LINGUISTIC LOOPHOLES IN LLM UNLEARNING: FROM A 174-LANGUAGE–SCRIPT BENCHMARK TO COVERAGE-AWARE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unlearning a fact in one language does not guarantee its removal in others as changing the query or even the requested answer language can reopen seemingly forgotten knowledge—a cross-lingual loophole. The most straightforward solution to this challenge – unlearning in all languages – is neither scalable nor desirable as it amplifies damage to unrelated model capabilities. We introduce the task of language budgeted multilingual unlearning with the goal of selecting a subset of languages that maximize cross-lingual erasure. To study this task we introduce CROSS-LINGUAL UNLEARNING TENSOR, an unlearning benchmark that spans 174 language–script pairs and 25 atomic paraphrase types to examine when forgetting generalizes across linguistic expressions of the same knowledge. We further propose COVER, which selects source languages to maximize predicted COVERage of languages receiving no forget supervision; enabling unlearning on a language budget. Surprisingly, we find naively selecting strong individual sources does not reliably compose into strong source sets motivating our development of COVER. At deployment COVER only requires benign calibration data and access to the frozen model. Across three model families and two disjoint forget sets, COVER reduces mean held-out residual access by 7.8–27.3% relative to uniform source selection. We find these gains extend beyond synthetic benchmarks to real news documents in low-resource languages settings using human translated data from the Low Resource Languages for Emergent Incidents (LORELEI) corpus. [1]

1 INTRODUCTION

Large language models can memorize and reproduce data seen during pretraining and fine tuning, including private content, copyrighted material, or knowledge that later becomes unsafe (Carlini et al., 2021; Lukas et al., 2023; Lee et al., 2023; Chang et al., 2023; Li et al., 2024). Guardrails and inference time interventions can mitigate unwanted, but they provide limited protection when users have white-box access to model parameters. Retraining from scratch without the offending data offers the strongest possible remedy, but its typically impractical given the resources required. This motivates LLM unlearning: update a model trained on data $\mathcal{D}$ so that its behavior on a forget set $\mathcal{F} \subseteq \mathcal{D}$ approximates a model trained without $\mathcal{F}$ while preserving utility on retained data $\mathcal{R}$.

As frontier language models improve in multilingual generation and understanding, so does the surface area for accessing memorized knowledge

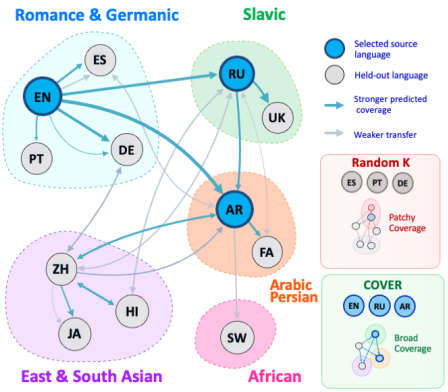


Figure 1: **Language-budgeted unlearning.** COVER selects K source languages (blue) to cover held-out languages (grey).

[1] Review CROSS-LINGUAL UNLEARNING TENSOR here

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

across languages (Zhao et al., 2024; Qi et al., 2023). This creates a potential challenge for Machine Unlearning (MU), as removal in one language may not guarantee forgetting in another. Recent works have documented this cross-lingual failure mode (Farashah et al., 2026; Xiang et al., 2026; Choi et al., 2024; Lu & Koehn, 2025), but existing studies remain limited in scope, covering a limited set of languages which tend to be high resource.

We release CROSS-LINGUAL UNLEARNING TENSOR, a multilingual unlearning dataset spanning 174 language–script pairs. Through the release of CROSS-LINGUAL UNLEARNING TENSOR, we look to study the dynamics of language and unlearning from five perspectives: the language the fact is learned in, the language(s) used to unlearn it, the language of the evaluation query, the language of the scored answer and how various same language paraphrases differ in terms of eliciting unlearned knowledge.

CROSS-LINGUAL UNLEARNING TENSOR shows that unlearning in one language can leave the same facts accessible through other query–answer routes. Full parallel forget data are often unavailable, and translating each request into every evaluation language increases cost. In our experiments, all-language supervision also increases damage to retained facts and multilingual generation (Table 21).

We therefore study *language budgeted multilingual unlearning*. Forget supervision is available only in a source subset $\mathcal{L}_{\text{src}} \subset \mathcal{L}$, with $|\mathcal{L}_{\text{src}}| = K$. We evaluate forgetting in held-out languages $\mathcal{L}_{\text{held}} = \mathcal{L} \setminus \mathcal{L}_{\text{src}}$, alongside retained knowledge and multilingual capabilities.

We introduce COVER (COVerage-guided source selection for multilingual ERasure), which uses benign calibration data to select sources for expected held-out coverage before applying an unlearning objective. It requires no held-out forget examples at deployment. COVER improves held-out forgetting and can reduce generated disclosure more than selecting the three strongest individual sources. Furthermore, we validate the performance of COVER on non-synthetic data and human generated translations through the Low Resource Languages for Emergent Incidents (LORELEI) corpus of news documents (Strassel & Tracey, 2016). On Aya (Salamanca et al., 2026) and Qwen (Yang et al., 2025) model families, source selection developed on CROSS-LINGUAL UNLEARNING TENSOR improve cross-lingual forgetting on our constructed LORELEI document removal task spanning eight languages. We summarize our contributions as follows:

1. Constructed CROSS-LINGUAL UNLEARNING TENSOR, a dataset with 174 language–script pairs (diverse scripts and resource levels), 25 atomic paraphrase types, translated paraphrases and evaluation on query and answer routes separately, enabling a systematic study of how language and wording effect unlearning.
2. Formalized the language-budgeted multilingual unlearning setting where we need to select K source languages for forgetting that generalizes to held-out languages while preserving retained knowledge and multilingual capabilities.
3. We show that meeting the removal criterion on a supervised route does not guarantee multilingual forgetting and characterize residual access across query and answer languages routes.
4. We propose COVER, which selects sources without held-out forget examples at deployment and reduces held-out residual access across existing benchmarks and real-world low-resource settings.

2 PROBLEM FORMULATION AND EVALUATION PROTOCOL

Unlearning Let M be an LM trained on distribution D. In unlearning, we are given a *forget set* $\mathcal{F} \subset D$ and a *retain set* $\mathcal{R} \subset D$, with the goal of producing an unlearned model M' that behaves as if it had been trained on $D \setminus \mathcal{F}$. We use $\mathcal{R}$ to measure drift from capabilities we wish the model to retain (Cao & Yang, 2015b).

Multilingual Unlearning Let $x = (q_\ell, a_\ell, s)$ denote a sample with query language q_ℓ, target answer language a_ℓ, and unlearning target s, where s groups paraphrases that query the same underlying fact. Under a language budgeted unlearning request, forget examples are available only in source languages $\mathcal{L}_{\text{src}} \subset \mathcal{L}$, where $\mathcal{L}$ is a set of all languages ℓ, $|\mathcal{L}_{\text{src}}| = K$, while evaluation covers both $\mathcal{L}_{\text{src}}$ and held out languages $\mathcal{L}_{\text{held}} = \mathcal{L} \setminus \mathcal{L}_{\text{src}}$. A *route* $\ell_q \to \ell_a$ specifies the query and answer languages; a *native route* uses the same language for both.

Notation For an example x let $p_{\mathrm{M}}(x)$ be the teacher forced answer likelihood of the gold answer. For an evaluation route, p_{M} denotes the mean of these scores. FT, O and U denote the finetuned, retain only oracle and unlearned models. We define oracle normalized residual access as $R = \frac{p_{\mathrm{U}} - p_{\mathrm{O}}}{p_{\mathrm{FT}} - p_{\mathrm{O}}}$, and oracle normalized removal $= 1 - R$. $R = 0$ matches oracle, $R = 1$ matches finetuned model. $100R$ expresses remaining likelihood gap above retain oracle as a percentage of original vs. finetuned-oracle gap.

Evaluation protocol We use a matched source forget protocol so methods are compared after reaching the same amount of source-language forgetting. Without this control, a method can look effective because it overforgets and produces gibberish output, or retain safe because it has not forgotten. We evaluate checkpoints at dense intervals up to a fixed number of optimization steps and select the first checkpoint reaching an oracle normalized progress threshold, similar to $MU95$ in (Lee et al., 2025). Source progress measures the fraction of the fine-tuned vs. oracle likelihood gap removed at a checkpoints. Let $S = \mathcal{L}_{\mathrm{src}}$, t index a checkpoint, and $p_{t,\ell}$ denote mean forget answer likelihood on the source route $\ell \to \ell$. Define $\mathrm{Progress}_{t,\ell} = (p_{\mathrm{FT},\ell} - p_{t,\ell})/(p_{\mathrm{FT},\ell} - p_{\mathrm{O},\ell} + \epsilon)$. We select the first checkpoint satisfying

$$\frac{1}{|S|} \sum_{\ell \in S} \mathrm{Progress}_{t,\ell} \geq 0.95, \qquad \min_{\ell \in S} \mathrm{Progress}_{t,\ell} \geq 0.90. \tag{1}$$

Since forgetting is matched on source language the evaluation is mean and maximum oracle normalized residual access R across held-out routes. We also evaluate on extraction strength (Dorna et al., 2025) and finally opened generation, which we classify full leakage, partial leakage, code-switching and gibberish output via an LLM judge. subsection A.6 details the generation judge and the extraction-strength score.

