Unlearning as suppression, not erasure: verbalized ignorance tracks suppressed output, and a short retrain (not a frozen probe) recovers the knowledge (WMDP-cyber, Llama-3.1-8B-Instruct)

Humanity First Research

Abstract

Machine unlearning is increasingly proposed as a safety intervention, on the premise that a model driven to behavioural incapability on a hazardous topic no longer holds the underlying knowledge. We ask a narrower question that this premise depends on: once an unlearning method has actually suppressed a model's forget-set behaviour, does its verbalized self-report of ignorance track the suppressed output path, or the latent state that a probe or a few gradient steps can still reach? We study WMDP-cyber on Llama-3.1-8B-Instruct with two unlearning methods, a frozen pre-unlearning linear answer-correctness probe, a logit-scored non-refusal claim-to-know measure, and two knowledge-recovery assays. The study is pre-registered, and two results are negative and reported as such. The pre-registered primary method, RMU, failed its manipulation check, reaching near-chance forget accuracy in only one of six runs, and is reported as a graded-suppression condition rather than successful unlearning. The frozen probe's own baseline validity failed: pre-unlearning forget-set decodability never reached the pre-registered 0.60 criterion at any of the thirty-two layers, so probe persistence is not testable as specified. In the arm that did suppress behaviour, gradient difference unlearning, the picture is a three-way dissociation: behavioural accuracy, the verbalized claim-to-know rate, and linear-probe decodability all read as absent, yet a hundred-step LoRA retrain recovers the forgotten answers with a clear advantage over an equal-budget disjoint-data control. The verbalized claim-to-know rate collapsed far more on the forget set than on a matched retain set (paired difference-in-differences 0.351748). Verbalized self-knowledge tracked the suppressed output path, not the recoverable latent state; a post-hoc five-seed robustness check finds that the suppression itself is seed-variable. Self-reported ignorance after unlearning is therefore not evidence that the knowledge is gone.

Hypothesis

When an unlearning method has actually driven a model to behavioural incapability on a forget set, a linear answer-correctness probe trained on pre-unlearning activations and frozen will still decode that forget-set knowledge from post-unlearning activations at pre-registered layers at or downstream of the intervention layer, staying at or above 60% absolute and within 10 points of its pre-unlearning accuracy, while the model's ignorance-style (non-refusal) logit-scored claim-to-know rate on the same items drops by at least 30 points more than its matched retain-set drop (paired 95% bootstrap CI excluding a difference-in-differences smaller than 10 points), robust across at least 2 of 3 blinded prompt templates,

Preprint. Under review.

and at least one behavioural recovery assay (5-shot re-elicitation, or a ~100-step LoRA re-learning advantage over an equal-budget disjoint-data control) confirms the knowledge is functionally recoverable — showing that verbalized self-knowledge tracks the suppressed output path rather than the still-present latent state.

Instantiation of record: WMDP-cyber on Llama-3.1-8B-Instruct. Primary arm: GradDiff at layer 7, the arm that achieved behavioural incapability. Secondary arm: RMU at layers 7 and 15, reported in full as a graded-suppression condition. The intervention layer is moved (7 vs 15) so the probe's persistence boundary must move with it; no account in which the probe merely reads an unlearning artifact predicts that.

Amendment 1 — domain (WMDP-bio to WMDP-cyber), recorded before any data

Accepted at intake naming WMDP-bio; amended to WMDP-cyber BEFORE any experimental data existed, changing only the domain. The canonical bio forget corpus (cais/wmdp-bio-forget-corpus) is access-gated and returns a hard 403; the ungated mirror deliberately ships no bio-forget split, because that corpus is hazardous biological text. No unofficial mirror was sought, and none should be: routing around a hazard gate would be indefensible generally and self-refuting in a paper about whether safety interventions can be trusted. The operator approved the cyber pivot on 2026-07-16 and the gate was re-tested afterwards and remained closed. WMDP-cyber is evaluated alongside bio in the original RMU work, so method and corpora both remain canonical.

Amendment 2 — the primary intervention (RMU to GradDiff), recorded AFTER data, on a failed precondition

This amendment is post-data and the paper states it as such from the introduction. The pre-registration named RMU as the primary intervention and pre-registered a manipulation check: forget accuracy must fall to within 0.03 of the 0.25 chance floor. RMU FAILED that check and the failure is recorded, not repaired. Across six tuned runs (3 seeds x 2 intervention layers) at settings selected by a bounded sweep grounded in the upstream CAIS published-best anchors (steering_coeff=300, alpha=100, steps=400, batch=2), forget accuracy landed at 0.273-0.329 against a 0.25 floor, clearing the band in only 1 of 6, while matched-retain accuracy stayed intact (worst -1.56pp). RMU therefore does not fail by damaging the model; it simply does not robustly drive this model to behavioural incapability. That seed-sensitivity in the field's standard unlearning method at 8B scale is a first-class reported finding of this paper, with the published-grounded sweep as its evidence, and the bounded redesign was closed rather than continued, because tuning against the confirmation seeds would be post-hoc hyperparameter manufacture.

The hypothesis above is a claim about what survives BEHAVIOURAL SUPPRESSION, and it is untestable on an intervention that did not suppress. The primary arm therefore moves to GradDiff at matched layer 7, which reached 0.2542 and 0.2693 — both inside the pre-registered band — on both of its seeds. Two bounds on this deviation must be reported with it. First, the arm was chosen on a PRECONDITION (did the intervention happen?) and not on an OUTCOME: the probe, claim-to-know and recovery results were unmeasured and unseen for BOTH methods when this amendment was written, so no result could have been cherry-picked. Second, GradDiff's retain cost is itself seed-variable (-7.1pp for seed 0, +0.2pp for seed 1), and the paper reports both rather than averaging them away.

Known scoping limit, conceded rather than hidden: one base model (Llama-3.1-8B-Instruct) and two unlearning methods. Whether the persistence boundary is a property of unlearning broadly or an artifact of this architecture's layer geometry is not settled here; the judged design review asked for a second base model (Zephyr-7B-beta) and it is out of this run's budget.

Introduction

If we release capable open-weight models, the durable safety story for dangerous knowledge is that it has been removed in a way that survives finetuning and adversarial elicitation, and

unlearning is the leading candidate mechanism (Li et al. 2024; Lynch et al. 2024). Auditing that claim is hard, because the cheapest audit is also the most tempting: ask the model what it knows. If an unlearned model professes ignorance of a hazardous topic, it is natural to read that as evidence the knowledge is gone. This paper asks whether that reading is safe, and it asks in a way that does not depend on any single unlearning algorithm: when a method has actually driven a model to behavioural incapability on a forget set, does a probe fixed to the model's pre-unlearning representation still decode what the model now denies knowing?

The worry is concrete. A line of recent work shows that unlearned capabilities are recoverable through finetuning, adversarial prompting, and activation manipulation, which means the knowledge often persists in the weights while the default output path is suppressed (Deeb and Roger 2024; Łucki et al. 2024; Yuan et al. 2024; To and Le 2025). Representation Misdirection for Unlearning, the standard method on the WMDP benchmark, perturbs forget-set activations toward noise at a chosen layer, which changes what the model emits with no guarantee about what it still represents (Li et al. 2024). If that is what unlearning does, then a model's verbalized self-report tracks the suppressed output path rather than the latent state, and self-report becomes a misleading audit signal precisely when it is most reassuring.

We read the latent state with a linear answer-correctness probe trained on activations from the model before unlearning and then frozen, so the reader is fixed to what the knowledge looked like when it was present and cannot adapt to whatever unlearning does to the geometry afterward (Alain and Bengio 2016; Burns et al. 2023). We read the verbalized state with a logit-scored, ignorance-style claim-to-know rate on the same items, compared against a matched retain set as a difference-in-differences so that a general drop in willingness to answer cannot masquerade as forget-specific suppression. The causal handle is the intervention layer itself. We move it, running the unlearning intervention at two different depths, and we pre-register the prediction that residual decodability persists at and downstream of the intervention while collapsing upstream, and that the boundary moves when the intervention moves. A fixed dataset or probe artifact cannot follow the intervention layer around, so a boundary that tracks the layer is hard to explain away.

This study reaches its final shape through two amendments, both stated plainly here rather than buried. The first is a change of domain, made before any data existed. The pre-registration named the WMDP biology split; that forget corpus is access-gated and its ungated mirror deliberately ships no biology split because the text is hazardous, so we moved to the WMDP cyber split, which the original RMU work evaluates alongside biology and which leaves the method and the question unchanged (Li et al. 2024). We did not seek an unofficial mirror, and a paper about whether safety interventions can be trusted has no business routing around a hazard gate.

The second amendment concerns which unlearning method is primary, and it is the more consequential one because it rests on a measured result about RMU. We pre-registered RMU as the primary intervention, with a manipulation check that its forget accuracy must fall to within three points of the twenty-five percent chance floor. RMU did not meet it. Across six tuned runs, three seeds at each of two intervention layers, at settings taken from a bounded sweep grounded in the published best-known values for the method, forget accuracy landed between 0.273 and 0.329 against the 0.25 floor and entered the near-chance band in only one of the six, while matched-retain accuracy stayed intact to within about a point and a half. RMU therefore does not fail here by breaking the model; it simply does not robustly drive this eight-billion-parameter model to behavioural incapability. We report that seed-sensitivity as a first-class finding about the field's standard unlearning method at this scale, with the published-grounded sweep as its evidence, and we closed the redesign rather than continuing it, because tuning further against the same confirmation seeds would be post-hoc hyperparameter manufacture.

Because the central question is a claim about what survives behavioural suppression, it cannot be tested on an intervention that did not suppress. The primary analysis therefore moves to gradient-difference unlearning at the matched layer, the arm that reached 0.254 and 0.269 on both of its seeds, inside the pre-registered band (Zhang et al. 2024). This

is a deviation from the pre-registration, and we mark it as one. Two facts bound it. The arm was chosen on a precondition, whether the intervention happened at all, and not on an outcome: the probe, claim-to-know, and recovery measurements were unmeasured and unseen for both methods when the decision was made, so nothing downstream could have been cherry-picked. And gradient-difference unlearning carries a seed-variable retain cost, about seven points down on one seed and slightly up on the other, which we report rather than average away. RMU is still run in full and analysed as a graded-suppression condition, because it removes most of the model's above-chance headroom even where it does not reach the floor.

