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)

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.

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?

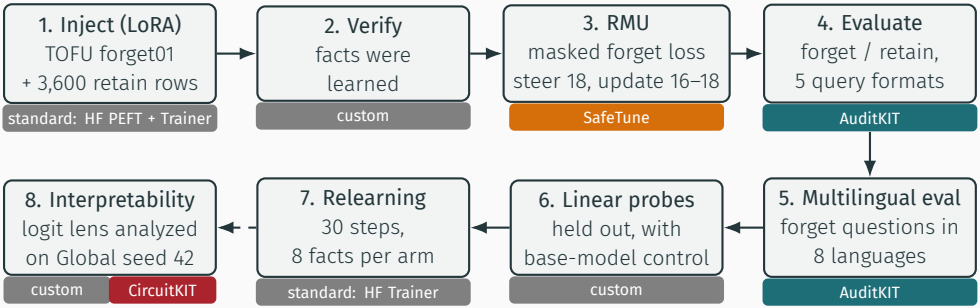
RQ2 Is the knowledge still recoverable? Do held-out probes, a logit-lens readout of the answer token, or relearning give evidence of residual knowledge?

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.

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.

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.

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

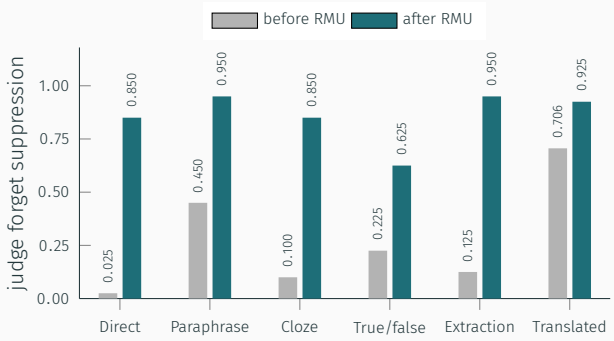
- `qwen3.5:35b` primary, run locally; Kimi K3 re-grades 15%

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).

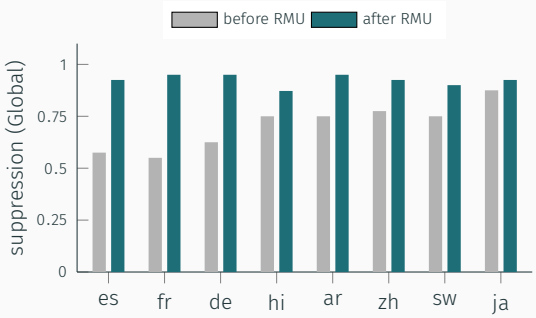
Results: behavioral suppression on Global (RQ1)



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: 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.

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.

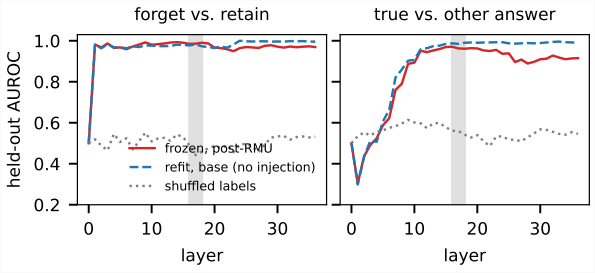
Leakage under format change

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.

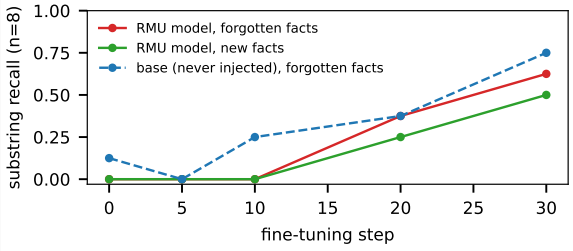
Results: are the facts still decodable? Probes (RQ2)



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: 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.

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.

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.

Where the same settings fail (RQ3)

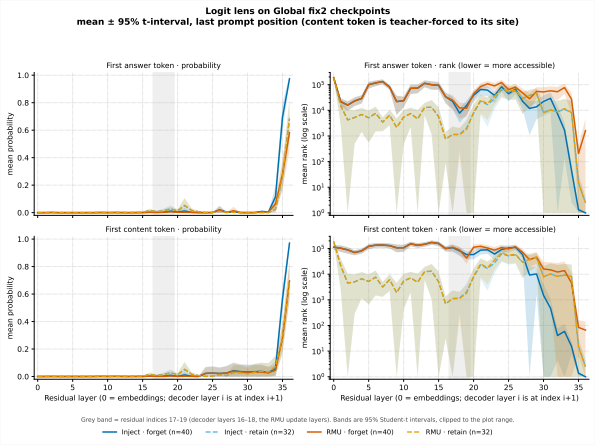
Tiny Aya Base (not instruction-tuned)

- 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.

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.

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.

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.

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.