


Figure 22: FIM for NPO Across Layers. All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

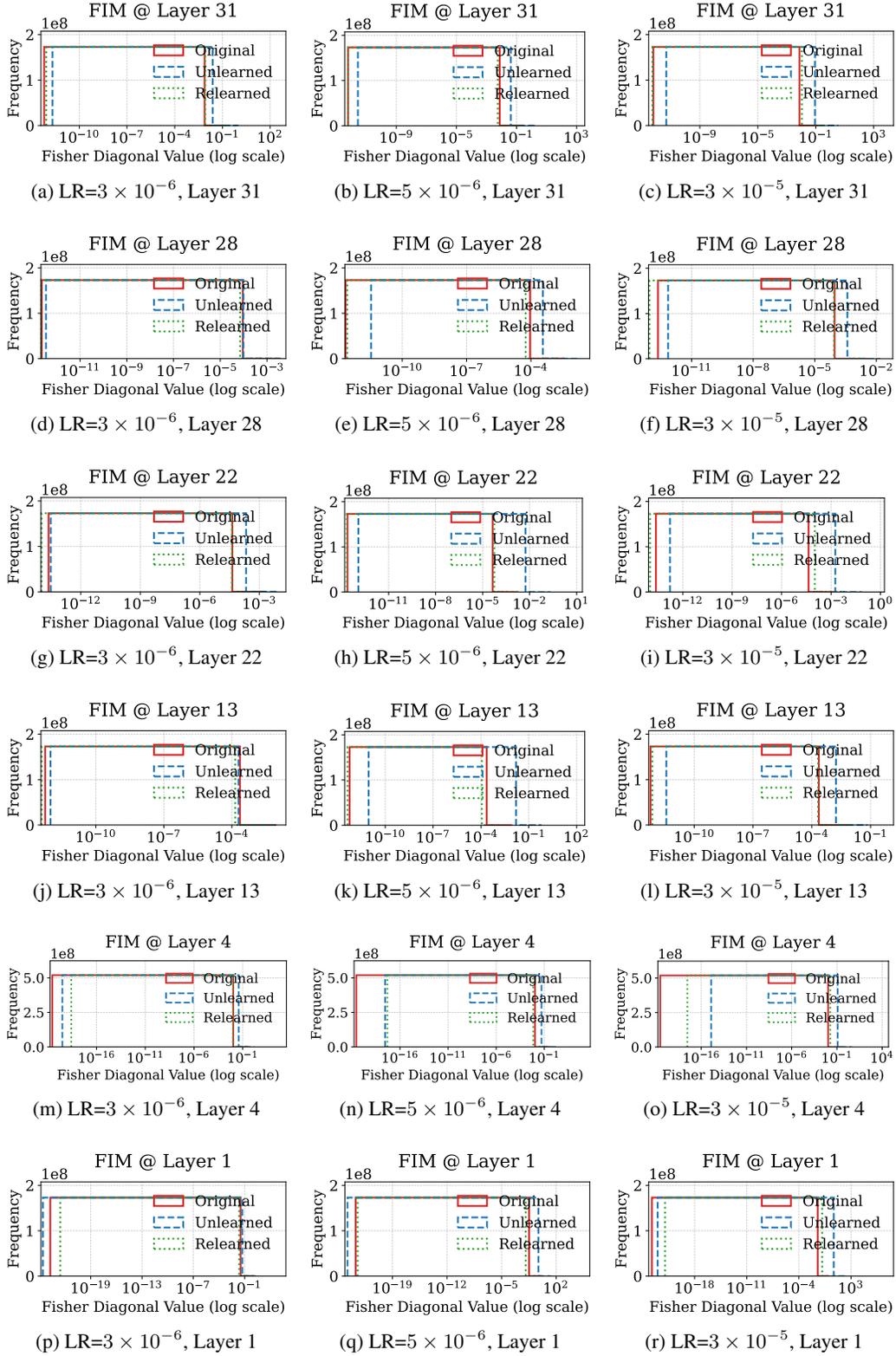


Figure 23: FIM for NPO+KL Across Layers. All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

