Unlearning vs. Hiding: Does Unlearning Remove Knowledge or Just Suppress Its Expression?

Mackenzie O'Brien
Mac Labs AI

Lexsi AI Safety and Alignment Hackathon, Track 2b (problem statement 7)

Abstract

Unlearning should remove designated knowledge, but lower accuracy on direct questions does not show the knowledge is gone. We audit Representation Misdirection for Unlearning (RMU) on Tiny Aya, a 3.35B multilingual family, with SafeTune and AuditKIT. We inject fictitious authors from the Task of Fictitious Unlearning (TOFU) benchmark (`forget01`, 40 pairs) into Tiny Aya Global with Low-Rank Adaptation (LoRA), unlearn with RMU at layer 18 (updating layers 16–18), and repeat on Earth, Fire, Water and Base.

Method contribution. Standard RMU did nothing here. A 360-token system preamble, plus attention-sink tokens, dominates the forget loss, which stayed flat near 3.5. Four unmasked settings left forget ROUGE-L (answer overlap) at 0.92–0.97, versus 0.97 injected. Masking the forget loss to first-user-turn tokens ($\alpha = 100$, steering coefficient 20, 80 steps) lowers it to 0.32.

Results. Judge forget suppression (the fraction of injected facts an output does not reveal) is graded by `qwen3.5:35b` with a seeded 15% Kimi K3 re-grade. On Global (mean of seeds 42–44) it rises from 0.03 to 0.92 on direct questions, 0.33 to 0.95 on paraphrases and 0.08 to 0.95 on extraction prompts, and from 0.71 to 0.93 across eight machine-translated languages. Earth, Fire and Water behave alike (direct 0.875, 0.875 and 0.975 after RMU). Judged retain accuracy falls 1.6–9.4 points (Global mean 0.49 to 0.44). Format changes leak: the pooled cloze and true/false bucket reaches only 0.79, driven by true/false (0.70). On Base (not instruction-tuned), a format fix let injection pass, but the same settings moved direct suppression only from 0.00 to 0.10 and forget ROUGE-L from 1.000 to 0.913.

We do not claim erasure. After 30 fine-tuning steps the RMU model recovers 5, 7 and 7 of 8 forgotten facts across Global seeds 42–44, against 4, 4 and 6 of 8 new facts on the same model and 7, 7 and 4 of 8 for an uninjected base. The gap is one to three facts, too small to call significant. Held-out probes separate forget from retain questions at AUROC 0.92–1.00 from layer 1 up (0.87–1.00 on Base), as does a never-injected base, so they reflect surface differences, not stored knowledge. On Global (seed 42), a teacher-forced logit lens shows RMU lowers but does not remove the first content token probability (0.971 to 0.696, n=40). We read the results as output-level suppression on the four instruction-tuned variants, with weak evidence of residual knowledge, and as a failure of the same settings on Base. On Aya Expanse 8B the same settings gave direct suppression of only 0.20 and cut judged retain accuracy from 0.91 to 0.67, so we claim no cross-model generality. Cohen's kappa is uninformative when labels are one-sided, even if the judges agree, so we report raw and prevalence-robust agreement. Small samples and three Global seeds limit every conclusion. We found no published paper on unlearning Tiny Aya.

1 Introduction

Unlearning is proposed for privacy, copyright and hazardous-knowledge removal, but if it only hides knowledge, a deployer who checks direct prompts alone gets false assurance. Earlier work already shows that unlearned information is often recoverable [1, 2, 3]; we treat that as motivation, not as our finding.

We study one pipeline end to end. We teach a small multilingual model facts it cannot know from pretraining, unlearn them with Representation Misdirection for Unlearning (RMU) [4], and ask three questions.

RQ1. *How complete is behavioral forgetting?* How far does a judge-based forget metric rise under direct, paraphrased, reformatted, extraction-style and translated questions, and what does it cost on retain and general-capability questions?

RQ2. *Is the knowledge still recoverable?* Do held-out probes, a logit-lens readout of the answer token, or relearning give evidence of residual knowledge?

RQ3. *Do the answers differ across Tiny Aya variants, RMU settings and one larger model?*

Why the resulting model is meaningfully better for usage. Out of the box, RMU left Tiny Aya Global unchanged (Table 1). With a masked forget loss it suppresses the injected facts under paraphrase,

extraction and translation at a retain cost of 1.6–9.4 points. The audit shows what that does not guarantee: format-change leakage remains, and 30 fine-tuning steps recover most forgotten facts. It also does not transfer by default: on Tiny Aya Base the same settings barely suppressed direct questions (0.10 after RMU; Table 5), and on Aya Expanse 8B they suppressed them only partly and cost as much on retain facts (Table 7).

Contributions. This is a measurement and method-fix study, and we do not claim to discover shallow unlearning (Section 2). In order of importance, we contribute four things. (i) *Masked-forget-loss RMU.* On Tiny Aya the shared system preamble and attention-sink tokens dominate the forget loss, so unmasked RMU barely forgets; masking those positions fixes it on Tiny Aya with no other change to the SafeTune loop (Section 3), though the same masked settings only partly worked on Aya Expanse 8B. (ii) *One protocol across Tiny Aya variants.* We compare Global [5] (three seeds) with Earth, Fire, Water and Base (one seed each) under one protocol; Base required a prompt-format fix, described below, before the injection gate passed. An informal web search on 14 November 2026 found no published paper on unlearning Tiny Aya, but it did find concurrent, unreviewed repositories that apply Negative Preference Optimization (NPO) and RMU to Tiny Aya (Section 2); we have not run a systematic literature search (Section 7). We add one cross-model case, Aya Expanse 8B, run once with the Tiny Aya settings and no tuning. Probes and relearning ran on all five Tiny Aya variants, while the logit lens, activation patching and CircuitKIT circuit discovery ran on Global seed 42 only (CircuitKIT output is not analyzed in this version). (iii) *Multilingual leakage paired with internal tests.* We score eight-language queries next to held-out probes (with a base-model control), relearning and a teacher-forced logit-lens readout of the answer token. This follows up Farashah et al. [6], whose analysis is behavioral. It differs from the Unlearning Depth Score (UDS) work [7], which evaluates on forget10 of the Task of Fictitious Unlearning (TOFU) benchmark (Lee et al., Sec. 4.1) and neither tests other languages nor uses linear probes. (iv) *Mid-depth RMU, relearning and judge-agreement reporting.* We apply RMU at layers 16–18 of 36 instead of the library's early default (layers 5–7), measure a per-step relearning curve, and report agreement with statistics that are insensitive to label prevalence.

2 Related Work

Benchmarks and methods. TOFU [8] supplies fictitious author profiles with known ground truth, and we use its `forget01`, `retain99`, `world_facts` and `real_authors` splits. RMU [4] steers intermediate activations on forget data toward a random vector while anchoring retain activations; we use the method, not the Weapons of Mass Destruction Proxy (WMDP) data. Negative Preference Optimization (NPO) [9] treats the forget set as negative preference pairs to avoid the collapse of gradient-ascent unlearning. Who's Harry Potter [10] is the Eldan and Russinovich 2023 method that fine-tunes a model to forget the Harry Potter books. We do not run either method here.

Unlearning often hides rather than removes. ConceptVectors [1] finds that existing methods barely change the parameter vectors that encode a concept and mostly suppress them at inference, and that ablating those vectors removes the knowledge. For RMU, fine-tuning on 10 unrelated examples or removing activation directions recovers most hazardous capability [2]. Unlearning changes how knowledge is retrieved, mainly through final-layer multilayer perceptron (MLP) coefficients, rather than what is stored [11]. Other routes recover it too: quantization [12], tampering (state-of-the-art unlearning is undone within 16 fine-tuning steps [13]), benign relearning [14], activation steering [15] and, for an extraction attack, hidden states [16]. Defenses exist. Distilling an unlearned model into a fresh student [17] and localizing edits to factual-recall components [18] both resist relearning.

Counterpoint. In an eight-method study, RMU and Tampering Attack Resistance (TAR) resisted straightforward prompt attacks while Erasure of Language Memory (ELM) did not [19]; a clean result here would fit that.

Multilingual unlearning. Farashah et al. [6] study Aya-Expanse 8B on TOFU translated into ten languages (en, fr, ar, ja, ru, fa, ko, hi, he, id), with a 1% forget set. They unlearn in each source language with Gradient Difference (GradDiff), GradDiff-KL and NPO and measure transfer to the others, using behavioral metrics. Forgetting stays mostly in the unlearning language with limited, asymmetric transfer, and syntactic distance best predicts transfer. Cross-lingual unlearning has also been studied as an adaptive-weighting method [20]. In multilingual models fine-tuned on parallel data, a model may answer in the wrong language after unlearning, which breaks reference-based metrics [21]. Our test is a follow-up, not a new idea. It differs in using RMU rather than gradient-based objectives, Tiny Aya (3.35B), English-only injection and unlearning (so the question is whether it leaks into other languages), and internal recoverability and relearning next to behavioral scores.

Concurrent work on Tiny Aya. We found no published paper on unlearning Tiny Aya. We did find unreviewed repositories that appear to be entries to the same hackathon and apply SafeTune to it. One applies RMU

for hazardous biology (WMDP-bio) at layers 7–9, chosen with CircuitKIT, and audits 4-bit files, including on Earth and Fire [22]. Another applies NPO to a TOFU fine-tune of Tiny Aya Global and studies 4-bit compression [23]. Our setting differs (injected TOFU facts, a masked RMU loss, multilingual queries, and probes, relearning and a logit lens). We did not compare against them or check whether the system-preamble problem of Section 3 arises in their setups, and we describe them only from their public pages.

Measuring depth. Layer-wise activation patching as a depth score for unlearning, with a logit lens as comparison, is the UDS [7]. Activation patching and the logit lens are established tools for testing whether knowledge persists internally [24, 25]. We run both on Global but do not present either as a new way to measure unlearning depth, and our patching covers one author and a single-token measure. UDS evaluates on TOFU forget10 (Lee et al., Sec. 4.1) and does not test other languages or use linear probes [7]. We also use linear probes [26] and Low-Rank Adaptation (LoRA) [27]. CircuitKIT with edge attribution patching with integrated gradients (EAP-IG) [28, 29] ran on Global, but its circuits are not analyzed in this version and are left to future work.

Caveat on injected facts. Deeb and Roger [3] recover most pre-unlearning accuracy by fine-tuning on related facts. They also note that evaluations on information learned during an additional fine-tuning phase, as our LoRA injection of TOFU facts is, may overestimate how resistant unlearning is.

3 Method

Model and data. Our model is `CohereLabs/tiny-aya-global` [5], an instruction-tuned multilingual (70 languages) model with 3.35B parameters, 36 decoder layers and $d_{\text{model}} = 2048$. We run it in bfloat16 with its chat template; Earth, Fire and Water share the architecture. Data come from TOFU [8], and `forget01` (40 pairs, two fictitious authors) is the forget set. Injection and the RMU retain data use the first 400 of the 3,960 `retain99` rows (20 authors); the relearning control draws its new facts from held-out authors beyond that slice. We score retain accuracy on the first 64 retain rows. General capability is TOFU `world_facts` and `real_authors`, 40 items each for Global seed 42 and 20 each for every other run; we never inject or unlearn these sets, and we score them before and after RMU. Paraphrases are the official `forget01_perturbed` questions.