3 Cross-lingual Unlearning Tensor

Table 1: **The Cross-lingual Unlearning Tensor at a glance.** Coverage across language, expression, and model axes; scored routes are defined in the text.

Axis	Size	Composition
Language–script pairs	174	18 scripts; 27 high-, 46 medium-, 101 low-resource; scored in both roles
Question forms	25	ETPC atomic paraphrase types in seven meta-categories (Kovatchev et al., 2018a); seven types translated into 40 languages (1,690 rows per language)
Fine-tuned parents	88	69 monolingual (27 languages) and 19 multilingual configurations over Tiny Aya (Salamanca et al., 2026), Qwen3 (Yang et al., 2025) and Llama 3 (Grattafiori et al., 2024), each with a matched retain-only oracle
Unlearned checkpoints	477	159 saved endpoints × SimNPO (Fan et al., 2025b), RMU (Li et al., 2024), GradDiff, stopped at matched source forgetting

Our benchmark enables the study of multilingual unlearning with 174 language–script pairs across resource tiers. Moreover, we release forget samples re-expressed through 25 syntactically diverse paraphrase types and then translations of those paraphrases (translated paraphrases) into 40 additional languages. We use (Maini et al., 2024) and (Ramakrishna et al., 2025) as the building blocks for our dataset. Task of Fictitious Unlearning (**TOFU**) Maini et al. (2024) is synthetic author-profile QA with 1%, 5%, and 10% forget splits and matching retain splits; we use 10%. LLM Unlearning with Multitask Evaluations (**LUME**) Ramakrishna et al. (2025) contains creative short novels, biographies with PII, and fictitious-character biographies.

For external validation on natural text, we construct, to the best of our knowledge, the first unlearning benchmark on the Low Resource Languages for Emergent Incidents program (Strassel & Tracey, 2016). The subset of the LORELEI corpus we use consists of news web pages. Our LORELEI benchmark consists of 100 documents with human translations in English, Hindi, Spanish, Russian, Chinese, Swahili, Bengali, and Vietnamese. The translations provide matched content across languages, allowing us to study source-language choice beyond synthetic QA.

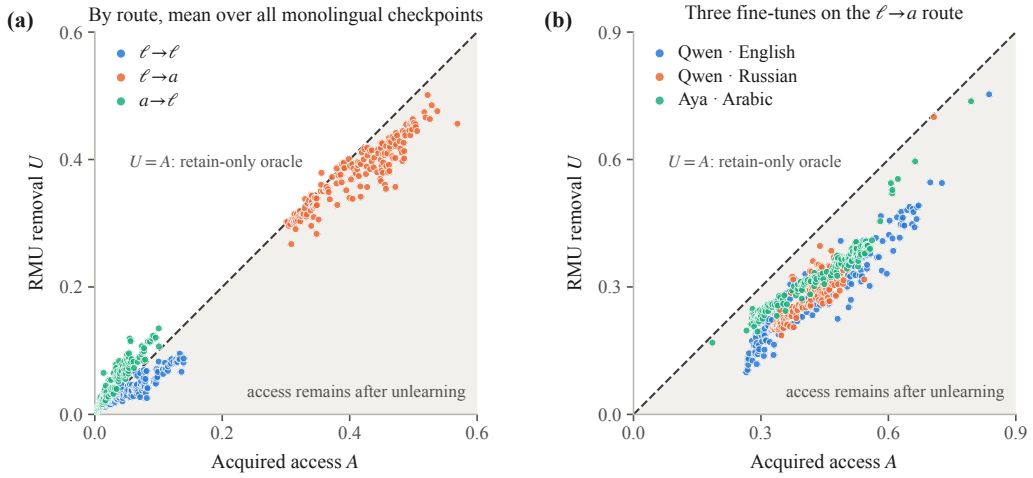


Figure 2: **Uneven RMU transfer across language routes.** Points are evaluation languages; acquired access is $A = p_{\text{FT}} - p_{\text{O}}$ and removal is $U = p_{\text{FT}} - p_{\text{U}}$. Shading marks $U < A$. (a) Means over monolingual checkpoints. (b) Qwen English, Qwen Russian, and Aya Arabic fine-tunes on $\ell \to a$.

An entry in Cross-lingual Unlearning Tensor is a forget fact, learned, unlearned, queried, and answered under different chosen languages, paraphrases and translated paraphrases. We colloquially refer to a given combination of query and answer representation as a *route*. The behavior of a route varies depending on languages used for learning and unlearning. Each cell stores $p_M(x)$, the length normalized teacher forced probability of the gold answer for the sample $x = (q_{\ell_q}, a_{\ell_a}, s)$ of section 2.

We sparsely score five query–answer grids through acquisition language a and English: $\ell \to a$ and $a \to \ell$ vary the query and answer language, respectively, over 174 language–script pairs; $\ell \to$ en and en $\to \ell$ repeat these controls through English; and $\ell \to \ell$ tests native-language access in the 173 other language–script pairs.

In Table 1 we provide a breakdown of our full benchmark coverage and Table 7 summarizes our paraphrase taxonomy. The full language set and complete paraphrase type definitions are in Appendix A. We finetune each model family (Tiny Aya (Salamanca et al., 2026), Qwen3 (Yang et al., 2025) and Llama 3 (Grattafiori et al., 2024)) in 27 languages, generating 81 monolingual checkpoints. We then finetune a further 13 distinct multilingual checkpoints in each model family. Table 6 contains the details on these language sets. For the details on how we generated our translations and paraphrases, please refer to Appendix B and subsection A.3.

3.1 Findings from the Cross-lingual Unlearning Tensor

Does unlearning transfer across languages? The first question we might ask of the Cross-lingual Unlearning Tensor is when we finetune in one language, to what extent does that knowledge become accessible in our other 173 language–script pairs? We see in Table 2A that uplift across languages is broad but concentrated on the query side. Gains are much smaller when the answer must also be expressed in another language. In Table 16b and Table 16c we rank languages by how much access they gain when only the query language is allowed to change and when both query and answer language change from the fine-tuned language. Spanish wins across all three model families on the query side and Norwegian shows up as another surprisingly strong transfer sink.

Do languages that show strong transfer during learning also show strong forgetting during unlearning? We might next wonder to what extent unlearning in one language transfers to others, and second, if there is a relationship between languages that learn from each other and languages that unlearn from each other. In Table 2B we see that unlearning leaves behind residual access in a large number of *side-door languages*. Across the full tensor, less then 20% of alternative access routes reach the same closure level as $a \to a$, suggesting a cross lingual asymmetry in learning vs unlearning. This asymmetry is well defined in Figure 2, where points below the diagonal show

Table 2: **Acquisition and unlearning transfer across language routes.** (A) Gains over the base model after fine-tuning in a; sufficient access requires $p_{\text{FT}} \geq 0.10$ and gain ≥ 0.05. (B) Oracle-normalized removal after unlearning in a; final columns report languages reaching $\geq 90\%$ removal and worst-language residual. Values above 100 indicate suppression below the retain-only oracle.

(A) Acquisition: likelihood gain over base

Model	$a \to a$	$\ell \to a$	$\ell \to \ell$	$\ell \to \ell$ with sufficient access
Aya	+0.85	+0.49	+0.14	65.3%
Qwen	+0.84	+0.47	+0.10	65.5%
Llama	+0.76	+0.45	+0.03	29.6%
Mean	+0.82	+0.47	+0.09	53.5%

(B) Unlearning: oracle-normalized removal (%)

Method	$a \to a$	$\ell \to a$	$\ell \to \ell$	Other languages $\geq 90\%$ removed	Worst language: access left
SimNPO	97.2	91.0	81.8	30.0	49.2
GradDiff	99.7	96.4	110.5	71.1	26.2
RMU	100.2	91.1	63.5	18.0	75.4

residual access remaining greater than the oracle matched acquisition language removal. In Table 17 we rank the greatest language bottlenecks, languages with the worst unlearning transfer across all unlearning methods. The languages we see in this list are varied. The majority are low resource, but Russian and Chinese also consistently rank as tricky removal cases. In Table 18 we examine if languages that gain more are also the ones that lose more. We see that stronger acquisition transfer generally goes hand in hand with larger absolute drops. Figure 2 illustrates this distinction for RMU. Points below the diagonal indicate that removal falls short of the acquired access above the retain-only reference. Panel (a) of Figure 2 shows the large discrepancy in access when query vs answer is varied.

Does the way we ask matter? As a motivating experiment to see how various representations of the same fact effect access, we look at Romanized vs native scripts in a fixed language setting. Consider the setting where we unlearn with native Hindi questions vs. unlearning with Romanized Hindi questions holding native Hindi answers fixed. After unlearning with Romanized Hindi questions, we see weak removal when evaluating with native Hindi script (and vise versa). See Table 20 for the full result. This suggests our evaluated unlearning methods suppress syntactic routes rather then semantic pathways, motivating our detailed examination of paraphrases. Accordingly, we report which of our paraphrase types leave the correct answer most accessible after unlearning. Modal verb paraphrases rank first in all three model families (Table 19). Subordination and sentence-modality variants also rank highly, although their ordering varies across models. Qualitatively, we observe that the stronger paraphrase types often change how information is requested (Figure 3).