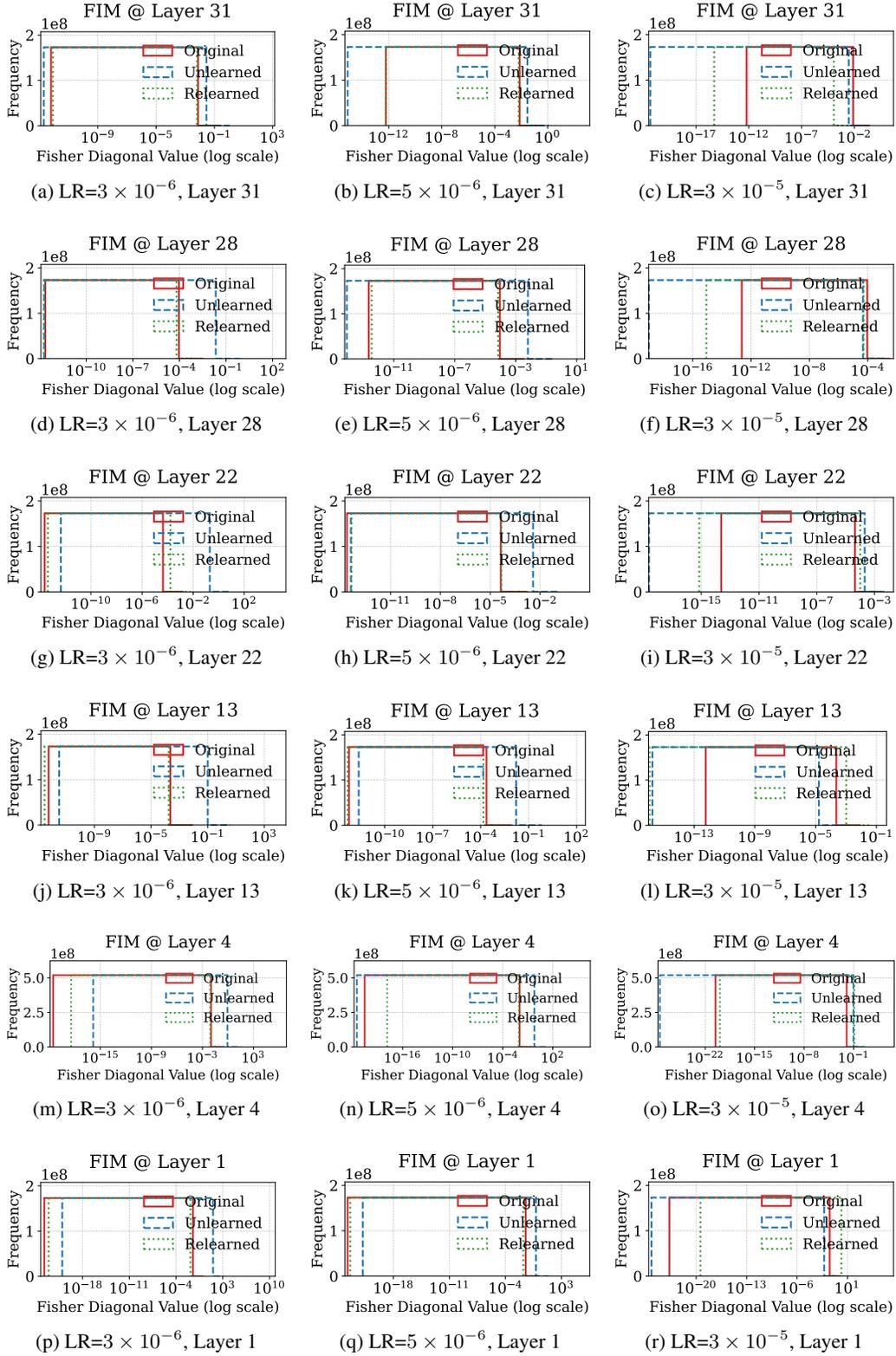


Figure 24: FIM for Rlable Across Layers. All plots are for the simple task on Yi-6B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 100$.

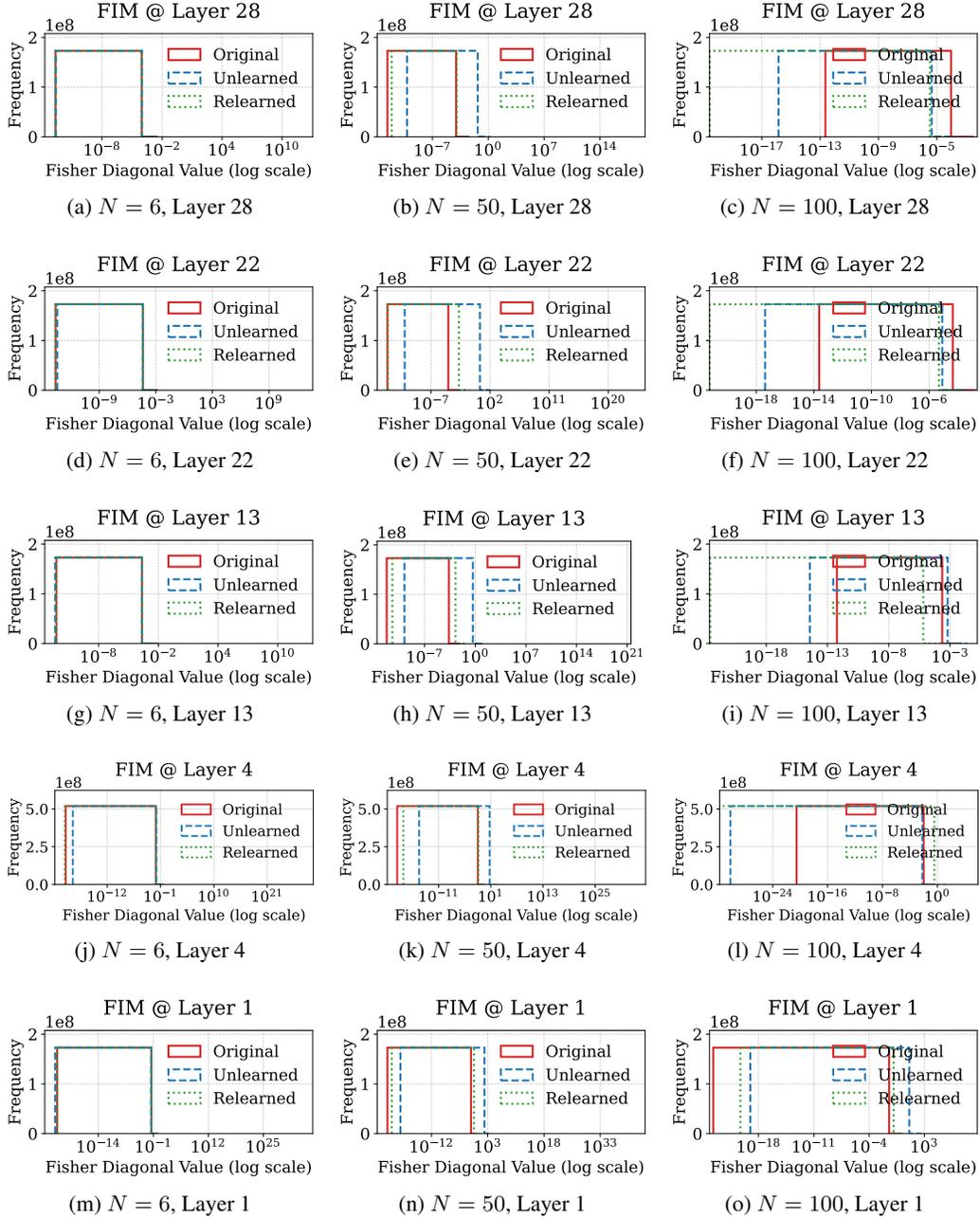


Figure 25: FIM for GA Across Layers. Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

