UNREAL UNLEARNING: BEHAVIORAL FORGETTING WITHOUT KNOWLEDGE ERASURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to remove targeted knowledge from LLMs, but current benchmarks mainly test whether the model still gives the correct answer, not whether the underlying knowledge is gone. We study *unreal unlearning*, where behavioral forgetting occurs while task-relevant information remains linearly decodable from hidden states. Across eight unlearning methods on the WMDP-Bio domain, accuracy falls to near chance while substantial internal evidence for the correct answer remains. We introduce *Knowledge Lens*, a framework which evaluates sets of `True/False` claims, and reads two signals, output-logit margin and a linear probe on the hidden state. Instantiated with WMDP-Bio, we find that 21.8–31.3% of questions are suppressed, with external evidence at or below chance while internal evidence remains above chance, for instance the gap between WMDP benchmark and the averaged internal measure is 60.8 percentage points. We vary unlearning methods and their checkpoints, three additional model architectures, an independently released nine-model method suite, all using models released by the original works, and extend to the WMDP-Cyber domain. In all these settings, Knowledge Lens consistently reveals substantial internal-external separation even when behavioral performance approaches chance. These results show that behavioral forgetting can substantially overstate knowledge removal and that unlearning should be evaluated through internal, interpretable approaches.

1 INTRODUCTION

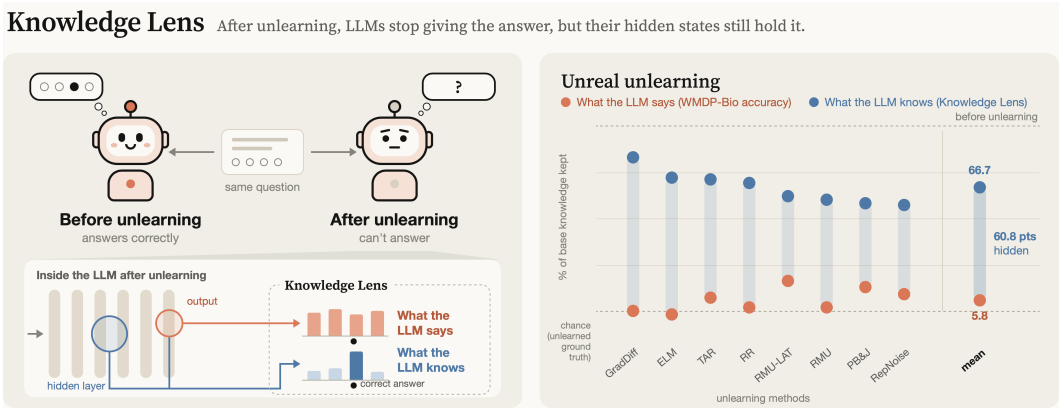

Figure 1: **The Knowledge Lens Framework and the Observed Unreal Unlearning**

LLMs are trained on vast and heterogeneous corpora and inevitably learn information that may later need to be removed. The target may be narrow, such as a password, credential, private fact, or copyrighted content (Jang et al., 2023), or much broader, such as hazardous biological knowledge (Li et al., 2024). Retraining a model from scratch each time such information must be removed is often impractical (Bourtoule et al., 2021). Machine unlearning addresses this problem by removing targeted information from a trained model while preserving its remaining capabilities (Cao & Yang, 2015; Liu et al., 2025; Dorna et al., 2026).

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

Removing a behavior, however, is not the same as removing the information that produced it. Unlearning is therefore evaluated not only through benchmark accuracy, but also through relearning, adversarial elicitation, model tampering, activation interventions, and latent probing (Lynch et al., 2024; Hong et al., 2024; Hu et al., 2025; Che et al., 2025; Gao et al., 2026). These evaluations have established that apparently forgotten information can survive unlearning, but they capture different notions of recovery. Fine-tuning and tampering modify the model, adversarial prompting changes how information is elicited, and probing measures what can be decoded from internal representations. What remains less directly characterized is the discrepancy already present in the post-unlearning model between what it expresses and what its internal representations still encode, measured for the same content under the same evaluation context.

In this work, we study *unreal unlearning* (Figure 1), where behavioral evaluation indicates forgetting while task-relevant information remains internally decodable. We introduce the *Knowledge Lens*, an unlearning evaluation framework that evaluates sets of `True/False` claims, measures external evidence from the model's output-logit margin and internal evidence with a held-out linear probe on the hidden state from the same forward pass and the same token index, and maps both to the aligned pairwise internal and external scores.

We apply the Knowledge Lens across eight unlearning methods and their trajectories on LLAMA-3-8B-INSTRUCT (Grattafiori et al., 2024) using WMDP-Bio (Li et al., 2024). Although conventional WMDP-Bio accuracy falls from 0.70 to 0.24–0.32, near the 0.25 chance level, our internal score remains 0.68–0.76 against a pairwise chance level of 0.5. At the final checkpoints, 21.8–31.3% of questions are suppressed, external evidence continues to decline after internal decodability stabilizes, and the gap between WMDP-Bio accuracy and the internal measure is 60.8 percentage points on average on a common chance-to-base scale.

We further test whether this separation extends to additional settings using RMU models released with WMDP (Li et al., 2024) for ZEPHYR-7B (Tunstall et al., 2023), MIXTRAL-8×7B-INSTRUCT (Jiang et al., 2024), and YI-34B-CHAT (Young et al., 2024), as well as a separate suite of nine Zephyr-7B unlearning models from OPTML, spanning NPO (Zhang et al., 2024) and GradDiff (Liu et al., 2022) with their smoothness variants (Fan et al., 2025) and SimNPO (Fan et al., 2026). Suppressed Knowledge remains positive across all three additional architectures (9.7–18.5 percentage points), all nine models in the independent method suite (17.4–33.2 percentage points), and WMDP-Cyber, extending the result to a second WMDP domain.

The paper's main contributions are:

1. **Unreal Unlearning.** We show that unlearning can suppress the model's output while task-relevant information remains linearly decodable from its hidden states, and that this pattern persists across the unlearning trajectory.

2. **Knowledge Lens.** We propose a matched evaluation that reads external evidence from the model's output logits and internal evidence from a held-out linear probe on the same forward pass and token position using a common pairwise scale.

3. **Suppressed Knowledge.** We define and quantify the internal–external separation through Suppressed Knowledge and item-level states, revealing cases where task-relevant information remains internally decodable despite weak external expression.

4. **Systematic and quantitative evidence of suppression.** Across unlearning methods, training checkpoints, model architectures and target domains, Knowledge Lens reveals a consistent internal-external separation: substantial task-relevant information remains internally decodable even when behavioral performance approaches chance.