Fact injection (English only). We inject, and later unlearn, facts in English only; translations serve only as queries. The injection is a LoRA [27] fine-tune on the shuffled union of `forget01` (each pair shown 4 times per epoch) and the 400-row retain slice (loss on answer tokens). It uses rank 16, $\alpha = 32$, dropout 0.05, all attention and MLP projections, 5 epochs, learning rate 2×10^{-4}, batch 8 with 2 accumulation steps, sequence length 512 and bfloat16. We merge the adapter into the full weights before RMU and apply RMU to the merged weights. Injection is partial on retain facts (Section 5).

RMU with a masked forget loss. SafeTune's RMU trains the MLP output projections of the update layers on the following loss, written with mean squared error (MSE),

$$\mathcal{L} = \text{MSE}\big(h^{\ell}_{\text{forget}},\, c\,u\big) + \alpha\ \text{MSE}\big(h^{\ell}_{\text{retain}},\, h^{\ell}_{\text{retain,frozen}}\big), \tag{1}$$

with h^{ℓ} the residual activation at steering layer ℓ, u a fixed random unit vector, c the steering coefficient, and the frozen reference a copy of the starting model. SafeTune averages the forget MSE over every position, and that is where Tiny Aya breaks it. Every example starts with the same system preamble of about 360 tokens. Roughly 83% of positions in a forget batch are therefore identical to retain positions, and the first two positions (beginning-of-sequence, BOS, and start-of-turn, attention sinks) have a residual norm near 23. On one batch of 8 forget rows at layer 18 on the injected model, the forget MSE was 3.50, of which the two sink positions contributed about 2.5. The loss therefore cannot fall without damaging the retain term, and the gradient on the roughly 60 tokens that carry the fact is diluted about sevenfold. Table 1 shows four unmasked settings that left the forget loss flat and the model unchanged, as measured by forget ROUGE-L recall (ROUGE-L is the longest-common-subsequence variant of Recall-Oriented Understudy for Gisting Evaluation, ROUGE). We keep SafeTune's loop, hooks, parameter selection and AdamW and change one thing. We take the forget MSE only over non-pad tokens from the first user-turn token on (including the start-of-turn token). The retain term still covers all positions, so the preamble and sinks stay pinned to the frozen model. We use steering layer 18, update layers $\{16, 17, 18\}$, $\alpha = 100$, $c = 20$, learning rate 5×10^{-4} and 80 steps (5 forget batches and 50 retain batches, cycled). The library default is layer 7 with update layers $\{5, 6, 7\}$.

Behavioral metrics. We decode greedily (up to 64 new tokens). Our headline metric is `judge_forget_suppression` $= 1-$ the primary judge's match rate on forget prompts, that is, the fraction of injected facts the output does not reveal (higher is better). We report it per query format and per language, including translated prompts. The secondary

Setting	α	lr	Loss	F-RL	F-sub	R-RL
Injected, no RMU	–	–	–	0.967	0.825	0.662
Unmasked	1200	5e−5	3.52→3.54	0.974	0.825	0.658
Unmasked	1200	5e−4	3.52→3.55	0.963	0.750	0.628
Unmasked	100	5e−4	3.52→3.49	0.921	0.675	0.632
Masked	100	5e−4	0.88→0.43	**0.322**	**0.000**	0.619

Table 1: RMU diagnostics on the injected Global checkpoint (layer 18, 80 steps, 40 forget and 64 retain questions, direct prompts, no judge). Loss is the forget loss from first to last step; F-RL and F-sub are forget ROUGE-L recall and substring recall, and R-RL is retain ROUGE-L. The base model's forget ROUGE-L is 0.357. One run per setting; these runs were used to set α and the learning rate, on the same 40 forget questions that are later evaluated.

metrics are `judge_retain_accuracy` on the retain slice and `judge_utility_accuracy` on `world_facts` and `real_authors` (the judge's match rate), both before and after RMU. We also report substring recall of the full reference answer (lowercased). A failed judgment is missing, not zero; we exclude it from the rate and report n_{judged} and n_{missing}. The query formats are *direct*, *paraphrase*, *cloze*, *true/false* (a statement containing the correct answer, target "True"), *extraction* ("Tell me everything you know about the following topic" with the TOFU question as topic) and *translation* (eight languages). The pipeline scores cloze and true/false as one 80-item bucket, and we split it using item-level verdicts. We constructed the cloze, true/false and extraction prompts for this study, and we judged the base model on direct prompts only.

Judging and agreement. The primary judge is `qwen3.5:35b`, served locally through Ollama (temperature 0, thinking off, JSON verdict with a boolean `match`). A secondary judge, `moonshotai/kimi-k3` (Kimi K3) through OpenRouter, re-grades a seeded random 15% sample of the same cases with the same rubric. We report raw agreement, Cohen's κ, prevalence-adjusted bias-adjusted kappa (PABAK, $2p_o - 1$) and Gwet's first-order agreement coefficient (AC1), all computed from the paired verdicts stored with each run. Cohen's κ is a poor summary when almost all labels fall in one class: it can be near zero or undefined even if the judges agree on nearly every item, because the chance-agreement term then approaches the observed agreement (the kappa paradox). That is why we also report raw agreement and the two prevalence-robust measures, PABAK and AC1, which stay high when the judges agree.

Multilingual leakage. We machine-translate forget questions and answers into es, fr, de, hi, ar, zh, sw and ja with the local helper model, which is the primary judge's model (`qwen3.5:35b`), and use them only as queries. A seeded review sheet per language was written, but no native speaker has checked the translations (Section 7).

Linear probes. At each of the 37 residual positions we take the activation at the last non-padding token. A logistic probe (standardized features, balanced classes) then separates the 40 forget questions from the 64 scored retain questions. All numbers are held out, from a seeded stratified 70/30 split repeated 5 times for Global seed 42 and 3 times for every other run. We report the mean and standard deviation of the area under the receiver operating characteristic curve (AUROC). *Frozen* probes are fit on the post-injection model and applied after RMU, whereas *refit* probes are fit on each model's own training rows. The controls are a refit probe on the base model before injection and a probe fit on shuffled labels. A *content* probe separates a question followed by its true answer from the same question followed by another forget answer. Forget-versus-retain probes can partly reflect question topic.

Relearning. We fine-tune for 30 steps (learning rate 2×10^{-4}, batch size 1, one run per arm) and score at steps $\{0, 5, 10, 20, 30\}$ for Global seed 42 and $\{0, 10, 30\}$ for every other run. The *RMU-old* arm trains the post-RMU model on 8 forgotten facts. The *RMU-new* arm trains a fresh copy of the same model on 8 new fictitious TOFU authors held out from the forget and retain sets (rows 420–427 of `retain99`, beyond the 400-row slice). The *base-old* arm trains the never-injected base model on the same 8 forgotten facts; it is an acquisition-from-scratch reference, not a matched control. We also score the old-fact arms on held-out paraphrases from `forget01_perturbed`, while the new facts have none. RMU-old against RMU-new on trained questions is the comparison that isolates suppression, because both arms share weights, schedule and scoring. Substring match is scored at every step, and the judge runs at the first and last step only.

Logit lens. On Global we project each of the 37 residual positions to the vocabulary with the output head, for the injected and RMU checkpoints, both with the injected checkpoint's tokenizer. Residual 0 is the embeddings and residual $i + 1$ the output of decoder layer i, so RMU's edit of decoder layers 16–18 acts at residuals 17–19. Items are the 40 `forget01` facts and a 32-item retain control. We score the probability and rank of the answer's first content token (openers such as "The" and "B" are skipped and teacher-forced into the prompt) and, separately, of its first answer token, with 95% Student-t intervals over items.

4 Experimental Setup

Pipeline and runs. The pipeline runs inject → verify → unlearn → eval → probes → interp → relearn, and it logs every query under a stable case identifier, so before and after scores pair the same questions. The

Global run reuses the injected checkpoint of an earlier run whose RMU step was unmasked; we copied its pre-RMU numbers and recomputed everything after RMU. Earth, Fire and Water ran the full pipeline once each with seed 42. Global seeds 43 and 44 also ran the full pipeline from a fresh injection; the seed sets the trainer, the injection shuffle and the RMU sampling, and the judge seed is fixed. Seeds 43 and 44 differ from the seed-42 Global run (and from Earth, Fire and Water) in running the primary judge only, so they have no agreement statistics, and in having no interpretability stage; like Earth, Fire and Water they score 20 utility questions per set, relearning at steps $\{0, 10, 30\}$ and 3 probe splits. Tiny Aya Base failed the injection gate twice because of a prompt-format bug (Section 7). A corrected run with a `plain_template` option, which applies the training format (`User:`/`Assistant:` text) at injection, evaluation and unlearning, passed the gate and then ran unlearn, evaluation, probes and relearning, with the primary judge only, 20 utility questions per set, relearning at steps $\{0, 10, 30\}$, 3 probe splits and no interpretability stage. The logit lens is analyzed on Global seed 42 only (Section 5.2), and Appendix A reports an activation-patching rerun with controls on 10 facts of one author. Head-ablation and CircuitKIT artifacts exist for Global but are not analyzed in this version, and we leave them to future work.

Aya Expanse 8B run. As a cross-model case we ran one job on `CohereLabs/aya-expanse-8b` (about 8B parameters, 2.4 times Tiny Aya's 3.35B; 32 layers, $d_{\text{model}} = 4096$) with inject, verify, unlearn, eval and relearn, and no probes or interpretability stages. Injection, RMU and evaluation settings are Tiny Aya Global's, carried over unchanged. Nothing was tuned on Expanse: layer 18 with update layers 16–18 (18 of 32 layers here, 18 of 36 on Tiny Aya), $\alpha = 100$, $c = 20$, learning rate 5×10^{-4}, 80 steps and the content mask. The mask starts at token 1 here and at token 359 on Tiny Aya. Only the model, 20 instead of 40 utility questions, relearning steps $\{0, 10, 30\}$ and the omitted stages differ (Table 6). No sweep ran. The secondary judge did not run either, so Expanse has no agreement statistics. A host hang interrupted the job during relearning (exit code 137) after inject, verify, unlearn and eval had finished; relearning alone was rerun from the same RMU checkpoint and exited normally. The runs are `20261114T061138Z-inject+verify+unlearn+eval+relearn-6d3be84b` and `20261114T151251Z-relearn-6d3be84b`.