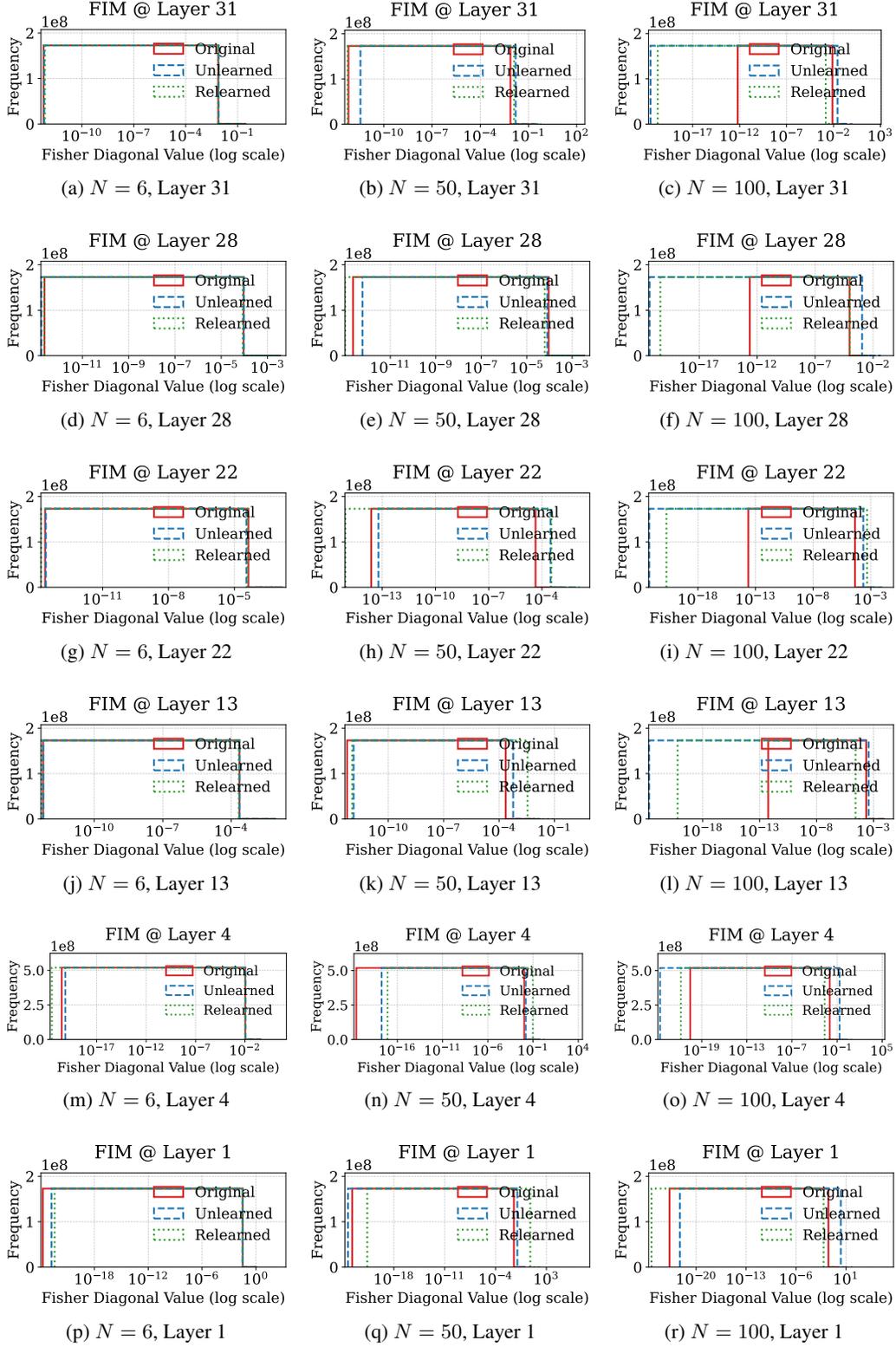


Figure 26: FIM for GA+GD Across Layers. Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

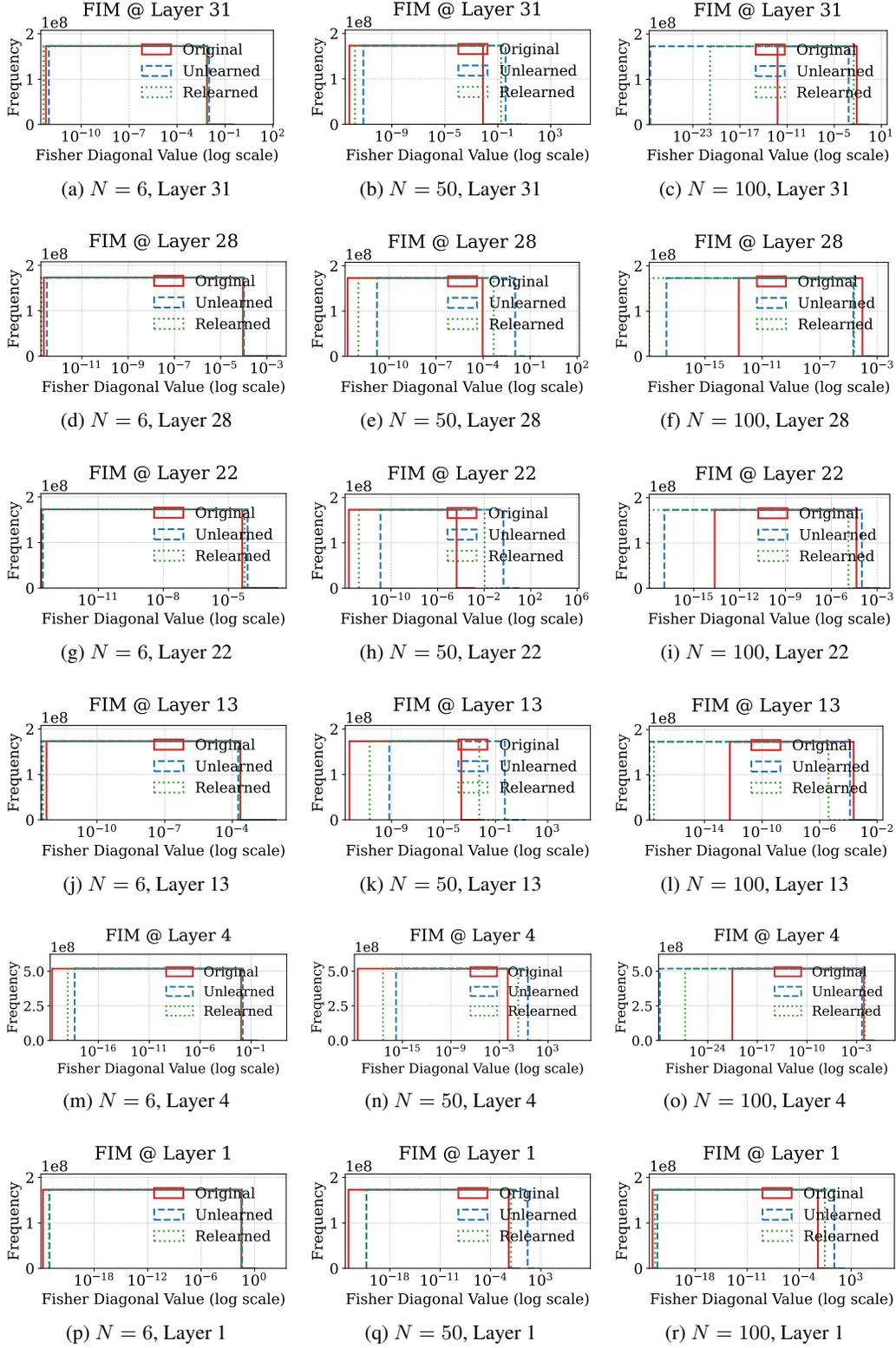


Figure 27: FIM for GA+KL Across Layers. Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