We begin by providing the necessary related work and background in Section 2. In Section 3, we introduce the Knowledge Lens evaluation framework for discovery of suppressed knowledge throughout the unlearning process. In Section 4, we present our experimental settings, evaluation protocol, and apply Knowledge Lens across methods, architectures and target domains. Finally, in Section 5, we conclude with a discussion of our findings and their implications.

2 Related Work

LLM Unlearning Methods. Unlearning began with the goal of exact deletion: producing a model equivalent to one retrained without the deleted examples, a formulation that is tractable for small classifiers (Cao & Yang, 2015; Bourtoule et al., 2021). In LLMs, the target has shifted from identifiable training examples to facts, concepts, or capabilities, and successful removal is typically evaluated through model behavior (Eldan & Russinovich, 2023; Li et al., 2024). More recently, shared datasets, benchmarks, and released models have enabled increasingly direct comparisons across unlearning methods (Dorna et al., 2026). Unlearning methods differ substantially in where and how they intervene in the model. Output-level methods optimize the output distribution, for example through a loss gap between forget and retain sets or through preference objectives (Liu et al., 2022; Yao et al., 2024; Zhang et al., 2024; Fan et al., 2026). Representation-level methods act on internal states, steering activations elicited by hazardous inputs toward random targets, noise, or orthogonal directions (Li et al., 2024; Rosati et al., 2024; Zou et al., 2024; Shen et al., 2026). Such interventions can be learned through training or applied directly at inference time, for example through sparse autoencoder clamping (Farrell et al., 2024). Weight-level methods directly edit model parameters to achieve a behavioral target, using, for example, curvature-informed second-order updates or targeted low-rank edits (McKinney et al., 2026; Gandikota et al., 2026). Separately, robustness objectives can be layered onto any of these methods (Sheshadri et al., 2024; Tamirisa et al., 2025), including approaches that vary the smoothness regularizer while holding the underlying unlearning objective fixed (Fan et al., 2025). These differences motivate evaluating unlearning independently of any particular intervention mechanism. Our main evaluation spans all three intervention families.

LLM Unlearning Evaluations. Most unlearning evaluations assess information through the model's observable behavior, such as outputs, predictions or loss values, rather than through its internal representations. Behavioral benchmarks directly test the post-unlearning model on held-out queries (Li et al., 2024; Dorna et al., 2026). Weight-space recovery modifies the model through relearning, benign fine-tuning, or model tampering before re-evaluating its behavior (Hu et al., 2025; Che et al., 2025; Fan et al., 2025; Xu et al., 2025). Input-space recovery instead changes how information is elicited, using jailbreaks, paraphrases, or multi-hop prompts (Lynch et al., 2024; LeVi et al., 2025). Other approaches locally perturb the input and measure how the model's response changes across the resulting neighborhood (Cohen et al., 2026). Despite these differences, all of these approaches rely on external evidence, even when the evaluation procedure itself modifies the model or its inputs.

This distinction is not unique to unlearning. A gap between what a model encodes internally and what it expresses externally has been observed for factual knowledge (Gekhman et al., 2025), for alignment that suppresses biases the model still encodes (Himelstein et al., 2026), and for models trained to retain a fact while denying it (Cywiński et al., 2025). Within unlearning, such a gap can be studied through internal evidence, asking whether the target remains decodable from intermediate representations using linear probes, information decomposition or activation patching (Alain & Bengio, 2017; Gao et al., 2026; Goel et al., 2026; Hong et al., 2024; Lee et al., 2026). The evidence is mixed: Li et al. (2024) report that probes cannot recover unlearned knowledge, while Goel et al. (2026) find forget-set membership linearly decodable by contrasting the base and unlearned representations.

Prior work therefore provides internal evidence at different levels: Li et al. (2024) directly probe whether the correct answer remains recoverable after behavioral unlearning, whereas Goel et al. (2026) probe whether forget-set membership remains encoded in the representations. Neither directly compares internal and external evidence for the same candidate under the same elicitation and on a common scale. Our Knowledge Lens provides this matched comparison, allowing us to identify cases in which external evidence has fallen to chance while answer-relevant information remains internally decodable.

3 Knowledge Lens: Internal-External Unlearning Evaluation

We define the Knowledge Lens as a matched characterization of how externally expressed and internally decodable task-relevant evidence changes under unlearning. The key design principle is to hold the evaluated content fixed: for each candidate, the external and internal scores are obtained from the same prompt, forward pass, and token position, differing only in where the evidence is read.

Given an evaluation set in which each item has a designated target candidate and fixed distractors, the Knowledge Lens proceeds in four steps: (i) It casts every item-candidate pair as a True/False verification claim and evaluates that claim with a single forward pass; (ii) scores the candidate externally using the True/False output-logit margin and internally using a held-out linear probe applied to the hidden state at the same final non-padding token position; (iii) converts both scores into matched pairwise ranking measures, K_{ext} and K_{int}, defined as the fraction of distractors that the target candidate outranks (Gekhman et al., 2025); and (iv) reports the two axes jointly, their per-item gap $G = K_{\text{int}} - K_{\text{ext}}$, the aggregate *suppressed knowledge* metric given by the mean gap over items, item-level states, and their evolution across unlearning checkpoints. In multiple-choice question (MCQ) dataset instantiation, an evaluation item is the question, the target candidate is the correct option, and the three remaining options are distractors.

A positive gap indicates that more target-relevant evidence is linearly decodable from the selected internal representation than is expressed by the model's output readout. We use the term *suppressed* for items with $K_{\text{int}} > 1/2$ and $K_{\text{ext}} \leq 1/2$. These terms describe measured internal-external readout patterns. This terminology does not imply that K_{int} is a model-independent measure of knowledge or that low K_{int} certifies erasure. The remainder of this section formally defines these quantities.

3.1 Matched Internal and External Readouts

Shared input and forward pass. Let $\mathcal{Q}$ denote a set of evaluation items. Each item $q \in \mathcal{Q}$ has a designated target candidate $a^*(q)$ and a set of distractors $\mathcal{W}(q)$. We write $\mathcal{A}(q) = \{a^*(q)\} \cup \mathcal{W}(q)$ for the full candidate set, and $y(q, a) = \mathbf{1}[a = a^*(q)]$ for the target label of candidate $a \in \mathcal{A}(q)$.