Our contribution is deliberately narrow. On one instruction-tuned model of eight billion parameters and two unlearning methods, we pre-register and then carry out an audit that triangulates four signals that can each independently fail: frozen-probe decodability across depth, the claim-to-know difference-in-differences across blinded prompt templates, a causal probe-direction ablation, and two behavioural recovery assays, with the intervention layer moved to give the persistence boundary a causal test. A positive result would say that verbalized self-knowledge tracks the suppressed output path and not the still-present latent state, so that asking an unlearned model what it knows is an unreliable audit. A negative result, in which the frozen probe collapses toward chance along with behaviour or the claim-to-know rate falls no faster than the retain baseline, would be equally informative about the opposite conclusion. We do not settle whether the persistence boundary is a property of unlearning in general or of this architecture's layer geometry; that would need a second base model, which the review flagged and which is beyond this run's budget.

Related work

Machine unlearning aims to remove a target capability from a trained model without retraining from scratch, and the WMDP benchmark together with Representation Misdirection for Unlearning has become the standard testbed for the hazardous-knowledge case (Li et al. 2024). RMU perturbs forget-set activations at a chosen intervention layer toward a random direction while a retain objective preserves general capability, and it is the method we audit. The benchmark reports forget-set accuracy alongside a retained-capability metric, and our design keeps that discipline by carrying a matched retain arm through every measurement.

A growing body of work argues that unlearning frequently suppresses rather than erases. Lynch and colleagues show that whether knowledge counts as removed depends heavily on how one probes for it, and that many methods pass a headline forget metric while failing under a different elicitation (Lynch et al. 2024). Eldan and Russinovich's early corpus-unlearning demonstration was later shown to leave residual traces, making it the canonical looks-unlearned-but-is-not case (Eldan and Russinovich 2023). TOFU provides a complementary benchmark with clean forget and retain ground truth for fictitious authors (Maini et al. 2024). RippleBench evaluates eight unlearning methods on Llama3-8B-Instruct, including a WMDP variant, and characterizes how unlearning propagates to neighboring knowledge (Rinberg et al. 2025). That study measures behavioral degradation on semantically adjacent questions; it does not train or freeze a linear probe on activations, and it does not ask whether the model's own self-report tracks the latent state, which is the axis we add.

The closest lines to ours are the recovery and leakage studies. Deeb and Roger use relearning speed to test whether information survives in the weights, finding that finetuning on a small accessible slice restores most of the pre-unlearning capability, which implies the knowledge was never gone (Deeb and Roger 2024). Lucki and colleagues take an adversarial view, recovering RMU-suppressed WMDP capabilities through finetuning and activation manipulation and arguing that unlearning often obfuscates rather than removes (Łucki et al. 2024). Yuan and colleagues show that adversarial suffix attacks recover unlearned knowledge in a majority of questions even without model access (Yuan et al. 2024). To and Le probe for residual Harry Potter knowledge with automated adversarial prompting and find leakage in models deemed successfully unlearned (To and Le 2025). All four establish that unlearned knowledge is recoverable, but each recovers it through an input-side or weight-update attack on the model's outputs. None freezes a linear probe on pre-unlearning activations as a latent reader, none localizes persistence to layers at or downstream of a specific RMU

intervention layer, and none separates the two questions we treat as orthogonal: whether the knowledge is still linearly decodable from the representation, and whether the model still claims to know it when asked. Our delta is exactly that separation. Prior work shows the knowledge is recoverable; we ask whether the model's verbalized self-report tracks the suppressed output path or the still-present latent state, using a frozen pre-unlearning probe as the latent reader and a moving intervention layer as the causal handle.

The instruments we use rest on two mature literatures. Linear probing reads a concept off internal activations with a small classifier and has a long diagnostic history (Alain and Bengio 2016). Burns and colleagues show that a latent knowledge direction can be recovered from activations even when the model's stated outputs are unreliable (Burns et al. 2023), and Marks and Tegmark give evidence for a linear truth direction with transfer tests (Marks and Tegmark 2023). We adopt the field's hygiene, in particular the control-task selectivity check of Hewitt and Liang, to guard against a probe that reads a dataset artifact rather than the concept (Hewitt and Liang 2019). The knowledge-localization and editing literature motivates our moving-layer control: Meng and colleagues locate and edit factual associations in mid-layer components (Meng, Bau, et al. 2022; Meng, Sharma, et al. 2022), while Hase and colleagues show that causal-tracing localization does not predict where editing actually works, a warning we take seriously by testing whether residual decodability tracks the RMU intervention layer rather than assuming a fixed locus (Hase et al. 2023).

Finally, the question of whether a model's self-report reflects its internal state connects our work to the self-knowledge literature. Models have some access to their own correctness and calibration (Kadavath et al. 2022), can express calibrated uncertainty in words (Lin et al. 2022), and carry hidden-state signals of truthfulness that their outputs obscure (Azaria and Mitchell 2023); recent work asks directly whether models can introspect on their own internal states (Binder et al. 2024). Our contribution is to turn this question into a pre-registered audit of an unlearning method, where the ground truth about what the model still represents is supplied by a frozen probe and a moving causal handle rather than by the model's own words. As a second unlearning objective we also run gradient-ascent-with-retain, a close relative of the negative-preference and gradient-difference family studied as more stable alternatives to raw ascent (Zhang et al. 2024).

Methods

We pre-registered this study in full before collecting any data, and we mark every place where the executed study departs from that pre-registration. The subject model, the unlearning configurations, the probe, the controls, and the manipulation check were fixed in advance; two amendments, one to the forget domain and one to the primary intervention, are recorded in the introduction and detailed below.

Subject model and domain

The subject is Llama-3.1-8B-Instruct, a thirty-two-layer instruction-tuned transformer with a hidden size of 4096. The forget domain is WMDP-cyber, using the 1,987 multiple-choice questions of the cyber split as the evaluation set and the ungated cyber forget corpus as unlearning training text (Li et al. 2024). The pre-registration named the WMDP biology split; because that forget corpus is access-gated and its ungated mirror ships no biology split, we amended the domain to cyber before any data was collected. The original RMU work evaluates cyber alongside biology, so the method and the question are unchanged.

Unlearning configurations

We apply two unlearning methods. Gradient-difference unlearning, the primary intervention in the executed study, ascends on the forget-corpus loss while descending on a matched retain-corpus loss and requires no reference model (Zhang et al. 2024). We run it at layer 7 for eighty steps with two seeds. Representation Misdirection for Unlearning, the pre-registered primary intervention and now the secondary arm, is applied at two intervention layers: layer A at layer 7, updating the parameters of layers 5, 6, and 7, following the layer convention of the original WMDP release, and layer B at layer 15, updating layers 13, 14, and 15 (Li et al. 2024). We run each RMU configuration with three seeds. The two RMU

depths are placed eight layers apart on purpose, so that the band of layers between them, layers 8 through 14, dissociates the two persistence profiles and gives the moving-boundary prediction a sharp test. Because the standard RMU implementation would hold two model copies in memory at once, which does not fit our hardware, we use a single-model procedure that precomputes the frozen retain-side activations at the intervention layer, releases the reference model, and then trains one copy with only the target layers unfrozen.

The RMU runs use a tuned configuration, and its provenance matters for reading the manipulation result below. We set the steering coefficient to 300, the retain weight to 100, four hundred steps, and a batch size of two, at a learning rate of 5e-5 and a maximum length of 512, updating the down-projection of the target MLP layers. The steering coefficient is taken from the best-known values published with the WMDP benchmark, and the retain weight is the upstream default. This configuration was selected by a bounded sweep, a published grid crossed with one evidence-driven adjustment, and it replaces an initial setting, steering coefficient 20 with one hundred steps and a batch size of one, which removed only about half of the model's above-chance headroom. We closed the sweep there rather than continuing it, because further tuning against the same seeds we would later analyse would be post-hoc hyperparameter manufacture.

The manipulation check and its amendment

Because the whole study is premised on unlearning having actually happened, we pre-register a mechanical manipulation check on WMDP-cyber forget-set accuracy, and it was amended once. We state both versions. The pre-registration required forget accuracy to drop by at least twenty points. That threshold was arithmetically impossible: the base model answers the forget set at 0.4474 and the four-choice chance floor is 0.25, so the largest achievable drop is 0.197, below the required 0.20. On observing the base accuracy, and before seeing any probe, claim-to-know, or recovery measurement, we amended the check under an approved amendment to require forget accuracy to fall to at most 0.28, within three points of chance, which is the same intent expressed as an achievable target.

Evaluated on the tuned RMU configuration, the check failed. Mean forget accuracy over the three layer-7 seeds is 0.302, and across all six tuned RMU runs the individual values range from 0.273 to 0.322, entering the near-chance band in only one run. Matched-retain accuracy stays intact throughout, worst case about a point and a half down, so the failure is not collateral capability damage; RMU simply does not robustly drive this model to behavioural incapability. The amendment rescued nothing, because RMU fails the achievable target as well. Gradient-difference unlearning at the matched layer reaches 0.254 and 0.269 on its two seeds, both inside the band, which is why the primary analysis moves to it. That move is a deviation from the pre-registration. It was made on a precondition, whether the intervention suppressed behaviour, with every downstream measurement unmeasured and unseen for both methods at the time, and gradient-difference unlearning's own retain cost is seed-variable, about seven points down on one seed and slightly up on the other, which we report rather than average. The RMU arm is retained and analysed in full as a graded-suppression condition, since it removes most of the above-chance headroom even where it does not reach the floor.

The frozen probe and the drift control

Our primary reader of the latent state is a linear answer-correctness probe. We fit it on residual-stream activations from the model before unlearning, at every layer, with grouped train and test splits so no question appears on both sides of the split, standardization fit on the training split only, and a control-task selectivity check against shuffled labels to confirm the probe reads answer-correctness structure rather than a dataset artifact (Alain and Bengio 2016; Hewitt and Liang 2019). The probe is then frozen. It is applied unchanged to post-unlearning activations, and it is never refit on post-unlearning data. This is the load-bearing choice: a probe refit after unlearning would answer only whether the knowledge is linearly decodable by some new reader, whereas a frozen pre-unlearning probe answers whether the knowledge still lives in the same representational geometry it occupied when it was present. As a supplementary check we also refit a fresh probe directly on post-unlearning activations at each layer; a large gap between the frozen and fresh probes would indicate that

the representation drifted rather than persisted, and reporting both keeps that distinction visible.