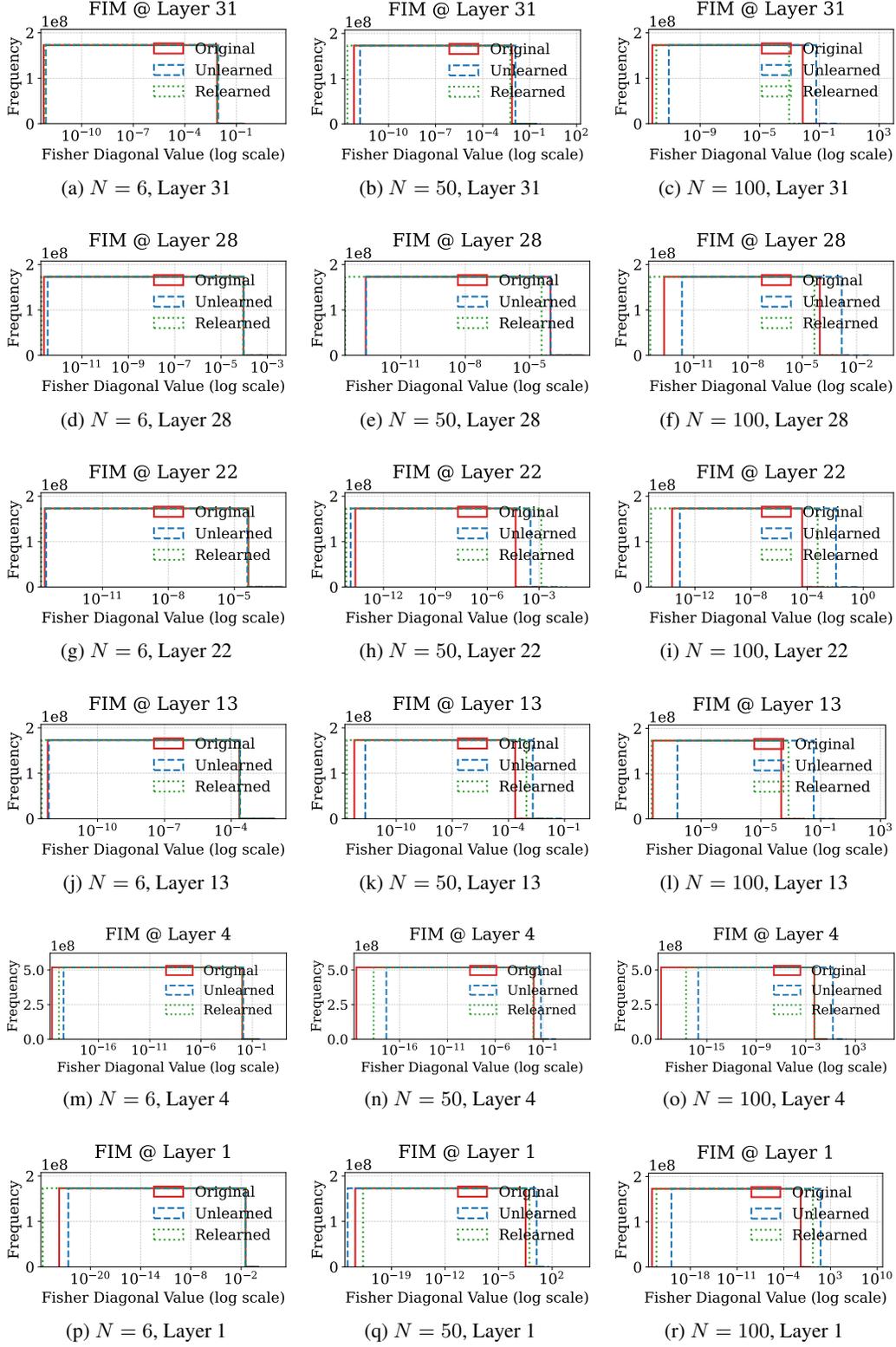


Figure 28: FIM for NPO Across Layers. Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

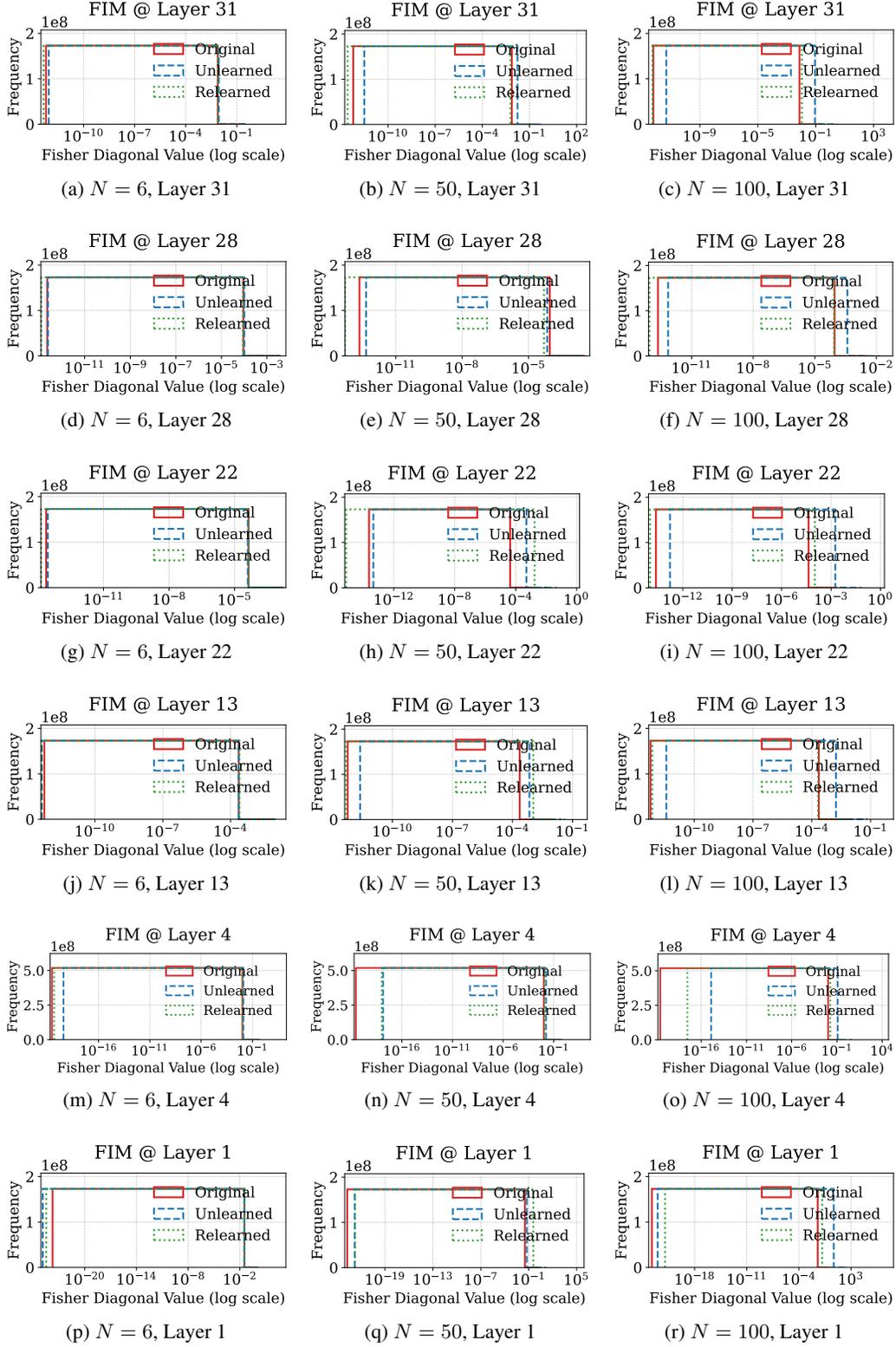


Figure 29: FIM for NPO+KL Across Layers. Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