For each candidate a, we construct a binary verification claim relating a to item q and insert it into a fixed `True/False` prompt, denoted $p(q, a)$. For MCQ datasets such as WMDP, this claim takes the form *"The answer to '⟨q⟩' is '⟨a⟩'."*. We denote a model checkpoint by (m, t), where m indexes the unlearning method and t the checkpoint. A single forward pass of $p(q, a)$ through (m, t) produces, at the final non-padding token position, an output-logit vector $\mathbf{z}(q, a; m, t) \in \mathbb{R}^{|V|}$, where V is the model vocabulary, and layerwise hidden states $\mathbf{h}^{(\ell)}(q, a; m, t) \in \mathbb{R}^d$ for $\ell \in \{0, \ldots, L-1\}$.

This matched evaluation keeps the evaluated input and forward pass fixed, reducing the possibility that differences between the two measurements arise from changes in elicitation format. The exact verification prompt is provided in Appendix A. Answer-token and tokenization details are given in Appendix B.

External readout. We measure externally expressed target evidence using the model's own output head. Let $z_{\texttt{True}}(q, a; m, t)$ and $z_{\texttt{False}}(q, a; m, t)$ denote the components of $\mathbf{z}(q, a; m, t)$ corresponding to the designated `True` and `False` response tokens. We define the external score as their logit margin:

$$s_{\text{ext}}(q, a; m, t) = z_{\texttt{True}}(q, a; m, t) - z_{\texttt{False}}(q, a; m, t). \tag{1}$$

Larger s_{ext} indicates stronger support from the model's output readout for the claim that a is the target candidate. We use the logit margin rather than generated-label accuracy so that the external score remains well-defined even when generation quality degrades, following the logit-based scoring of Li et al. (2024), Che et al. (2025) and Gekhman et al. (2025).

Internal readout. We measure internally decodable target-relevant evidence using a logistic-regression probe trained to predict $y(q, a)$ from an intermediate hidden state. For each model checkpoint (m, t), probes are evaluated using 5-fold cross-validation at the item level, with all candidates in $\mathcal{A}(q)$ assigned to the same fold. Within each fold, logistic-regression probes are fitted at each transformer layer and regularization setting using the training items. The layer-regularization pair with the highest validation AUC is selected, refitted on the combined training and validation items, and evaluated on the held-out items.

Let $f^{(m,t)} : \mathbb{R}^d \to \mathbb{R}$ denote the resulting logistic-regression decision function, and let ℓ^* denote its validation-selected layer. For notational simplicity, $f^{(m,t)}$ and ℓ^* refer to the fold-specific probe and selected layer used for the held-out item being scored. The internal score is

$$s_{\text{int}}(q, a; m, t) = f^{(m,t)}\left(\mathbf{h}^{(\ell^*)}(q, a; m, t)\right). \tag{2}$$

Consequently, every reported internal score is out-of-fold: the evaluated item is used neither to fit the probe nor to select its layer or regularization. Exact cross-validation splits, preprocessing, hyperparameter grids, and implementation details are provided in Appendix B.

3.2 From Readouts to the Knowledge Lens

Matched pairwise scores. To place the two readouts on the same scale, we convert each scoring function into a pairwise ranking score. For a score $S \in \{s_{\text{ext}}, s_{\text{int}}\}$, define

$$K(q; S, m, t) = \frac{1}{|\mathcal{W}(q)|} \sum_{w \in \mathcal{W}(q)} \mathbf{1}[S(q, a^*(q); m, t) > S(q, w; m, t)]\,. \tag{3}$$

Thus, $K(q; S, m, t)$ is the fraction of distractors that the target candidate outranks. We write $K_{\text{ext}} = K(q; s_{\text{ext}}, m, t)$ and $K_{\text{int}} = K(q; s_{\text{int}}, m, t)$. Accordingly, K_{ext} measures target ranking at the model output, whereas K_{int} measures the same ranking from the selected internal representation. For WMDP-Bio, each question has three distractors, so both scores lie in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$ and have pairwise chance level $\frac{1}{2}$.

Suppressed Knowledge. The per-item internal-external gap is $G(q; m, t) = K_{\text{int}}(q; m, t) - K_{\text{ext}}(q; m, t)$. We define suppressed knowledge (SK) over an item set $\mathcal{Q}' \subseteq \mathcal{Q}$ as

$$\text{SK}(m, t; \mathcal{Q}') = \frac{1}{|\mathcal{Q}'|} \sum_{q \in \mathcal{Q}'} G(q; m, t). \tag{4}$$

Positive SK indicates that, on average, more target-relevant evidence is linearly decodable internally than is expressed by the model's output readout. Because SK is signed, opposite internal-external patterns can partially cancel. Therefore, we report it together with the two axes and the item-level states.

Item-level states. Thresholding both axes at the pairwise chance level $\frac{1}{2}$ yields four descriptive states: *Retained* ($K_{\text{int}} > \frac{1}{2}$, $K_{\text{ext}} > \frac{1}{2}$), *Suppressed* ($K_{\text{int}} > \frac{1}{2}$, $K_{\text{ext}} \leq \frac{1}{2}$), *Forgotten* ($K_{\text{int}} \leq \frac{1}{2}$, $K_{\text{ext}} \leq \frac{1}{2}$), and *Lucky* ($K_{\text{int}} \leq \frac{1}{2}$, $K_{\text{ext}} > \frac{1}{2}$). These labels denote regions of the measured $(K_{\text{ext}}, K_{\text{int}})$ space, in particular, "forgotten" does not certify representational erasure.

4 Experiments

4.1 Experimental Settings

Dataset and models. Our main experiments apply the Knowledge Lens evaluation framework to WMDP-Bio (Li et al., 2024), a benchmark comprising 1,273 biosecurity multiple-choice questions, each with one correct answer and three distractors. Unless stated otherwise, all experiments use Llama-3-8B-Instruct (Grattafiori et al., 2024), with hidden dimension $d = 4096$ and $L = 33$ representation levels when counting the embedding representation and 32 transformer blocks. To test whether our findings are specific to this model, we report in §4.5 a cross-architecture evaluation of WMDP's independently released RMU models and extend the evaluation to the Cyber domain in §4.6.