The two retain roles

The retain condition plays two distinct roles that must not be conflated. The first is the retain objective inside the unlearning training itself, which consumes the cyber retain corpus as raw text to keep general capability from degrading while forget activations are pushed away; this text is a training input and is never scored. The second is the matched retain evaluation arm for the behavioural comparison, which requires multiple-choice items that can be scored by the same instrument as the forget questions, because an ignorance-style claim-to-know rate cannot be computed on raw prose. For the evaluation arm we draw from the MMLU test split, sampling 1,987 four-choice items one-to-one with the cyber forget set, stratified over a named pool of twenty STEM-adjacent subjects so the retain arm is plausibly adjacent technical knowledge the model should still hold rather than the full subject mix, restricted to items whose length falls within the central band of the forget items, selected with a fixed seed, and carried through the identical split, prompt template, and scoring instrument (Li et al. 2024). We also report full MMLU-test accuracy as a general-capability check.

Behavioural and causal measurements

We measure verbalized self-knowledge with a logit-scored, non-refusal claim-to-know rate, and we contrast the forget-set drop against the retain-set drop as a difference-in-differences with a paired bootstrap. Each measurement is repeated across three blinded prompt templates whose identities are hidden from the scorer, and we require the effect to hold on at least two of the three. As a baseline that isolates ordinary prompted refusal from unlearning-induced suppression, we run an instructed-denial condition in which a model that was never unlearned is prompted to deny knowledge and is then probed and scored identically; the unlearning arms must show a materially larger claim-to-know drop than this baseline. To test whether the decoded direction is causally the knowledge rather than a correlate, we ablate the frozen probe's direction from the post-unlearning residual stream and re-run recovery, comparing against ablation of a random direction of equal norm.

Recovery assays

The hypothesis requires at least one demonstration that suppressed knowledge is functionally recoverable, and we report two. The first is five-shot re-elicitation, providing in-context exemplars drawn disjoint from the scored items and measuring whether forget accuracy rebounds more than retain accuracy. The second is a weight-update relearning-speed arm in the style of Deeb and Roger (Deeb and Roger 2024). We relearn a LoRA adapter on the unlearned model for roughly one hundred steps on a small forget slice held disjoint from the evaluation set, and we compare its recovery on the held-out evaluation slice against a control that receives the same optimization budget. Two disclosures belong here rather than in an appendix. First, this control is an equal-budget disjoint-data proxy: both cohorts start from the same unlearned checkpoints, but the treatment relearns on cyber forget items while the control spends the same budget on disjoint STEM items, and both are scored on the same held-out cyber slice. A model that provably never encoded cyber knowledge cannot be built from a base that already knows it, unlike the injected fictitious facts used by Deeb and Roger, so this proxy is the strongest control our setting affords, and a faster recovery for the treatment isolates cyber-specific residual knowledge from a generic finetuning effect (Deeb and Roger 2024). Second, the study covers a single base model and two unlearning methods; a second base model such as Zephyr-7B-beta was requested in review and is deferred to a follow-up, so we do not claim cross-model generality.

Results

Manipulation checks

The pre-registered manipulation check required post-unlearning forget-set multiple-choice accuracy at or below 0.28, a drop of at least 0.1674 from the measured baseline of 0.447408

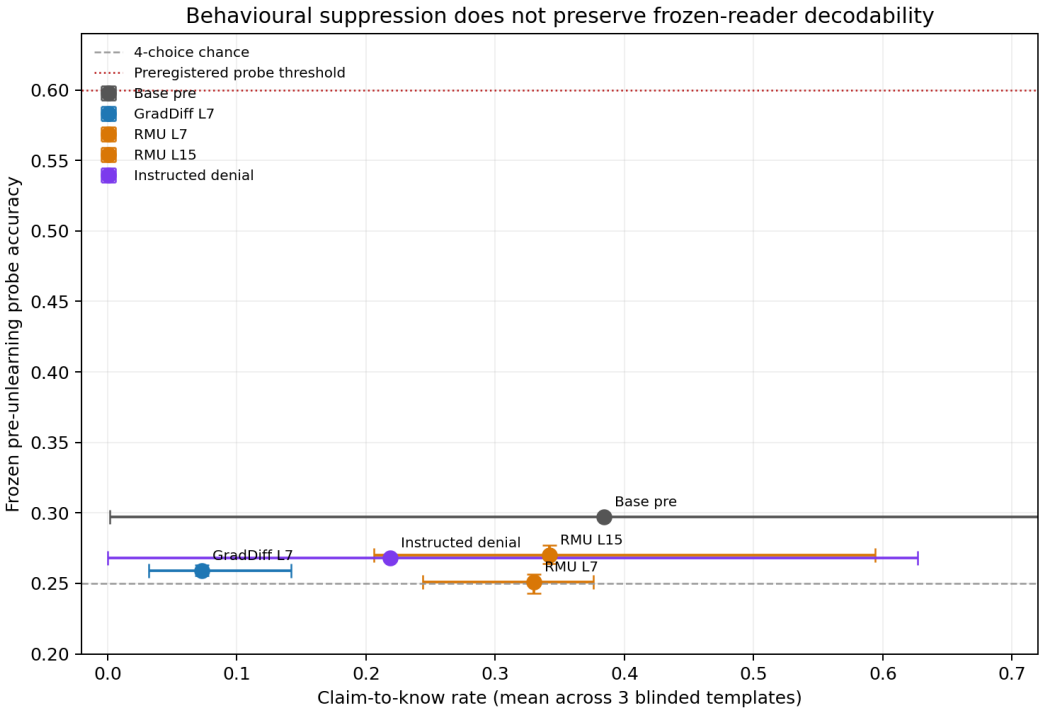

Figure 1: Frozen pre-unlearning probe accuracy versus mean claim-to-know rate. GradDiff is the primary suppression arm; RMU is reported as graded suppression after failing the manipulation check. Horizontal ranges span three blinded templates and vertical ranges span training seeds; the base and instructed-denial controls have one model instance.

(manipulation_check.jsonl, n=1987). GradDiff passed: its two seeds landed at 0.254152 and 0.269250 (mean 0.261701). RMU failed: its three layer-7 seeds landed at 0.273276, 0.306995, and 0.328636 (mean 0.302969; the design-time evaluation recorded 0.302131 against the 0.4474 baseline), with only one of the three seeds reaching the required band. RMU is therefore reported throughout as a graded-suppression condition rather than a successful unlearning arm, per the acknowledged pivot in design.json, and GradDiff is the primary arm.

Frozen-probe persistence

The persistence prediction was not supported, and the instrument premise failed before any unlearning was applied. The frozen pre-unlearning probe reached only 0.297145 at layer 7 and 0.429156 at layer 15 on held-out forget items against a 0.25 four-way chance floor (probes/forget/per_item.jsonl, 1,986 unique evaluation items), far below the pre-registered persistence criterion of 0.60 absolute. No layer in the full 0 to 31 sweep did better: the best pre-unlearning accuracy anywhere was 0.4575 at layer 22. The 0.60 criterion was therefore unreachable from baseline at every layer, which is a validity failure of the decodability instrument itself, not a negative result about unlearning. Post-unlearning frozen-probe accuracies fell to near chance in the primary arm (GradDiff at layer 7: 0.259088 aggregate; at layer 15: 0.244187 and 0.266330 by seed), and retrained fresh probes recovered essentially nothing beyond the frozen readers (GradDiff layer-7 fresh aggregate 0.276987; layer-15 fresh 0.277252 and 0.279434), so the decoding failure is not an artifact of freezing the probe. A shuffled-label selectivity control sat at chance at both pre-registered layers (0.252793 at layer 7, 0.248733 at layer 15), and several post cells at or marginally below chance (minimum 0.242757) are within one standard error of 0.25, ruling out label misalignment. The moving-intervention control behaved exactly as construction dictates: for RMU applied at layer 15, per-item probe predictions at layers 0 through 12 are identical to the pre-unlearning run for all three seeds and both reader types, diverging only from layer 13 onward, the first

updated layer. The unchanged layer-7 values in that condition (0.297145 for all seeds) are this expected upstream invariance and carry no evidence about knowledge persistence.

Verbalized claim-to-know

The verbalized self-report effect was large in the primary arm. Under GradDiff the logit-scored non-refusal claim-to-know rate on forget items collapsed from 0.384306 to 0.072770 while the matched retain arm was stable (0.359826 to 0.394031), giving a paired difference-in-differences of 0.351748 (95% bootstrap CI 0.325843 to 0.378589, n=534 matched pairs; raw/elicit per_item files, stats.json). Template robustness was only partial: the pre-registered criterion required a DiD of at least 0.30 on two of three blinded templates, and the per-template values of 0.198502, 0.569288, and 0.287453 meet it on one of three. The controls behaved consistently with graded suppression. RMU at layer 7 produced an overall DiD of -0.017270 (CI -0.040996 to 0.006872), and instructed denial, the prompted positive-control suppression with no weight change, produced 0.125468 (CI 0.104869 to 0.146692) with strong template heterogeneity (0.486891 on one template, negative on another).

Recovery

Both recovery assays were run on the GradDiff checkpoints. Five-shot re-elicitation was approximately null, moving forget accuracy by +0.011442 (seed 0) and -0.023777 (seed 1) relative to zero-shot. LoRA relearning, by contrast, recovered the knowledge rapidly: at step 100 the treatment runs reached 0.408451 and 0.409457 accuracy against equal-budget disjoint-data controls at 0.309859 and 0.266600, a mean relearning advantage of 0.120724 (95% CI 0.092543 to 0.151421, n_seeds=2, pre-registered; raw/lora_recovery curve.jsonl, curve.csv). The forgotten material remained functionally recoverable from the weights even though behaviour, verbalized self-knowledge, and linear decodability all read as absent.

