Unlearning vs. Hiding

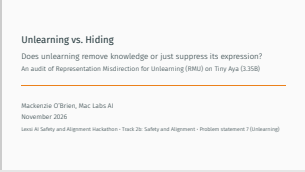
Does unlearning remove knowledge or just suppress its expression?

An audit of Representation Misdirection for Unlearning (RMU) on Tiny Aya (3.35B)

Mackenzie O'Brien, Mac Labs AI

November 2026

Lexsi AI Safety and Alignment Hackathon · Track 2b: Safety and Alignment · Problem statement 7 (Unlearning)



- Title follows the paper: "Unlearning vs. Hiding: Does Unlearning Remove Knowledge or Just Suppress Its Expression?"
- Solo project, Track 2b (Safety and Alignment), problem statement 7 (Unlearning).
- One line: we audit Representation Misdirection for Unlearning (RMU) on Tiny Aya, a 3.35B multilingual model family, using Lexsi's SafeTune and AuditKIT.
- Short answer up front: output-level suppression on the four instruction-tuned Tiny Aya variants, with weak evidence of residual knowledge; the same settings failed on Base and only partly worked on Aya Expanse 8B. We do not claim erasure.

The problem: removing knowledge or hiding it?

Why it matters

- Unlearning is proposed for privacy, copyright and hazardous-knowledge removal.
- If it only *hides* knowledge, a deployer who checks direct prompts alone gets false assurance.
- Prior work shows unlearned information is often recoverable: our motivation, not our finding.

What we do

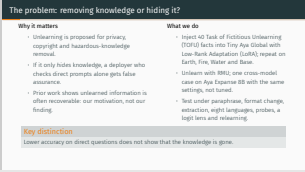
- Inject 40 Task of Fictitious Unlearning (TOFU) facts into Tiny Aya Global with Low-Rank Adaptation (LoRA); repeat on Earth, Fire, Water and Base.
- Unlearn with RMU; one cross-model case on Aya Expanse 8B with the same settings, not tuned.
- Test under paraphrase, format change, extraction, eight languages, probes, a logit lens and relearning.

Key distinction

Lower accuracy on direct questions does not show that the knowledge is gone.

Unlearning vs. Hiding

The problem: removing knowledge or hiding it?



- Motivation from the paper's introduction: privacy, copyright and hazardous-knowledge removal.
- Recoverability of unlearned information is prior work (cited in the paper), not our discovery.
- Setup in one sentence: teach a small multilingual model facts it cannot know from pretraining (Task of Fictitious Unlearning, TOFU, forget01 split, 40 question-answer pairs about two fictitious authors) with Low-Rank Adaptation (LoRA), then unlearn them with Representation Misdirection for Unlearning (RMU).

Research questions and contributions

Research questions (RQ)

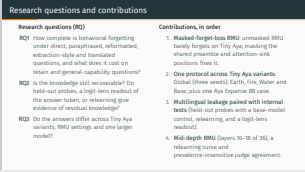
RQ1 How complete is behavioral forgetting under direct, paraphrased, reformatted, extraction-style and translated questions, and what does it cost on retain and general-capability questions?

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?

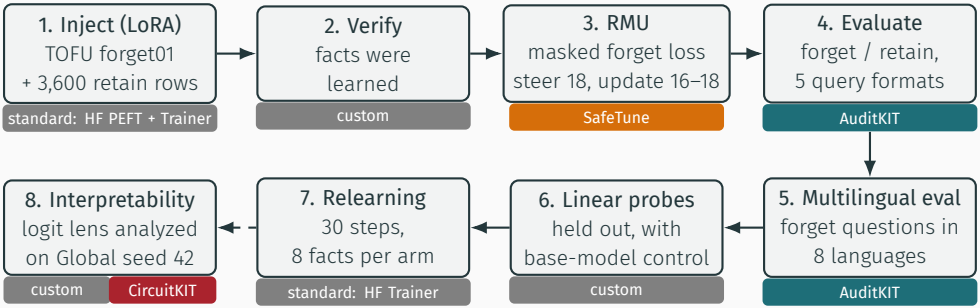
Contributions, in order

1. **Masked-forget-loss RMU**: unmasked RMU barely forgets on Tiny Aya; masking the shared preamble and attention-sink positions fixes it.
2. **One protocol across Tiny Aya variants**: Global (three seeds), Earth, Fire, Water and Base, plus one Aya Expanse 8B case.
3. **Multilingual leakage paired with internal tests** (held-out probes with a base-model control, relearning, and a logit-lens readout).
4. **Mid-depth RMU** (layers 16–18 of 36), a relearning curve and prevalence-insensitive judge agreement.



- The three research questions (RQ1 to RQ3) are the paper's.
- This is a measurement and method-fix study; we do not claim to discover shallow unlearning.
- Contribution 1 is the main one: on Tiny Aya the shared system preamble and attention-sink tokens dominate the forget loss, so unmasked RMU barely forgets.
- Contribution 4: the library default is layers 5–7; we apply RMU at layers 16–18 of 36.