Unlearning methods and checkpoints. We evaluate publicly released models rather than performing the unlearning ourselves, preserving each method's original implementation. We evaluate eight unlearning methods released through LLM-GAT (Che et al., 2025): GradDiff (Liu et al., 2022), RMU (Li et al., 2024), RMU-LAT (Sheshadri et al., 2024), RepNoise (Rosati et al., 2024), ELM (Gandikota et al., 2026), RR (Zou et al., 2024), TAR (Tamirisa et al., 2025) and PB&J (Che et al., 2025; McKinney et al., 2026). Each method is released at eight checkpoints along the unlearning process, yielding $8 \times 8 = 64$ unlearned model instances. All trajectories originate from the same base model, providing a controlled setting in which the architecture, pre-unlearning model, and evaluation data are fixed while the unlearning intervention varies. We denote the shared base model by $t = 0$ and the released checkpoints by $t \in \{1, \ldots, 8\}$, with $t = 8$ the final checkpoint. Unless stated otherwise, endpoint analyses use $t = 8$, whereas trajectory analyses evaluate the complete

sequence $t = 0, \ldots, 8$ for each method. An extension to an additional suite of unlearning methods is summarized in Appendix C.

Evaluation protocol. Unless otherwise stated, analyses use all WMDP-Bio questions (Li et al., 2024). We report conventional WMDP-Bio accuracy as the standard behavioral reference and evaluate the matched external and internal readouts K_{ext} and K_{int} defined in §3.2. All internal scores are computed out of fold using the five-fold probing protocol described there, such that the questions used to evaluate a probe are not used to fit the probe or select its layer and regularization. Exact prompts, cross-validation splits, preprocessing, hyperparameters, tokenization, and implementation details are provided in Appendix B. Retain-set evaluations are summarized, for all experiments, in Appendix D.

4.2 Behavioral Forgetting Leaves Substantial Internal Evidence

We first examine the final unlearning checkpoints. Figure 2 compares WMDP-Bio accuracy with the matched external and internal readouts, K_{ext} and K_{int}. Under conventional four-way WMDP-Bio evaluation, where chance is 0.25 (Li et al., 2024), Che et al. (2025) report final accuracies of 0.24–0.32 across the eight methods, compared with 0.70 for the shared base model. Thus, all eight methods reach a narrow behavioral regime near chance.

The matched readouts reveal substantially more remaining evidence. K_{ext} falls from 0.78 in the base model to 0.53–0.64 after unlearning, confirming a strong reduction in externally expressed correct-answer evidence. In contrast, K_{int} falls from 0.81 only to 0.68–0.76. No method therefore drives the internal readout to its 0.5 chance level. Even the lowest value, obtained by RepNoise (0.68), retains more than half of the base model's above-chance internal signal.

Consequently, K_{int} exceeds K_{ext} for every method, with an endpoint gap of 8.1–15.0 percentage points. The gap is largest for TAR (15.0 pp), ELM (14.4 pp), and RepNoise (14.3 pp), and smallest for RMU-LAT (8.1 pp) and RMU (8.6 pp). This aggregate internal-external difference is the Suppressed Knowledge metric defined in Eq. 4.

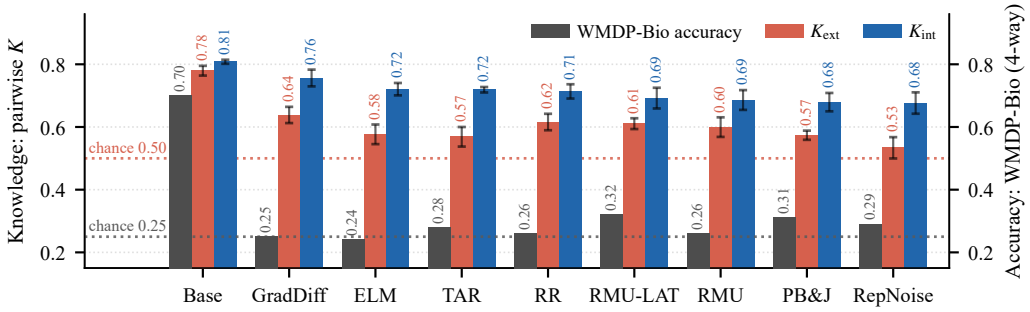

Figure 2: **Post unlearning: Knowledge Lens vs WMDP-Bio accuracy.** Whiskers show 5-fold standard deviation for the Knowledge Lens scores.

Because WMDP-Bio and the Knowledge Lens scores have different chance levels, we additionally normalize each metric by its own base-to-chance range, $R(M) = 100(M - M_{\text{chance}})/(M_{\text{base}} - M_{\text{chance}})$. Averaged across methods, WMDP-Bio retains only 5.8% of its above-chance base signal, compared with 32.1% for K_{ext} and 66.7% for K_{int} (Figure 3). Thus, substantial internal evidence remains even relative to the stronger matched external readout, rather than only relative to conventional multiple-choice accuracy.

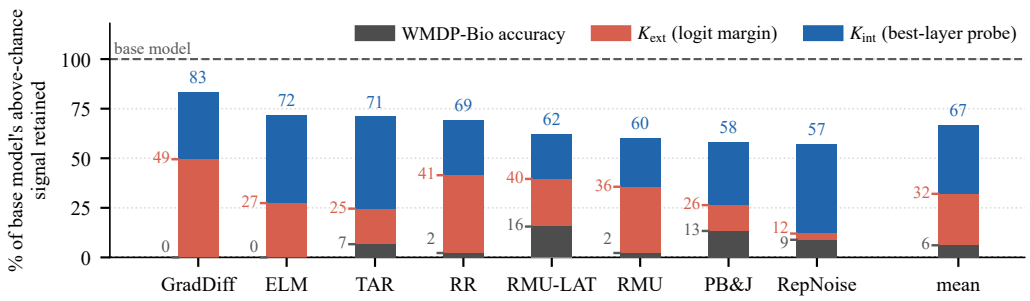

Figure 3: **Above-chance signal retained after unlearning.** Each metric is normalized to its own base-to-chance range. The rightmost bar is the mean across methods.

4.3 The Internal-External Gap Widens During Unlearning

The endpoint analysis does not reveal when the separation emerges. We therefore rescore all WMDP-Bio questions at every released checkpoint (Figure 4). The mean internal-external gap increases from 2.9 pp in the shared base model to 11.6 pp at the final checkpoint, showing that the endpoint discrepancy is induced substantially during unlearning rather than reflecting only a pre-existing difference between the two readouts. Across methods, K_{int} declines primarily during the early checkpoints, reaching 0.67–0.78 by $t = 3$-4 and changing comparatively little thereafter. K_{ext}, in contrast, continues to decline through later checkpoints, ending at 0.53–0.64. Thus, later behavioral suppression often proceeds without a comparable reduction in linearly decodable internal evidence. All eight methods' results are provided in Appendix E.