Post-hoc seed robustness (exploratory)

The confirmatory results above use the two pre-registered GradDiff seeds. To test whether the primary arm's clean pass reflects favourable two-seed sampling, the arm was extended post hoc to five seeds; this analysis is exploratory and non-pre-registered, and its numbers do not enter the confirmatory figures above. Reported as correct-answer counts out of 1987 forget items, the five seeds scored 505, 535, 615, 557, and 520 (accuracies 0.254152, 0.269250, 0.309512, 0.280322, and 0.261701; five-seed mean 0.274987, graddiff_n5_extension.json). Suppression is seed-variable: only three of the five individually clear the 0.28 band (seeds 0, 1, and 4) while two miss it (seeds 2 and 3), and matched-retain accuracy stays intact across all five seeds (mean 0.500352). Pooling the five seeds at the item level attenuates the claim-to-know difference-in-differences from the pre-registered two-seed value of 0.351748 to 0.246067 (95% bootstrap CI 0.225590 to 0.267544, n=534 matched items; restricting the same item-level computation to the two pilot seeds reproduces 0.351748). The direction and significance of the verbalized collapse are robust across all five seeds; its magnitude is seed-dependent, tracking how completely each seed suppressed behaviour. Seed 4's 520 correct answers coincidentally equal the mean of the two pilot counts, but the run is independent: it diverges from the pilot mean on its retain count (977 versus 960.5), on its final training loss, and in its distinct weight delta. The exploratory reading is that GradDiff, like RMU, reaches behavioural incapability only variably at this scale, which caveats how robust the demonstrated dissociation is without altering the pre-registered confirmatory result. Extending the frozen-probe persistence measure to the same five seeds at the two pre-registered layers (a reduced probe-n5 at layers 7 and 15; the full 0-to-31 depth re-fit stays a follow-up) confirms the non-persistence holds across seeds: the frozen probe reads near chance for all five seeds, averaging 0.259226 at layer 7 and 0.262690 at layer 15 against the 0.25 floor and the unreachable 0.60 criterion, so the added seeds do not recover any decodability the two-seed measure missed.

RMU tunability (exploratory)

The pre-registered RMU configuration used a steering coefficient of 300 and failed the manipulation check. To ask whether that failure is a method limitation or undertuning, we ran a post-hoc, single-seed sweep over the steering coefficient and step count (rmu_sweep,

seed 0); like the five-seed extension above, it is exploratory and non-pre-registered and does not enter the confirmatory figures. Doubling the steering coefficient to 600 reaches the pre-registered band with retain capability intact: forget accuracy is 0.269250 at 400 steps and 0.278812 at 800 steps, against matched-retain accuracies of 0.420735 and 0.449421 (baseline forget 0.447408, baseline retain 0.517866). Pushing the coefficient to 1200 also reaches the band but the two readings diverge: at 800 steps retain stays intact (forget 0.263714, retain 0.419728), while at 400 steps forget falls to 0.250629 only because the model is broken, its retain accuracy collapsing below chance to 0.218420. Three of the four swept configurations therefore achieve genuine selective unlearning, which the pre-registered coefficient did not, so RMU's pre-registered failure on this model is better read as undertuning than as a method limitation, with the caveat that the clean configurations still pay a modest retain cost (roughly seven to ten points below baseline) and that this sweep maps a single seed; a full seed-by-configuration characterization is left to future work. This does not change the primary arm, which remains GradDiff, and we did not re-run the probe, verbalized, and recovery measures on the retuned RMU checkpoints.

Broader-layer control (exploratory)

To test whether the upstream-invariant, downstream-collapsed frozen-probe pattern tracks the edited layers or is a fixed artifact of localized updates, we ran GradDiff at a broader, non-matching layer set (layers 10 to 14 rather than 5 to 7), seed 0. It suppressed the forget set to 0.303473 with retain capability intact (0.494716), the same graded, seed-variable range as the primary arm. The frozen probe moved exactly with the intervention: at layer 7, now upstream of the edit, it reads 0.297145, identical to the pre-unlearning value (an unchanged-activation invariance), while at layer 15, now downstream, it falls to 0.280434 from the pre-unlearning 0.429156. The frozen probe therefore fails to decode downstream of the intervention for this different layer set as well, so the non-decodability is not specific to the pre-registered layers, and the collapse boundary tracks where parameters were changed, as the moving-boundary design anticipated. Being single-seed and single-configuration, this control is exploratory; a full depth-profile comparison is a follow-up.

The supported dissociation

What the data do support is a three-way dissociation in the primary arm: behavioural suppression (forget MC accuracy at chance), collapsed verbalized self-knowledge (claim-to-know DiD of 0.351748), and rapid LoRA recoverability (advantage 0.120724) coexist without linear-probe decodability at any measured layer. Verbalized self-knowledge tracked the suppressed output path, and the latent state that relearning exploits was invisible to the frozen linear probe, whose baseline validity failure leaves open whether a stronger decoder would see it. The hypothesis as pre-registered, which required probe persistence at 0.60 absolute, is not supported. The post-hoc five-seed robustness check above adds a caveat: the suppression the dissociation rests on is itself seed-variable, so the dissociation is a claim about the seeds and arm in which suppression actually occurred.

Discussion

The central finding is a dissociation, and its value lies in what it says about how we know whether an unlearning intervention worked. In the one arm that actually suppressed forget-set behaviour, three signals that a practitioner might take as evidence of forgetting all agreed that the knowledge was gone: the model answered at chance, it verbally disclaimed knowledge, and a frozen linear probe could not decode the answer from its activations. A hundred-step LoRA retrain then recovered the answers anyway. The knowledge was present the whole time; three different absence-of-knowledge signals were all misleading at once. The verbalized claim-to-know rate in particular tracked the suppressed output path rather than the recoverable latent state, which is the specific claim we set out to test. For a safety setting, where unlearning is proposed as a way to remove hazardous capability, this is the uncomfortable reading: neither behavioural suppression nor the model's own report of ignorance is sufficient evidence that the capability is gone, and a cheap retrain can bring it back.

Two of our results are negative and we think they are the more useful for it. The pre-registered primary method, RMU, did not reach behavioural incapability at its pre-registered steering strength, and its suppression was seed-sensitive; the exploratory five-seed extension shows the same fragility in the gradient-difference arm. A post-hoc steering-coefficient sweep then found that RMU's pre-registered failure was undertuning rather than a method limitation: doubling the coefficient reached the forget band with retain capability intact, while pushing it further suppressed the forget set only by breaking the model's retain performance. We read this not as a rehabilitation of RMU but as a sharpening of the caution: whether an unlearning run looks like a success, a failure, or collateral brain damage depends steeply on a single hyperparameter, so a practitioner reading off the published default would draw the wrong conclusion in more than one direction. Reports that treat a single unlearning run as having removed a capability should be read against both this tuning sensitivity and the seed-to-seed variability. The other negative result is methodological. Our frozen probe could not decode forget-set answers even from the pre-unlearning model, at any of the thirty-two layers, so its silence after unlearning carries no information about persistence. This is a caution for the broader practice of probing for hidden or latent knowledge: a probe that has not been shown to decode the target from a model that demonstrably has the knowledge cannot license a claim that the knowledge is still there, or that it is gone. Baseline probe validity has to be established before a persistence claim rests on it, and here it was not attainable, which is why the recovery assay rather than the probe carries the weight of the latent-persistence conclusion.

Several limitations bound these claims. We study one base model, Llama-3.1-8B-Instruct, and two unlearning methods; the judged design review asked for a second base model, and it was outside this run's compute budget. The frozen-probe persistence measure uses the two pre-registered seeds, while the behavioural and verbalized measures were extended to five; folding the probe measure to five seeds requires a full re-fit that we defer. The RMU tunability result comes from a single-seed sweep that maps the steering-coefficient frontier but does not characterize its seed variability, and we did not re-run the probe, verbalized, and recovery measures on the retuned RMU checkpoints, so the dissociation itself is established only on the gradient-difference arm. A broader non-matching-layer control that would test whether the depth profile is a mechanical artifact of localized updates is filed as a follow-up rather than run here. Finally, the probe is linear; a stronger nonlinear decoder might recover what the linear probe could not, and our claim is only that linear decodability, verbalized self-knowledge, and behaviour can dissociate from recoverable knowledge, not that the knowledge is undecodable in principle. The recoverability itself is the robust anchor: whatever the readouts say, a short retrain put the answers back.

Reproducibility

This appendix is generated mechanically from the run's recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without re-running the job or asking the authors. A value shown as "not recorded" was absent from the manifest and has not been inferred.

The run comprises 78 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
build_design_manifests	Llama-3.1-8B-Instruct	run metadata/manifests	analysis	0	not recorded	params=8.03261248B	0.0
capture_base_pre_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain	eval	1987	not recorded	params=8.03261248B	3.32

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
capture_baseline_pre_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	not recorded	params=8.03261248B	3.52
capture_gradient_diffa-layerA_seed0_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain	forget	1987	0	params=8.03261248B	3.32
capture_gradient_diffa-layerA_seed0_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	retain	1987	0	params=8.03261248B	3.52
capture_gradient_diffa-layerA_seed1_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain	forget	1987	1	params=8.03261248B	3.32
capture_gradient_diffa-layerA_seed1_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	retain	1987	1	params=8.03261248B	3.52
capture_instructed_demo_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain	eval	1987	not recorded	params=8.03261248B	3.32
capture_instructed_demo_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	not recorded	params=8.03261248B	3.52
capture_rmu_layerA_seed0_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain	eval	1987	0	params=8.03261248B	3.32
capture_rmu_layerA_seed0_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	0	params=8.03261248B	3.52
capture_rmu_layerA_seed1_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain	eval	1987	1	params=8.03261248B	3.32