RMU layer sweep. We ran five masked configurations on Global from the same injected checkpoint and judged each on a seeded subset of 20 forget and 20 retain direct questions (same items throughout). We set layer 18 and steering 20 before any sweep (mid-depth, not tuned) and set α and the learning rate in the diagnostics of Table 1. The main run started before the masked sweep finished. Four earlier unmasked sweep points (layers 7 and 18, $\alpha \in \{600, 1200, 1800\}$, learning rate 5×10^{-5}) scored 0.05 on the same subset, meaning no forgetting.

Compute. We ran on a single workstation GPU with 128 GB unified memory, about 2.1 hours per Tiny Aya variant job (Earth 127.5, Fire 128.9 and Water 127.3 minutes; Global seeds 43 and 44 took 177.0 and 130.2 minutes); the Base unlearn–eval–probes–relearn job took 60.6 minutes after a separate validation injection; the Expanse job was cut off by a host hang after 113 minutes and its relearning rerun took 11 minutes. The secondary judge made 1,074 logged calls over all runs, including the sweeps and smoke tests, at a mean of \$0.00106 per call (\$1.14 in total); Global seeds 43 and 44, Base and Expanse made none.

5 Results

Injection. Before RMU the judge reveals the injected fact on 97.5% of Global direct forget questions (suppression 0.025), while the base model before injection reveals it on 5% (suppression 0.95, $n = 40$). Retain facts are injected only partly: judged retain accuracy is 0.0625 for the base model ($n = 64$) and 0.516 after injection. The retain cost below is therefore measured on facts the model knew about half the time. The injection gate (forget ROUGE-L recall at least 0.80) passed in every completed run (Global seeds 42–44: 0.974, 0.971, 0.961; Earth 0.950, Fire 0.957, Water 0.922; Base 1.000, after the `plain_template` fix), and judged retain accuracy after injection is 0.516, 0.469 and 0.484 on Global seeds 42–44 (Earth 0.562, Fire 0.531, Water 0.516, Base 0.688).

Behavioral suppression. On Tiny Aya, RMU suppresses the injected facts on every format we tried (Aya Expanse 8B is the exception, below). On Global seed 42, `judge_forget_suppression` rises from 0.025 to 0.85 on direct questions, from 0.45 to 0.95 on paraphrases and from 0.125 to 0.95 on extraction prompts ($n = 40$ each; Wilson 95% intervals after RMU 0.71–0.93 for direct and 0.83–0.99 for paraphrase and extraction), and from 0.706 to 0.925 on the pooled translated prompts ($n = 320$; 0.89–0.95; Table 2). Substring recall of the full reference answer falls from 0.825 to 0 on direct questions, and Earth, Fire and Water show the same pattern (Water 0.775 to 0). Global seeds 43 and 44 show it too (direct suppression 0.026 to 0.975 and 0.050 to 0.925; Table 3), so the seed-42 direct figure of 0.85 is the lowest of the three seeds (mean 0.917). Suppression after RMU is higher on paraphrase and extraction prompts than on direct questions, which we did not expect and cannot

	Global (seed 42)		Earth		Fire		Water	
Judge metric	before	after	before	after	before	after	before	after
Forget suppression, `judge_forget_suppression` (higher is better)								
Direct (40)	0.025	**0.850**	0.050	**0.875**	0.075	**0.875**	0.075	**0.975**
Paraphrase (40)	0.450	**0.950**	0.325	**0.950**	0.400	**0.875**	0.275	**1.000**
Cloze (40)	0.100	0.850	0.075	0.825	0.175	0.850	0.100	0.950
True/false (40)	0.225	0.625	0.175	0.625	0.075	0.475	0.200	0.575
Cloze + T/F pooled (80)	0.163	**0.738**	0.125	**0.725**	0.125	**0.663**	0.150	**0.763**
Extraction (40)	0.125	**0.950**	0.050	**0.925**	0.075	**0.900**	0.075	**0.925**
Translated, 8 languages (320)	0.706	**0.925**	0.716	**0.934**	0.709	**0.925**	0.697	**0.953**
Accuracy on non-forget facts (judge)								
Retain (64)	0.516	0.422	0.562	0.516	0.531	0.500	0.516	0.484
World facts (40 / 20 / 20 / 20)	0.875	0.850	0.900	0.900	0.900	0.900	0.900	0.900
Real authors (40 / 20 / 20 / 20)	0.825	0.775	1.000	1.000	0.950	0.950	0.900	0.850
Substring recall of the reference answer (secondary; lower is more forgetting)								
Direct	0.825	0.000	0.825	0.000	0.725	0.025	0.775	0.000
Paraphrase	0.100	0.000	0.100	0.000	0.075	0.000	0.075	0.000
Extraction	0.475	0.000	0.525	0.025	0.375	0.025	0.425	0.000
Cloze + T/F pooled	0.663	0.325	0.738	0.300	0.725	0.400	0.625	0.300

Table 2: Judge forget suppression ("before" is the injected model before RMU), judged accuracy on retain and general-capability questions, and substring recall. Item counts in parentheses. Global is seed 42 (Table 3 gives seeds 43 and 44); the other variants are single runs. All items were judged except one Global post-RMU translated item ($n_{\text{missing}} = 1$; Hindi, 39 of 40 judged); every other cell has $n_{\text{missing}} = 0$. No confidence intervals; with 40 questions one item moves a rate by 2.5 points (Wilson 95% interval for Global direct after RMU: 0.71–0.93).

explain. Translation leaves little headroom, because the injected model already withheld the fact on 71% of translated prompts before RMU (Spanish 0.575, French 0.55, Japanese 0.875; Table 4), so the pooled gain is 0.22, not 0.92. The substring metric cannot help there (0.00 before RMU), and we rely on the judge.

Seed spread. Table 3 lists Global seeds 42, 43 and 44 with the mean and range; with $n = 3$ we report neither a standard deviation nor a confidence interval. After RMU the suppression is 0.85 or more on direct questions and 0.91 or more on paraphrase, extraction and translated prompts in all three seeds, and seed 42 is the lowest on direct, cloze, true/false and the pooled bucket. Three seeds change three statements we first made from seed 42 alone. (1) The retain cost is 9.4, 4.7 and 1.6 points, so the 9.4 of seed 42 is the largest, and seed 44's 1.6 points is one question of 64. (2) True/false suppression after RMU ranges from 0.625 to 0.750, and in seed 43 true/false substring recall after RMU is 0.875, against 0.65 and 0.60 in seeds 42 and 44. (3) The relearning comparisons change (paragraph on relearning below). The other conclusions hold in all three seeds: high suppression on direct, paraphrase, extraction and translated prompts, a modest retain cost, and the pooled cloze and true/false bucket as the weakest format.

Variants. Earth, Fire and Water follow Global within a few points on most formats (Table 2). Water is higher than Global seed 42 on direct (0.975), paraphrase (1.000) and translated prompts (0.953) and lower on true/false (0.575). After RMU, Hindi is the weakest language in Global seed 42, Earth and Fire (0.872, 0.875, 0.825), and in Fire, the South Asian variant [5], its suppression did not rise (0.850 to 0.825). Hindi is not the weakest elsewhere: it scores 0.950 in Water, where Arabic is lowest (0.875), and the weakest language is Arabic (0.850) in Global seed 43 and Hindi (0.850) in seed 44. We therefore do not read Hindi as systematically weak, and with 40 questions per language we do not interpret the Fire result. Fire's true/false suppression after RMU (0.475) is the lowest cell in Table 2.

Tiny Aya Base. The earlier injection failures were a prompt-format bug in our code, not a failure to learn (Section 7). With a `plain_template` option, a validation injection under the original recipe (5 epochs, forget repeats 4) passed the gate at forget ROUGE-L recall 1.000 (substring recall 0.975; retain ROUGE-L 0.751). The full run reused that adapter and applied the same RMU settings as the instruction-tuned variants (layer 18, updates at 16–18, $\alpha = 100$, steering coefficient 20, 80 steps, content mask). It did not suppress the facts. Direct `judge_forget_suppression` rose only from 0.000 to 0.100 ($n = 40$; 4 of 40 withheld; Wilson 95% interval 0.04–0.23), against 0.85–0.975 after RMU on Global, Earth, Fire and Water (Tables 2 and 5). Forget ROUGE-L recall fell from 1.000 to 0.913, and substring recall from 0.975 to 0.600, not to 0. On the

	Seed 42	Seed 43	Seed 44	Mean	Range after
Forget suppression, `judge_forget_suppression` *(before → after RMU)*					
Direct (40)	0.025→0.850	0.026→0.975	0.050→0.925	0.034→0.917	0.850–0.975
Paraphrase (40)	0.450→0.950	0.275→0.950	0.275→0.950	0.333→0.950	0.950–0.950
Cloze (40)	0.100→0.850	0.000→0.900	0.050→0.900	0.050→0.883	0.850–0.900
True/false (40)	0.225→0.625	0.075→0.750	0.325→0.725	0.208→0.700	0.625–0.750
Cloze + T/F pooled (80)	0.163→0.738	0.038→0.825	0.188→0.813	0.129→0.792	0.738–0.825
Extraction (40)	0.125→0.950	0.050→0.975	0.050→0.925	0.075→0.950	0.925–0.975
Translated, 8 languages (320)	0.706→0.925	0.706→0.934	0.706→0.919	0.706→0.926	0.919–0.934
Accuracy on non-forget facts (judge)					
Retain (64)	0.516→0.422	0.469→0.422	0.484→0.469	0.490→0.438	0.422–0.469
World facts (40 / 20 / 20)	0.875→0.850	0.900→0.850	0.850→0.850	0.875→0.850	0.850–0.850
Real authors (40 / 20 / 20)	0.825→0.775	1.000→1.000	1.000→0.950	0.942→0.908	0.775–1.000
Retain change (points)	−9.4	−4.7	−1.6	−5.2	−9.4 to −1.6
Substring recall (before → after RMU)					
direct	0.825→0.000	0.800→0.000	0.775→0.000	0.800→0.000	0.000–0.000
cloze + T/F pooled	0.663→0.325	0.800→0.438	0.700→0.300	0.721→0.354	0.300–0.438
Relearning, judge match out of 8 (step 0 → step 30)					
RMU-old, trained	0→5	1→7	1→7	0.7→6.3	5–7
Base-old, trained	2→7	2→7	2→4	2.0→6.0	4–7
RMU-new, trained	0→4	0→4	0→6	0.0→4.7	4–6
Injected forget ROUGE-L (gate 0.80)	0.974	0.971	0.961	0.969	0.961–0.974

Table 3: Global under seeds 42, 43 and 44, one cell per seed and metric (before → after RMU). Seed 42 is the run of Table 2. Seeds 43 and 44 started from a fresh injection, used the primary judge only, scored 20 world-facts and 20 real-authors questions, and scored relearning at steps 0, 10 and 30; the seed-42 pre-RMU numbers come from the reused earlier injection. Mean is the unweighted mean over the three seeds of the before and after values, and Range is the minimum and maximum of the three after values. With $n = 3$ we give no standard deviation or interval. One pre-RMU direct judgment failed in seed 43 (39 judged) and one post-RMU Hindi judgment in seed 42 (319 judged); every other cell has $n_{\text{missing}} = 0$. The seeds share the same 40 forget questions and the same 8 + 8 relearning facts, so they vary training and RMU randomness, not the questions.

run's own pre-RMU evaluation, retain ROUGE-L is 0.741, against the 0.751 recorded at the injection gate, and 0.714 after RMU. Paraphrase suppression moved from 0.275 to 0.450, extraction from 0.050 to 0.100, and the eight-language pool from 0.569 to 0.619. Cloze rose from 0.000 to 0.150; true/false from 0.200 to 0.225, while true/false substring recall increased from 0.675 to 0.725. Judged retain accuracy was 0.688 before RMU and 0.703 after ($n = 64$), a one-question rise, and world-facts and real-authors scores were unchanged or slightly higher ($n = 20$ each). The primary judge scored every item ($n_{\text{missing}} = 0$); the secondary judge was not run, as in Global seeds 43 and 44. After RMU, Spanish remains the weakest language (0.300) and Swahili the strongest (0.875); Hindi and Arabic fall slightly (0.800 to 0.725 and 0.825 to 0.750; Table 4). Because suppression never took hold, the relearning curve is not evidence of hidden knowledge: at step 0 the RMU model already judged-matched 7 of 8 trained forgotten facts, then 8 of 8 at step 30, against 2 to 8 for the never-injected base and 0 to 7 for new facts (Table 11). We read this as a failure of the same masked-RMU recipe on the model that is not instruction-tuned, not as a test of erasure.