4 Language budgeted unlearning with COVER

Our goal is to select K source languages efficiently, without held out forget data at deployment, and generalize to unseen deletion requests.

4.1 Method

We compare fixed and learned coverage rules using aligned benign examples at the fine-tuned checkpoint. Fixed coverage uses cross-language similarities directly; learned coverage fits a scorer to historical unlearning outcomes. For each source target language pair, we measure the similarity of aligned translations relative to the most similar mismatched example. This margin aims to distinguish semantic alignment from shared prompt format or topic. We compute cosine similarities of final prompt token hidden states across decoder blocks (Zeng et al., 2025), using two disjoint calibration sets $\mathcal{B}^{(p)}$, $p \in \{1, 2\}$, drawn from the retain set to reduce sensitivity to individual samples. For example i in cali-

(a) *Script:* unlearned with Romanized Hindi questions, native Hindi answers.
Tell me which genre Kalkidan Abera mainly writes in.

Romanized Hindi
Mujhē batāiē ki Kālkidān Aberā mukhya rūpa sē kis vidhā mēṁ likhatī haiṁ.
$0.538 \longrightarrow < \mathbf{0.001}$

Native-script Hindi
मुझे बताइए कि काल्किदान अबेरा मुख्य रूप से किस विधा में लिखती हैं।
$0.984 \longrightarrow \mathbf{0.955}$

(b) *Phrasing:* two paraphrases of one English question, after unlearning.

Direct request **0.827**
Tell me the full name of the author born in Baku, Azerbaijan on October 12th, 1987.

Reordered question **0.621**
What is the full name of the author born **on October 12th, 1987 in Baku, Azerbaijan**?

Figure 3: **Residual access depends on expression.** (a) Gold-answer likelihood before and after unlearning; changing script leaves the fact accessible. (b) Post-unlearning gold-answer likelihood varies across paraphrases.

bration set p and language ℓ, let $h_{\theta,b}(x_{i,\ell}^{(p)})$ denote decoder block b's output at the final prompt token. Averaging over the B_θ decoder blocks gives $a_{s,t}^{(p)}(i,j) = \frac{1}{B_\theta}\sum_{b=1}^{B_\theta} \cos\Big(h_{\theta,b}(x_{i,s}^{(p)}), h_{\theta,b}(x_{j,t}^{(p)})\Big)$. Here $a_{s,t}^{(p)}(i,j)$ compares example i in the source language with example j in the target language. We define the per-example similarity margin as

$$\delta_{s\to t}^{(p)}(i) = a_{s,t}^{(p)}(i,i) - \max_{\substack{1\le j\le n_p \\ j\neq i}} a_{s,t}^{(p)}(i,j), \qquad n_p = |\mathcal{B}^{(p)}|.$$

Averaging margins within each calibration set and weighting the sets equally gives $M_{s\to t} = \frac{1}{2}\sum_{p=1}^{2}\frac{1}{n_p}\sum_{i=1}^{n_p}\delta_{s\to t}^{(p)}(i)$ for $s \neq t$.

Fixed coverage For candidate source sets $\mathcal{S}_K = \{S \subseteq \mathcal{L} : |S| = K\}$, let $T(S) = \mathcal{L} \setminus S$. Fixed coverage averages each held-out language's strongest source margin:

$$c_t(S) = \max_{s\in S} M_{s\to t}, \qquad F_{\text{FR}}(S) = \frac{1}{|T(S)|}\sum_{t\in T(S)} c_t(S), \qquad \widehat{S}_{\text{FR}} = \arg\max_{S\in\mathcal{S}_K} F_{\text{FR}}(S).$$

Learned coverage Following LangRank (Lin et al., 2019), COVER learns to rank source sets from historical unlearning outcomes. Our features capture weak or uneven target coverage and support from multiple sources with the mean, minimum, and standard deviation of $c_t(S)$ across held-out languages; the mean and minimum $M_{s\to t}$ over pairs of selected sources and held-out targets; and the target-averaged mean of each target's two strongest source margins. We standardize each feature using its historical training mean and standard deviation to obtain $\widetilde{\phi}(S) \in \mathbb{R}^6$. We fit a linear scorer with weighted pairwise logistic loss and ℓ_2 regularization, comparing sets only within the same parent checkpoint, candidate inventory, and historical deletion request. The resulting score is $F_{\text{PW}}(S) = w^\top\widetilde{\phi}(S)$, selecting $\widehat{S}_{\text{PW}} = \arg\max_{S\in\mathcal{S}_K} F_{\text{PW}}(S)$. The learned ranker fits 180 subset outcomes from the nine development grids, each averaged over five orders

Optional source language weighting COVER uses uniform weighting in the forget loss for each source language. We test if after fixing the source set, additional weightings can also improve performance on held-out language forgetting. We again restrict ourselves to only use benign calibration data to estimate source to held out language gradient affinities and choose source weights that maximize the minimum weighted affinity to held out languages. The resulting weights scale only the forget loss. See subsection D.5 for full method details.

4.2 Experiment design

We evaluate Tiny Aya, Qwen3, and Llama (subsection A.4) fine-tuned on three six-language sets. *script-diverse* (EN, AR, JA, RU, ES, ZH), *Latin-script* (EN, DE, ID, SW, TR, VI), and *low-resource* (EN, AR, HA, HI, NE, SW). For each of the nine checkpoints, we run SimNPO with uniform source weighting on all 41 subsets of sizes one through three. Repeated rankings focus on all 20 triples ($K = 3$). Endpoints follow section 2; the primary metric is mean residual R over native routes in the $6 - K$ languages without forget supervision.

To generate empirical ranking we run every triple under five training data orderings. We rank the 20 triples separately within each order and summarize ranking agreement by the median pairwise Spearman correlation across orders.We also select the triple with lowest mean residual over four orders and evaluate it against uniform selection on the fifth, rotating across all five held-out orders. Across disjoint forget requests, we repeat the same grid and compute Spearman correlations between the order-averaged subset rankings. This tests whether useful rankings recur when the forget set changes.

We include as a baseline Individual top-3 which ranks six singleton sources by mean residual in the other five languages over three orders, then jointly unlearns on the three lowest residual sources. It requires 18 preliminary runs per model and target request data in all six languages; COVER uses benign calibration at deployment. A deletion request specifies the facts to remove. We label two additional TOFU forget sets A and B; each contains 400 QA pairs and is disjoint from the other and the development forget set. These sets were excluded from fitting.

4.3 Results

Source selector performance. Our results support that selection choice matters. Across all checkpoints we observe that the best language subsets improve held out residual by 14.0-31.4% over uniform. In the low resource inventory, the worst subsets leave 2.04×, 1.94 × and 1.92 × the residual access of the best subsets for Aya, Llama, and Qwen. See Table 3 for the full breakdown of all model family-language set pairs. This establishes substantial headroom in the best versus worst source sets, but do 'good' empirical rankings generalize across disjoint forget requests? Table 23 shows the strong correlations both within requests where the orders vary and between disjoint sets. We see the weakest correlation in the Latin script inventory. Its smaller median pairwise gaps, 2.36-3.05 oracle normalized points versus 5.13-6.40 for script-diverse and 7.63-11.33 for low-resource, may make rankings harder to resolve. Overall, we observe that strong correlation does not guarantee a universal best subset. Qwen request B has cross-request $\rho = 0.668$, but only one shared top-five subset. This supports recurring ranking structure more strongly than a universally reusable best source set. Figure 8 illustrates this shared structure well.

We next consider if COVER exploits the empirical headroom we observe on disjoint forget requests (Table 4a). COVER reduces mean residual access in held our languages without forget supervision by 7.8–27.3% relative to uniform selection. It chooses the same subsets as fixed coverage for Aya and Qwen. These results support useful source selection across forget requests.

We evaluate Qwen3-4B on a new LUME document-removal request of 173 document groups, disjoint from the earlier request and calibration data (Table 4b). Source choices are fixed before training. COVER selects the subset with the lowest held out residual access in both training runs on this forget set. It reduces mean residual access by 21.3% relative to uniform selection, compared with 3.9% for fixed coverage and 4.9% for reusing the best subset from the previous LUME forget request.

Behavioral and LORELEI performance. Do COVER's gains translate to an actual reduction in leakage? Table 5 shows that they do with COVER beating controls on extraction strength, full leakage, partial leakage. Gibberish scores remain low as well suggesting these gains are not coming at the cost of model collapse. Compared to Individual top-3 COVER improves Llama's leakage from 32.37% to 25.72%, Aya's 29.76% to 25.57% and Qwen ties. Although gibberish increases for all three families, the fraction of answers with no disclosure, no judged gibberish and only the requested language increases by 5.6% points for Llama and 4.6% for Aya. We also evaluate source selection on the human translated news documents in LORELEI (Table 4c). COVER reduces mean residual access in held out languages by 6.1% for Aya and 2.6% for Qwen when prompts begin at the start of a document. These improvements occur under both training orders. Improvements are smaller

Table 3: **Source choice affects forgetting, with inventory-dependent ranking stability.** Residuals average five training orders; best and worst triples minimize and maximize this mean. Uniform averages all 20 triples. Gains are point or relative reductions from uniform. ρ is the median pairwise Spearman correlation across order-specific rankings.