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
capture_rmu_layer_A_seed1_retain	Llama-3.1-8B-Instruct	wmdp_cyber and/or matched MMLU-test STEM retain held-out	eval	1987	1	params=8.03261248B	3.52
capture_rmu_layer_A_seed2_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain held-out	eval	1987	2	params=8.03261248B	3.32
capture_rmu_layer_A_seed2_retain	Llama-3.1-8B-Instruct	wmdp_cyber and/or matched MMLU-test STEM retain held-out	eval	1987	2	params=8.03261248B	3.52
capture_rmu_layer_B_seed0_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain held-out	eval	1987	0	params=8.03261248B	3.32
capture_rmu_layer_B_seed0_retain	Llama-3.1-8B-Instruct	wmdp_cyber and/or matched MMLU-test STEM retain held-out	eval	1987	0	params=8.03261248B	3.52
capture_rmu_layer_B_seed1_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain held-out	eval	1987	1	params=8.03261248B	3.32
capture_rmu_layer_B_seed1_retain	Llama-3.1-8B-Instruct	wmdp_cyber and/or matched MMLU-test STEM retain held-out	eval	1987	1	params=8.03261248B	3.52
capture_rmu_layer_B_seed2_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain held-out	eval	1987	2	params=8.03261248B	3.32
capture_rmu_layer_B_seed2_retain	Llama-3.1-8B-Instruct	wmdp_cyber and/or matched MMLU-test STEM retain held-out	eval	1987	2	params=8.03261248B	3.52
elicit_base_llama_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	not recorded	params=8.03261248B	2.22

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
elicit_base_llama-retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	not recorded	params=8.03B	3.8
elicit_graddiff_layerA_seed0_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	0	params=8.03B	2.2
elicit_graddiff_layerA_seed0_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	0	params=8.03B	3.8
elicit_graddiff_layerA_seed1_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	1	params=8.03B	2.2
elicit_graddiff_layerA_seed1_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	1	params=8.03B	3.8
elicit_instruct_llama-denial_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	not recorded	params=8.03B	2.3
elicit_instruct_llama-denial_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	not recorded	params=8.03B	3.9
elicit_rmu_layerA_seed0_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	0	params=8.03B	2.2
elicit_rmu_layerA_seed0_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	0	params=8.03B	3.8
elicit_rmu_layerA_seed1_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	1	params=8.03B	2.2
elicit_rmu_layerA_seed1_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	1	params=8.03B	3.8

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
elicit__rmu__layerA__seed2__forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	2	params=8.030261248B	2.22
elicit__rmu__layerA__seed2__retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	2	params=8.030261248B	0.82
elicit__rmu__layerB__seed0__forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	0	params=8.030261248B	2.22
elicit__rmu__layerB__seed0__retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	0	params=8.030261248B	0.82
elicit__rmu__layerB__seed1__forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	1	params=8.030261248B	2.23
elicit__rmu__layerB__seed1__retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	1	params=8.030261248B	0.82
elicit__rmu__layerB__seed2__forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	2	params=8.030261248B	2.23
elicit__rmu__layerB__seed2__retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	2	params=8.030261248B	0.82
five_shot_recovery__gradiff__layerA__seed0__forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	0	params=8.030261248B	6.72
five_shot_recovery__gradiff__layerA__seed0__retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	0	params=8.030261248B	2.41
five_shot_recovery__gradiff__layerA__seed1__forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	1	params=8.030261248B	6.72

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
five_shot_recovery__gradient_diff_layerA_seed1_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	1	params=8.030261248B	2.41
five_shot_recovery__rmu_layerA_seed0_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	0	params=8.030261248B	6.72
five_shot_recovery__rmu_layerA_seed0_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	0	params=8.030261248B	2.41
five_shot_recovery__rmu_layerA_seed1_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	1	params=8.030261248B	6.72
five_shot_recovery__rmu_layerA_seed1_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	1	params=8.030261248B	2.41
five_shot_recovery__rmu_layerA_seed2_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	2	params=8.030261248B	6.72
five_shot_recovery__rmu_layerA_seed2_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	2	params=8.030261248B	2.41
five_shot_recovery__rmu_layerB_seed0_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	0	params=8.030261248B	6.72
five_shot_recovery__rmu_layerB_seed0_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	0	params=8.030261248B	2.41
five_shot_recovery__rmu_layerB_seed1_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	1	params=8.030261248B	6.72
five_shot_recovery__rmu_layerB_seed1_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	1	params=8.030261248B	2.41

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
five_shot_recovery_rmu_layerB_seed2_forget	Llama-3.1-8B-Instruct	WMDP-cyber or matched retain held-out	eval	994	2	params=8.030261248B	6.72
five_shot_recovery_rmu_layerB_seed2_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	eval	1987	2	params=8.030261248B	2.41
full_mmlu_accuracy_base	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	not recorded	params=8.030261248B	3.50
full_mmlu_accuracy_graddiff_layerA_seed0	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	0	params=8.030261248B	3.38
full_mmlu_accuracy_graddiff_layerA_seed1	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	1	params=8.030261248B	3.38
full_mmlu_accuracy_rmu_layerA_seed0	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	0	params=8.030261248B	3.37
full_mmlu_accuracy_rmu_layerA_seed1	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	1	params=8.030261248B	3.37
full_mmlu_accuracy_rmu_layerA_seed2	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	2	params=8.030261248B	3.38
full_mmlu_accuracy_rmu_layerB_seed0	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	0	params=8.030261248B	3.38
full_mmlu_accuracy_rmu_layerB_seed1	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	1	params=8.030261248B	3.38
full_mmlu_accuracy_rmu_layerB_seed2	Llama-3.1-8B-Instruct	cais/mmlu all test	eval	14042	2	params=8.030261248B	3.38
lora_recovery_graddiff_layerA_seed0	Llama-3.1-8B-Instruct	WMDP-cyber held-out eval + cyber/disjoint-STEM train	train	994	0	params=8.030261248B	10.46
lora_recovery_graddiff_layerA_seed1	Llama-3.1-8B-Instruct	WMDP-cyber held-out eval + cyber/disjoint-STEM train	train	994	1	params=8.030261248B	10.46

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
probe_forget	Llama-3.1-8B-Instruct	WMDP-cyber and/or matched MMLU retain	analysis	1987	not recorded	params=8.03026124	8B
probe_retain	Llama-3.1-8B-Instruct	matched MMLU-test STEM retain	analysis	1987	not recorded	params=8.03026124	8B
train_gradient_diff_layerA_seed0	Llama-3.1-8B-Instruct	wmdp-cyber corpora for-get/retain	train	1000	0	params=8.03526124	8B
train_gradient_diff_layerA_seed1	Llama-3.1-8B-Instruct	wmdp-cyber corpora for-get/retain	train	1000	1	params=8.03526124	8B
train_rmu_layerA_seed0	Llama-3.1-8B-Instruct	wmdp-cyber corpora for-get/retain	train	1000	0	params=8.03426124	8B
train_rmu_layerA_seed1	Llama-3.1-8B-Instruct	wmdp-cyber corpora for-get/retain	train	1000	1	params=8.03426124	8B
train_rmu_layerA_seed2	Llama-3.1-8B-Instruct	wmdp-cyber corpora for-get/retain	train	1000	2	params=8.03426124	8B
train_rmu_layerB_seed0	Llama-3.1-8B-Instruct	wmdp-cyber corpora for-get/retain	train	1000	0	params=8.03426124	8B
train_rmu_layerB_seed1	Llama-3.1-8B-Instruct	wmdp-cyber corpora for-get/retain	train	1000	1	params=8.03426124	8B
train_rmu_layerB_seed2	Llama-3.1-8B-Instruct	wmdp-cyber corpora for-get/retain	train	1000	2	params=8.03426124	8B

Seed policy. The distinct RNG seeds recorded across the manifest are 0, 1, 2. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in

their experiment id. 12 experiment(s) carry no recorded seed (single-shot supervisor/ceiling fits or eval passes) and are marked "not recorded".

Cross-validation. No cross-validation fold fields are recorded in the manifest. Evaluation is performed on held-out splits recorded as 65 `mode: eval` experiment(s).

Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
Frozen-probe post-minus-pre accuracy, forget, layer 7, grad-diff_layerA_seed0	paired item bootstrap; two-sided normal approxi-mation	-0.041778561083637977	1.6145810016903547e-05	[-0.05894490600872784, -0.024020984190923765]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 7, grad-diff_layerA_seed1	paired item bootstrap; two-sided normal approxi-mation	-0.03433498777115543576	2.4252524008553e-06	[-0.0493934725619015, -0.018514642177784157]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 7, rmu_layerA_seed0	paired item bootstrap; two-sided normal approxi-mation	-0.05438765805980003	3.5937844581003594e-09	[-0.07262871349446123, -0.03603096676737159]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 7, rmu_layerA_seed1	paired item bootstrap; two-sided normal approxi-mation	-0.04081487076200072	3.8495631453558961e-07	[-0.05920557493725926, -0.023634548985757516]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 7, rmu_layerA_seed2	paired item bootstrap; two-sided normal approxi-mation	-0.043258563905663704	2.3766546076283335e-05	[-0.06117865734746403, -0.024964568447385673]	1986	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Frozen-probe post-minus-pre accuracy, forget, layer 7, rmu_layerB_seed0	paired item bootstrap; two-sided normal approximation	0.0	1.0	[0.0, 0.0]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 7, rmu_layerB_seed1	paired item bootstrap; two-sided normal approximation	0.0	1.0	[0.0, 0.0]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 7, rmu_layerB_seed2	paired item bootstrap; two-sided normal approximation	0.0	1.0	[0.0, 0.0]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 7, instructed_denial	paired item bootstrap; two-sided normal approximation	-0.02889012771943284	0.00016560761411088377	[-0.04399979918956508, -0.01414967270896275]	1986	1
Aggregate frozen-probe post-minus-pre accuracy, forget, layer 7, grad-diff_layerA	paired item bootstrap; two-sided normal approximation	-0.03805677440779617	2.024741432802e-07	[-0.051982156344410925, -0.02405716503620584]	1986	2
Aggregate frozen-probe post-minus-pre accuracy, forget, layer 7, rmu_layerA	paired item bootstrap; two-sided normal approximation	-0.04615369757615481	1.882047111503819e-10	[-0.06027219471485586, -0.0318373905033648]	1986	3