5.1 A second model: Aya Expanse 8B

On the one larger model we ran, the Tiny Aya RMU settings suppressed the forget facts only partly and cost about as much on retain facts as they gained on forget questions. This is one run (seed 42), with 40 forget and 64 retain questions, 20 utility questions per set and the primary judge only. Table 6 lists how the setup differs from Global, and Table 7 gives every cell.

Injection. Injection worked on Expanse. Forget ROUGE-L recall rises from 0.321 for the base model to 0.998 and forget substring recall from 0.000 to 0.925 (the verify gate is 0.80 on ROUGE-L; $n = 40$). Retain ROUGE-L recall rises from 0.357 to 0.892 and retain substring recall from 0.000 to 0.656 ($n = 64$). The primary judge finds the injected fact in 3 of 40 direct forget answers of the base model (0.075) and in 40 of 40 after injection, and it scores retain answers correct on 7 of 64 for the base model (0.109) and 58 of 64 after injection (0.906). Retain facts were therefore learned far better than on Global, whose judged retain accuracy after injection is 0.516, so the retain cost below is measured on facts the model knew nine times in ten, not half the time.

	Global (seed 42)		Earth		Fire		Water		Base	
Language	bef.	aft.	bef.	aft.	bef.	aft.	bef.	aft.	bef.	aft.
es	0.575	0.925	0.625	0.950	0.675	0.875	0.600	0.975	0.225	0.300
fr	0.550	0.950	0.600	0.975	0.675	0.950	0.625	0.950	0.400	0.525
de	0.625	0.950	0.500	0.925	0.575	0.975	0.550	1.000	0.325	0.475
hi	0.750	0.872*	0.700	0.875	0.850	0.825	0.775	0.950	0.800	0.725
ar	0.750	0.950	0.875	0.925	0.700	0.900	0.825	0.875	0.825	0.750
zh	0.775	0.925	0.850	0.950	0.700	0.975	0.750	0.975	0.500	0.600
sw	0.750	0.900	0.725	0.950	0.675	0.925	0.700	0.950	0.775	0.875
ja	0.875	0.925	0.850	0.925	0.825	0.975	0.750	0.950	0.700	0.700

Table 4: `judge_forget_suppression` by question language (40 questions each), before and after RMU; Global is seed 42. Translations are machine-generated. *One of 40 Hindi judgments failed (39 judged). Every Base cell has $n_{\text{missing}} = 0$.

Judge metric	before	after
Forget suppression, `judge_forget_suppression` (higher better)		
Direct (40)	0.000	0.100
Paraphrase (40)	0.275	0.450
Cloze (40)	0.000	0.150
True/false (40)	0.200	0.225
Cloze + T/F pooled (80)	0.100	0.188
Extraction (40)	0.050	0.100
Translated (320)	0.569	0.619
Non-forget accuracy (judge)		
Retain (64)	0.688	0.703
World facts (20)	0.850	0.850
Real authors (20)	0.850	0.900
Substring recall (lower is better)		
Direct	0.975	0.600
Paraphrase	0.150	0.075
Extraction	0.850	0.600
Cloze	0.900	0.500
True/false	0.675	0.725
Cloze + T/F pooled	0.788	0.613
Forget ROUGE-L (direct)	1.000	0.913
Retain ROUGE-L (direct)	0.741	0.714

Table 5: Tiny Aya Base, one run, before and after RMU with the same settings as the instruction-tuned variants. Item counts in parentheses; $n_{\text{missing}} = 0$ in every cell. The secondary judge was not run. Forget ROUGE-L at the injection gate was 1.000 (retain 0.751); the before column uses the run's own pre-RMU evaluation.

Forgetting by format. After RMU, `judge_forget_suppression` on direct questions rises only from 0.000 to 0.200 (Wilson 95% interval 0.11–0.35), against 0.025 to 0.850 on Global, while direct substring recall falls from 0.975 to 0.500. Paraphrase suppression rises from 0.275 to 0.400 (0.26–0.55; substring 0.125 to 0.100) and extraction from 0.000 to 0.100 (0.04–0.23; substring 0.900 to 0.575). The judge changed from match to no match on 8 of 40 direct, 6 of 40 paraphrase and 4 of 40 extraction questions, and in the other direction on 0, 1 and 0. Every after-RMU suppression on these formats is far below Global's 0.85–0.95, and the intervals are wide at $n = 40$.

Format change on Expanse. The pooled cloze and true/false bucket (80 items) goes from 0.025 to 0.088 on the judge and from 0.925 to 0.700 on substring recall, and the split shows two different stories. On cloze the judge moves from 0.000 to 0.100 and substring recall from 0.850 to 0.400 (judge flips 4 of 40 from match to no match). On true/false the judge moves from 0.050 to 0.075 and substring recall stays at 1.000 before and after (2 flips one way, 1 the other). As on Global, a true/false prompt contains the correct statement, so we cannot say how much of the true/false figure is leakage, and the pooled substring drop comes from cloze alone.

Translation on Expanse. Translation is uninformative here, as on Global. The injected Expanse model already withholds the fact on 67% of translated prompts before RMU (suppression 0.672, $n = 320$), and RMU raises that to 0.738 (0.69–0.78), a gain of 0.066. The headroom varies by language, from Swahili at 1.000 both before and after to Spanish at 0.425 before and 0.625 after (Table 7). The substring metric starts near zero (0.006 before and 0.003 after, $n = 320$; 2 of 320 and 1 of 320 hits) because the reference answer for a translated question is the full translated reference sentence, and the metric needs it verbatim in the output. A free-form answer rarely reproduces a whole sentence, even when the judge counts it as revealing the fact (judge match 0.328 before RMU), so we rely on the judge.

	Tiny Aya Global (seed 42)	Aya Expanse 8B
Model	`tiny-aya-global`	`aya-expanse-8b`
Parameters; decoder layers; d_{model}	3.35B; 36; 2048	about 8B (2.4 times Global); 32; 4096
RMU steering layer; update layers	18; 16–18	18; 16–18
Steering layer as a fraction of depth	$18/36 = 0.50$	$18/32 = 0.56$
α; c; learning rate; steps	100; 20; 5×10^{-4}; 80	identical
Forget-loss mask start (logged)	token 359	token 1
Injection: LoRA rank 16, $\alpha = 32$, 5 epochs, 400 retain rows, forget pairs $\times 4$	as stated	identical
Layer and α sweep	five masked configurations	none
Utility questions per set	40	20
Relearning scored at steps	0, 5, 10, 20, 30	0, 10, 30
Probes; logit lens; patching	probes on all four variants, lens and patching on Global	not run
Secondary judge (15% re-grade)	ran	did not run
Seeds	42	42

Table 6: Setup of the Aya Expanse 8B run against Tiny Aya Global. The RMU, injection and evaluation settings are the same configuration values; the rows that differ are the model, the logged mask start, the utility subset size, the relearning grid and the stages that were not run. Nothing was tuned on Expanse.

Retain facts were hurt as much as forget facts. Judged retain accuracy falls from 0.906 to 0.672 (Wilson 95% interval 0.55–0.77 after), a 23.4-point drop, against a 20.0-point gain in direct forget suppression. On the substring metric retain recall falls 50.0 points (0.719 to 0.219) and direct forget recall 47.5 points (0.975 to 0.500). Item by item, the judge flipped 16 of 64 retain questions from correct to incorrect (one went the other way), against 8 of 40 direct forget questions flipped from match to no match. In absolute points, then, RMU removed slightly more judged accuracy from retain facts than it removed forget answers, and RMU did not separate the two on this model. With 40 and 64 questions the 3.4-point difference between the two judged changes is within noise, so we claim that the effects are of the same size, not that retain suffered more. World facts and real authors stay at judged accuracy 1.000 before and after ($n = 20$ each, Wilson interval 0.84–1.00, so a small cost would not show), and real-author substring recall moves from 0.95 to 0.90.

What we cannot attribute. The Expanse run used Tiny Aya's settings unchanged, and nothing was tuned on Expanse, so we do not attribute the weak effect to model size. Two differences are confounded with size, and we did not test either. The steering layer is 18 of 32 decoder layers on Expanse and 18 of 36 on Tiny Aya (0.56 against 0.50 of depth), and the logged forget-loss mask starts at token 1 on Expanse against token 359 on Tiny Aya. Hyperparameters were also set on Tiny Aya's evaluated questions (Section 7). No layer, α or coefficient sweep ran on Expanse, so we do not know whether other settings would forget more or damage retain facts less.

Relearning on Expanse. The relearning run (8 facts, steps 0, 10 and 30; Table 8) is uninformative, and we do not use it as evidence for or against residual knowledge. At step 0 the RMU model still reproduces 7 of the 8 forget answers verbatim (substring 0.875, against 0.500 on all 40 direct questions), so little was suppressed on these 8 facts. At step 10 every arm scores 0 of 8 on substring recall. The generations at that step are degenerate: 12 of the 16 never-injected base-arm outputs are a single period, three of the 8 RMU-new outputs repeat a word several times in a row ("Ethiopia Ethiopia Ethiopia"), and the RMU-old outputs answer different questions with similar off-target text, some garbled ("Promise by bythe the the Seine"). The judge did not run at step 10, and we do not know why the step-10 generations degrade. At step 30 the forget facts and the new facts on the same model score the same on substring recall (4 of 8, 0.500; judge 5 of 8 against 4 of 8), and the never-injected base reaches 2 of 8 (judge 3 of 8). On the 8 held-out paraphrases of the forget facts the RMU model scores 0.250, 0.000 and 0.125 at steps 0, 10 and 30. RMU-old falls from 0.875 to 0.500 on its own trained questions over the 30 steps, so the run shows no recovery on them. With an RMU model that already answered most of these facts, a step-10 collapse in every arm and a forget-versus-new gap of zero, the run cannot separate suppression from erasure.