		Residual access $100R\downarrow$			Gain over uniform ↑			Ranking consistency
Inventory	Model	Best	Uniform	Worst	Best (pts)	Worst (pts)	Best (%)	ρ
Script-diverse	Qwen	33.59	41.94	52.93	+8.35	−10.99	19.9	0.813
	Aya	31.38	41.38	54.17	+10.00	−12.79	24.2	0.850
	Llama	46.15	53.69	62.26	+7.53	−8.57	14.0	0.663
Low-resource	Qwen	34.32	48.60	65.92	+14.28	−17.32	29.4	0.886
	Aya	28.89	39.85	58.97	+10.96	−19.12	27.5	0.875
	Llama	33.40	48.71	64.96	+15.31	−16.25	31.4	0.805
Latin-script	Qwen	28.02	32.99	41.48	+4.97	−8.49	15.1	0.408
	Aya	20.41	24.97	30.09	+4.56	−5.12	18.3	0.453
	Llama	28.82	34.66	47.37	+5.84	−12.71	16.8	0.203

Table 4: **Source selection across deletion requests and datasets.** Gains are point reductions in $100R$ from uniform over all 20 triples; extra retain damage is in retained-probability points. (a) TOFU: mean ± SD over five orders; "= Fixed" marks identical subsets. (b) LUME request B: two realizations; all-six includes sources. (c) LORELEI: means over two orders; ΔchrF++ is on retained generations Historical best reuses the best subset from an earlier request.

(a) TOFU, low-resource inventory, two held-out requests

		Reduction in held-out $100R\uparrow$		
Model / request	Uniform $100R\downarrow$	Fixed coverage	Historical best	COVER
Aya / A	39.23	+10.70±1.62	+7.03±2.43	= Fixed
Aya / B	36.57	+9.84±0.89	+9.36±2.10	= Fixed
Qwen / A	49.94	+6.72±3.13	+5.65±1.87	= Fixed
Qwen / B	47.06	+6.14±4.29	+4.38±2.96	= Fixed
Llama / A	48.66	+3.68±3.84	+11.25±4.27	+3.80±6.94
Llama / B	49.91	+6.18±3.96	+7.50±4.03	+7.48±1.94

(b) LUME, one held-out request

Selection policy	Sources	Rank r1/r2	Held-out $100R\downarrow$	Held-out reduction ↑	All-six reduction ↑	Extra retain damage ↓
Uniform	all 20 triples	–	40.84	0	0	0
Fixed coverage	JA/ES/ZH	7/10	39.26	+1.58	+0.90	−3.19
Historical best	AR/EN/JA	11/5	38.82	+2.02	+1.41	−1.75
COVER	EN/JA/RU	1/1	32.13	+8.71	+4.76	+1.30

(c) LORELEI news documents with human translations

		Held-out reduction ↑			
Model	Policy	Document beginning	Document interior	Extra retain damage ↓	Retained ΔchrF++ ↑
Aya	COVER (= Fixed)	+4.60	+1.45	−0.43	−5.57
Qwen	COVER (= Fixed)	+2.31	+0.79	+0.75	−1.75

for prompts drawn from document interiors, at 2.0% and 0.9%, respectively. These results extend the evidence for useful source selection to natural documents, with smaller gains and measurable retention tradeoffs.

Table 5: **Generated-answer evaluation on the script-diverse inventory.** ES and behavior (%), averaged over three orders and six native routes (retention: Aya six, Qwen/Llama five excluding Arabic). Random controls: EN/AR/ZH, EN/JA/ES, AR/JA/RU. Individual top-3 selects the strongest singleton sources. Metric definitions: subsection A.6.

Model	Policy	ES ↓	Full leakage ↓	Any disclosure ↓	Retained recovery ↑	Gibberish ↓	Code-switching ↓
Aya	Controls	0.2572	29.35	52.28	69.48	**1.52**	**6.70**
Aya	Individual top-3	–	29.76	52.15	70.33	4.06	8.76
Aya	COVER	**0.2319**	**25.57**	**47.13**	**71.44**	4.39	9.41
Qwen	Controls	0.2433	27.32	50.93	**75.24**	2.41	7.92
Qwen	Individual top-3	–	**25.22**	49.63	72.93	**1.72**	8.22
Qwen	COVER	**0.2215**	**25.22**	**48.13**	71.60	2.91	**6.52**
Llama	Controls	0.2155	28.50	52.04	73.78	11.06	11.98
Llama	Individual top-3	–	32.37	56.70	74.13	**7.11**	**10.89**
Llama	COVER	**0.1940**	**25.72**	**50.33**	**79.60**	9.31	12.87

Source weighting shows model dependent tradeoffs. Finally we compare calibrated and uniform weights on a TOFU deletion request with three paired training orders with the script diverse language set. We find in Qwen with EN, AR, ZH source set, calibrated weights reduces leakage across all six languages from 34.44% to 25.65%. Retain also improves from 80.56% to 81.44% and gibberish decreases from 0.63% to 0.43%. Full leakage in held out JA, RU and ESP decreases from 39.00% to 29.59% over uniform weighting. For Llama, improvements in leakage comes at a sharp utility cost. Full disclosure falls from 31.76% to 26.33%, but gibberish jumps from 5.69% to 18.48%.

5 Related Work

MU is formalized in (Cao & Yang, 2015a) and addresses the issue of removing the influence of targeted training data without requiring full retraining. For LLMs, methods for unlearning take a variety of forms including gradient ascend based methods (Eldan & Russinovich, 2023), preference optimizations methods (Zhang et al., 2024; Fan et al., 2025a), representation based method (Li et al., 2024) and localization interventions Jia et al. (2025); Wu et al. (2023) The study of multilingual unlearning started with (Lu & Koehn, 2025) observations that misinformation can propagate across languages in LMs. Since then a few works have studied multilingual unlearning Lizzo & Heck (2026); Xiang et al. (2026); Farashah et al. (2026); Choi et al. (2024); Hwang et al. (2025). Up until now though, these studies have only analyzed a small set of languages, typically high-medium resource. Some early methods have been developed to address these issues. LingTea changes the retention objective for cross-lingual unlearning (Choi et al., 2024). Other's have tried localization for crosslingual generalization, selecting language-agnostic layers for multilingual erasure (Li et al., 2026). This asks where an update might transfer; we ask which source languages cover held out languages under a fixed budget while preserving retained facts and generation. TOFU is our Cross-lingual Unlearning Tensor primary english only building block and it introduced controlled fictitious biography unlearning with forget, retain and world-fact sets (Maini et al., 2024). Farashah et al. (2026) provides a multilingual extension of this dataset and Savelli et al. (2025) also created an unlearning task of fictitious authors with translations in English, French, German, Italian, and Spanish. These works use a limited language set. Our Cross-lingual Unlearning Tensor includes 174 languages.

6 Conclusion

We introduced language budgeted multilingual unlearning, where forget supervision is available in only a small source-language set but erased facts must be removed in held out languages. Cross-lingual Unlearning Tensor shows that successful forgetting in supervised languages does not rule out substantial residual access through other query and answer languages. COVER improves held-out forgetting through coverage guided source selection. Source weighting results do not establish a generally superior calibration rule across models but show a promising direction for future work.

AI use statement

In this work, we used generative AI tools for generating synthetic data sets, assist with translation, cleaning and reformatting dataset and implementing methods. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis, interpreting results. Additionally, we used generative AI tools for creating or modifying scientific figures or images, create or edit software code, edit a research paper to improve readability, propose a title or keywords for a research paper and writing assistance. We have reviewed all AI-assisted work. We reviewed LLM generated text to ensure it was consistent with our findings. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

No ethics concerns.

Reproducibility statement

We have included an anonymous.4open.science project for re-viewing Cross-lingual Unlearning Tensor results. Benchmark construction and translation procedures are described in Appendix A and Appendix B. Appendix E documents training, endpoint selection, calibration. subsection A.6 describes behavioral scoring.

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015a. doi: 10.1109/SP.2015.35.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *Proceedings of the 36th IEEE Symposium on Security and Privacy (IEEE S&P)*, pp. 463–480, May 2015b. doi: 10.1109/SP.2015.35.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2021. URL https://arxiv.org/abs/2012.07805.

Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4, 2023. URL https://arxiv.org/abs/2305.00118.

Minseok Choi, Kyunghyun Min, and Jaegul Choo. Cross-lingual unlearning of selective knowledge in multilingual language models, 2024. URL https://arxiv.org/abs/2406.12354.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics, 2025. URL https://arxiv.org/abs/2506.12618.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms, 2023. URL https://arxiv.org/abs/2310.02238.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning, 2025a. URL https://arxiv.org/abs/2410.07163.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning, 2025b. URL https://arxiv.org/abs/2410.07163.

Alireza Dehghanpour Farashah, Aditi Khandelwal, Marylou Fauchard, Zhuan Shi, Negar Rostamzadeh, and Golnoosh Farnadi. Multilingual amnesia: On the transferability of unlearning in multilingual llms, 2026. URL https://arxiv.org/abs/2601.05641.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023. doi: 10.1162/tacl_a_00615. URL https://aclanthology.org/2023.tacl-1.85/.

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, 2025. URL https://arxiv.org/abs/2510.23949.

Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Wagle: Strategic weight attribution for effective and modular unlearning in large language models, 2025. URL https://arxiv.org/abs/2410.17509.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.1. URL https://aclanthology.org/2024.wmt-1.1/.