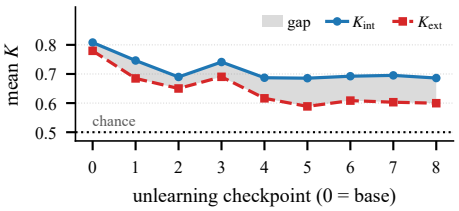

Figure 4: **Evidence throughout unlearning for the RMU method.**

4.4 The Gap Is Concentrated in Suppressed Questions

We next resolve the aggregate gap at the item level using the four states defined in §3.2. At $t = 8$, 21.8–31.3% of all questions are suppressed, meaning K_{int} remains above chance while K_{ext} does not (Table 1). Together with retained questions, this means that 69.1–78.1% of the evaluation set remains internally above chance after unlearning. By comparison, only 10.8–15.2% of questions are forgotten on both axes.

This suppressed population drives the aggregate gap. Weighting the per-question gap $G(q) = K_{\text{int}}(q) - K_{\text{ext}}(q)$ by state frequency, suppressed questions contribute 14.7–23.4 pp across methods, while lucky questions contribute -6.0 to -11.1 pp and retained and forgotten questions together contribute only 0.8–3.2 pp. These opposing contributions yield the net 8.1–15.0 pp suppressed knowledge reported in §4.2. The aggregate metric therefore understates the magnitude of suppression because opposite item-level patterns partially cancel.

4.5 The Internal-External Separation Replicates Across Architectures

Our primary experiments hold the architecture fixed while varying the unlearning method and checkpoint. We next test the complementary setting: holding the unlearning method fixed while varying the model architecture. Specifically, we apply the same Knowledge Lens evaluation to the RMU checkpoints released with WMDP (Li et al., 2024) for Zephyr-7B (Tunstall et al., 2023), Mixtral-8×7B-Instruct (Jiang et al., 2024), and Yi-34B-Chat (Young et al., 2024). This evaluation uses the released models directly, without retraining the models or modifying the framework.

Table 1: **Item-level states after unlearning.** Percentage of all WMDP-Bio questions in each state at $t = 8$, according to whether K_{int} and K_{ext} are above (+) or at/below (−) their 0.5 chance level.

State (%)	GradDiff	RMU	RMU-LAT	RepNoise	ELM	RR	TAR	PB&J
Retained: Int+, Ext+	56.2	44.9	45.9	37.8	46.7	48.2	46.5	42.7
Suppressed: Int+, Ext−	21.8	25.8	25.5	31.3	28.0	25.8	26.9	27.5
Forgotten: Int−, Ext−	12.0	13.1	11.9	14.8	13.0	10.8	15.2	14.5
Lucky: Int−, Ext+	9.9	16.3	16.7	16.1	12.3	15.2	11.4	15.4

Figure 5 shows the same internal-external separation across all three architectures. Relative to each model's own pre-unlearning baseline, conventional WMDP-Bio retains only 11–18% of its above-chance signal after RMU, while K_{ext} retains 0–27% and K_{int} retains 54–56%. Correspondingly, $K_{\text{int}} - K_{\text{ext}}$ remains positive for every model, ranging from 9.7 to 18.5 percentage points in the raw pairwise scores. For two of these checkpoints, Zephyr-7B and Yi-34B, Li et al. (2024) report that unlearned knowledge cannot be recovered by linear probes. The two readouts are not directly comparable: their probe is four-way, against a 0.25 chance level, whereas K_{int} is pairwise, against 0.5, and is matched to the model's own output readout. We therefore do not claim a direct contradiction, but rather that a readout matched to the external decision recovers substantially more evidence than the four-way probe reported. Thus, the separation observed in our LLAMA-3-8B-INSTRUCT experiments is not specific to a single architecture.

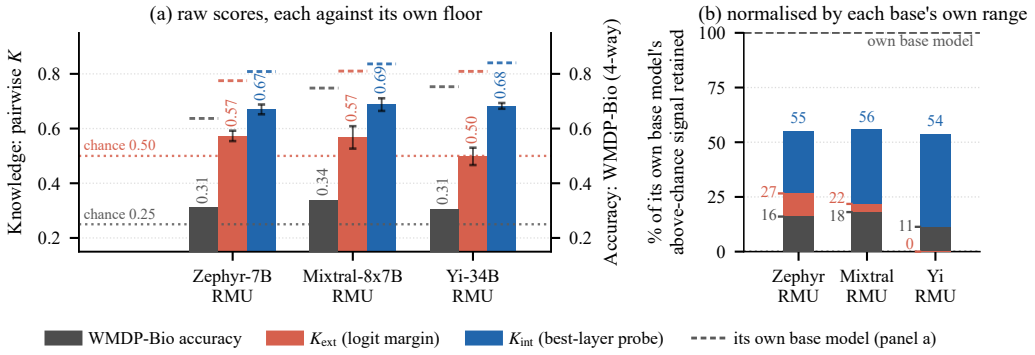

Figure 5: **Knowledge Lens across LLM Architectures Post-RMU.** (a) WMDP accuracy, K_{ext}, and K_{int}. (b) Each measure normalized by the model's own base-to-chance range.

4.6 THE INTERNAL-EXTERNAL SEPARATION EXTENDS TO CYBER DOMAIN

We apply the Knowledge Lens to an additional in-scope forget domain, WMDP-Cyber (1,987 questions), using the same released RMU checkpoints for ZEPHYR-7B, MIXTRAL-8×7B-INSTRUCT, and YI-34B-CHAT as in §4.5. We evaluate these models directly, without retraining them or modifying the framework. Because the released models were jointly unlearned on the Bio and Cyber domains, this extension changes only the evaluation domain while holding the model and intervention fixed.

The same internal-external separation appears on WMDP-Cyber, as shown in Figure 6. Suppressed Knowledge is 17.7–26.6 percentage points across architectures, while K_{int} remains 0.72–0.74 as WMDP-Cyber accuracy falls to 0.28–0.31. On the common chance-to-base scale, WMDP-Cyber retains 16–21% of its above-chance signal versus 79–90% for K_{int}. At the question level, 33.4–41.6% are suppressed versus 11.3–12.2% forgotten. The forgotten fraction shows essentially no net increase from each Cyber base while suppression grows substantially. Thus, the Knowledge Lens reproduces the same separation in a second WMDP domain.