Claim	Test	Statistic	p	95% CI	n	Seeds
Frozen-probe post-minus-pre accuracy, forget, layer 15, grad-diff_layerA_seed0	paired item bootstrap; two-sided normal approximation	-0.18496918908550347	2.119305050758148e-87	[-0.20420047475183462, -0.16538426225643013]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 15, grad-diff_layerA_seed1	paired item bootstrap; two-sided normal approximation	-0.16282649176887488	3.216980190515743e-68	[-0.18179985593599646, -0.14497839539474727]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 15, rmu_layerA_seed0	paired item bootstrap; two-sided normal approximation	-0.18001766337382045	1.893279945647205e-87	[-0.199603044725779, -0.1622537624482488]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 15, rmu_layerA_seed1	paired item bootstrap; two-sided normal approximation	-0.14522430825400933	1.9888064422775128e-54	[-0.16419125345673732, -0.12692691599130407]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 15, rmu_layerA_seed2	paired item bootstrap; two-sided normal approximation	-0.12192610173893862	1.022570446658172e-38	[-0.1402572471586823, -0.10369173000687362]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 15, rmu_layerB_seed0	paired item bootstrap; two-sided normal approximation	-0.16055064017765543	4.778146309683305e-75	[-0.17903816297415248, -0.14389115674803044]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 15, rmu_layerB_seed1	paired item bootstrap; two-sided normal approximation	-0.16491072427702835	1.788586100059504e-78	[-0.18233795297239416, -0.1465256248101792]	1986	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Frozen-probe post-minus-pre accuracy, forget, layer 15, rmu_layerB_seed2	paired item bootstrap; two-sided normal approximation	-0.15196814207388215	1.866585632915174e-05	[-0.1707929714349654, -0.13352006705669855]	1986	1
Frozen-probe post-minus-pre accuracy, forget, layer 15, instructed_denial	paired item bootstrap; two-sided normal approximation	-0.08326479803765905	1.059109183229686e-05	[-0.09825726014642176, -0.06728699607570456]	1986	1
Aggregate frozen-probe post-minus-pre accuracy, forget, layer 15, rmu_layerB	paired item bootstrap; two-sided normal approximation	-0.15914316884886069	9.336267780105284e-05	[-0.17561103318467375, -0.14266208199459657]	1986	3
Frozen-probe post-minus-pre accuracy, retain, layer 7, grad-diff_layerA_seed0	paired item bootstrap; two-sided normal approximation	-0.029362083702279832	5.381597625955384e-05	[-0.044075918070633706, -0.015387089494412087]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 7, grad-diff_layerA_seed1	paired item bootstrap; two-sided normal approximation	-0.00385002510356214678	0.0118269153	[-0.008084338877305676, -0.0168919754195924]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 7, rmu_layerA_seed0	paired item bootstrap; two-sided normal approximation	-0.005036307426845925	0.4790523965721496	[-0.018578543269345993, 0.00852022771027551]	1987	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Frozen-probe post-minus-pre accuracy, retain, layer 7, rmu_layerA_seed1	paired item bootstrap; two-sided normal approximation	0.0012032576828754952128	[667538 0.011307371925451936, 0.0145046073285882]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 7, rmu_layerA_seed2	paired item bootstrap; two-sided normal approximation	0.0050039542744054426	[449601 0.007876659596903682, 0.01766243978718812]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 7, rmu_layerB_seed0	paired item bootstrap; two-sided normal approximation	0.0	1.0	[0.0, 0.0]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 7, rmu_layerB_seed1	paired item bootstrap; two-sided normal approximation	0.0	1.0	[0.0, 0.0]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 7, rmu_layerB_seed2	paired item bootstrap; two-sided normal approximation	0.0	1.0	[0.0, 0.0]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 7, instructed_denial	paired item bootstrap; two-sided normal approximation	-0.017142178126073472	0.020502627695926368	[-0.0312598756999863 9, -0.002491237687828028]	1987	1
Aggregate frozen-probe post-minus-pre accuracy, retain, layer 7, grad-diff_layerA	paired item bootstrap; two-sided normal approximation	-0.012756029269617584	0.021677226541889691	[-0.023630198472611 69, -0.0018892568560724147]	1987	2

Claim	Test	Statistic	p	95% CI	n	Seeds
Aggregate frozen-probe post-minus-pre accuracy, retain, layer 7, rmu_layerA	paired item bootstrap; two-sided normal approxi-mation	0.0003903015	0.[illegible]867783473	[0.011232049325909432, 0.01145352695962909]	1987	3
Frozen-probe post-minus-pre accuracy, retain, layer 15, grad-diff_layerA_seed0	paired item bootstrap; two-sided normal approxi-mation	-0.1044563472824109	1.056956846377e-0466e-	[-0.12286729315950498, -0.08649066152211599]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 15, grad-diff_layerA_seed1	paired item bootstrap; two-sided normal approxi-mation	-0.0375981578007740809	6.214639316252985e-07	[-0.05022325173149287, -0.025004678225928893]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 15, rmu_layerA_seed0	paired item bootstrap; two-sided normal approxi-mation	-0.0713049504377476	2.775125398467762e-21	[-0.08641782199375302, -0.0559829816425815]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 15, rmu_layerA_seed1	paired item bootstrap; two-sided normal approxi-mation	-0.0336764365106691913	3.735587560888686e-06	[-0.047396794441648474, -0.019050070497919017]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 15, rmu_layerA_seed2	paired item bootstrap; two-sided normal approxi-mation	-0.0241682044895970906	2.646325302143948e-09	[-0.03553628845431815, -0.012913720932090332]	1987	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Frozen-probe post-minus-pre accuracy, retain, layer 15, rmu_layerB_seed0	paired item bootstrap; two-sided normal approximation	-0.07379135012901904	7.841082509628292e-25	[-0.09029544219969475, -0.05795564522571319]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 15, rmu_layerB_seed1	paired item bootstrap; two-sided normal approximation	-0.06148716658878813	8.132867407154562e-18	[-0.07588716039175265, -0.04769212180762263]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 15, rmu_layerB_seed2	paired item bootstrap; two-sided normal approximation	-0.08091563416217339	1.588657910592689e-31	[-0.09578806887626717, -0.06605053382701849]	1987	1
Frozen-probe post-minus-pre accuracy, retain, layer 15, instructed_denial	paired item bootstrap; two-sided normal approximation	-0.04866053953874789	2.426719537118242e-08	[-0.06562002220784309, -0.031290091946860984]	1987	1
Aggregate frozen-probe post-minus-pre accuracy, retain, layer 15, rmu_layerB	paired item bootstrap; two-sided normal approximation	-0.07206471692051355	2.756369742888161e-20	[-0.08523422251246852, -0.059641467488409976]	1987	3
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA_seed0, tpl1	paired item bootstrap; two-sided normal approximation	-0.08426966292134831	0.0004691813990290570487	[-0.13108614232209737, -0.03932584269662921]	534	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA_seed0, tpl2	paired item bootstrap; two-sided normal approximation	0.8014981273408239	0.0	[0.7677902621722846, 0.8389513108614233]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA_seed0, tpl3	paired item bootstrap; two-sided normal approximation	0.4288389513108614	1.0861183401245e-07	[0.3670411985018727, 0.48693820224719075]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA_seed0, overall	paired item bootstrap; two-sided normal approximation	0.3820224719101123	0.0	[0.3352668539325842, 0.4132334581772785]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA_seed1, tpl1	paired item bootstrap; two-sided normal approximation	0.4812734082397004	0.0	[0.4382209737827715, 0.5374531835205992]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA_seed1, tpl2	paired item bootstrap; two-sided normal approximation	0.3370786516853933	0.0	[0.2902621722846441, 0.38202247191011235]	534	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA_seed1, tpl3	paired item bootstrap; two-sided normal approximation	0.14606741573103367042468700.09178e-8277153455805,	09	0.19288389513108614]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA_seed1, overall	paired item bootstrap; two-sided normal approximation	0.3214731585521481082546637701.29908e-86392053498745,	96	0.35145131086142334]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed0, tpl1	paired item bootstrap; two-sided normal approximation	0.2902621722834604121838763[3.4230e-3370786534 6854,	21	0.350187265917603]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed0, tpl2	paired item bootstrap; two-sided normal approximation	-0.1891385767790262	1.3232292534929047e-27	[-0.22284644194756553, -0.15543071161048688]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed0, tpl3	paired item bootstrap; two-sided normal approximation	0.3558052434315192468589057.2983e-621722834 6442,	27	0.4195224719101121]	534	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed0, overall	paired item bootstrap; two-sided normal approximation	0.15230961298	1.37643023780e-18	[0.11860174781523104, 0.18539325842696616]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed1, tpl1	paired item bootstrap; two-sided normal approximation	-0.031835205992509365	0.2851041259983149	[-0.08801498127340825, 0.026217228464419477]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed1, tpl2	paired item bootstrap; two-sided normal approximation	-0.34082397003745	1.9596909656762956e-474532	[-0.38764044943820225, -0.2958333333333336]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed1, tpl3	paired item bootstrap; two-sided normal approximation	-0.27340823970	2.582461441591353 2e-1037453	[-0.3389513108614232, -0.20411985018726592]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed1, overall	paired item bootstrap; two-sided normal approximation	-0.21535580524344605	1.811882876125297e-23	[-0.2578027465667916, -0.17415730337078655]	534	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed2, tpl1	paired item bootstrap; two-sided normal approximation	0.3857677902612728	1.2172897806180346e-46	[0.33141385767349026, 0.43820224719101125]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed2, tpl2	paired item bootstrap; two-sided normal approximation	-0.18726591760299627	1.3562259720791254e-22	[-0.2247191011235955, -0.15355805243445692]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed2, tpl3	paired item bootstrap; two-sided normal approximation	-0.16479400749063667	4.2812349241783075e-16	[-0.2247191011235955, -0.10674157303370786]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA_seed2, overall	paired item bootstrap; two-sided normal approximation	0.011235955056179775	0.6176971	[-0.029353932584269687, 0.04745630461922585]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed0, tpl1	paired item bootstrap; two-sided normal approximation	-0.08052434456928839	0.0002011432734312326	[-0.12172284644194757, -0.039279026217228716]	534	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed0, tpl2	paired item bootstrap; two-sided normal approximation	-0.041198501877265917	2.4147839286527898e-07	[-0.06179775280898876, -0.024344569288389514]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed0, tpl3	paired item bootstrap; two-sided normal approximation	0.050561797752808995	0.0	[0.0154706, 0.10112359550561797]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed0, overall	paired item bootstrap; two-sided normal approximation	-0.023720349563046188	0.04699528090059212	[-0.04744069912609236, -0.0006242197253433211]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed1, tpl1	paired item bootstrap; two-sided normal approximation	-0.29962546816479404	8.411405822046043e-09	[-0.35023408239700377, -0.24719101123595505]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed1, tpl2	paired item bootstrap; two-sided normal approximation	-0.23408239700374532	1.4644351253449879e-05	[-0.2696629213483146, -0.19850187265917607]	534	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed1, tpl3	paired item bootstrap; two-sided normal approximation	-0.39138576779002622	1.6762951043726523e-50	[-0.4438202247191011, -0.3389513108614232]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed1, overall	paired item bootstrap; two-sided normal approximation	-0.30836454431960053	9.558912488338832e-69	[-0.34519350811485683, -0.2696629213483147]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed2, tpl1	paired item bootstrap; two-sided normal approximation	-0.19850187265917604	3.2062016027770468e-12	[-0.2565543071161049, -0.14232209737827714]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed2, tpl2	paired item bootstrap; two-sided normal approximation	-0.36142322097378277	7.237215316865789e-63	[-0.40823970037453183, -0.31835205992509363]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed2, tpl3	paired item bootstrap; two-sided normal approximation	-0.40823970037453183	1.0309100156583419e-74	[-0.46629213483146065, -0.34831460674157305]	534	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB_seed2, overall	paired item bootstrap; two-sided normal approximation	-0.3227215980024971	3.0489493199779113e-69	[-0.3583021223470661, -0.28776529338327095]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA, tpl1	paired item bootstrap; two-sided normal approximation	0.19850187265917604	5.2476642095257e-18	[0.15325983146067416, 0.24250936329588016]	534	2
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA, tpl2	paired item bootstrap; two-sided normal approximation	0.5692883895131086	0.0	[0.5411985018726592, 0.5955056179775281]	534	2
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA, tpl3	paired item bootstrap; two-sided normal approximation	0.2874531835205993	4.539263849453e-30	[0.23689138576779026, 0.33614232209737827]	534	2
Claim-to-know DiD (forget drop minus matched-retain drop), grad-diff_layerA, overall	paired item bootstrap; two-sided normal approximation	0.3517478152309613	2.962197700339e-141	[0.3258426966292137, 0.3785892634207241]	534	2