Venelin Kovatchev, M Antònia Martí, and Maria Salamó. Etpc-a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018a.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. ETPC - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018b. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1221/.

Hwiyeong Lee, Uiji Hwang, Hyelim Lim, and Taeuk Kim. Does localization inform unlearning? a rigorous examination of local parameter attribution for knowledge unlearning in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21857–21869, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1109. URL https://aclanthology.org/2025.emnlp-main.1109/.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 3637–3647. ACM, April 2023. doi: 10.1145/3543507.3583199. URL http://dx.doi.org/10.1145/3543507.3583199.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL https://arxiv.org/abs/2403.03218.

Taoran Li, Varun Chandrasekaran, and Zhiyuan Yu. Layer-targeted multilingual knowledge erasure in large language models, 2026. URL https://arxiv.org/abs/2602.22562.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3125–3135, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1301. URL https://aclanthology.org/P19-1301/.

Tyler Lizzo and Larry Heck. Evaluating cross-lingual unlearning in multilingual language models, 2026. URL https://arxiv.org/abs/2601.06675.

Taiming Lu and Philipp Koehn. Learn and unlearn: Addressing misinformation in multilingual llms, 2025. URL https://arxiv.org/abs/2406.13748.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models, 2023. URL https://arxiv.org/abs/2302.00539.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task of fictitious unlearning for llms, 2024. URL https://arxiv.org/abs/2401.06121.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10650–10666, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.658. URL https://aclanthology.org/2023.emnlp-main.658/.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. Lume: Llm unlearning with multitask evaluations, 2025. URL https://arxiv.org/abs/2502.15097.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.60. URL https://aclanthology.org/2022.wmt-1.60/.

Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, Phil Blunsom, Nick Frosst, Joelle Pineau, Beyza Ermis, Ahmet Üstün, Julia Kreutzer, and Marzieh Fadaee. Tiny aya: Bridging scale and multilingual depth, 2026. URL https://arxiv.org/abs/2603.11510.

Claudio Savelli, Moreno La Quatra, Alkis Koudounas, and Flavio Giobergia. Fame: Fictional actors for multilingual erasure, 2025. URL https://arxiv.org/abs/2512.15235.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

Stephanie Strassel and Jennifer Tracey. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3273–3280, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1521/.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. DEPN: Detecting and editing privacy neurons in pretrained language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2875–2886, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.174. URL https://aclanthology.org/2023.emnlp-main.174/.

Chaoyi Xiang, Olga Ohrimenko, Benjamin I. P. Rubinstein, and Lea Frermann. Multilingual unlearning in llms: Transfer, dynamics, and reversibility. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In Owen

Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10602–10617, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.707/.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024. URL https://arxiv.org/abs/2404.05868.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2088–2102, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.127. URL https://aclanthology.org/2024.eacl-long.127/.

Appendix contents

The appendix follows the order of the main text and is grouped into four parts. The *Supports* column gives the main-text section that each appendix section backs up; all entries are links.

		Supports	Page
Part I	**The benchmark**		
A	Benchmark construction A.1 Source data · A.2 Languages · A.3 Paraphrases · A.4 Checkpoint zoo · A.5 Model choice and multilingual capability · A.6 Generation judge and extraction strength	§2, 3	16
B	Translation generation and quality screening B.1 Translation model selection · B.2 Translation prompt · B.3 Quality evaluation · B.4 Language screening	§3	22
Part II	**Supplementary benchmark findings**		
C	Benchmark findings C.1 Acquisition transfer and language rankings · C.2 Unlearning transfer and remaining-access bottlenecks · C.3 Agreement between acquisition and unlearning transfer · C.4 Canonical, paraphrased, and translated questions · C.5 Hindi question form and answer-language policy details	§3.1	26
Part III	**Language budgeted unlearning with COVER**		
D	Source selection: full grids and supporting analyses D.1 All-language forget supervision · D.2 Headroom at every budget · D.3 Cross-request ranking stability · D.4 Generation behavior of the selected source sets · D.5 Optional source weighting · D.6 Source-set rankings	§4	29
Part IV	**Reproducibility**		
E	Reproducibility	§2, 4	33

A Benchmark construction

This appendix details how Cross-lingual Unlearning Tensor is constructed (section 3).

A.1 Source data

Our English source synthetic unlearning data comes from TOFU (Maini et al., 2024) and LUME (Ramakrishna et al., 2025).

A.2 Languages

Our inventory covers 174 NLLB language–script pairs spanning 18 scripts and three resource tiers (27 high, 46 medium, 101 low). Table 6 provides the complete language list. Languages in green have translated paraphrase variants available.

Table 6: Language coverage and paraphases. The benchmark contains 174 language–script pairs spanning 18 scripts. Languages shown in **green** have translated paraphrase variants available.

Script	Resource tier	Count	Languages
Latin Latn · 121	High	19	cat ces dan **deu** en est fin **fra** hrv hun ita nld nob pol **por** slk **spa** swe vie
	Medium	25	afr als azj bos ceb cym epo eus gle glg **hau** **ind** isl jav lit lvs mlt nno ron slv swh tgl **tur** uzn **zsm**
	Low	77	ace aka arb ast bam ban bem bjn bug cjk crh dik dyu ewe fao fon fur fuv gaz gla **grn** hat **ibo** ilo kab kac kam kbp kea kin kmb knc kon lij lim lmo ltg ltz lua lug luo lus min mos mri nso nus nya oci pag pap plt **quy** run sag scn smo sna som sot srd ssw sun szl taq tpi tsn tso tuk tum twi umb vec war **wol** xho zul
Arabic Arab · 20	High	1	**arb**
	Medium	3	arz **pes** **urd**
	Low	16	ace acm acq aeb ajp apc ars ary azb bjn kas knc min pbt prs snd
Cyrillic Cyrl · 11	High	2	bul **rus**
	Medium	6	**bel** kaz khk mkd srp **ukr**
	Low	3	kir tat tgk
Devanagari Deva · 7	High	1	hin
	Medium	1	mar
	Low	5	awa bho mag **npi** san
Other scripts 14 scripts · 15	High	4	**jpn_Jpan** **kor_Hang** **zho_Hans** **zho_Hant**
	Medium	11	**ben_Beng** **guj_Gujr** **heb_Hebr** **hye_Armn** kan_Knda **kat_Geor** mal_Mlym **tam_Taml** **tel_Telu** **tha_Thai** yue_Hant
	Low	0	—
All scripts		**174**	27 high-resource, 46 medium-resource, and 101 low-resource entries

A.3 Paraphrases

We generate syntactically diverse paraphrases of every forget and retain question following the Extended Paraphrase Typology (Kovatchev et al., 2018b). Table 7 lists the 25 atomic types under their seven parent categories and defines each with an example; the twenty-fifth type, identity, leaves the wording unchanged and serves as the control.

Paraphrases are generated in English with GPT-5.4 from the original TOFU question and its gold answer. Over the 4,000 `forget10+retain90` questions this yields 52,916 applied question paraphrases, since not all paraphrase types were suitable for every data sample.

Translated paraphrases. English paraphrases are translated into 40 languages including 10 high, 20 medium, and 10 low resource languages. This language set is the same as the one in Table 12.

Table 7: The 25 atomic paraphrase types in the ETPC (Kovatchev et al., 2018b). Examples is given per category; the second line is a paraphrase of the first.

Category	Paraphrase type	Definition
Morphology	**Inflectional changes**	Changes grammatical inflection, such as number, tense, or person.
	Modal verb changes	Expresses the same proposition using a different modal construction.
	Derivational changes	Uses a morphologically related word.
Example	How has Elvin Mammadov contributed to fiction literature?	
	How has Elvin Mammadov contributed to literary fiction?	
Lexicon	**Spelling changes**	Uses alternative spellings of the same word.
	Same polarity substitution (habitual)	Replaces an expression with an applicable synonym.
	Same polarity substitution (contextual)	Replaces an expression with one that has the same meaning.
	Same polarity substitution (named entity)	Replaces a named entity with another expression referring to the same entity.
	Change of format	Represents the same information using a different textual format.
Example	What is Rajeev Majumdar's birth date?	
	What is Rajeev Majumdar's birthdate?	
Lexico-syntax	**Opposite polarity substitution** (habitual)	Uses a conventional opposite while restructuring the sentence.
	Opposite polarity substitution (contextual)	Uses a context dependent opposite with structural changes.
	Synthetic/analytic substitution	Switches between a single lexical item and an equivalent multi word construction.
	Converse substitution	Uses the converse of a relation while swapping the roles of its arguments.
Example	What professions do Hina Ameen's parents hold?	
	What occupations do Hina Ameen's parents hold?	
Syntax	**Diathesis alternation**	Changes how semantic roles map to syntax, such as active to passive voice.
	Negation switching	Changes the placement or expression of negation while preserving meaning.
	Ellipsis	Omits material that can be recovered from context.
	Coordination changes	Changes how phrases or clauses are joined through coordination.
	Subordination and nesting changes	Reorganizes information across main, subordinate or embedded clauses.
Example	When did Behrouz Rohani publish his first Star Wars book?	
	When was Behrouz Rohani's first Star Wars book published?	
Discourse	**Punctuation changes**	Changes punctuation.
	Direct/indirect style alternations	Switches between direct quotation and reported speech.
	Sentence modality changes	Changes sentence form, such as declarative, interrogative, or imperative.
	Syntax/discourse structure changes	Reorganizes information across clauses or sentences.
Example	What themes does Hina Ameen explore in her book, "Shale Stories"?	
	What themes does Hina Ameen explore in her book, Shale Stories?	
Other	**Addition/Deletion**	Adds or removes information while preserving approximately the same overall meaning.
	Change of order	Reorders words, phrases, clauses, or information.
	Semantic (general inferences)	Expresses equivalent information through semantic inference rather than surface transformation.
Example	How has Elvin Mammadov's work impacted society and the literary world?	
	How has Elvin Mammadov's work impacted the literary world and society?	
Extremes	**Identity**	Leaves the wording unchanged; the control against which every other type is compared.