Utility cost. Judged retain accuracy falls by 9.4 points on Global seed 42 (0.516 to 0.422; Wilson 95% intervals 0.40–0.63 before and 0.31–0.54 after, $n = 64$), 4.7 on Earth, 3.1 on Fire and 3.1 on Water (0.516 to 0.484), and by 4.7 and 1.6 points on Global seeds 43 and 44 (0.469 to 0.422 and 0.484 to 0.469; Table 3). Over the

	Judge		Substring		Wilson 95% interval
	before	after	before	after	of the judge rate after
Forget questions					
Direct (40)	0.000	0.200	0.975	0.500	0.11–0.35
Paraphrase (40)	0.275	0.400	0.125	0.100	0.26–0.55
Cloze (40)	0.000	0.100	0.850	0.400	0.04–0.23
True/false (40)	0.050	0.075	1.000	1.000	0.03–0.20
Cloze + T/F pooled (80)	0.025	0.088	0.925	0.700	0.04–0.17
Extraction (40)	0.000	0.100	0.900	0.575	0.04–0.23
Translated, 8 languages (320)	0.672	0.738	0.006	0.003	0.69–0.78
es (40)	0.425	0.625	0.000	0.000	0.47–0.76
fr (40)	0.475	0.550	0.000	0.000	0.40–0.69
de (40)	0.525	0.550	0.025	0.025	0.40–0.69
hi (40)	0.875	0.925	0.000	0.000	0.80–0.97
ar (40)	0.775	0.775	0.000	0.000	0.63–0.88
zh (40)	0.625	0.675	0.000	0.000	0.52–0.80
sw (40)	1.000	1.000	0.000	0.000	0.91–1.00
ja (40)	0.675	0.800	0.025	0.000	0.65–0.90
Other facts					
Retain (64)	0.906	0.672	0.719	0.219	0.55–0.77
World facts (20)	1.000	1.000	1.000	1.000	0.84–1.00
Real authors (20)	1.000	1.000	0.950	0.900	0.84–1.00

Table 7: Aya Expanse 8B, one run (seed 42), before and after RMU with the Tiny Aya settings carried over unchanged. Item counts in parentheses; $n_{\text{missing}} = 0$ in every cell. On forget rows the judge column is `judge_forget_suppression` (higher is more forgetting) and the substring column is recall (lower is more forgetting); on other rows the judge column is judged accuracy. The pooled bucket is cloze plus true/false. Per-language rows are the translated prompts split by language. The interval is a Wilson 95% interval on the after-RMU judge rate (suppression on forget rows, accuracy on the others), with n as stated.

	Substring			Judge	
Arm and questions	s0	s10	s30	s0	s30
RMU-old, trained	0.875	0.000	0.500	0.875	0.625
RMU-old, paraphrase	0.250	0.000	0.125	0.500	0.125
Base-old, trained	0.000	0.000	0.250	0.250	0.375
Base-old, paraphrase	0.125	0.000	0.125	0.250	0.375
RMU-new, trained	0.000	0.000	0.500	0.000	0.500

Table 8: Relearning on Aya Expanse 8B (8 facts per arm, $n = 8$ questions per cell; one run). Substring recall at steps 0, 10 and 30; the judge ran at steps 0 and 30 only. New facts have no paraphrases.

three Global seeds the mean drop is 5.2 points (range 1.6–9.4), and seed 44's drop is one question of 64. With 64 retain questions the drop in any single run is within sampling noise (intervals of about ±0.12). It has the same sign in all six runs, so we report it as a real but small cost; across the six runs the range is 1.6–9.4 points. Global seed 42 world facts fall 2.5 points (0.875 to 0.850) and real authors 5.0 points (0.825 to 0.775), $n = 40$ each. On 20 items each, Earth and Fire are unchanged, Water world facts are unchanged (0.900) and its real authors fall from 0.900 to 0.850 (one question), seed 43 world facts fall from 0.900 to 0.850 with real authors at 1.000 before and after, and seed 44 world facts stay at 0.850 with real authors from 1.000 to 0.950.

Residual leakage under format change. Format changes still leak. Substring recall on the pooled cloze and true/false bucket stays at 0.325 after RMU on Global seed 42 (seeds 43 and 44: 0.438 and 0.300; Earth 0.300, Fire 0.400, Water 0.300), against at most 0.025 on direct, paraphrase and extraction prompts. Splitting the bucket (Table 2) shows that cloze recall is 0 after RMU while true/false carries the leak (substring recall 0.65; judge suppression 0.625, i.e. judge match 0.375, so the two metrics disagree on this format). Cloze recall is also 0 after RMU in seeds 43 and 44 and in Water, and true/false substring recall is 0.875, 0.600 and 0.600 there (judge suppression 0.750, 0.725 and 0.575). A true/false prompt contains the correct statement, so a model that never learned the fact would accept some of them. We did not evaluate the base model on this format and cannot say how much of the 0.625 is leakage. The pooled bucket (0.74, $n = 80$; Wilson 95% interval 0.63–0.82) is the weakest format, true/false is its weaker half (0.625, $n = 40$; 0.47–0.76), and we report it as residual leakage. Across the three Global seeds the pooled bucket is 0.74–0.83 and true/false 0.63–0.75, and the pooled bucket is the weakest format in every run.

Judge agreement. On the 16 paired items of the base-model pass the judges agree on 15 (raw 0.9375),

Judge agreement	n	raw	κ	PABAK / AC1
Global, base-model pass	16	0.938	0.000	0.875 / 0.934
Global, before RMU	100	0.970	0.939	0.940 / 0.941
Global, after RMU	99	0.929	0.829	0.859 / 0.880
Earth, before + after	188	0.936	0.861	0.872 / 0.882
Fire, before + after	188	0.872	0.738	0.745 / 0.753
Water, before + after	188	0.941	0.877	0.883 / 0.889

Table 9: Agreement between `qwen3.5:35b` and `moonshotai/kimi-k3` on the seeded 15% sample. The base-model pass has 15 of 16 pairs labeled "no match" by the primary judge and none "match", so $\kappa = 0$ although the judges disagree on one item. The Earth, Fire and Water base-model passes (16 pairs each) give raw agreement 0.938, 0.875 and 0.875 with $\kappa = 0$. Global seeds 43 and 44 and Expanse have no secondary judge.

Configuration	Suppr.	Retain
L7, $\alpha = 100$, $c = 20$	1.00	0.55
L18, $\alpha = 50$, $c = 20$	1.00	0.40
L18, $\alpha = 100$, $c = 20$ (main)	0.85	0.45
L18, $\alpha = 200$, $c = 20$	0.50	0.60
L18, $\alpha = 100$, $c = 40$	1.00	0.50

Table 10: Layer sweep on Global: judge suppression (Suppr.) on direct forget questions and judged retain accuracy, seeded subset of 20 forget and 20 retain questions. Not comparable to Table 2.

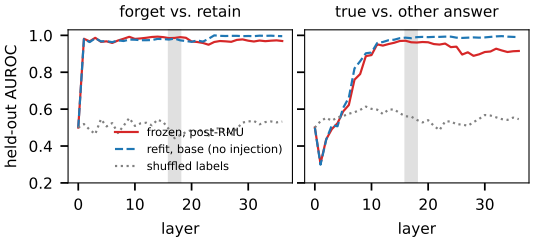


Figure 1: Held-out probe AUROC by layer on Global (mean over 5 splits; shaded: update layers 16–18). A probe fit on the injected model and applied after RMU, a probe refit on the base model before injection, and shuffled labels. Layer 0 is the embedding output.

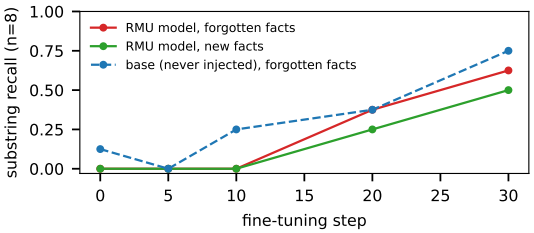


Figure 2: Relearning on Global seed 42: substring recall on the 8 trained questions at each step. Judge match at step 30 is 5/8 (RMU, forgotten facts), 4/8 (RMU, new facts) and 7/8 (base, forgotten facts).

yet Cohen's κ is 0.0: the primary judge never says "match" there, so chance agreement equals observed agreement. PABAK is $2 \times 0.9375 - 1 = 0.875$ and AC1 is 0.934. On the evaluation passes, where both labels occur, κ is 0.83–0.94 on Global and 0.74–0.88 on Earth, Fire and Water (Table 9). Disagreements mostly run one way. In Global, all 10 of 199 disagreements have the secondary judge saying "match" and the primary not (Earth 9 against 3, Fire 19 against 5, Water 7 against 4), so primary-judge suppression may be overstated by a few points.

Sweep. Two other configurations scored better than the chosen one on this subset: layer 7 with $\alpha = 100$ (1.00 and 0.55) and layer 18 with $c = 40$ (1.00 and 0.50; Table 10). We do not claim the main configuration is optimal. We fixed it before the masked sweep finished and chose it without this subset. With 20 items per cell, one item moves a rate by 5 points, so the subset cannot separate these configurations. Every masked configuration forgot at least partly (0.50–1.00), while the unmasked sweep points all scored 0.05, so the mask matters more than the layer on this subset.

Probes. Held-out forget-versus-retain probes decode at AUROC 0.98 from layer 1 on Global seed 42 (accuracy 0.96, before RMU) and, after RMU, stay between 0.949 and 0.993 across layers 1–36 on that run (Figure 1). Over all six instruction-tuned runs (Global seeds 42, 43 and 44, Earth, Fire and Water; mean over 5 splits for seed 42 and 3 for the others) the frozen post-RMU AUROC ranges from 0.920 (Global seed 44, layer 23) to 1.000 (Water, layer 3). The per-run ranges are 0.949–0.993, 0.939–0.997, 0.920–0.996, 0.949–0.996, 0.965–0.996 and 0.946–1.000 in the order Global seeds 42, 43, 44, Earth, Fire, Water, with the lowest values at layers 22–23 in every run. On Base (3 splits) the frozen post-RMU range is 0.865 (layer 15) to 0.997 (layer 24), and that same 0.865 minimum is already present before RMU at layer 15, so the dip is again not an RMU effect. The same frozen probes score as low as 0.909–0.937 at their weakest layer before RMU, so the mid-layer dip is not an RMU effect. The base model before injection, which never saw the facts, does the same (base-model refit probes range 0.949–1.000 over layers 1–36 across the six runs): a refit probe reaches 0.981 at layer 1, 0.977 at layer 16 and 0.973 at layer 18; shuffled labels sit at 0.46–0.53. Earth, Fire and Water agree (base-model refit AUROC 0.994, 0.994 and 0.991 at layer 1), as do Global seeds 43 and 44 (0.994 each, 3 splits). High scores from layer 1, equal on the base model, point to lexical differences between the question sets, such as fictitious names, not to stored facts. The content probe agrees (0.96 after RMU at layer 18, 0.99 for the base-model refit). The probe results are uninterpretable as knowledge retention: they do not show that facts survive RMU, and a lower score would not have shown erasure.