Pipeline



HF PEFT: Hugging Face Parameter-Efficient Fine-Tuning library. Logit lens analyzed on Global seed 42; CircuitKIT (EAP-IG) ran on Global, not analyzed in this version.
Grading: `qwen3.5:35b` primary judge; Kimi K3 re-grades a seeded 15% sample.



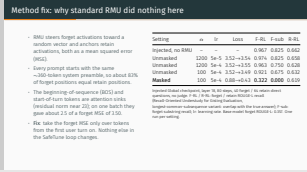
- The paper's pipeline order is inject, verify, unlearn, evaluate, probes, interpretability, relearn; every query is logged under a stable case identifier, so before and after scores pair the same questions.
- 1: Low-Rank Adaptation (LoRA) injection of TOFU forget01 plus a 3,600-row retain99 slice, using standard tooling (Hugging Face Parameter-Efficient Fine-Tuning, HF PEFT, with the Transformers Trainer). The adapter is merged before RMU.
- 2: a custom check confirms the facts were learned.
- 3: RMU with a masked forget loss in Lexsi's SafeTune; steering layer 18, update layers 16–18.
- 4–5: scoring with Lexsi's AuditKIT: direct, paraphrase, cloze, true/false and extraction prompts, plus translation into 8 languages.
- 6: custom held-out linear probes with a base-model control.
- 7: relearning with the standard Transformers Trainer, 30 steps.
- 8: the logit lens is analyzed on Global seed 42; an activation-patching rerun is reported in the paper's appendix. Head-ablation and CircuitKIT (edge attribution patching with integrated gradients, EAP-IG) artifacts exist for Global but are not analyzed in this version.

Method fix: why standard RMU did nothing here

- RMU steers forget activations toward a random vector and anchors retain activations, both as a mean squared error (MSE).
- Every prompt starts with the same ∼360-token system preamble, so about 83% of forget positions equal retain positions.
- The beginning-of-sequence (BOS) and start-of-turn tokens are attention sinks (residual norm near 23); on one batch they gave about 2.5 of a forget MSE of 3.50.
- **Fix**: take the forget MSE only over tokens from the first user turn on. Nothing else in the SafeTune loop changes.

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

Injected Global checkpoint, layer 18, 80 steps, 40 forget / 64 retain direct questions, no judge. F-RL / R-RL: forget / retain ROUGE-L recall (Recall-Oriented Understudy for Gisting Evaluation, longest-common-subsequence variant: overlap with the true answer); F-sub: forget substring recall; lr: learning rate. Base model forget ROUGE-L: 0.357. One run per setting.



- SafeTune's Representation Misdirection for Unlearning (RMU) trains the multilayer perceptron (MLP) output projections of the update layers; the forget term is a mean squared error (MSE) toward a scaled random vector, the retain term an MSE to a frozen copy.
- SafeTune averages the forget MSE over every position. On Tiny Aya the shared preamble and the two attention-sink positions (beginning-of-sequence, BOS, and start-of-turn) dominate it, so the gradient on the roughly 60 tokens that carry the fact is diluted about sevenfold.
- Four unmasked settings left the forget loss flat near 3.5 and forget ROUGE-L recall (Recall-Oriented Understudy for Gisting Evaluation, longest-common-subsequence variant) at 0.92–0.97, against 0.97 for the injected model.
- Masking lowers forget ROUGE-L to 0.32. The retain term still covers all positions.
- Caveat: these diagnostic runs set alpha and the learning rate on the same 40 forget questions that are later evaluated.

Experimental setup

Model and data

- Tiny Aya Global, 3.35B, 70 languages, 36 layers; Earth, Fire and Water share the architecture; Base is not instruction-tuned
- Cross-model case: Aya Expanse 8B (32 layers), same RMU settings, not tuned
- Forget: TOFU forget01, 40 pairs, two fictitious authors
- Retain: first 3,600 retain99 rows for training; 64 rows scored
- Utility: TOFU world_facts and real_authors (40 on Global seed 42, 20 otherwise)

Tooling

- SafeTune: RMU
- AuditKIT: scoring
- CircuitKIT: EAP-IG on Global, not analyzed in this version
- Standard: HF PEFT + Trainer
- Custom: verify, probes