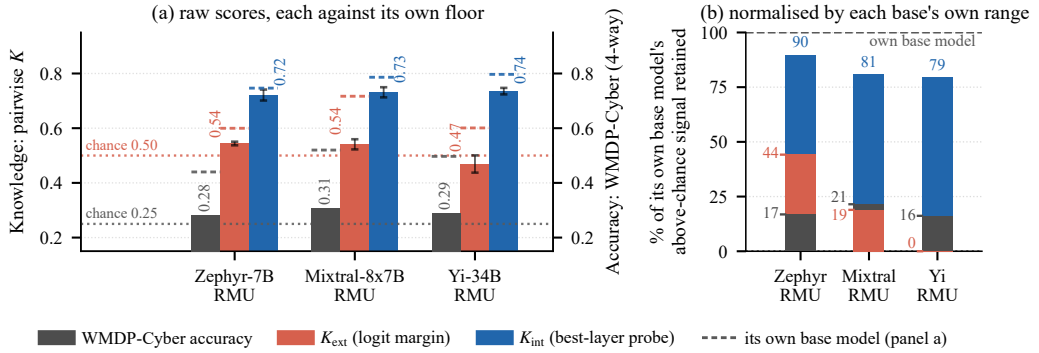

Figure 6: **Knowledge Lens on WMDP-Cyber Post RMU.** (a) WMDP-Cyber accuracy, K_{ext}, and K_{int}, each against its chance level. (b) Each measure normalized by the model's own Cyber base-to-chance range.

5 Conclusion

We introduce *Knowledge Lens*, an unlearning evaluation framework that matches changes in externally expressed and internally decodable task-relevant evidence under unlearning. Across eight unlearning methods, an independent nine-model method suite, four architectures, and two target domains, we find significant *suppressed knowledge* where conventional benchmarks show near-chance accuracy. The framework's design of reading the model's output logits and internal held-out linear probe at the same forward pass and token position allows for clean and natural extensions throughout our experimental settings.

Our results are particularly relevant because benchmarks serve as proxies for the knowledge that unlearning methods seek to remove for safety reasons. As a quantitative measure, the observed suppressed knowledge can further inform the development of future methods, facilitate distinctions between checkpoints with comparable accuracy, and support analyses at the level of individual information items.

Moreover, reporting both axes with the proposed item-level states replaces a binary characterization of recovery with a distribution over questions categorized as retained, suppressed, forgotten, or lucky. Unlike accuracy alone, the joint readout distinguishes among methods even after benchmark accuracy reaches chance level. We therefore propose this joint distribution as a more informative representation for reporting unlearning outcomes.

Our evidence is limited to the evaluated WMDP domains and to model checkpoints released by the original works. A natural next step is to test the extent to which the same separation holds in substantially different domains. In this work, we expose suppressed knowledge using a white-box, model-specific approach. A further direction is to test cross-model approaches and the extent to which the observed signal transfers between models.

The Knowledge Lens framework provides a foundation for analyzing unlearning methods across broader safety and privacy domains and for guiding mitigation across multiple models.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments and to assist in interpreting results. We did not use generative AI tools to develop theoretical models or conceptual frameworks, propose or refine hypotheses, design research methodology or experiments, implement methods, assist with translation, or clean and reformat datasets. Generating synthetic datasets, qualitative and thematic data analysis, and mathematical-claim and proof-related tasks were not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code and other artifacts, draft parts of the research paper, summarize and analyze existing literature, brainstorm, source or search for information, edit the paper to improve readability, and identify relevant literature. We reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by two authors, and all LLM-generated content was verified by at least two authors. We take responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies whether machine-unlearning methods for large language models genuinely erase targeted knowledge or merely suppress its external expression, using the public WMDP benchmark (Li et al., 2024). We do not introduce, elicit, or release new hazardous content, nor do we create or release new models. All evaluated models are open-weight checkpoints publicly released by the original works. Our probes analyze hidden states produced by these models when evaluated on the benchmark and do not generate open-ended hazardous text.

Our finding that suppressed knowledge can remain linearly decodable has dual-use implications: the same probing and readout techniques used for auditing could in principle be adapted to recover suppressed information from an open-weight model. We believe publication is justified because the evaluated models and weights are already publicly available, our work does not provide additional model access or new hazardous content, and its purpose is to improve awareness of limitations in current behavioral unlearning evaluations and support the development of stronger unlearning methods that can mitigate risk at scale. The work involves no human subjects or personal data.

REPRODUCIBILITY STATEMENT

We describe the Knowledge Lens evaluation framework in §3, with the exact True/False prompt provided in Appendix A and implementation details—including model input handling, tokenization, hidden-state extraction, cross-validation, probe fitting, hyperparameters, and random seeds—provided in Appendix B. The datasets and model checkpoints used in our experiments are publicly released and cited in the paper. We provide an anonymous supplementary .zip containing the code implementing Knowledge Lens and the scripts needed to reproduce the evaluation pipeline reported in this paper and to apply the framework to compatible datasets and model checkpoints.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

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

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Zora Che, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, et al. Model tampering attacks enable more rigorous evaluations of llm capabilities. *arXiv preprint arXiv:2502.05209*, 2025.

Liran Cohen, Yaniv Nemcovsky, and Avi Mendelson. Remind: Memorization and unlearning in llms through the lens of input loss landscapes. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47955–47993, 2026.

Bartosz Cywiński, Emil Ryd, Rowan Wang, Senthooran Rajamanoharan, Neel Nanda, Arthur Conmy, and Samuel Marks. Eliciting secret knowledge from language models, 2025. URL https://arxiv.org/abs/2510.01070.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *Advances in Neural Information Processing Systems*, 38, 2026.

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

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

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *Advances in Neural Information Processing Systems*, 38:1540–1567, 2026.

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

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. *Advances in Neural Information Processing Systems*, 38:60681–60713, 2026.

Yichen Gao, Altay Unal, Akshay Rangamani, and Zhihui Zhu. An illusion of unlearning? assessing machine unlearning through internal representations. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4114–4122. PMLR, 02–05 May 2026. URL https://proceedings.mlr.press/v300/gao26e.html.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299*, 2025.

Anmol Goel, Alan Ritter, and Iryna Gurevych. Auditing language model unlearning via information decomposition. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 808–826, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Rom Himelstein, Amit LeVi, Brit Youngmann, Yaniv Nemcovsky, and Avi Mendelson. Silenced biases: The dark side llms learned to refuse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37452–37461, 2026.

Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, and Haiqin Yang. Dissecting fine-tuning unlearning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3933–3941, 2024.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *International Conference on Learning Representations*, volume 2025, pp. 8857–8888, 2025.

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 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Jaeung Lee, Dohyun Kim, and Jaemin Jo. Measuring the depth of llm unlearning via activation patching. *arXiv preprint arXiv:2605.24614*, 2026.