Judge match, step 30 (of 8)	RMU-old	Base-old	RMU-new
Global seed 42, trained	5	7	4
Global seed 43, trained	7	7	4
Global seed 44, trained	7	4	6
Earth, trained	6	6	5
Fire, trained	6	8	5
Water, trained	7	6	5
Base, trained	8	8	7
Global seed 42, paraphrases	3	3	–
Global seed 43, paraphrases	6	5	–
Global seed 44, paraphrases	3	4	–
Earth, paraphrases	2	4	–
Fire, paraphrases	6	2	–
Water, paraphrases	3	3	–
Base, paraphrases	7	6	–

Table 11: Relearning after 30 steps, judge match out of 8 facts. At step 0 the trained-question counts are 0, 2 and 0 for Global seed 42 (RMU-old, base-old, RMU-new), 2, 2 and 0 for Earth, 1, 2 and 0 for Global seeds 43 and 44, Fire and Water, and 7, 2 and 0 for Base. All runs use the same 8 forgotten and 8 new facts, so the rows are not independent replicates.

Relearning. On Global seed 42, the RMU model recovers 5 of 8 forgotten facts after 30 steps (judge match 0.625, from 0 of 8), against 4 of 8 for new facts on the same model (0.5). A never-injected base model trained on the same facts reaches 7 of 8 (0.875; 0.53–0.98) (Figure 2, Table 11). The comparison that isolates suppression is RMU-old against RMU-new on trained questions, and the other five runs change how it reads. On the judge the gap is positive in all six runs: one fact for Global seed 42 (5 against 4), Earth (6 against 5), Fire (6 against 5) and Global seed 44 (7 against 6), two for Water (7 against 5) and three for Global seed 43 (7 against 4). On substring recall the gap is mixed: RMU-old minus RMU-new is $+1$, $+5$, -1, $+2$, $+3$ and $+1$ facts for Global seed 42, seed 43, seed 44, Earth, Fire and Water. We do not present any single gap as significant (the Wilson intervals at $n = 8$ overlap), and the six runs reuse the same 8 forgotten and 8 new facts, so they are not independent replicates. The fair reading is that the RMU model relearns old facts slightly faster than new ones: weak evidence of residual knowledge at best, consistent in sign on the judge. The comparison with the base model is mixed. We initially expected the never-injected base to learn the same facts at least as fast as the RMU model in every variant; the data do not show that. On the judge the base reaches more forgotten facts than the RMU model in Global seed 42 (7 against 5) and Fire (8 against 6), the same number in Global seed 43 (7) and Earth (6), and fewer in Global seed 44 (4 against 7) and Water (6 against 7); on substring recall the RMU model is ahead of the base in four of the six runs. In two runs the RMU model recovered the forgotten facts faster than learning from scratch, which fits residual knowledge but which one 8-fact sample per run cannot confirm. Held-out paraphrases are mixed (Table 11). The base arm already answers 2 of 8 trained questions before training, so it is an acquisition reference, not a matched control. Base does not join the six-run comparison: after RMU the model already judged-matched 7 of 8 trained forgotten facts at step 0, then 8 of 8 at step 30, so there was little suppression to recover from. Substring recall on those trained questions was 0.75 at step 0, 0.00 at step 10 and 1.00 at step 30 ($n = 8$); we report the dip but do not interpret it.

5.2 Logit lens at the output token

The curves coincide through residual 16 (Figure 3) because RMU changes no weight below decoder layer 16; this holds by construction and is not a finding. Before the last layers the lens is uninformative for forget and retain facts alike. At residual 16 the mean probability is 0.0023 on forget facts (mean rank about 10^5) and 0.016 on retain facts (mean rank 697). At residual 30 it is still near 0.03 for forget facts in both models, and it rises sharply only in the last layers (residual 18: 0.00263 injected against 0.00074 RMU; residual 19: 0.0086 against 0.0037). So a low mid-network readout is not evidence that the answer cannot be decoded there. The lens can neither confirm nor rule out an internal representation, and a tuned lens [25] or probes would be needed. Our probes cannot settle it either, because of the surface-feature confound.

The substantive result is at the output (residual 36). The injected model puts mean probability 0.971 [0.946, 0.995] on the correct first content token (rank 1 on all 40 items). RMU lowers this to 0.696 [0.567, 0.824] but does not remove it. The token is still rank 1 on 30 of 40 items (median rank 1), like the retain control (0.696 injected, 0.695 under RMU, $n = 32$). On the 34 facts the judge scored as suppressed the mean is still 0.648 [0.502, 0.793] (rank 1 on 24 of 34). The 6 still matching reach 0.968 (too few to interpret; Figure 4). This is the model's own next-token distribution under teacher forcing, not a hidden one. It fits suppressed rather than erased: the answer is often still the most likely next token after the prompt, while greedy generation on the 34 suppressed facts does not reveal them.

The measure is single-token and teacher-forced, and the first-content-token rule matters: it skips an opener on 18 of 40 forget items ("B" before "asil" on 13). On the first answer token, without a prefix, RMU gives 0.582 [0.442, 0.722] (rank 1 on 25 of 40; injected 0.974). The gap comes from those 18 items, where it gives 0.285 on the opener against 0.538 on the content token (0.825 on the other 22). Mean ranks (66 content, 1616 answer) are dominated by a few items: one opener at rank 64,290 gives most of 1616, which is 9.3 without it. We therefore lead with probability, median rank and the fraction at rank 1, computed from per-item ranks at residual 36. Items cluster on two authors ("Nik" 13 and "asil" 13 of

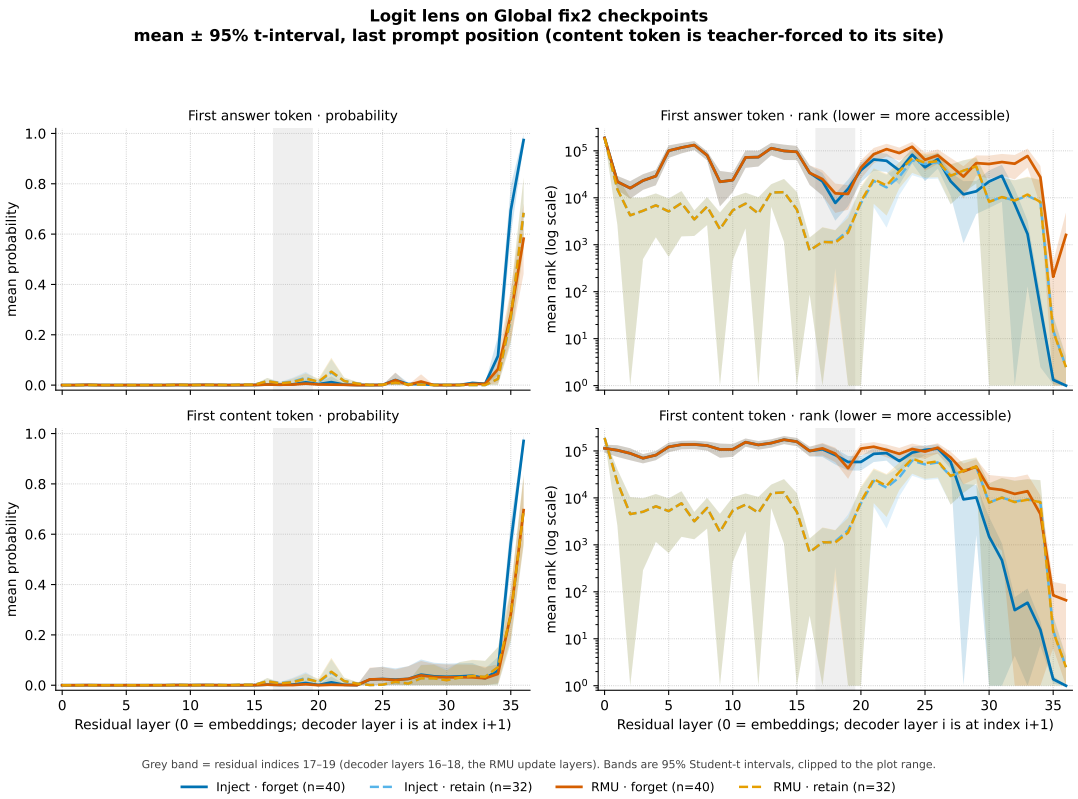


Figure 3: Logit lens on Global (method in Section 3): mean over items with 95% Student-t intervals ($n = 40$ forget, $n = 32$ retain); blue injected, orange RMU, solid forget, dashed retain. Top row first answer token, bottom row first content token; left probability, right mean rank on a log scale (dominated by a few items, see text). Grey band: residuals 17–19, the RMU update layers. Curves coincide through residual 16 by construction.

40 content tokens; RMU probability 0.939 and 0.385), so intervals over items are optimistic.

6 Discussion

What the results support. With a masked forget loss, RMU suppresses the injected facts across the four instruction-tuned Tiny Aya variants (and three seeds of Global), and the suppression holds under paraphrase, extraction prompts and translation into eight languages, at a utility cost of 1.6–9.4 points on retain facts. That is the claim we make, and it is confined to those instruction-tuned Tiny Aya checkpoints. On Tiny Aya Base, which is not instruction-tuned, the same settings moved direct suppression only from 0.00 to 0.10 and left forget ROUGE-L at 0.913, so the recipe failed there. On the one other model we ran, Aya Expanse 8B, the same settings suppressed direct questions only partly (0.20) and cost about as much on retain facts (23.4 points) as they gained on forget questions (20.0 points), so we make no cross-model claim. The claim does not establish erasure either. Format-change leakage lacks a base-model reference, relearning is not consistently faster or slower than learning from scratch (on the judge the RMU model is ahead of the base in two of six runs, behind in two and level in two), and the RMU-old against RMU-new gap is one to three facts at $n = 8$, positive in all six runs. The evidence is consistent with suppression rather than erasure but does not distinguish them.

Aya Expanse 8B. The Expanse run weakens any claim that the masked-RMU recipe is general. With Tiny Aya's settings carried over, direct suppression reached 0.20 against 0.85 on Global, and retain facts lost 23.4 points of judged accuracy against a 20.0-point forget gain (8 of 40 forget and 16 of 64 retain questions flipped), so RMU was neither strong nor selective on that model. We cannot say why. Model size is confounded with the layer index (18 of 32 layers against 18 of 36), with the mask start (token 1 against token 359) and with the absence of any tuning, and Expanse's retain facts were also learned far better than Global's (0.906 against 0.516), a further difference we did not isolate. The relearning run adds nothing either way. We therefore treat Expanse as a documented partial failure of the transferred settings, not as evidence about scale or about residual knowledge.