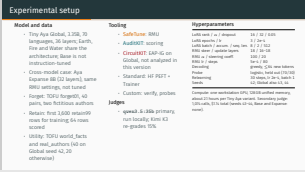
Judges

- `qwen3.5:35b` primary, run locally; Kimi K3 re-grades 15%

Hyperparameters

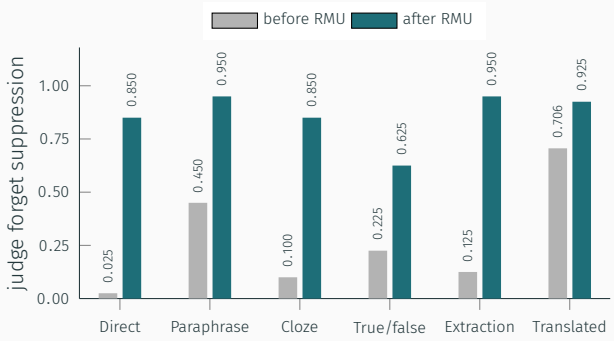
LoRA rank / α / dropout	16 / 32 / 0.05
LoRA epochs / lr	3 / 2e-4
LoRA batch / accum. / seq. len.	8 / 2 / 512
RMU steer / update layers	18 / 16–18
RMU α / steering coeff.	100 / 20
RMU lr / steps	5e-4 / 80
Decoding	greedy, $\leq$64 new tokens
Probe	logistic, held out (70/30)
Relearning	30 steps, lr 2e-4, batch 1
Seeds	42; Global also 43, 44

Compute: one workstation GPU, 128 GB unified memory, about 2.1 hours per Tiny Aya variant. Secondary judge: 1,074 calls, $1.14 total (seeds 43–44, Base and Expanse: none).



- Models: Tiny Aya Global, instruction-tuned, 70 languages, 3.35B parameters, 36 decoder layers, run in bfloat16 with its chat template; the same protocol on Earth, Fire, Water and Base. One cross-model case on Aya Expanse 8B (about 8B parameters, 32 layers) with the Tiny Aya settings carried over unchanged.
- Data: TOFU forget01 (40 pairs, two fictitious authors); the first 3,600 of 3,960 retain99 rows (leaving 18 authors held out for the relearning control); retain accuracy on the first 64 retain rows; paraphrases are the official forget01_perturbed questions.
- Tooling: SafeTune for RMU, AuditKIT for scoring; CircuitKIT ran edge attribution patching with integrated gradients (EAP-IG) on Global but is not analyzed in this version. Low-Rank Adaptation (LoRA) injection uses Hugging Face Parameter-Efficient Fine-Tuning (HF PEFT) with the Transformers Trainer; relearning uses the Trainer.
- Headline metric: judge forget suppression, one minus the primary judge's match rate on forget prompts, that is, the fraction of injected facts the output does not reveal. Missing judgments are counted, not scored as zero.
- Judges: qwen3.5:35b served locally (temperature 0, JavaScript Object Notation, JSON, verdict); Kimi K3 re-grades a seeded 15% sample with the same rubric. 1,074 logged secondary-judge calls at a mean of $0.00106 per call, $1.14 total. Global seeds 43 and 44, Base and Expanse used the primary judge only.
- Seed 42 for every run, and seeds 42, 43 and 44 for Global.

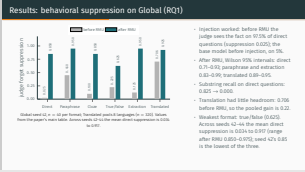
Results: behavioral suppression on Global (RQ1)



Global seed 42, $n = 40$ per format; Translated pools 8 languages ($n = 320$). Values from the paper's main table. Across seeds 42–44 the mean direct suppression is 0.034 to 0.917.

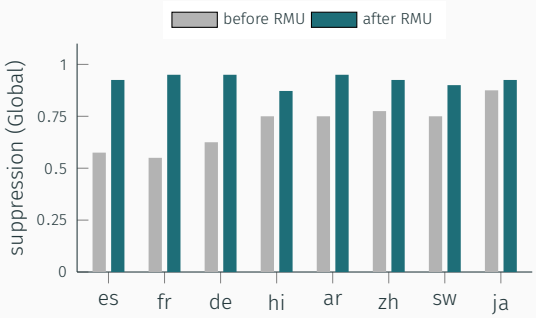
- Injection worked: before RMU the judge sees the fact on 97.5% of direct questions (suppression 0.025); the base model before injection, on 5%.
- After RMU, Wilson 95% intervals: direct 0.71–0.93; paraphrase and extraction 0.83–0.99; translated 0.89–0.95.
- Substring recall on direct questions: $0.825 \rightarrow 0.000$.
- Translation had little headroom: 0.706 before RMU, so the pooled gain is 0.22.
- Weakest format: true/false (0.625). Across seeds 42–44 the mean direct suppression is 0.034 to 0.917 (range after RMU 0.850–0.975); seed 42's 0.85 is the lowest of the three.