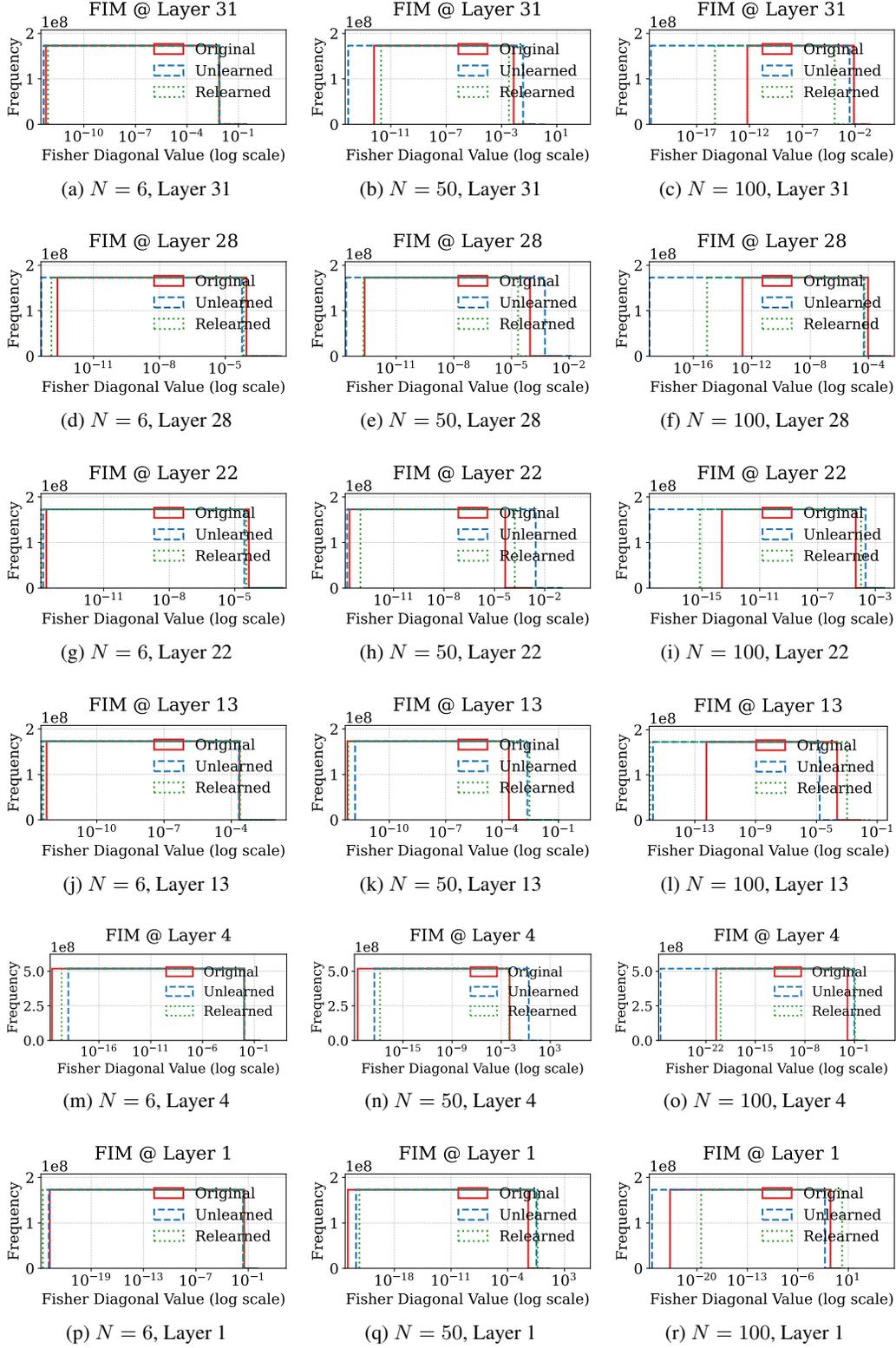


Figure 30: FIM for Rlable Across Layers. Simple task on Yi-6B with fixed learning rate $LR = 3 \times 10^{-5}$ and varying unlearning requests $N \in \{6, 50, 100\}$.