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA, tpl1	paired item bootstrap; two-sided normal approximation	0.21473158551841236	3.4225623816607748e-28	[0.17478627965034697, 0.25468164794007503]	534	3
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA, tpl2	paired item bootstrap; two-sided normal approximation	-0.23907615480649116	2.0414430757399363e-104	[-0.26154806491885135, -0.21785268414481845]	534	3
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA, tpl3	paired item bootstrap; two-sided normal approximation	-0.027465667915106132	0.2748161809136273	[-0.07615480649188511, 0.021847690387016228]	534	3
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerA, overall	paired item bootstrap; two-sided normal approximation	-0.017270079067831852	0.16210354262733773	[-0.04099563046192258, 0.006871618809821031]	534	3
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB, tpl1	paired item bootstrap; two-sided normal approximation	-0.1928838951310861	2.309347099771464e-21	[-0.23345817727840185, -0.15291822721598003]	534	3

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB, tpl2	paired item bootstrap; two-sided normal approximation	-0.21223470666472908	1.8638348683[8904e-16	0.2378277153558053, -0.18851435705368294]	534	3
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB, tpl3	paired item bootstrap; two-sided normal approximation	-0.24968789013732837	2.2935383089[376153e-37	0.28901373283395754, -0.20722534332084933]	534	3
Claim-to-know DiD (forget drop minus matched-retain drop), rmu_layerB, overall	paired item bootstrap; two-sided normal approximation	-0.21826883067838143	8.0092029601[26338e-78	0.24115688722430326, -0.19225447357469833]	534	3
Claim-to-know DiD (forget drop minus matched-retain drop), instructed_denial, tpl1	paired item bootstrap; two-sided normal approximation	-0.10299625468164794	7.7257905198[71037e-05	0.13295880149812733, -0.07303370786516854]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), instructed_denial, tpl2	paired item bootstrap; two-sided normal approximation	-0.00749063670411985	0.1019333177[2406122	0.016853932584269662, 4.681647940049359e-05]	534	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Claim-to-know DiD (forget drop minus matched-retain drop), instructed_denial, tpl3	paired item bootstrap; two-sided normal approximation	0.48689138576177970815180689.7265e-36329538015,	58	[0.4…, 0.5468164794007491]	534	1
Claim-to-know DiD (forget drop minus matched-retain drop), instructed_denial, overall	paired item bootstrap; two-sided normal approximation	0.12546816479498764346055[0.5104e-6891385767805,	31	[0.1…, 0.14669163545568042]	5374	1
Full-MMLU post-minus-pre accuracy, grad-diff_layerA_seed0	paired item bootstrap; two-sided normal approximation	-0.08951716279	3.1430720307879245e-1032232	[-0.09799173906850876, -0.08175295541945592]	14042	1
Full-MMLU post-minus-pre accuracy, grad-diff_layerA_seed1	paired item bootstrap; two-sided normal approximation	-0.01915681526	3.8332249392933156e-18480273	[-0.02499643925366757, -0.01360205098988748]	14042	1
Full-MMLU post-minus-pre accuracy, rmu_layerA_seed0	paired item bootstrap; two-sided normal approximation	-0.09343398376	1.2009560899461727e-1289672	[-0.10183734510753453, -0.08566977638513032]	14042	1
Full-MMLU post-minus-pre accuracy, rmu_layerA_seed1	paired item bootstrap; two-sided normal approximation	-0.03446802444	2.2172101800243284e-2793477	[-0.04066372311636519, -0.02820111095285572]	14042	1
Full-MMLU post-minus-pre accuracy, rmu_layerA_seed2	paired item bootstrap; two-sided normal approximation	-0.01495513459	1.9252012203468246e-06211365	[-0.020225039168209658, -0.009898874804158951]	14042	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Full-MMLU post-minus-pre accuracy, rmu_layerB_seed0	paired item bootstrap; two-sided normal approximation	-0.016308218208535252	1.8916650013903052e-08	[-0.022005412334425295, -0.010822888477424878]	14042	1
Full-MMLU post-minus-pre accuracy, rmu_layerB_seed1	paired item bootstrap; two-sided normal approximation	-0.020153824241561032	1.6350132832118925e-11	[-0.026278307933342827, -0.01424298532972511]	14042	1
Full-MMLU post-minus-pre accuracy, rmu_layerB_seed2	paired item bootstrap; two-sided normal approximation	-0.032616436422070504	2.4244343218204286e-22	[-0.03931241988320752, -0.02577980344680245]	14042	1
RMU manipulation post-minus-pre forget MC, rmu_layerA_seed0	paired item bootstrap; two-sided normal approximation	-0.17413185703096124	1.3205541945417412e-07	[-0.20181177654755914, -0.14544539506794163]	1987	1
RMU manipulation post-minus-pre forget MC, rmu_layerA_seed1	unpaired difference in proportions (normal CI; item-level pairing unavailable)	-0.14041265842298767	2.7428369178006854e-06	[-0.17023406989247242, -0.11059124701350292]	1987	1
RMU manipulation post-minus-pre forget MC, rmu_layerA_seed2	unpaired difference in proportions (normal CI; item-level pairing unavailable)	-0.11877200007743866	9.927576405878796e-14	[-0.1488479291527507, -0.0886960709960225]	1987	1
RMU layerA mean manipulation post-minus-pre forget MC	unpaired difference in proportions (normal CI; item-level pairing unavailable)	-0.14443883299327576	1.9171190352927143e-20	[-0.1742092968485436, -0.11466836914800792]	1987	3