Results: behavioral suppression on Global (RQ1)



- Judge forget suppression on Global seed 42 rises from 0.025 to 0.85 on direct questions, 0.45 to 0.95 on paraphrases and 0.125 to 0.95 on extraction prompts (n = 40 each), and from 0.706 to 0.925 on the pooled translated prompts (n = 320). Across seeds 42–44 the means are 0.034 to 0.917 (direct), 0.333 to 0.950 (paraphrase), 0.075 to 0.950 (extraction) and 0.706 to 0.926 (translated).
- Wilson 95% intervals after RMU, as reported in the paper: 0.71–0.93 (direct), 0.83–0.99 (paraphrase and extraction), 0.89–0.95 (translated). With 40 questions one item moves a rate by 2.5 points.
- Suppression after RMU is higher on paraphrase and extraction than on direct questions; the paper says this was not expected and is not explained.
- Cloze 0.100 to 0.850 and true/false 0.225 to 0.625 are split from one 80-item bucket using item-level verdicts.
- One Global post-RMU translated item was not judged (Hindi, 39 of 40).

Results: multilingual leakage (RQ1)

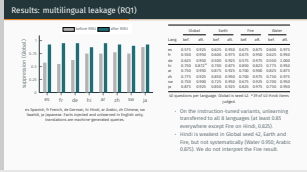


es Spanish, fr French, de German, hi Hindi, ar Arabic, zh Chinese, sw Swahili, ja Japanese. Facts injected and unlearned in English only; translations are machine-generated queries.

	Global		Earth		Fire		Water	
Lang.	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
fr	0.550	0.950	0.600	0.975	0.675	0.950	0.625	0.950
de	0.625	0.950	0.500	0.925	0.575	0.975	0.550	1.000
hi	0.750	0.872*	0.700	0.875	0.850	0.825	0.775	0.950
ar	0.750	0.950	0.875	0.925	0.700	0.900	0.825	0.875
zh	0.775	0.925	0.850	0.950	0.700	0.975	0.750	0.975
sw	0.750	0.900	0.725	0.950	0.675	0.925	0.700	0.950
ja	0.875	0.925	0.850	0.925	0.825	0.975	0.750	0.950

40 questions per language. Global is seed 42. *39 of 40 Hindi items judged.

- On the instruction-tuned variants, unlearning transferred to all 8 languages (at least 0.85 everywhere except Fire on Hindi, 0.825).
- Hindi is weakest in Global seed 42, Earth and Fire, but not systematically (Water 0.950; Arabic 0.875). We do not interpret the Fire result.



- On the instruction-tuned variants, unlearning transferred to all 8 languages (at least 0.85 everywhere except Fire on Hindi, 0.825).
- Hindi is weakest in Global seed 42, Earth and Fire, but not systematically (Water 0.950; Arabic 0.875). We do not interpret the Fire result.

- Forget questions machine-translated into Spanish, French, German, Hindi, Arabic, Chinese, Swahili and Japanese, used only as queries; a native-speaker check of the translations is still pending.
- On the instruction-tuned variants, English-only unlearning of English-injected facts transferred to all eight languages, with judged suppression after RMU of at least 0.85 in every language of every run except Fire on Hindi (0.825). Hindi is not systematically weak (Water 0.950; Arabic is lowest there at 0.875). With 40 questions per language we do not interpret the Fire result.
- Before RMU the injected model already withheld the fact on 71% of translated prompts (Spanish 0.575, French 0.55, Japanese 0.875 on Global), so headroom is small.
- In Fire, the South Asian variant, Hindi suppression did not rise (0.850 to 0.825). Base is the exception to the transfer: after RMU, six of the eight languages remain below 0.75, and Spanish is 0.300.
- Farashah et al. (Multilingual Amnesia) report mostly language-specific forgetting, but the setups differ: they unlearn in each language with gradient-based objectives on an 8B model.

Results: utility cost and format-change leakage (RQ1)

Judged accuracy on non-forget facts

	Global		Earth		Fire		Water	
	bef.	aft.	bef.	aft.	bef.	aft.	bef.	aft.
Retain (64)	0.516	0.422	0.562	0.516	0.531	0.500	0.516	0.484
World facts	0.875	0.850	0.900	0.900	0.900	0.900	0.900	0.900
Real authors	0.825	0.775	1.000	1.000	0.950	0.950	0.900	0.850

Global is seed 42 (40 utility questions); Earth, Fire and Water use 20 each.

- Retain falls 9.4 points on Global seed 42 (Wilson 95%: 0.40–0.63 before, 0.31–0.54 after), 4.7 on Earth, 3.1 on Fire and 3.1 on Water. Across seeds 42–44 the mean is 5.2 points (range 1.6–9.4).
- Any single run is within sampling noise, but the drop has the same sign in all six instruction-tuned runs, so we report it as a real but small cost.

Leakage under format change

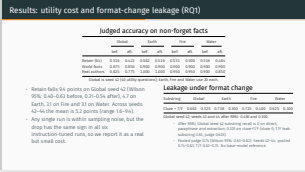
Substring	Global		Earth		Fire		Water	
Cloze + T/F	0.663	0.325	0.738	0.300	0.725	0.400	0.625	0.300

Global seed 42; seeds 43 and 44 after RMU: 0.438 and 0.300.