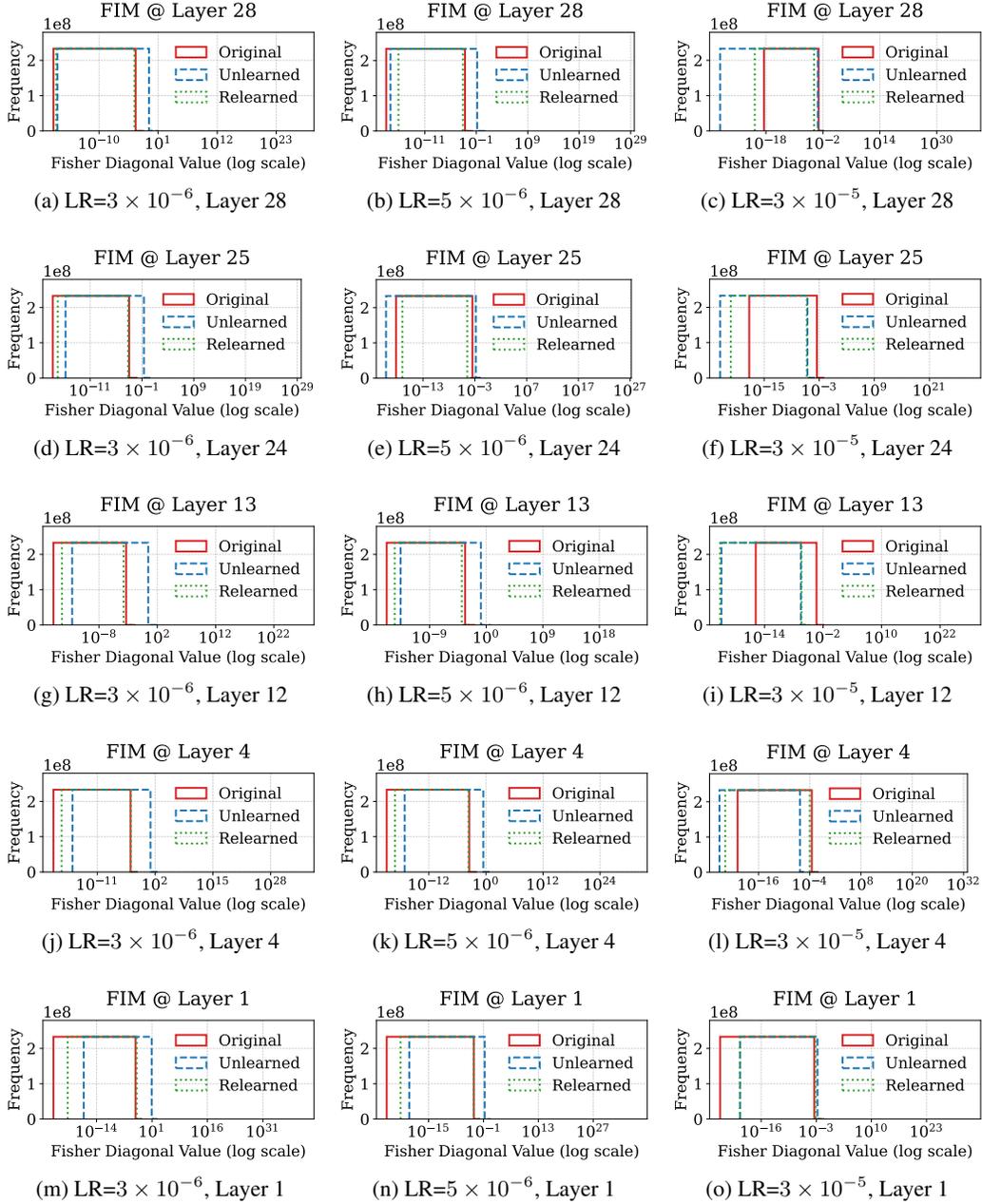


Figure 31: FIM for GA Across Layers. All plots are for the complex task on Qwen2.5-7B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 6$.

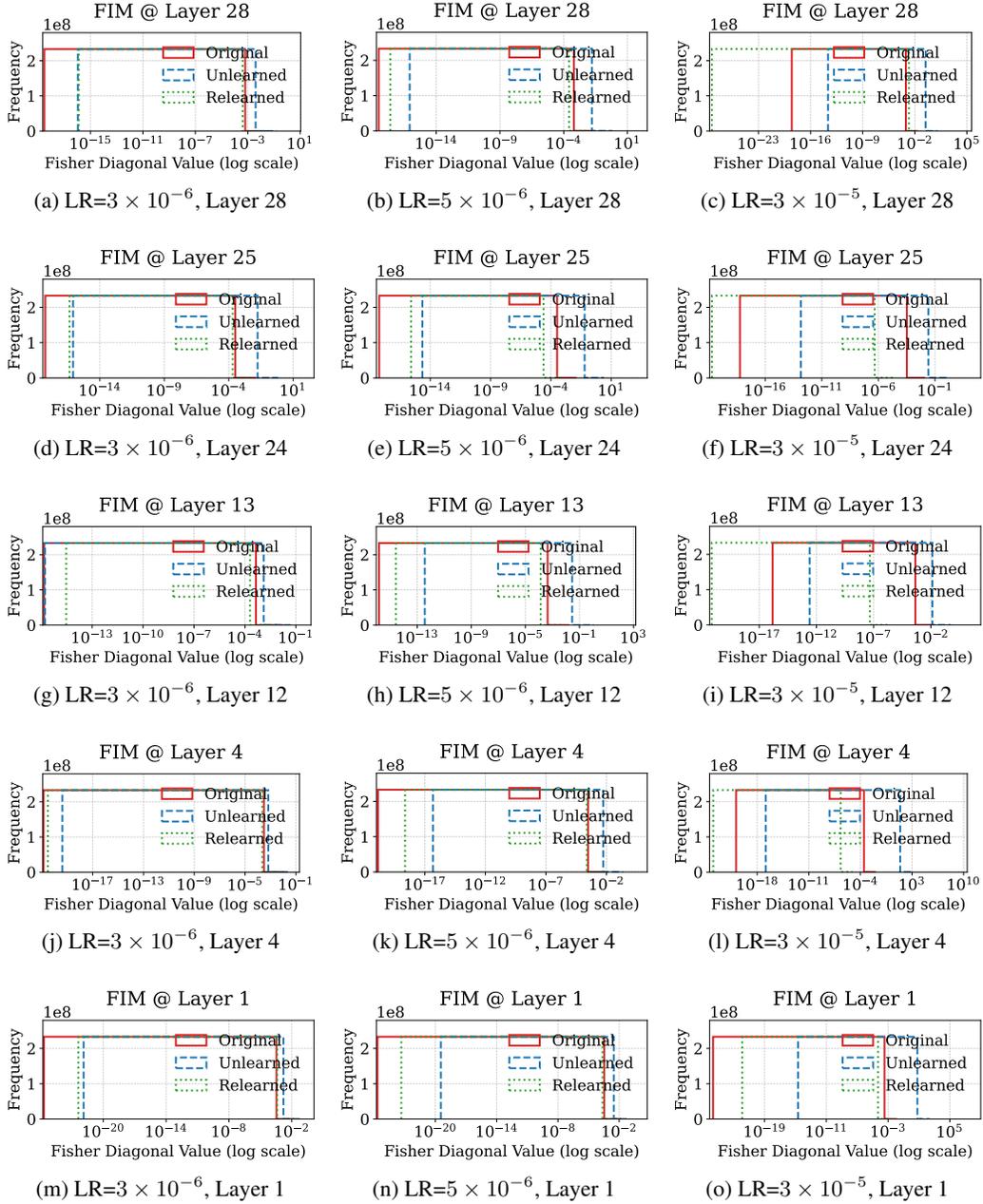


Figure 32: FIM for NPO Across Layers. All plots are for the complex task on Qwen2.5-7B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 6$.