Claim	Test	Statistic	p	95% CI	n	Seeds
Five-shot causal ablation, grad-diff_layerA_seed0, forget	paired item bootstrap; two-sided normal approximation	-0.001006036217303823	0.5638336991363709	[-0.005030181086519115, 0.002012072434607646]	994	1
Five-shot re-elicitation minus zero-shot, grad-diff_layerA_seed0, forget	paired item bootstrap; two-sided normal approximation	-0.006036217303822937	0.7055780950797044	[-0.03822937625754527, 0.026156941649899398]	994	1
Five-shot causal ablation, grad-diff_layerA_seed0, retain	paired item bootstrap; two-sided normal approximation	-0.004024144869215292	0.47961070956341545	[-0.014109657947686115, 0.007042253521126761]	994	1
Five-shot re-elicitation minus zero-shot, grad-diff_layerA_seed0, retain	unpaired difference in proportions (normal CI; item-level pairing unavailable)	0.0661736029[illegible]	[illegible]05380[illegible]	[0.02[illegible]34574, 0.10416606936087099]	994	1
Five-shot causal ablation, grad-diff_layerA_seed1, forget	paired item bootstrap; two-sided normal approximation	0.0	1.0	[0.0, 0.0]	994	1
Five-shot re-elicitation minus zero-shot, grad-diff_layerA_seed1, forget	paired item bootstrap; two-sided normal approximation	-0.02716297786720322	0.007037777309232268	[-0.04627766599597585, -0.007042253521126761]	994	1
Five-shot causal ablation, grad-diff_layerA_seed1, retain	paired item bootstrap; two-sided normal approximation	-0.004024144869215292	0.39383357953598075	[-0.013078470824949699, 0.005030181086519115]	994	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Five-shot re-elicitation minus zero-shot, grad-diff_layerA_seed1, retain	unpaired difference in proportions (normal CI; item-level pairing unavailable)	0.009295828385612780239432	0.398342	[-0.028723709026504007, 0.047315365797728326]	994	1
Five-shot causal ablation, rmu_layerA_seed0, forget	paired item bootstrap; two-sided normal approximation	-0.001006036217303823	0.8658373607209519	[-0.013078470824949699, 0.01006036217303823]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerA_seed0, forget	paired item bootstrap; two-sided normal approximation	0.003018108651759138636499	0.745875	[-0.014084507042253521, 0.02112676056338028]	994	1
Five-shot causal ablation, rmu_layerA_seed0, retain	paired item bootstrap; two-sided normal approximation	-0.02112676056338028	0.05187574086803867	[-0.043259557344064385, 0.0]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerA_seed0, retain	unpaired difference in proportions (normal CI; item-level pairing unavailable)	-0.003773001477991811	0.8459897573766171	[-0.041845335242145065, 0.034299332286161444]	994	1
Five-shot causal ablation, rmu_layerA_seed1, forget	paired item bootstrap; two-sided normal approximation	-0.019114688128772636	0.3988385780268293	[-0.06338028169014084, 0.02716297786720322]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerA_seed1, forget	unpaired difference in proportions (normal CI; item-level pairing unavailable)	-0.06152264423053749	0.00032851181859147767	[-0.09509559136456, -0.027949697096514986]	994	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Five-shot causal ablation, rmu_layerA_seed1, retain	paired item bootstrap; two-sided normal approximation	-0.009054325955734407	0.3657561810413149	[-0.02819416498993963 4, 0.01006036217303823]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerA_seed1, retain	unpaired difference in proportions (normal CI, item-level pairing unavailable)	0.016845383270478328	0.0555769035167	[-0.0211838727643678 67, 0.05487463931381400 5]	994	1
Five-shot causal ablation, rmu_layerA_seed2, forget	paired item bootstrap; two-sided normal approximation	0.009054325955734407	0.0772190087183	[-0.0027188128772635815, 0.04426559356136821]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerA_seed2, forget	unpaired difference in proportions (normal CI, item-level pairing unavailable)	-0.08618141126705003	5.422314554736072e-07	[-0.11989216200412284, -0.05247066051797722]	994	1
Five-shot causal ablation, rmu_layerA_seed2, retain	paired item bootstrap; two-sided normal approximation	0.005030181086515921	0.1658496495	[-0.01006036217303823, 0.02112676056338028]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerA_seed2, retain	unpaired difference in proportions (normal CI, item-level pairing unavailable)	0.011816217026242726	0.0310263630361	[0.0026231730331909356, 0.04986416438439008]	994	1
Five-shot causal ablation, rmu_layerB_seed0, forget	paired item bootstrap; two-sided normal approximation	-0.004024144869215292	0.8481554970078566	[-0.04527162977867203, 0.03624245472837008]	994	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Five-shot re-elicitation minus zero-shot, rmu_layerB_seed0, forget	unpaired difference in proportions (normal CI, item-level pairing unavailable)	-0.0494542391727148	0.004618327855082665	[-0.08367481000002912, -0.015233668345400478]	994	1
Five-shot causal ablation, rmu_layerB_seed0, retain	paired item bootstrap; two-sided normal approximation	0.0	1.0	[-0.01006036217303823, 0.01006036217303823]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerB_seed0, retain	unpaired difference in proportions (normal CI, item-level pairing unavailable)	0.03546339603256715	0.0014054076211	[0.0025499443795024973, 0.07347673644464513]	994	1
Five-shot causal ablation, rmu_layerB_seed1, forget	paired item bootstrap; two-sided normal approximation	-0.006036217303822937	0.7265093737134546	[-0.04024144869215292, 0.02716297786720322]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerB_seed1, forget	unpaired difference in proportions (normal CI, item-level pairing unavailable)	-0.07460768170540138	1.4410806685589312e-05	[-0.10832084051184226, -0.040894522888960505]	994	1
Five-shot causal ablation, rmu_layerB_seed1, retain	paired item bootstrap; two-sided normal approximation	0.005030181086351193	0.3785577150644	[-0.006036217303822937, 0.01609657947686117]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerB_seed1, retain	unpaired difference in proportions (normal CI, item-level pairing unavailable)	0.021376381697257185	0.0005767579824	[0.016663947135180192, 0.05941671053027856]	994	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Five-shot causal ablation, rmu_layerB_seed2, forget	paired item bootstrap; two-sided normal approximation	-0.05432595573440644	0.005821960641952407	[-0.09154929577464789, -0.015065392354124886]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerB_seed2, forget	unpaired difference in proportions (normal CI; item-level pairing unavailable)	-0.05849996880746225	0.0005725355070122626	[-0.0917897134733097, -0.025210224141614793]	994	1
Five-shot causal ablation, rmu_layerB_seed2, retain	paired item bootstrap; two-sided normal approximation	[illegible]	[illegible]	[-0.007042253521126761, 0.014084507042253521]	994	1
Five-shot re-elicitation minus zero-shot, rmu_layerB_seed2, retain	unpaired difference in proportions (normal CI; item-level pairing unavailable)	[illegible]	[illegible]	[-0.012110395572033417, 0.06391285854705764]	994	1
LoRA relearning treatment-minus-disjoint control at step 100, grad-diff_layerA_seed0	paired item bootstrap; two-sided normal approximation	[illegible]	[illegible]	[illegible], 0.1398390342052314]	994	1
LoRA relearning treatment-minus-disjoint control at step 100, grad-diff_layerA_seed1	paired item bootstrap; two-sided normal approximation	[illegible]	[illegible]	[illegible], 0.18008048289738432]	994	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Primary-arm LoRA relearning advantage at step 100, mean across GradDiff seeds	paired item bootstrap; two-sided normal approximation	0.12072434602[illegible]	[illegible]	[illegible]	[illegible]	2
Claim-to-know difference-in-differences (forget minus matched retain), GradDiff primary arm, n_seeds=5	paired item bootstrap; two-sided normal approximation	0.24606741574[illegible]	[illegible]	[illegible]	[illegible]	5

Compute. Total recorded GPU time across all experiments is 3.2971 GPU-hours on a single-GPU workstation / NVIDIA GeForce RTX 5090.

Code and data availability. A public code repository URL is not recorded in `project.yaml` (`links.github`). The per-example data of record that backs every reported number is provided under `results/real/` in the project repository: `curve.csv`, `experiments.json`.

References

Alain, Guillaume, and Yoshua Bengio. 2016. "Understanding intermediate layers using linear classifier probes." *arXiv Preprint arXiv:1610.01644*. https://arxiv.org/abs/1610.01644.

Azaria, Amos, and Tom Mitchell. 2023. "The Internal State of an LLM Knows When It's Lying." *arXiv Preprint arXiv:2304.13734*. https://arxiv.org/abs/2304.13734.

Binder, Felix J., James Chua, Tomek Korbak, et al. 2024. "Looking Inward: Language Models Can Learn About Themselves by Introspection." *arXiv Preprint arXiv:2410.13787*. https://arxiv.org/abs/2410.13787.

Burns, Collin, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. "Discovering Latent Knowledge in Language Models Without Supervision." *arXiv Preprint arXiv:2212.03827*. https://arxiv.org/abs/2212.03827.

Deeb, Aghyad, and Fabien Roger. 2024. "Do Unlearning Methods Remove Information from Language Model Weights?" *arXiv Preprint arXiv:2410.08827*. https://arxiv.org/abs/2410.08827.

Eldan, Ronen, and Mark Russinovich. 2023. "Who's Harry Potter? Approximate Unlearning in LLMs." *arXiv Preprint arXiv:2310.02238*. https://arxiv.org/abs/2310.02238.

Hase, Peter, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. "Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge

Editing in Language Models." *arXiv Preprint arXiv:2301.04213*. https://arxiv.org/abs/2301.04213.

Hewitt, John, and Percy Liang. 2019. "Designing and Interpreting Probes with Control Tasks." *arXiv Preprint arXiv:1909.03368*. https://arxiv.org/abs/1909.03368.

Kadavath, Saurav, Tom Conerly, Amanda Askell, et al. 2022. "Language Models (Mostly) Know What They Know." *arXiv Preprint arXiv:2207.05221*. https://arxiv.org/abs/2207.05221.

Li, Nathaniel, Alexander Pan, Anjali Gopal, et al. 2024. "The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning." *arXiv Preprint arXiv:2403.03218*. https://arxiv.org/abs/2403.03218.

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. "Teaching Models to Express Their Uncertainty in Words." *arXiv Preprint arXiv:2205.14334*. https://arxiv.org/abs/2205.14334.

Łucki, Jakub, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. 2024. "An Adversarial Perspective on Machine Unlearning for AI Safety." *arXiv Preprint arXiv:2409.18025*. https://arxiv.org/abs/2409.18025.

Lynch, Aengus, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. "Eight Methods to Evaluate Robust Unlearning in LLMs." *arXiv Preprint arXiv:2402.16835*. https://arxiv.org/abs/2402.16835.

Maini, Pratyush, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. "TOFU: A Task of Fictitious Unlearning for LLMs." *arXiv Preprint arXiv:2401.06121*. https://arxiv.org/abs/2401.06121.

Marks, Samuel, and Max Tegmark. 2023. "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets." *arXiv Preprint arXiv:2310.06824*. https://arxiv.org/abs/2310.06824.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. "Locating and Editing Factual Associations in GPT." *arXiv Preprint arXiv:2202.05262*. https://arxiv.org/abs/2202.05262.

Meng, Kevin, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022. "Mass-Editing Memory in a Transformer." *arXiv Preprint arXiv:2210.07229*. https://arxiv.org/abs/2210.07229.

Rinberg, Roy, Usha Bhalla, Igor Shilov, Flavio P. Calmon, and Rohit Gandikota. 2025. "RippleBench: Capturing Ripple Effects Using Existing Knowledge Repositories." *arXiv Preprint arXiv:2512.04144*. https://arxiv.org/abs/2512.04144.

To, Bang Trinh Tran, and Thai Le. 2025. "Harry Potter is Still Here! Probing Knowledge Leakage in Targeted Unlearned Large Language Models via Automated Adversarial Prompting." *arXiv Preprint arXiv:2505.17160*. https://arxiv.org/abs/2505.17160.

Yuan, Hongbang, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. "Towards Robust Knowledge Unlearning: An Adversarial Framework for Assessing and Improving Unlearning Robustness in Large Language Models." *arXiv Preprint arXiv:2408.10682*. https://arxiv.org/abs/2408.10682.

Zhang, Ruiqi, Licong Lin, Yu Bai, and Song Mei. 2024. "Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning." *arXiv Preprint arXiv:2404.05868*. https://arxiv.org/abs/2404.05868.