- After RMU, Global seed 42 substring recall is 0 on direct, paraphrase and extraction; 0.325 on cloze+T/F (cloze 0; T/F leak: substring 0.65, judge 0.625).
- Pooled judge 0.74 (Wilson 95%: 0.63–0.82). Seeds 42–44: pooled 0.74–0.83, T/F 0.63–0.75. No base-model reference.

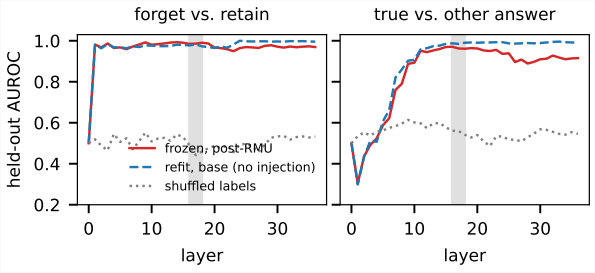
Unlearning vs. Hiding

Results: utility cost and format-change leakage (RQ1)



- Utility: 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). Across Global seeds 42–44 the mean drop is 5.2 points (range 1.6–9.4).
- Global seed 42 world facts fall 2.5 points and real authors 5.0 points (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).
- Retain facts were injected only partly (base model 0.0625, after injection 0.516), so the cost is measured on facts the model knew about half the time.
- Leakage: a true/false (T/F) prompt contains the correct statement, so a model that never learned the fact would accept some of them. Without a base-model reference on this format, we report it as residual leakage, not as a measured amount.
- Fire's true/false suppression after RMU (0.475) is the lowest cell in the paper's main table.

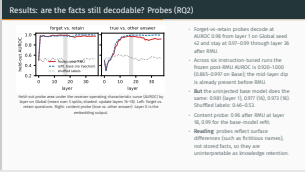
Results: are the facts still decodable? Probes (RQ2)



Held-out probe area under the receiver operating characteristic curve (AUROC) by layer on Global (mean over 5 splits; shaded: update layers 16–18). Left: forget vs. retain questions. Right: content probe (true vs. other answer). Layer 0 is the embedding output.

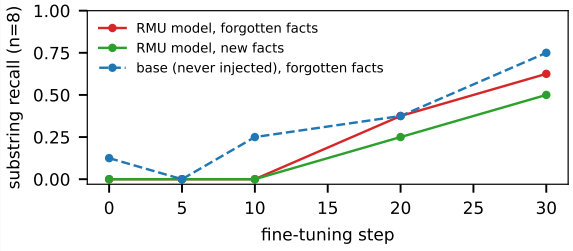
- Forget-vs-retain probes decode at AUROC 0.98 from layer 1 on Global seed 42 and stay at 0.97–0.99 through layer 36 after RMU.
- Across six instruction-tuned runs the frozen post-RMU AUROC is 0.920–1.000 (0.865–0.997 on Base); the mid-layer dip is already present before RMU.
- **But** the uninjected base model does the same: 0.981 (layer 1), 0.977 (16), 0.973 (18). Shuffled labels: 0.46–0.53.
- Content probe: 0.96 after RMU at layer 18, 0.99 for the base-model refit.
- **Reading**: probes reflect surface differences (such as fictitious names), not stored facts, so they are uninterpretable as knowledge retention.

└Results: are the facts still decodable? Probes (RQ2)



- At each of the 37 residual positions we take the last non-padding token's activation; a logistic probe separates the 40 forget from the 64 retain questions. Numbers are held out (70/30 split, repeated 5 times on Global seed 42, 3 times on every other run). Metric: area under the receiver operating characteristic curve (AUROC).
- Frozen probes are fit on the post-injection model and applied after RMU; refit probes are fit on each model's own rows. Controls: base model before injection, and shuffled labels.
- 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). On Base 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.
- The probes do not show that facts survive RMU, and a lower score would not have shown erasure. Probes that compare a fact with its counterfactual, with the base-model control, would be needed.

Results: relearning test (RQ2)

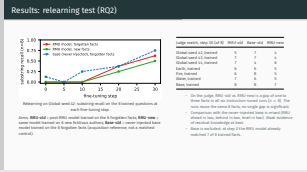


Relearning on Global seed 42: substring recall on the 8 trained questions at each fine-tuning step.

Arms: **RMU-old** = post-RMU model trained on the 8 forgotten facts; **RMU-new** = same model trained on 8 new fictitious authors; **Base-old** = never-injected base model trained on the 8 forgotten facts (acquisition reference, not a matched control).

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

- On the judge, RMU-old vs. RMU-new is a gap of one to three facts in all six instruction-tuned runs ($n = 8$). The runs reuse the same 8 facts; no single gap is significant.
- Comparison with the never-injected base is mixed (RMU ahead in two, behind in two, level in two). Weak evidence of residual knowledge at best.
- Base is excluded: at step 0 the RMU model already matched 7 of 8 trained facts.