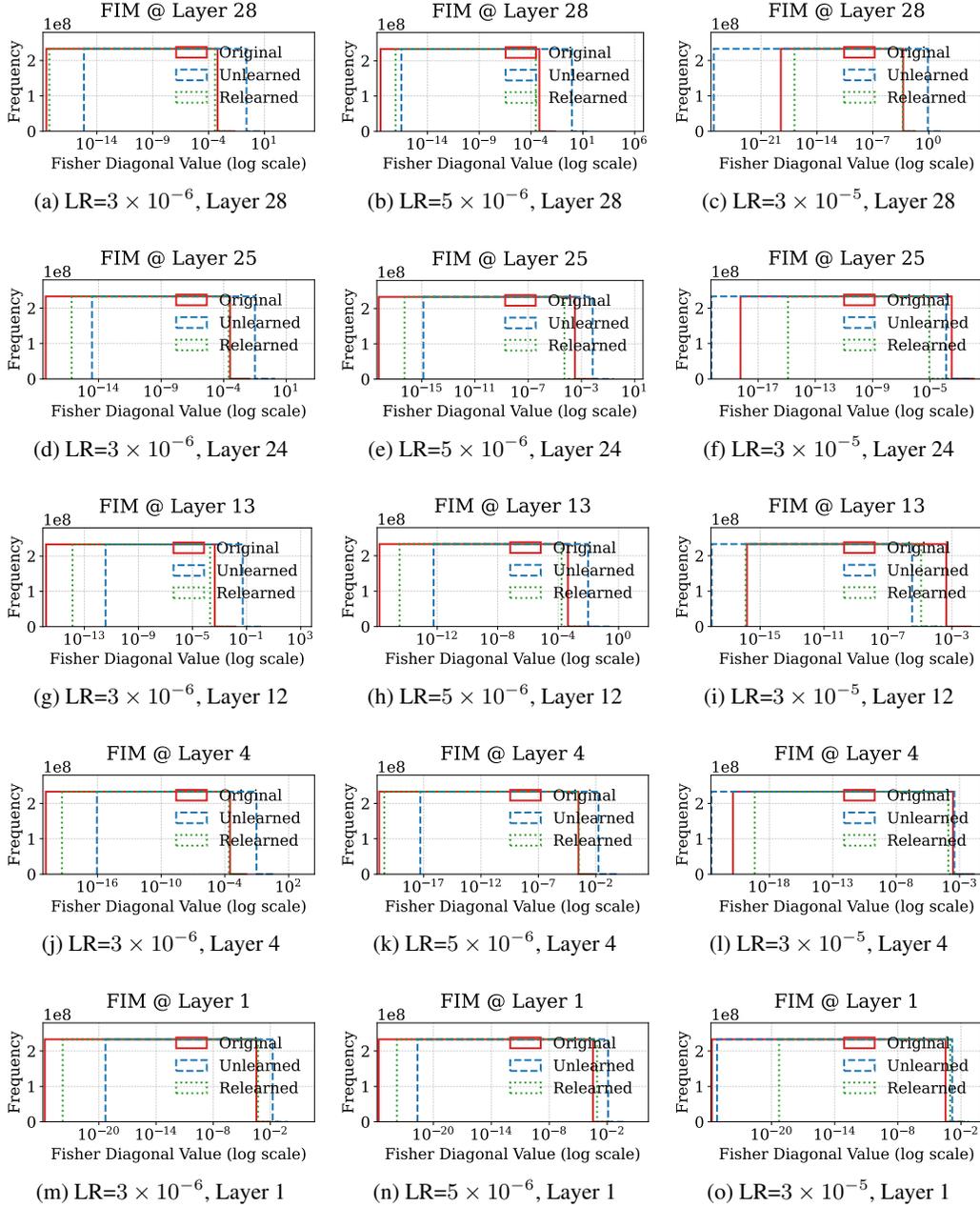


Figure 33: FIM for Rlable Across Layers. All plots are for the complex task on Qwen2.5-7B, using three learning rates $\{3 \times 10^{-6}, 5 \times 10^{-6}, 3 \times 10^{-5}\}$ and fixed $N = 6$.

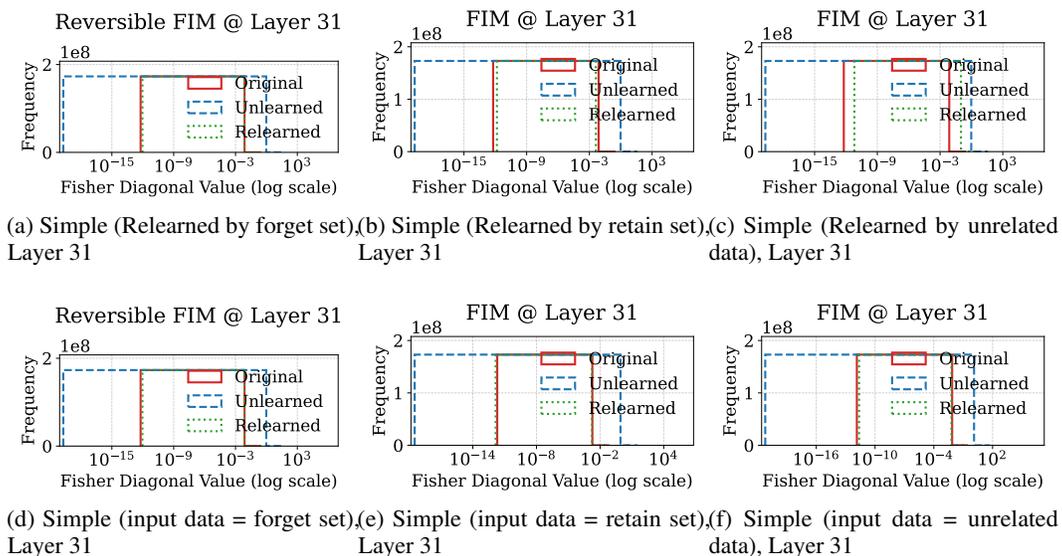


Figure 34: FIM in layer 31 under Varied Relearning and Evaluation Inputs on Yi-6B (Simple Task). (a–c): Relearning is performed using the forget set, retain set, or unrelated data respectively. (d–f): FIM is measured using the forget set, retain set, or unrelated data as evaluation input.