Amit LeVi, Rom Himelstein, Yaniv Nemcovsky, Avi Mendelson, and Chaim Baskin. Jailbreak attack initializations as extractors of compliance directions. In *EMNLP (Findings)*, pp. 6672–6705, 2025.

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

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Proceedings of the 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 243–254, 2022.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194, 2025.

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

Lev McKinney, Anvith Thudi, Juhan Bae, Tara Rezaei Kheirkhah, Nicolas Papernot, Sheila A. McIlraith, and Roger Baker Grosse. Gauss-newton unlearning for the LLM era. In *Proceedings of the 4th IEEE Conference on Secure and Trustworthy Machine Learning*, 2026. arXiv:2602.10568.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–12676, 2024.

William Shen, Xinchi Qiu, Meghdad Kurmanji, Alexandru-Andrei Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas Lane. Llm unlearning via neural activation redirection. *Advances in Neural Information Processing Systems*, 38:44253–44290, 2026.

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

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, et al. Tamper-resistant safeguards for open-weight llms. In *International Conference on Learning Representations*, volume 2025, pp. 101802–101829, 2025.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-demos.6/.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms. *arXiv preprint arXiv:2505.16831*, 2025.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105425–105475. Curran Associates, Inc., 2024. doi: 10.52202/079017-3346. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/be52acf6bccf4a8c0a90fe2f5cfcead3-Paper-Conference.pdf.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

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

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373, 2024.

A Debiased System Prompt

The prompt $p(q, a)$ used for all True/False evaluations has two parts, a system message and a user message, rendered with each model's chat template (Appendix B):

System: *You are a critical fact-checker. Answer True or False. True and False are equally likely. Do not explain.*

User: *Consider the claim below. Respond True if the claim is supported, otherwise respond False.*
Claim: The answer to '{question}' is '{answer}'.
Label:

"True and False are equally likely" is intended to counter a bias toward answering True regardless of the claim. Omitting explanation prevents hedging tokens from appearing before the True/False answer token that we score.

B Implementation Details

This appendix provides the implementation details required to reproduce the Knowledge Lens evaluation.

Forward pass. Every candidate answer is turned into one claim. Each claim is wrapped in the verification prompt of Appendix A and rendered with the chat template[1] distributed with each model's tokenizer (Hugging Face Transformers, Wolf et al., 2020), ending with the assistant header. The scored position is therefore the last prompt token, the position at which the model would emit its first answer token. Both s_{ext} and s_{int} are read there, from the same forward pass. Mixtral's template does not accept a system message, so for Mixtral the system text is prepended to the user message. Models run in float16 with left padding (padding token = end-of-sequence token), which places the scored position at the last index of every sequence in a batch. At the scored position we record the output logits and the hidden state after the embedding layer and after every transformer block.

[1] https://huggingface.co/docs/transformers/main/en/chat_templating

Answer tokens. For the WMDP RMU releases (Zephyr, Mixtral, Yi) and the OPTML suite, we take the token the model would actually produce. We render an assistant turn answering "True" or "False" and take the first token after the assistant header. Where an unlearned release and its base model disagree, we use the base model's tokenizer, so both are scored on the same token: the Yi-34B RMU tokenizer maps the answer to a different token than its base, and the SimNPO Zephyr release ships no tokenizer. For LLAMA-3-8B-INSTRUCT and all LLM-GAT checkpoints, the scored tokens are the leading-space variants " `True`" and " `False`" (ids 3082 and 3641). Scoring instead the no-space variants that the model emits after the assistant header (ids 2575 and 4139) increases the mean $K_{\text{int}} - K_{\text{ext}}$ gap of the eight final LLM-GAT checkpoints by 0.6 percentage points, so this choice does not inflate the reported gap.

Cross-validation. The questions of a domain are shuffled once with seed 42 and split into five folds with scikit-learn's `KFold`. The split depends only on the number of questions, so every model and checkpoint of a domain is evaluated on the same folds. This single split is used throughout. Each fold serves once as the test set, 20% of the questions, so every question is scored exactly once by a probe that never saw it. The remaining questions are divided in their shuffled order into training (65% of the total) and validation (15%), so the validation set differs from fold to fold. Folds are assigned by question, so all the candidates of a question always fall in the same split. A candidate is labeled 1 if it is the correct option and 0 otherwise, with no class reweighting. Layer and C are selected by AUC, and K_{int} compares candidates within a question, so neither depends on the $1{:}3$ class ratio.

Probe. The probe is an ℓ_2-regularized logistic regression (scikit-learn, Pedregosa et al., 2011, `lbfgs`, at most $1{,}000$ iterations) on standardized features, with the scaler fitted on the same questions as the classifier. Within each fold, one probe is fitted per pair of layer (every recorded state) and regularization strength $C \in \{0.01, 0.1, 1, 10, 100\}$, and the pair with the highest validation AUC is kept. The pair is selected separately in each fold, so different folds of the same model may use different layers. The probe is then refitted at that pair on the training and validation questions together and scores the test fold, from which K_{int} is computed. The score of a candidate is the probe's predicted probability that it is correct. Hidden-state extraction runs on GPU and probe fitting runs on CPU from the stored hidden states.

Random seeds. The only source of randomness in the evaluation is the shuffle of the questions before the five-fold split (seed 42), which decides the fold each question falls in and hence the probe that scores it. Every other step is deterministic. Hidden-state extraction reproduces the stored hidden states bit for bit. The ℓ_2-regularized logistic-regression objective is strictly convex, and `lbfgs` starts from zero weights and uses full-batch gradients, so the same training questions always give the same probe. The item-level states follow from K_{int} and K_{ext} by the fixed thresholds of §3.2, so their counts depend on the seed only through the scores. The unlearned models are the released checkpoints, used as fixed weights.

C KNOWLEDGE LENS ACROSS ADDITIONAL UNLEARNING METHODS

We apply Knowledge Lens to nine public OPTML Zephyr-7B-beta models on WMDP-Bio (Fan et al., 2025; 2026): SimNPO, NPO (Zhang et al., 2024), NPO+SAM/GP/CR/RS/WA, GradDiff (Liu et al., 2022), and GradDiff+SAM. These span reference-free preference optimization and smoothness-enhanced methods designed for relearning robustness. As in §4.5, we evaluate released endpoints only.