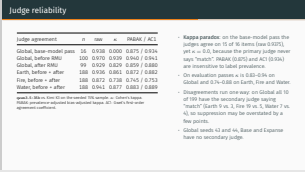
- 30 fine-tuning steps (learning rate 2e-4, batch size 1, one run per arm). RMU-old trains the post-RMU model on 8 forgotten facts; RMU-new trains a fresh copy of the same model on 8 new fictitious TOFU authors; base-old trains the never-injected base model on the 8 forgotten facts (an acquisition reference, not a matched control).
- On the judge the RMU-old against RMU-new gap is positive in all six instruction-tuned 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). The six runs reuse the same 8 forgotten and 8 new facts, so they are not independent replicates. 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 are mixed (RMU ahead in two, behind in two, level in two).
- 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.
- Held-out paraphrases are mixed. The Expanse relearning run is uninformative (the RMU model already reproduced 7 of 8 forget answers at step 0 on those facts).

Judge reliability

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

`qwen3.5:35b` vs. Kimi K3 on the seeded 15% sample. κ: Cohen's kappa. PABAK: prevalence-adjusted bias-adjusted kappa. AC1: Gwet's first-order agreement coefficient.

- **Kappa paradox**: on the base-model pass the judges agree on 15 of 16 items (raw 0.9375), yet $\kappa = 0.0$, because the primary judge never says "match". PABAK (0.875) and AC1 (0.934) are insensitive to label prevalence.
- On evaluation passes κ is 0.83–0.94 on Global and 0.74–0.88 on Earth, Fire and Water.
- Disagreements run one way: on Global all 10 of 199 have the secondary judge saying "match" (Earth 9 vs. 3, Fire 19 vs. 5, Water 7 vs. 4), so suppression may be overstated by a few points.
- Global seeds 43 and 44, Base and Expanse have no secondary judge.



- The primary judge is qwen3.5:35b (temperature 0, thinking off, JSON verdict); the secondary judge Kimi K3 re-grades a seeded random 15% sample with the same rubric.
- We report raw agreement, Cohen's kappa, prevalence-adjusted bias-adjusted kappa (PABAK, 2 times observed agreement minus 1) and Gwet's first-order agreement coefficient (AC1).
- Kappa becomes unstable when one judge gives almost one label: chance agreement approaches observed agreement. The Earth, Fire and Water base-model passes (16 pairs each) give raw 0.938, 0.875 and 0.875 with kappa 0.
- Because disagreements lean one way (primary stricter), primary-judge suppression may be overstated by a few points.

Across variants and RMU settings (RQ3)

Judge forget suppression by variant

	Global		Earth		Fire		Water	
Format (n)	bef.	aft.	bef.	aft.	bef.	aft.	bef.	aft.
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 langs (320)	0.706	0.925	0.716	0.934	0.709	0.925	0.697	0.953

Global is seed 42; Earth, Fire and Water are single runs. Seeds 43 and 44 are in the paper's seed table.

Masked RMU sweep on Global

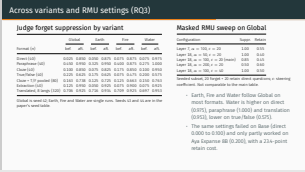
Configuration	Suppr.	Retain
Layer 7, $\alpha = 100$, $c = 20$	1.00	0.55
Layer 18, $\alpha = 50$, $c = 20$	1.00	0.40
Layer 18, $\alpha = 100$, $c = 20$ (main)	0.85	0.45
Layer 18, $\alpha = 200$, $c = 20$	0.50	0.60
Layer 18, $\alpha = 100$, $c = 40$	1.00	0.50

Seeded subset, 20 forget + 20 retain direct questions; c: steering coefficient. Not comparable to the main table.

- Earth, Fire and Water follow Global on most formats. Water is higher on direct (0.975), paraphrase (1.000) and translation (0.953), lower on true/false (0.575).
- The same settings failed on Base (direct 0.000 to 0.100) and only partly worked on Aya Expanse 8B (0.200), with a 23.4-point retain cost.

Unlearning vs. Hiding

Across variants and RMU settings (RQ3)



- Variants: Earth, Fire and Water follow Global within a few points on most formats. Global seeds 43 and 44 give a mean direct suppression of 0.034 to 0.917 (range after RMU 0.850–0.975). The same settings failed on Base and only partly worked on Aya Expanse 8B (next two results slides).
- Sweep: two configurations scored better than the chosen one on this subset: layer 7 with alpha 100 (1.00 and 0.55) and layer 18 with steering coefficient 40 (1.00 and 0.50). With 20 items per cell one item moves a rate by 5 points, so the subset cannot separate these configurations. We do not claim the main configuration is optimal; every masked setting forgot at least partly, unmasked points scored 0.05.
- The main configuration was fixed before the masked sweep finished. Four earlier unmasked sweep points all scored 0.05, so the mask matters more than the layer on this subset.