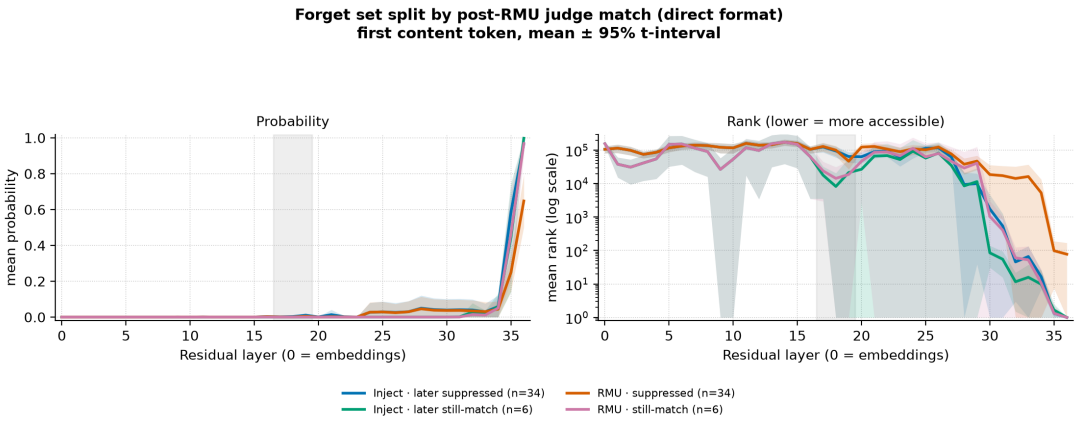


Figure 4: Forget facts split by the post-RMU judge match (direct format), first content token: 34 suppressed and 6 still matching. The group of 6 is too small to interpret. Left probability, right mean rank on a log scale, with 95% Student-t intervals.

How to read null internal results. WMDP reports linear probes only slightly above chance after RMU [4], yet fine-tuning on 10 unrelated examples or removing activation directions recovers most hazardous capability [2], so a null probe would not have shown erasure either. Our probes fail in the other direction. They are high before and after RMU and equally high on the base model, so they cannot speak to knowledge in either direction. A probe that separates question topics is not a probe for the fact. Probes that compare a fact with its counterfactual, together with the base-model control, would be needed. The logit lens adds an output-level result, not an internal one. It is uninformative before the last layers for forget and retain facts alike, and at the output it shows RMU lowering the teacher-forced probability of the first content token without removing it, which fits suppression rather than erasure. Our single-site patching restores only part of the first-token log-probability gap, inconsistently across layers and sites, so it neither localizes nor excludes an internal representation.

Consistency with prior work. Suppression that survives paraphrase and translation, leaks under some prompt formats and is largely regained within 30 fine-tuning steps (only slightly faster than learning new facts here) fits the retrieval-change accounts [1, 11] and the tampering and relearning results [13, 14]. RMU resisting straightforward prompt attacks [19] agrees with our direct, paraphrase and extraction results; our format-change leakage is a prompt change of a kind they did not report.

Multilingual. On the instruction-tuned Tiny Aya variants, English-only unlearning of English-injected facts transferred to all eight languages, with judged suppression after RMU at least 0.85 in every language of every run except Fire on Hindi (0.825). Base is the exception: after RMU, six of the eight languages remain below 0.75, and Spanish is 0.300. Farashah et al. [6] report mostly language-specific forgetting, but the setups differ: they unlearn in each language with gradient-based objectives on an 8B model. Headroom is small, since the injected model already withheld 70–72% of translated facts on Tiny Aya (67% on Expanse).

Practice. Check that the forget loss moves (a flat loss and unchanged ROUGE-L is a no-op, as here before masking), that evaluation uses the same prompt format as training (the Base injection failures were a format mismatch), score paraphrase, extraction, translation and format-change prompts with a judge, and report agreement with prevalence-robust statistics.

7 Limitations

- *Small samples, few seeds.* 40 forget questions from two authors, 64 retain questions, 20–40 utility questions, relearning on 8 facts per arm, a sweep on 20 and 20 items. Aya Expanse 8B has the same 40 forget and 64 retain questions, 20 utility questions and relearning on 8 facts. Only Global has more than one seed (42, 43, 44); Earth, Fire, Water, Base, Expanse and the sweep are single runs. With $n = 3$ we report per-seed values, the mean and the range (Table 3), not a standard deviation or a confidence interval, and the seeds share the same questions and relearning facts. Seeds 43 and 44 used the primary judge only and ran no interpretability stage. We computed no other confidence intervals except where stated.
- *Hyperparameters were tuned on the evaluated questions.* α and the learning rate were set in diagnostic runs scored on the same 40 forget questions, so the post-RMU numbers are not held-out with respect to

that choice. The config is not shown to be optimal (Table 10), and it was carried over unchanged to Base and to Aya Expanse 8B.

- *Injected facts.* An extra fine-tuning phase added the facts, which may make unlearning look more resistant than it is for pretraining knowledge [3], and retain facts were injected only partly (judged accuracy 0.52 on the instruction-tuned variants, 0.688 on Base).
- *English-only injection.* This asks a different question from within-language unlearning [6].
- *Judge.* The headline is a model judge. Agreement with a second judge is high on evaluation passes but disagreements lean one way (primary stricter), so suppression may be overstated by a few points, and Global seeds 43 and 44, Base and Expanse have no second judge. Translations are machine-generated by the primary judge's own model (`qwen3.5:35b`) and no native speaker has checked them, so some translated questions may be wrong or ambiguous and the same model both writes and grades the translated queries. A review sheet exists but was not completed, and the translated-prompt and per-language results should be read with that in mind.
- *Probes and relearning.* Probes are uninterpretable as knowledge retention (the base model scores the same) and may be confounded by topic. Relearning uses one schedule on 8 facts per arm, RMU-old and RMU-new differ by one to three facts, the new-fact arm has no paraphrases, and the base arm is an acquisition reference, not a matched control.
- *Logit lens and patching.* The lens covers Global seed 42 only, with a single-token, teacher-forced measure shaped by the first-content-token rule, two clustered authors and 6 still-matching items. Its mid-layer readouts are uninformative for forget and retain facts alike, so it cannot show whether an internal representation remains. The patching rerun covers 10 facts of one author with a single-token measure, shows no full-answer restoration and has no head- or component-level patching. Head ablation and CircuitKIT are not analyzed.
- *One cross-model case, untuned.* Aya Expanse 8B ran once with Tiny Aya's settings, no sweep, no probes or interpretability stages, the primary judge only and 20 utility questions. RMU was partial and not selective there, so Tiny Aya's results do not generalize to it, and we do not know why. The layer index (18 of 32 against 18 of 36) and the mask start (token 1 against 359) differ from Tiny Aya and are confounded with model size. Its relearning run is uninformative (degenerate step-10 generations; forget and new facts both 4 of 8 on substring recall at step 30).
- *Base (diagnosed format bug; RMU then failed).* Tiny Aya Base (not instruction-tuned) failed the injection gate twice, at forget ROUGE-L recall 0.703 and then 0.729 after doubling epochs (10) and forget repeats (8), against a gate of 0.80, and was not unlearned in those attempts. The cause was a prompt-format mismatch in our code, not a failure to learn. Base has no chat template; injection trained on text of the form `User: question` then `Assistant: answer`, but the gate and evaluation fed the bare question. A diagnostic rescoring of the saved second-attempt adapter, without training, gives forget ROUGE-L recall 0.722 with the bare question (18 of 40 questions at or above 0.80; substring recall 0.350) and 1.000 in the training format (40 of 40; substring recall 0.975). The 0.722 comes from a separate single-prompt script and differs by 0.007 from the 0.729 that the run's own gate gave for the same adapter and the same bare question; we did not trace that difference. We added a `plain_template` option that applies the training format at injection, evaluation and unlearning. A rerun with the original recipe (5 epochs, forget repeats 4) passed the gate at forget ROUGE-L recall 1.000 (substring recall 0.975; retain ROUGE-L 0.751) and then completed unlearning, evaluation, probes and relearning with the same RMU settings as the instruction-tuned variants. Direct `judge_forget_suppression` rose only from 0.000 to 0.100 and forget ROUGE-L recall fell only from 1.000 to 0.913 (Table 5), so the corrected run is a failure of that recipe on Base, not a missing result. Base used the primary judge only and had no interpretability stage.
- *Format-change leakage.* The base model was not evaluated on cloze or true/false prompts, so true/false leakage cannot be separated from accepting a statement placed in the prompt.
- *Literature search.* Our statement that no published paper unlearns Tiny Aya rests on an informal web search on 14 November 2026, not a systematic review. It found concurrent, unreviewed repositories (Section 2), and other work may exist that we did not find.
- *Coverage.* Global (three seeds), Earth, Fire and Water (one seed each), plus one Base run and one Expanse run; two models from one vendor; general capability is TOFU `world_facts` and `real_authors`, not broad benchmarks.

8 Ethics and Responsible Use

The forget set is synthetic: TOFU's fictitious authors contain no real personal data, and we use no hazardous-knowledge data. We report aggregate recoverability, not a recipe for extracting private information from deployed systems. A clean-looking result here would not certify that real-world unlearning works.

Licenses. TOFU, the source of our forget, retain and utility data, is MIT-licensed (its fictitious authors were generated with GPT-4). The Tiny Aya models and Aya Expanse 8B are licensed CC-BY-NC-4.0 and also require adherence to Cohere Labs' acceptable-use policy, which among other things prohibits generating synthetic data from the models for commercial purposes. Our released checkpoints, traces and generations derive from or contain outputs of those models, so we release them for non-commercial use only, under the same terms, and not under an open-source license. The translations are outputs of `qwen3.5:35b` (Apache 2.0 on its Hugging Face release; we did not check the Ollama packaging separately) from TOFU text. The judge verdicts come from that model and from Kimi K3, whose own license permits use and carries conditions aimed at large commercial model-as-a-service operators and attribution for very large products. We found no term in these licenses that prevents releasing the artifacts for non-commercial use. The license texts do not say whether derivatives of the gated Tiny Aya weights may be shared without a gate, and we have not asked Cohere, so we treat that as unconfirmed. We did not review OpenRouter's or the model providers' terms of service for outputs, and none of this is legal advice.