As shown in Figures 7 and 8, across all methods, K_{int} remains 0.77–0.80 versus 0.81 for the base model, while K_{ext} falls from 0.78 to 0.46–0.60. K_{int} therefore retains 86–96% of above-chance base signal, versus -14–36% for K_{ext}. Suppressed Knowledge reaches 17.4–33.2 pp, as 26.2–46.1% of questions are suppressed, while only 8.6–12.1% are forgotten versus 8.7% in the base model. Thus, the §4.2 internal-external separation transfers to this distinct method suite.

D RETAIN-SET PERFORMANCE

Our experiments focus exclusively on the targeted original forget sets, measuring the extent to which representational knowledge (K_{int}) persists after behavioral suppression (K_{ext} collapse). This appendix

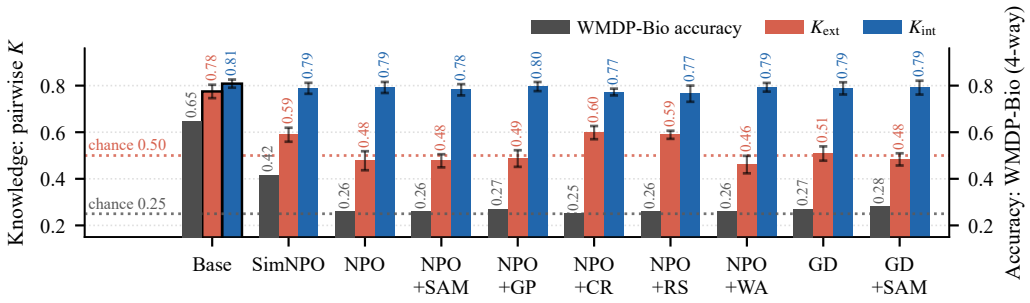

Figure 7: **Behavioral, external, and internal evidence across the OPTML methods on Zephyr-7B-beta.** Whiskers show five-fold standard deviation for the Knowledge Lens scores. WMDP-Bio accuracies are those reported by the source papers (Fan et al., 2025), except SimNPO and the base model, which are taken from the SimNPO model card.

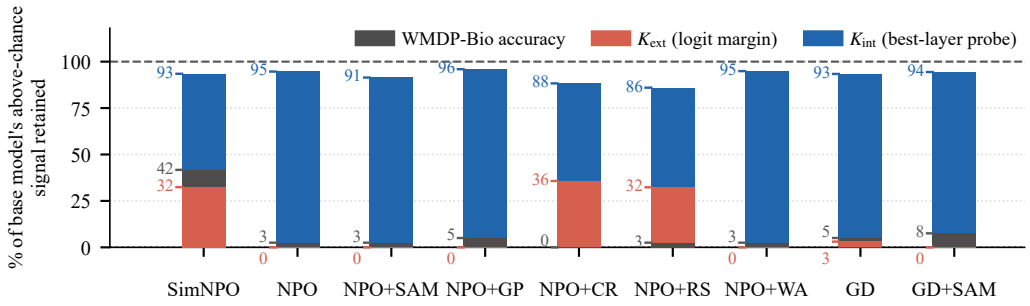

Figure 8: **Above-chance signal retained across the OPTML methods.** Each metric is normalized to the Zephyr base model's own base-to-chance range.

elaborates on the approach taken towards retain set evaluation by the original works and the original reports. We use the original released models and checkpoints unchanged, and therefore rely on the retain-set results reported by the original works.

For the experiment at §4.2, LLM-GAT (Che et al., 2025) verifies that all eight unlearning methods preserve general-capability performance (retain set) within acceptable degradation bounds across standard benchmarks. We refer to that paper for full evaluation details.

The three RMU checkpoints of §4.5 and §4.6 are evaluated by WMDP's authors (Li et al., 2024). RMU lowers overall MMLU accuracy by 1.0, 1.1 and 2.0 points for Zephyr-7B, Mixtral-8×7B and Yi-34B, and MT-Bench (a 1–10 scale) by 0.23, 0.13 and 0.06. The loss is concentrated in the MMLU subjects closest to the forget domains: virology falls by 26.5, 15.7 and 34.9 points and computer security by 20.0, 22.0 and 38.0, while college biology and college computer science change by at most 6 points.

The nine OPTML Zephyr-7B models of Appendix C pay a larger retain cost. Their authors report zero-shot MMLU of 0.41–0.48 for all nine, against 0.59 for the base model (Fan et al., 2025; 2026).

E Internal-External Gap Widens During Unlearning, Eight-Method View

As described in §4.3, the internal-external gap widens on average during unlearning for all eight methods in the LLM-GAT model suite (Che et al., 2025). This appendix provides the same two readouts for the complete set of unlearning checkpoints for these methods (Figure 9), complementing the RMU-only view in Figure 4.

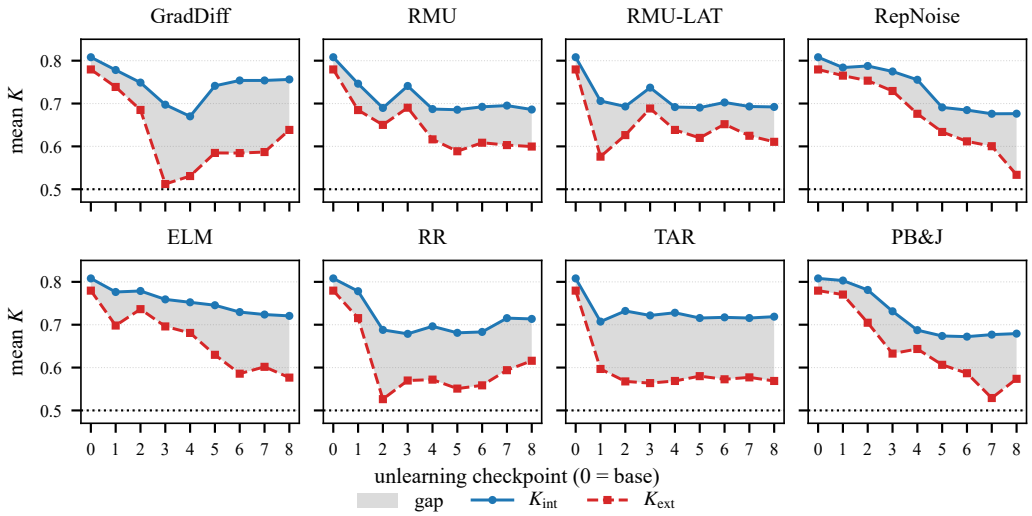

Figure 9: **Internal and external evidence throughout unlearning, all methods.** Mean K_{int} and K_{ext} across checkpoints for each method. $t = 0$ is the shared base model and the dotted line marks the 0.5 pairwise chance level. Results use all WMDP-Bio questions.