Where the same settings fail (RQ3)

Tiny Aya Base (not instruction-tuned)

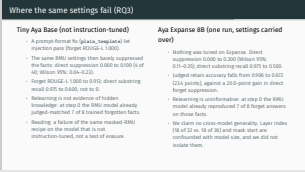
- A prompt-format fix (`plain_template`) let injection pass (forget ROUGE-L 1.000).
- The same RMU settings then barely suppressed the facts: direct suppression 0.000 to 0.100 (4 of 40; Wilson 95%: 0.04–0.23).
- Forget ROUGE-L 1.000 to 0.913; direct substring recall 0.975 to 0.600, not to 0.
- Relearning is not evidence of hidden knowledge: at step 0 the RMU model already judged-matched 7 of 8 trained forgotten facts.
- Reading: a failure of the same masked-RMU recipe on the model that is not instruction-tuned, not a test of erasure.

Aya Expanse 8B (one run, settings carried over)

- Nothing was tuned on Expanse. Direct suppression 0.000 to 0.200 (Wilson 95%: 0.11–0.35); direct substring recall 0.975 to 0.500.
- Judged retain accuracy falls from 0.906 to 0.672 (23.4 points), against a 20.0-point gain in direct forget suppression.
- Relearning is uninformative: at step 0 the RMU model already reproduced 7 of 8 forget answers on those facts.
- We claim no cross-model generality. Layer index (18 of 32 vs. 18 of 36) and mask start are confounded with model size, and we did not isolate them.

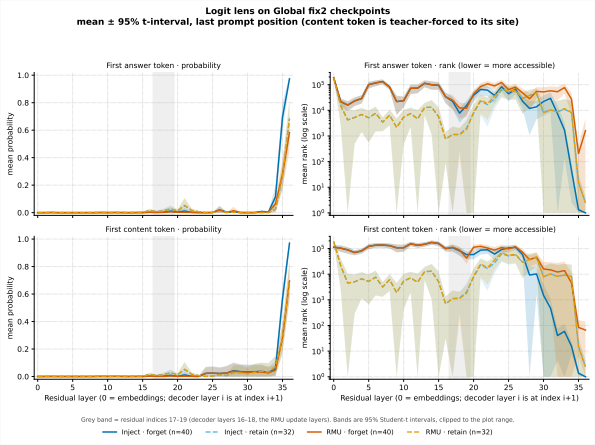
Unlearning vs. Hiding

Where the same settings fail (RQ3)



- Base: the earlier injection-gate failures were a prompt-format bug (Base has no chat template). After the fix, a run with the original recipe passed the gate at forget ROUGE-L 1.000 (substring recall 0.975) and then completed unlearning, evaluation, probes and relearning with the same RMU settings. 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. Judged retain accuracy was 0.688 before RMU and 0.703 after (a one-question rise). The secondary judge was not run.
- Expanse: injection worked (forget ROUGE-L 0.321 to 0.998; judged retain after injection 0.906, against 0.516 on Global). After RMU, paraphrase suppression is 0.275 to 0.400 and extraction 0.000 to 0.100. The retain cost and the forget gain are of the same size (16 of 64 retain questions flipped vs. 8 of 40 direct forget questions). World facts and real authors stay at judged accuracy 1.000 ($n = 20$). We do not attribute the weak effect to model size.

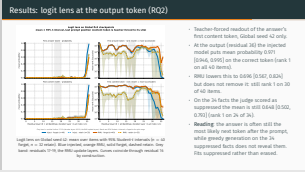
Results: logit lens at the output token (RQ2)



Logit lens on Global seed 42: mean over items with 95% Student-t intervals ($n = 40$ forget, $n = 32$ retain). Blue injected, orange RMU, solid forget, dashed retain. Grey band: residuals 17–19, the RMU update layers. Curves coincide through residual 16 by construction.

- Teacher-forced readout of the answer's first content token, Global seed 42 only.
- At the output (residual 36) the injected model puts mean probability 0.971 [0.946, 0.995] on the correct token (rank 1 on all 40 items).
- RMU lowers this to 0.696 [0.567, 0.824] but does not remove it: still rank 1 on 30 of 40 items.
- 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).
- **Reading**: 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. Fits suppressed rather than erased.

Results: logit lens at the output token (RQ2)

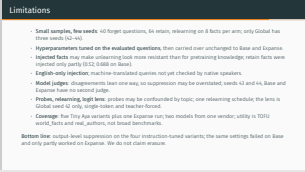


- Mid-network readouts are uninformative for forget and retain facts alike, so the lens cannot show whether an internal representation remains. A low mid-network probability is not evidence that the answer cannot be decoded there.
- The measure is single-token and teacher-forced. The first-content-token rule skips an opener on 18 of 40 forget items. Items cluster on two authors, so intervals over items are optimistic.
- An activation-patching rerun is reported in the paper's appendix (10 facts of one author, single-token measure, no full-answer restoration). Head-ablation and CircuitKIT are not analyzed in this version.