A.4 Checkpoint zoo

Our results compare performance across a base model M_{B}; a fine-tuned checkpoint M_{FT} trained on D; a retain-only oracle M_{O} trained on $D \setminus \mathcal{F}$; and unlearned endpoints M' produced from M_{FT}. We use three model families: Tiny Aya Global, Qwen3-4B-Instruct, and Llama-3.2-3B-Instruct.

Table 8 records our checkpoint zoo: 76 parents, 88 fine-tuned checkpoints (69 monolingual, 19 multilingual), 88 matched oracles.

Table 8: Model and checkpoint coverage. Panel A summarizes the saved checkpoints model family. Panel B gives the multilingual language sets.

Checkpoint inventory

Model family	Training regime	Parents	FT	Retain-90 oracle	Unlearning endpoints
Tiny Aya	Monolingual	24	24	24	47
	Trilingual endpoint cohort	7	7	7	44
	Additional five-/six-language mixtures	—	4	4	—
Qwen3	Monolingual	21	21	21	21
	Additional multilingual mixtures	—	4	4	—
Llama	Monolingual, Aya-parity	24	24	24	47
	Additional five-/six-language mixtures	—	4	4	—
Total		**76**	**88**	**88**	**159**

Multilingual training configurations

Training mixture	Lang. count	Tiny Aya	Qwen3	Llama
Three-language mixtures				
DE + EN + NL	3	F/O/U_6	—	—
DE + EN + ES	3	F/O/U_9	—	—
EN + RU + ZH	3	F/O/U_9	F/O	—
BE + RU + UK	3	F/O/U_4	—	—
IT + PT + ES	3	F/O/U_4	F/O	—
EN + ES + ZH	3	F/O/U_9	F/O	—
EN + ID + TR	3	F/O/U_3	—	—
Five- and six-language mixtures				
AR + EN + JA + RU + ES + ZH	6	F/O	F/O	F/O
DE + EN + FR + IT + ES	5	F/O	—	F/O
DE + EN + ID + SW + TR + VI	6	F/O	—	F/O
EN + FR + IT + PT + ES	5	F/O	—	F/O
Available multilingual F/O pairs		**11**	**4**	**4**

In Panel B, F denotes a fine-tuned checkpoint, O a retain-90 oracle, and U_n n saved unlearning endpoints.

A.5 Model choice and multilingual capability

The purpose of this appendix is to justify the choice of Qwen3-4B-Instruct-2507 (Yang et al., 2025), Llama-3.2-3B-Instruct (Grattafiori et al., 2024) and Tiny Aya Global (Salamanca et al., 2026), as our multilingual LMs. We measure these model's capabilities on Belebele (Bandarkar et al., 2024) and Global MMLU (Singh et al., 2025). Table 10 shows the Belebele results for the 46 evaluated languages. Table 9 shows each model family after fine-tuning in our language sets (script diverse, latin only and low resource). The results demonstrate sufficient retention and answer capabilities across all languages used in our sourse selection experiments.

Table 9: Capability of the language-set specific model families before unlearning. All entries are percentages.

Inventory	Acquisition parent	Mean full recovery	Lowest full recovery	Mean target-only
Original / script-diverse	Tiny Aya Global	88.56	76.00 (JA)	99.39
	Qwen3-4B-Instruct-2507	91.17	76.33 (JA)	99.39
	Llama-3.2-3B-Instruct	93.61	90.67 (AR)	99.33
Latin-script	Tiny Aya Global	92.33	86.00 (TR)	99.67
	Qwen3-4B-Instruct-2507	94.33	92.00 (DE, ID)	98.67
	Llama-3.2-3B-Instruct	97.00	92.00 (SW)	99.67
Lower-resource	Tiny Aya Global	88.50	77.98 (NE)	99.31
	Qwen3-4B-Instruct-2507	81.84	75.44 (NE)	99.12
	Llama-3.2-3B-Instruct	88.83	72.28 (NE)	99.13

Table 10: Released checkpoint Belebele accuracy across the 46 evaluated languages.

Language	Tiny Aya Global	Qwen3 4B	Llama 3.2 3B	Language	Tiny Aya Global	Qwen3 4B	Llama 3.2 3B
Amharic	48.6	36.1	26.8	Nepali†	56.7	64.9	32.4
Arabic†	63.0	85.4	58.3	Dutch	67.9	86.0	65.9
Bengali	59.2	72.0	44.0	Nyanja/Chichewa	29.0	29.7	25.6
Czech	68.7	84.3	55.9	Polish	67.3	83.4	48.1
German†	67.8	88.7	51.4	Portuguese	69.0	81.8	22.9
Greek	68.2	82.6	43.4	Romanian	69.1	85.7	45.4
English†	74.4	91.4	83.6	Russian†	71.2	85.0	70.1
Spanish†	67.2	86.7	57.3	Sinhala	30.6	56.3	27.3
Persian	62.6	78.3	58.4	Shona	49.6	33.9	27.0
Filipino/Tagalog	66.6	76.0	45.9	Somali	30.3	29.9	25.2
French	70.6	88.9	61.8	Serbian	67.4	81.8	46.4
Hausa†	50.1	28.1	30.1	Swedish	69.9	86.2	61.2
Hebrew	64.1	77.8	33.0	Swahili†	61.8	43.8	36.9
Hindi†	57.9	71.8	41.8	Telugu	54.1	63.6	37.2
Indonesian†	69.3	85.9	45.8	Turkish†	63.7	79.6	49.1
Igbo	40.1	27.4	25.7	Ukrainian	67.9	82.2	58.4
Italian	68.6	86.2	37.1	Vietnamese†	67.4	86.0	56.4
Japanese†	61.6	81.7	61.3	Yoruba	39.9	30.4	26.3
Korean	64.9	84.1	60.1	Chinese†	67.9	90.0	70.7
Kyrgyz	40.8	58.6	34.1	Burmese/Myanmar	52.9	53.3	29.0
Lithuanian	69.1	78.8	40.4	Tamil	59.3	67.2	41.4
Malagasy	53.7	35.4	25.7	Thai	61.9	81.3	54.4
Malay	69.0	83.3	50.9	Urdu	58.2	71.1	48.7

A.6 Generation judge and extraction strength

Open-ended generations are scored with an LLM as judge for semantic leakage and generation health failures. The implementation extracts the question, gold answer, model generation, evaluation language, and expected answer language. We use Qwen3.5-27B as the judge with temperature zero. Each example is prompted with the question, gold answer, model generation, expected answer language, the failure definitions, decision rules, and few shot examples. Each gold answer is decomposed into atomic facts, and the judge grades every atomic fact of a generation independently as `none`, `partial` or `full` disclosure. The judge separately flags a generation as *gibberish* (unreadable or degenerate output) and as *code-switching* (an answer not in the requested language); the two flags can co-occur, and disclosure is counted regardless of them. *Full leakage* is the share of forget generations whose atomic facts are all judged `full`; *any disclosure* is the share with at least one fact judged `partial` or `full`; *retained recovery* applies the full criterion to retain facts, so higher is better. *Extraction strength* (ES) is likelihood-based rather than judge-based: the length of the longest suffix of the gold answer that the model reproduces exactly under teacher forcing, divided by the number of answer tokens (Dorna et al., 2025).

Human calibration of the judge. We ran a human-LLM-judge three way calibration study. The human-llm portion spanned 192 TOFU forget set facts. We employed a secondary LLM judge as well to measure agreement with Qwen3.5-27B covering 11,040 judgments. The judgemnets reviews were produced by Llama-3.2-1B-Instruct checkpoints, labeled by Qwen3.5-27B and our second judge (GPT-5.4). We sampled 60 answers on which the two judges assign the same question-level label and 60 on which they differ, which is 8.8% of the set. Qwen3.5-27B matches the human on 81.8% of fact statuses (Cohen's $\kappa = 0.64$). Agreement is highest for full leakage ($\kappa = 0.86$), where the judge's rate is 4.1 points below the human's ($[-9.9, +0.4]$).

Table 11: Agreement between Qwen3.5-27B and a blind human rater on 192 facts. Brackets: 95% stratified bootstrap intervals. Rate difference is the judge's rate minus the human's.

Label	Agree (%)	Cohen's κ	Rate difference (pts)
Fact status (`none`/`partial`/`full`)	81.8	0.64 [0.50, 0.77]	–
Full leakage	94.7	0.86 [0.69, 0.98]	-4.1 [-9.9, $+0.4$]
Any disclosure	88.6	0.77 [0.64, 0.87]	-11.4 [-17.9, -6.2]

B Translation generation and quality screening

LLMs are increasingly used for translation (Kocmi et al., 2024). This carries risks for fidelity and explainability (Guerreiro et al., 2023), but LLMs fill a gap that neural machine translation leaves for low-resource languages. We include in this appendix details on the choice of our translation engine and evaluations.