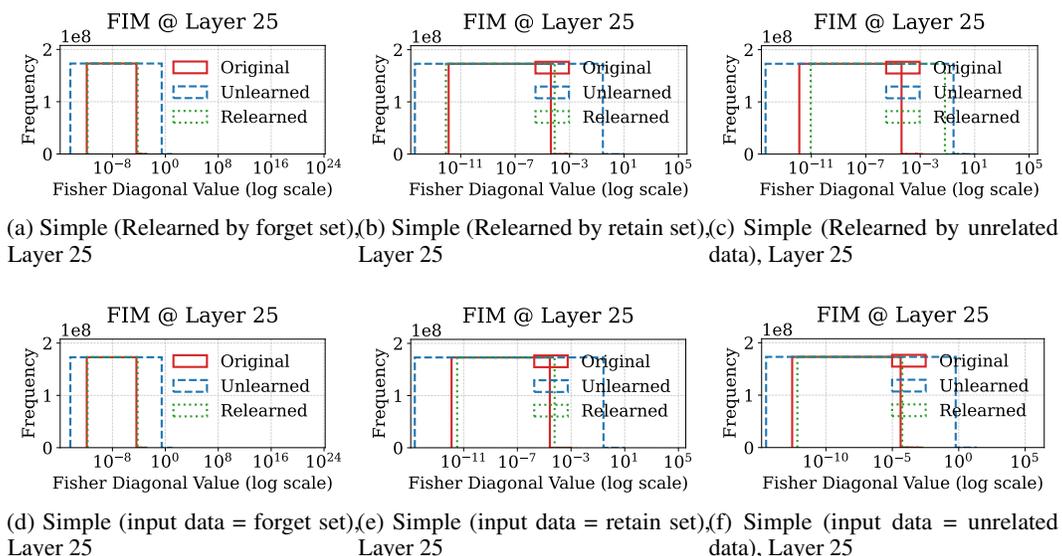
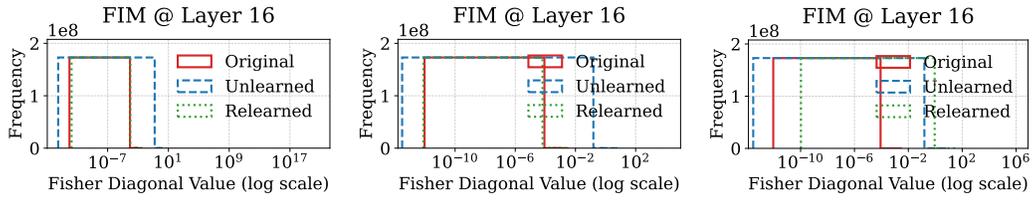
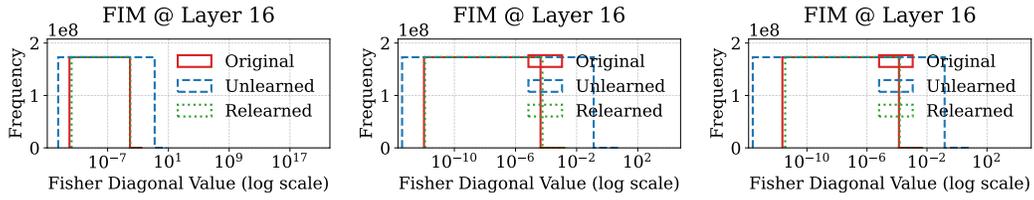


Figure 35: FIM in layer 25 under Varied Relearning and Evaluation Inputs on Yi-6B (Simple Task). (a–c): Relearning is performed using the forget set, retain set, or unrelated data respectively. (d–f): FIM is measured using the forget set, retain set, or unrelated data as evaluation input.

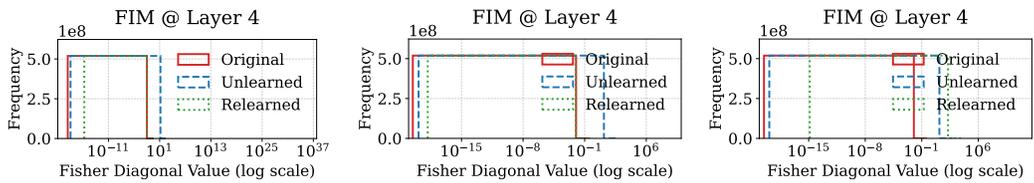


(a) Simple (Relearned by forget set), (b) Simple (Relearned by retain set), (c) Simple (Relearned by unrelated data), Layer 16

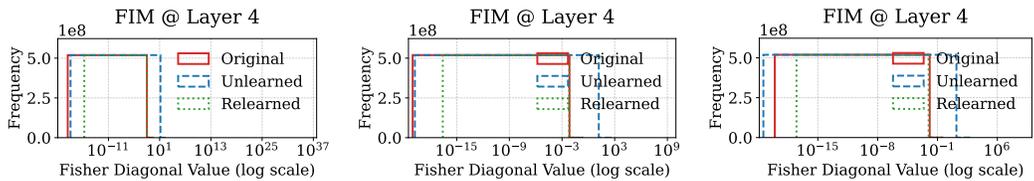


(d) Simple (input data = forget set), (e) Simple (input data = retain set), (f) Simple (input data = unrelated data), Layer 16

Figure 36: FIM in layer 16 under Varied Relearning and Evaluation Inputs on Yi-6B (Simple Task). (a–c): Relearning is performed using the forget set, retain set, or unrelated data respectively. (d–f): FIM is measured using the forget set, retain set, or unrelated data as evaluation input.



(a) Simple (Relearned by forget set), (b) Simple (Relearned by retain set), (c) Simple (Relearned by unrelated data), Layer 4



(d) Simple (input data = forget set), (e) Simple (input data = retain set), (f) Simple (input data = unrelated data), Layer 4

Figure 37: FIM in layer 4 under Varied Relearning and Evaluation Inputs on Yi-6B (Simple Task). (a–c): Relearning is performed using the forget set, retain set, or unrelated data respectively. (d–f): FIM is measured using the forget set, retain set, or unrelated data as evaluation input.

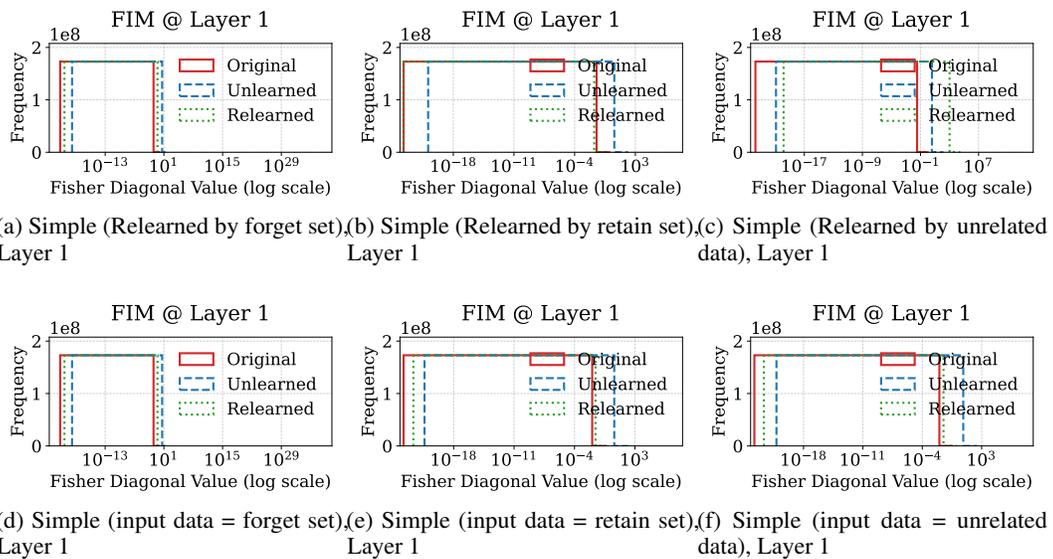


Figure 38: FIM in layer 1 under Varied Relearning and Evaluation Inputs on Yi-6B (Simple Task). (a–c): Relearning is performed using the forget set, retain set, or unrelated data respectively. (d–f): FIM is measured using the forget set, retain set, or unrelated data as evaluation input.