Limitations

- **Small samples, few seeds**: 40 forget questions, 64 retain, relearning on 8 facts per arm; only Global has three seeds (42–44).
- **Hyperparameters tuned on the evaluated questions**, then carried over unchanged to Base and Expanse.
- **Injected facts** may make unlearning look more resistant than for pretraining knowledge; retain facts were injected only partly (0.52; 0.688 on Base).
- **English-only injection**; machine-translated queries not yet checked by native speakers.
- **Model judges**: disagreements lean one way, so suppression may be overstated; seeds 43 and 44, Base and Expanse have no second judge.
- **Probes, relearning, logit lens**: probes may be confounded by topic; one relearning schedule; the lens is Global seed 42 only, single-token and teacher-forced.
- **Coverage**: five Tiny Aya variants plus one Expanse run; two models from one vendor; utility is TOFU world_facts and real_authors, not broad benchmarks.

Bottom line: output-level suppression on the four instruction-tuned variants; the same settings failed on Base and only partly worked on Expanse. We do not claim erasure.



- Confidence intervals were computed only where stated. With $n = 3$ Global seeds we report per-seed values, the mean and the range, not a standard deviation or a confidence interval, and the seeds share the same questions and relearning facts.
- The configuration is not shown to be optimal (see the sweep).
- English-only injection asks a different question from within-language unlearning.
- 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.
- Bottom line from the paper: with a masked forget loss, RMU suppresses the injected facts across the four instruction-tuned Tiny Aya variants (and three seeds of Global), holding under paraphrase, extraction prompts and translation into eight languages, at a utility cost of 1.6–9.4 points on retain facts. On Base the same settings moved direct suppression only from 0.00 to 0.10. On Aya Expanse 8B they reached only 0.20 and cost 23.4 points on retain facts. The claim does not establish erasure: format-change leakage lacks a base-model reference, relearning is not consistently faster or slower than learning from scratch, and the RMU-old against RMU-new gap is one to three facts at $n = 8$.

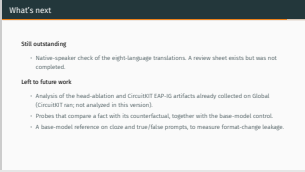
What's next

Still outstanding

- Native-speaker check of the eight-language translations. A review sheet exists but was not completed.

Left to future work

- Analysis of the head-ablation and CircuitKIT EAP-IG artifacts already collected on Global (CircuitKIT ran; not analyzed in this version).
- Probes that compare a fact with its counterfactual, together with the base-model control.
- A base-model reference on cloze and true/false prompts, to measure format-change leakage.



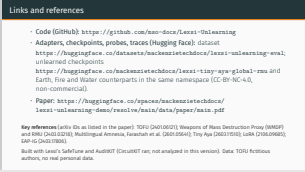
- Global seeds 43 and 44, Water, Base and Aya Expanse 8B have now been run and are in the paper and this deck. The native-speaker translation check is still outstanding; translations were written by the primary judge's own model.
- The logit lens is analyzed on Global seed 42 in this version; an activation-patching rerun is in the paper's appendix. Head-ablation and CircuitKIT edge attribution patching with integrated gradients (EAP-IG) artifacts still exist for Global and are not analyzed; we leave them to future work.
- The counterfactual-probe and base-model true/false items come from the paper's discussion and limitations.

Links and references

- **Code (GitHub):** `https://github.com/mso-docs/Lexsi-Unlearning`
- **Adapters, checkpoints, probes, traces (Hugging Face):** dataset `https://huggingface.co/datasets/mackenzietechdocs/lexsi-unlearning-eval`; unlearned checkpoints `https://huggingface.co/mackenzietechdocs/lexsi-tiny-aya-global-rmu` and Earth, Fire and Water counterparts in the same namespace (CC-BY-NC-4.0, non-commercial).
- **Paper:** `https://huggingface.co/spaces/mackenzietechdocs/lexsi-unlearning-demo/resolve/main/data/paper/main.pdf`

Key references (arXiv IDs as listed in the paper): TOFU (2401.06121); Weapons of Mass Destruction Proxy (WMDP) and RMU (2403.03218); Multilingual Amnesia, Farashah et al. (2601.05641); Tiny Aya (2603.11510); LoRA (2106.09685); EAP-IG (2403.17806).

Built with Lexsi's SafeTune and AuditKIT (CircuitKIT ran; not analyzed in this version). Data: TOFU fictitious authors, no real personal data.



- Code, dataset, checkpoints and the paper PDF are linked on this slide. The GitHub repository and Hugging Face artifacts are still private until publication.
- Releasing checkpoints depends on confirming un-gated republication under the base model's Creative Commons attribution non-commercial (CC-BY-NC) license and Cohere's acceptable-use policy; gated base weights are not redistributed.
- Ethics: the forget set is synthetic (TOFU fictitious authors); no real personal data and no hazardous-knowledge data.