9 Reproducibility

- Code at `github.com/mso-docs/Lexsi-Unlearning` (main branch), with configurations for the main run, the variants, the sweep, the Base job and the Expanse job.
- Data from TOFU (`forget01`, `retain99`, `forget01_perturbed`, `world_facts`, `real_authors`), plus translations, the review sheet and constructed prompts in the dataset `huggingface.co/datasets/mackenzietechdocs/lexsi-unlearning-eval`; each artifact's license is in Section 8.
- Seeds: 42 for every run and 42, 43 and 44 for Global. The seed sets the trainer, the injection shuffle and the RMU sampling; the judge seed is fixed (20261112).
- Artifacts are the unlearned checkpoints in the Global RMU model repo, with Earth, Fire and Water counterparts under the same namespace (CC-BY-NC-4.0, non-commercial), plus probe states, traces, generations and verdict logs in the same dataset. Gated base weights are not redistributed.
- Tools are SafeTune 0.1.18 (RMU), AuditKIT 1.1.11 (scoring) and CircuitKIT 0.2.3 (EAP-IG; run on Global, not analyzed in this version). They ran in one container image (built 2026-11-12, image ID sha256:a2ea1f2be522) with Python 3.12.3, PyTorch 2.15.0a0+875d815502 (NVIDIA build 26.10, CUDA 13.4), Transformers 5.22.2, PEFT 0.21.2, Accelerate 1.15.0, TRL 1.14.2, Datasets 4.7.0, Hugging Face Hub 1.33.0, scikit-learn 1.9.1 and NumPy 2.1.0. ROUGE-L recall is our own implementation and the judge clients use the Python standard library (`urllib`). Judges are `qwen3.5:35b` (Ollama; the server version was not recorded) and `moonshotai/kimi-k3` (OpenRouter).

References

[1] Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic test of unlearning using parametric knowledge traces. arXiv preprint arXiv:2406.11614, 2024. EMNLP 2025 Main.

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

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

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

[5] Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, et al. Tiny aya: Bridging scale and multilingual depth. arXiv preprint arXiv:2603.11510, 2026.

[6] Alireza Dehghanpour Farashah, Aditi Khandelwal, Marylou Fauchard, Zhuan Shi, Negar Rostamzadeh, and Golnoosh Farnadi. Multilingual amnesia: On the transferability of unlearning in multilingual LLMs. arXiv preprint arXiv:2601.05641, 2026.

[7] Jaeung Lee, Dohyun Kim, and Jaemin Jo. Measuring the depth of LLM unlearning via activation patching. arXiv preprint arXiv:2605.24614, 2026. EMNLP 2026 Main Conference (Oral).

[8] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. arXiv preprint arXiv:2401.06121, 2024.

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

[10] Ronen Eldan and Mark Russinovich. Who's Harry Potter? approximate unlearning in LLMs. arXiv preprint arXiv:2310.02238, 2023.

[11] Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, and Haiqin Yang. Dissecting fine-tuning unlearning in large language models. arXiv preprint arXiv:2410.06606, 2024. EMNLP 2024 Main (short paper).

[12] Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. arXiv preprint arXiv:2410.16454, 2024. ICLR 2025.

[13] Zora Che, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E. McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, Zikui Cai, Bilal Chughtai, Yarin Gal, Furong Huang, and Dylan Hadfield-Menell. Model tampering attacks enable more rigorous evaluations of LLM capabilities. arXiv preprint arXiv:2502.05209, 2025. TMLR.

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

[15] Atakan Seyitoğlu, Aleksei Kuvshinov, Leo Schwinn, and Stephan Günnemann. Extracting unlearned information from LLMs with activation steering. arXiv preprint arXiv:2411.02631, 2024.

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

[17] Bruce W. Lee, Addie Foote, Alex Infanger, Leni Shor, Harish Kamath, Jacob Goldman-Wetzler, Bryce Woodworth, Alex Cloud, and Alexander Matt Turner. Distillation robustifies unlearning. arXiv preprint arXiv:2506.06278, 2025. NeurIPS 2025 (Spotlight).

[18] Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. arXiv preprint arXiv:2410.12949, 2024.

[19] Yeonwoo Jang, Shariqah Hossain, Ashwin Sreevatsa, and Diogo Cruz. Prompt attacks reveal superficial knowledge removal in unlearning methods. arXiv preprint arXiv:2506.10236, 2025. COLM 2025 SoLaR Workshop.

[20] Minseok Choi, Kyunghyun Min, and Jaegul Choo. Cross-lingual unlearning of selective knowledge in multilingual language models. arXiv preprint arXiv:2406.12354, 2024. Findings of EMNLP 2024.

[21] Kyomin Hwang, Hyeonjin Kim, Seungyeon Kim, Sunghyun Wee, and Nojun Kwak. Uncovering the potential risks in unlearning: Danger of english-only unlearning in multilingual LLMs. arXiv preprint arXiv:2510.23949, 2025.

[22] itssaharsh. Relapse: forget the hazard, keep the rest, even in the 4-bit file. GitHub repository, accessed 14 November 2026, 2026.

[23] Muditkumar123. Compressing tiny aya to 4 bits: refusal shows no detected change, while a model after NPO unlearning loses retained-set overlap and its forget measures disagree. GitHub repository, accessed 14 November 2026, 2026.

[24] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. arXiv preprint arXiv:2309.16042, 2023.

[25] Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112, 2023.

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

[27] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[28] Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. arXiv preprint arXiv:2403.17806, 2024.

[29] Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. arXiv preprint arXiv:2310.10348, 2023.

A Activation patching rerun

Design. The source is the injected model (merged LoRA) and the target is the RMU model, both with the injected tokenizer. We score the teacher-forced first content token as in Section 5.2, at the final layer: the prompt plus any skipped opener (such as "B") goes through the RMU model once, and the score is the log probability of the content token at the last position. We write the injected model's output of decoder layer ℓ into that pass at chosen positions, on the prefill only (the hook skips passes of other lengths; an earlier version stayed active through `generate()` and crashed at a decode step). The recovered fraction is $(\log p_{\text{patched}} - \log p_{\text{RMU}})/(\log p_{\text{inject}} - \log p_{\text{RMU}})$. Positions are all positions of the scored sequence (prompt plus opener), the last prompt token, and the 9 name tokens of the author in the prompt ("subject"). All-position and last-token patches run at all 36 decoder layers and subject patches at 13. Decoder layer ℓ is residual $\ell+1$, and RMU edits decoder layers 16–18 (residuals 17–19). The 10 facts are those with a logit-lens gap of at least 1 nat, skipping id 0 (ids 17, 8, 19, 1, 13, 15, 16, 2, 4, 6), all about one author; no fact about the second author met the rule (largest gap 0.698 nats, id 31). An earlier sweep with a full-answer substring metric and no controls is superseded and not reported.

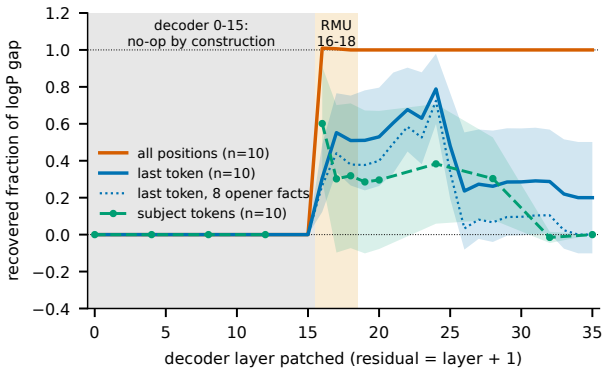


Figure 5: Activation patching from the injected model into the RMU model (Global): mean recovered fraction of the first-content-token log-probability gap over 10 facts, with 95% Student-t intervals; the dotted line is the last-token curve on the 8 facts with a teacher-forced opener. Subject tokens were measured at the 13 plotted layers. Left of layer 16 the patch changes nothing by construction. All positions is 1.000 by construction from decoder layer 18.

Cells that hold by construction. RMU changes no weight below decoder layer 16, so patching decoder layers 0–15 changes nothing: every condition is exactly 0.000 there, which is not evidence. An all-position patch at decoder layer 18 or later replaces the whole prefill state with the injected model's after the last edited layer, and the layers above are shared, so recovery is 1.000 there by construction; it checks the pipeline and is not a finding. Both controls agree. Patching the injected model's own activations back into itself (a self-patch control) at residual 19 (decoder layer 18), at all positions and at the last token, gives a mean absolute change in log probability of 0.0 with 10 of 10 facts within 0.05 nats. Patching injected activations into the RMU model at all positions at residual 19 gives a recovered fraction of 1.0 on 10 of 10 facts. The only non-trivial all-position cells are decoder layers 16 and 17, giving 1.010 [0.987, 1.032] and 1.006 [0.992, 1.021]: once the output of layer 16 or 17 is injected, RMU's remaining edited layers add

little for this measure. Opener handling matters for the other cells. The scored position is the opener for the 8 facts that have one and the last prompt token for ids 1 and 13, so a last-token patch touches it for those two facts only. At decoder layer 35 the last-token mean is 0.200 [−0.102, 0.502] because ids 1 and 13 give 1.0 and the other eight give 0.0, an artifact of the opener rule. Subject tokens never include the scored position, so decoder layer 35 gives 0.000 there too.

Single-site patches. Patching the last prompt token recovers part of the gap and only at some layers: 0.306 [0.122, 0.489] at decoder layer 16, 0.509 [0.266, 0.753] at layer 18, 0.789 [0.598, 0.979] at layer 24, and 0.236 to 0.291 over layers 26–32. The intervals include 0 at layers 26–30 and 33–35. On the 8 facts with an opener (a split we made after seeing the data) the last-token mean is 0.267 [0.089, 0.444] at layer 16, 0.379 [0.191, 0.568] at layer 18, 0.725 [0.508, 0.941] at layer 24, 0.067 [−0.019, 0.152] at layer 28 and 0.000 at layer 35. Subject tokens give 0.601 [0.292, 0.910] at layer 16, 0.319 [−0.073, 0.711] at layer 18, 0.383 [0.059, 0.707] at layer 24 and −0.015 at layer 32, with intervals that include 0 at layers 17–20 and 32. Per-fact values vary widely (subject tokens at layer 16 range from −0.08 to 1.00), and with 10 facts about one author we do not read the rise to layer 24 and the fall after it as structure. No single site we tried restores the gap fully or consistently, which neither localizes the answer nor shows it is absent.

Generations. As a secondary check we decode greedily for 32 new tokens from the prompt alone and test whether the first 6 content tokens of the answer appear. The injected model passes on 9 of 10 facts and the RMU model on 0 of 10. The two controls pass on 9 of 10, but the by-construction all-position control passes on 0 of 10, as does every all-position patch at the 7 layers we decoded (0, 16–19, 24, 35), and no single-site condition exceeds 1 of 10. We attribute this to decode steps that still pass through RMU's layers 16–18 (not tested), and make no claim of full-answer restoration.

Limits. The measure is a single teacher-forced token at the final layer, the cohort is 10 facts about one author, and we patch the full residual stream only, with no head- or component-level patching. The facts were selected for the largest drop in first-token log probability; the logit lens over all 40 facts still ranks the correct token first on 30. We ran the second-author facts (ids 31, 22, 30, 33) and 4 retain items (ids 0, 2, 3, 6) separately. Their gaps are near zero (0.07 to 1.31 and −0.056 to 0.005 nats), so their recovered fractions are unstable (all positions at decoder layer 16: 0.747 [−0.004, 1.498] and 0.050 [−3.070, 3.171]), and we do not pool them.