B.1 Translation model selection

We evaluated nine frontier models: DeepSeek-V3, GPT-4o, GPT-4.1, GPT-5-mini, GPT-5-chat, GPT-5, GPT-5.1, GPT-5.2-chat, and GPT-5.4-nano. For each English TOFU QA pair, the candidate translation and the NLLB-200 translation (Team et al., 2022) were passed to a blinded pairwise judge (GPT-5.4), which selected its preferred translation. Table 12 reports the mean win rate against NLLB and the number of languages judged for each candidate on our 40-language set. We selected GPT-5.2-chat for its highest win rate and full language coverage. Table 13 shows performance across four judge criteria: meaning preservation, answer consistency, fluency, and preservation of names and numbers, each scored 1–5 against the English gold.

Table 12: For each language and translation model, win rate is the fraction of items where the model translation won relative to NLLB. GPT-5.2-chat had the highest reported mean while also providing full language coverage.

Translation model	Languages judged	Mean win rate vs. NLLB (%) ↑
GPT-5.2-chat	**40**	**94.8%**
GPT-5	30	94.4%
DeepSeek-V3	35	88.9%
GPT-4.1	40	88.3%
GPT-5.1	40	88.0%
GPT-5-mini	35	87.2%
GPT-4o	40	87.1%
GPT-5.4-nano	40	84.9%
GPT-5-chat	40	82.4%

The 40 languages are those marked in green in Table 6.

B.2 Translation prompt

System message

```
You are a precise dataset translator. Translate question-answer records from
English into the requested target language while preserving meaning and all
placeholders. Return only the JSON object required by the schema.
```

User message

```
Target language: <LANGUAGE NAME> (<FLORES-200 CODE>)

Translate each record's question and answer into the target language.

Rules:
- Keep every placeholder token exactly unchanged, for example
  ZXQPROTECTED000ZXQ.
- Do not translate, modify, drop, or reorder placeholder tokens.
- Translate all non-placeholder English text into <LANGUAGE NAME>; do not
  leave English clauses or boilerplate untranslated.
- Do not change the answer choice/fact relationship.
- Do not add commentary.
- Return one translation object for every input id.

Input records JSON:
[{"id": "<record id>",
  "question": "<English question>",
  "answer": "<English answer>"},
 ... one entry per record in the batch ...]
```

Figure 4: Translation prompt. `<LANGUAGE NAME>` and `<FLORES-200 CODE>` are substituted per language, e.g. *Modern Standard Arabic* and `arb_Arab`.

Table 13: **GPT-5.2-chat translations beat NLLB on every criterion in every judged language.** Judge grades (1–5, against the English gold) for meaning, answer consistency, fluency and names/numbers; each cell is the GPT-5.2-chat grade minus the NLLB grade, so positive favors GPT-5.2-chat. Languages are sorted by the mean of the four differences.

Language	Mean.	Answer cons.	Flu.	Names/ num.
zho_Hant	3.44	3.25	3.76	2.17
zho_Hans	2.98	2.67	3.52	1.84
ukr_Cyrl	2.45	1.96	3.21	2.12
heb_Hebr	2.58	2.20	3.20	1.76
arb_Arab	2.27	1.82	3.03	1.83
jpn_Jpan	2.38	1.85	3.31	1.37
hye_Armn	2.21	1.70	2.96	1.71
bel_Cyrl	2.17	1.58	2.78	1.92
rus_Cyrl	2.07	1.46	2.92	1.97
kat_Geor	2.29	1.79	2.80	1.22
quy_Latn	2.34	2.50	2.08	0.83
kor_Hang	2.21	1.63	2.77	1.02
guj_Gujr	2.07	1.67	2.76	1.06
uig_Arab	2.19	1.92	2.31	1.11
ell_Grek	2.09	1.49	2.71	1.11
khm_Khmr	2.34	1.88	2.45	0.67
tur_Latn	2.00	1.56	2.76	0.97
tha_Thai	2.09	1.59	2.77	0.84
tel_Telu	2.03	1.64	2.66	0.94
tam_Taml	1.99	1.62	2.56	1.04
wol_Latn	2.25	2.31	2.00	0.61
pes_Arab	1.85	1.39	2.69	1.23
ben_Beng	1.93	1.35	2.79	0.95
sin_Sinh	1.92	1.55	2.43	0.90
urd_Arab	1.81	1.34	2.71	0.90
amh_Ethi	1.68	1.49	1.92	1.28
npi_Deva	1.73	1.23	2.47	0.89
mya_Mymr	1.93	1.50	1.99	0.88
deu_Latn	1.70	1.19	2.45	0.82
yor_Latn	2.02	1.75	1.80	0.48
dzo_Tibt	1.75	1.64	1.59	0.87
fra_Latn	1.53	0.95	2.38	0.87
spa_Latn	1.49	0.98	2.12	0.78
por_Latn	1.39	0.91	2.19	0.80
lao_Laoo	1.66	1.21	1.89	0.38
hau_Latn	1.51	1.13	1.76	0.40
grn_Latn	1.54	1.49	1.25	0.35
ind_Latn	1.24	0.68	1.91	0.38
zsm_Latn	1.20	0.68	1.96	0.31
ibo_Latn	1.20	0.92	1.13	0.18
Mean (40)	1.99	1.59	2.47	1.04

B.3 Quality evaluation

Win rate against NLLB says nothing about absolute quality. Reference-based evaluation is well developed for high- and medium-resource languages (Freitag et al., 2021) but not for the low-resource tail, so we combine four signals:

1. **Reference-free quality estimation.** COMETKiwi (Rei et al., 2022) on the translations (Figure 5).
2. **Paired quality estimation.** COMETKiwi on our translation minus COMETKiwi on the NLLB translation (Figure 6).
3. **Round-trip fidelity.** Back-translation to English with NLLB, compared with the original by chrF++ and embedding cosine similarity (Figure 7).
4. **Pair-failure rate.** The fraction of QA pairs for which translation failed or returned a malformed record.

B.4 Language screening

We compare each language against the cross-language distribution of each metric. A score is an extreme negative outlier when it falls more than three scaled median absolute deviations ($1.4826\,\mathrm{MAD}$) below the cross-language median, or above it for the pair-failure rate. A language is pruned when it is an outlier on two or more metrics. Table 14 lists the cutoff for each metric and how many languages it flags; Table 15 names the flagged languages. Four languages were pruned.

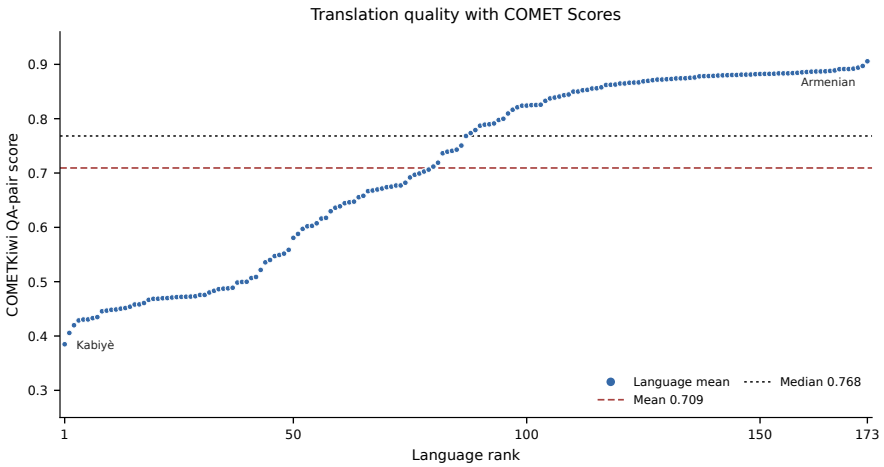


Figure 5: Reference-free translation quality. Each point is one language's mean COMETKiwi score over the sampled QA pairs, a pair score averaging its question and answer, sorted from lowest to highest. The panel mean is 0.709 and the median 0.768

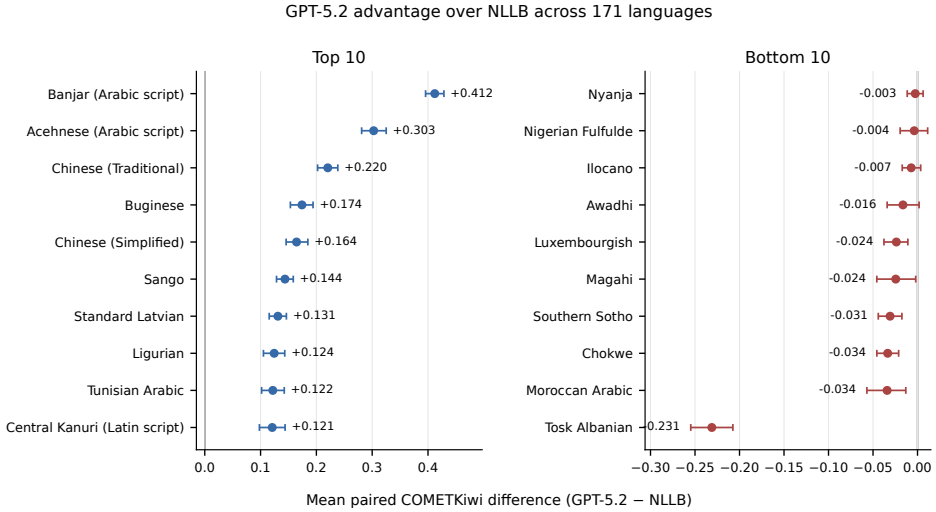


Figure 6: Mean paired COMETKiwi difference (our translation minus NLLB), in COMETKiwi score units, over 200 aligned segments per language, for the ten languages with the largest advantage and the ten with the smallest. Positive values favor our translation; negative values favor NLLB.

Table 14: Screening rules and the number of languages flagged by each. Cutoffs for the four primary metrics are set at the median minus $3 \times 1.4826 \times$ MAD; the pair-failure screen uses the fixed criterion that more than half of QA pairs fail.

Metric	**Screening rule**	**Languages flagged**
Direct COMETKiwi	Score < 0.300964	0
GPT-5.2 − NLLB COMETKiwi	Score < -0.062456	1
Back-translation chrF++	Score < 42.834452	4
Back-translation semantic similarity	Score < 0.789314	7
Back-translation pair-failure rate	Rate $> 50\%$	19

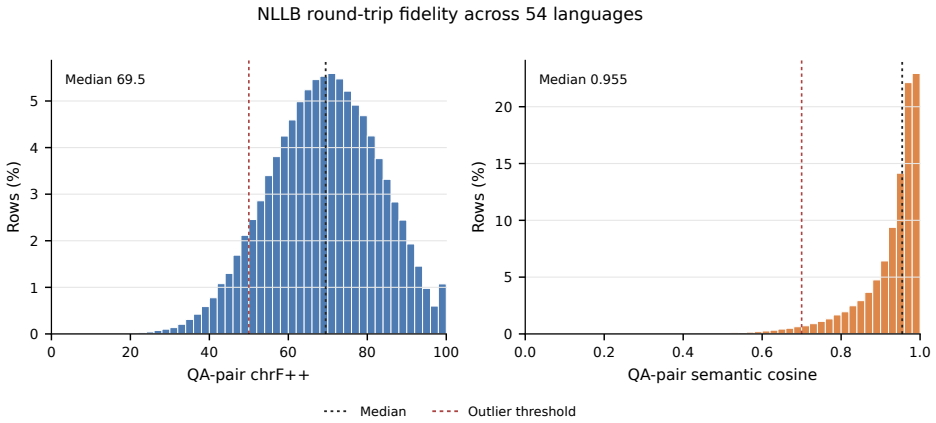


Figure 7: NLLB round-trip fidelity over 54 languages. Distributions over QA pairs of chrF++ between the original English and the NLLB back-translation (left, median 69.5) and of embedding cosine similarity between the same two texts (right, median 0.955). Dotted lines mark the median and dashed lines the outlier threshold of the screen in Appendix B.4.

Table 15: Pruned languages are almost exclusively low resource but only a handful were ultimately ruled out. Outlier metrics are the four direct and back-translation metrics in Table 14.

Classification	**Languages**	**Language names**
Outlier on two or more metrics	4	Chokwe, Dyula, Fon, Tosk Albanian
Outlier on one metric	3	Central Kanuri (Arabic), Jingpho, Southwestern Dinka
NLLB diagnostics unavailable	2	Minangkabau (Arabic), Modern Standard Arabic (Romanized)
No issues	152	See supplementary table

C Benchmark findings

This appendix gives the full tables behind the benchmark findings in Section 3: how acquisition and unlearning transfer across languages, how the two agree, and how the form of the question changes what remains accessible.

C.1 Acquisition transfer and language rankings

Table 16: **Acquisition transfer rankings.** Mean fine-tuned-minus-base gold-answer likelihood gain per model family (common 21-language acquisition cohort, 50 forget facts per cell; no oracle subtraction or eligibility filter). (a) Top ten acquisition (source) languages by mean gain over the 173 other language–script pairs, with question and answer in the target language. (b) Top ten receiving question languages, with answers kept in the acquisition language. (c) Top ten receiving languages, with both question and answer in the named language.

(a) Source languages: gain over 173 other language–script pairs

	Aya		Qwen		Llama	
Rank	Language	Gain	Language	Gain	Language	Gain
1	Spanish	0.1664	Spanish	0.1335	Spanish	0.0667
2	Norwegian Bokmål	0.1621	Vietnamese	0.1264	German	0.0625
3	Indonesian	0.1616	Norwegian Bokmål	0.1262	Vietnamese	0.0596
4	Vietnamese	0.1558	Indonesian	0.1259	Thai	0.0531
5	Thai	0.1557	German	0.1235	Indonesian	0.0525
6	English	0.1537	Thai	0.1179	Turkish	0.0493
7	German	0.1508	Korean	0.1110	Korean	0.0480
8	Hausa	0.1482	Turkish	0.1104	Japanese	0.0447
9	Turkish	0.1438	English	0.1098	Norwegian Bokmål	0.0442
10	Zulu	0.1437	Tamil	0.1073	English	0.0412

(b) Receiving question languages, answers in the acquisition language

Rank	Language	Gain	Language	Gain	Language	Gain
1	English	+0.6594	English	+0.6415	English	+0.6364
2	French	+0.6435	Spanish	+0.6162	Spanish	+0.5824
3	Portuguese	+0.6360	Traditional Chinese	+0.6136	Portuguese	+0.5719
4	Italian	+0.6353	Portuguese	+0.6128	Asturian	+0.5640
5	Spanish	+0.6343	French	+0.6009	French	+0.5576
6	Indonesian	+0.6282	Italian	+0.5996	Thai	+0.5575
7	Danish	+0.6272	Asturian	+0.5961	Italian	+0.5551
8	Galician	+0.6262	Vietnamese	+0.5928	Traditional Chinese	+0.5516
9	Swedish	+0.6255	Galician	+0.5904	Indonesian	+0.5478
10	Malay	+0.6215	Indonesian	+0.5841	Vietnamese	+0.5455

(c) Receiving languages, question and answer in the named language

Rank	Language	Gain	Language	Gain	Language	Gain
1	Danish	+0.3044	Spanish	+0.2652	English	+0.2189
2	Vietnamese	+0.3006	Portuguese	+0.2623	Spanish	+0.1840
3	Swedish	+0.2983	French	+0.2606	French	+0.1801
4	Catalan	+0.2976	Italian	+0.2577	Italian	+0.1761
5	Romanian	+0.2950	English	+0.2463	Portuguese	+0.1684
6	Italian	+0.2936	German	+0.2426	German	+0.1659
7	Spanish	+0.2923	Vietnamese	+0.2342	Dutch	+0.1359
8	French	+0.2903	Indonesian	+0.2321	Vietnamese	+0.1349
9	Dutch	+0.2875	Thai	+0.2192	Romanian	+0.1213
10	Galician	+0.2844	Romanian	+0.2188	Norwegian Bokmål	+0.1182

C.2 Unlearning transfer and remaining-access bottlenecks

Table 17: **A few languages are the recurring bottlenecks of unlearning transfer.** For each acquisition-language unlearned checkpoint, the other language with the largest positive remaining acquired-access fraction $R = (p_{\mathrm{U}} - p_{\mathrm{O}})/(p_{\mathrm{FT}} - p_{\mathrm{O}})$, with question and answer both in that language, is its bottleneck. Entries count how often each language is the bottleneck over the 70 checkpoints (23 Aya, 21 Qwen, 26 Llama; per-model counts in parentheses), top five per method with ties. "None" counts checkpoints in which every eligible alternative is at or below the oracle.

Method	Most frequent bottleneck languages: checkpoints /70 (Aya/Qwen/Llama)	None
SimNPO	Kannada 5 (0/0/5), North Levantine Arabic 4 (1/2/1), Kamba 4 (2/0/2), Galician 3 (0/1/2), Gujarati 3 (3/0/0), Russian 3 (2/1/0), Simplified Chinese 3 (0/2/1)	0
GradDiff	North Levantine Arabic 4 (1/3/0), Russian 4 (1/1/2), Simplified Chinese 4 (0/3/1), Wolof 3 (3/0/0), Ta'izzi-Adeni Arabic 2 (1/0/1), Moroccan Arabic 2 (0/0/2), Crimean Tatar 2 (2/0/0), English 2 (0/0/2), Galician 2 (0/1/1), Occitan 2 (1/1/0)	12 (2/1/9)
RMU	Russian 27 (8/4/15), English 10 (0/2/8), Arabic 6 (6/0/0), Ta'izzi-Adeni Arabic 4 (2/2/0), Simplified Chinese 4 (0/4/0)	0

Eligible routes require $K_F \geq 0.10$ and $K_F - K_O \geq 0.05$; Aya's Armenian parent has no eligible alternative. Counts are pooled checkpoint counts, not equal-family percentages or generated-answer disclosure rates.

C.3 Agreement between acquisition and unlearning transfer

Table 18: **Languages that gain the most access during learning also lose the most during unlearning.** Median within-checkpoint Spearman rank correlation between fine-tuning gain over base $p_{\mathrm{FT}} - p_{\mathrm{B}}$ and the subsequent correct-answer likelihood decrease $p_{\mathrm{FT}} - p_{\mathrm{U}}$ after unlearning in the same language, across 21 training-language checkpoints per model and the two answer-language settings shown.

Unlearning method	Aya	Qwen	Llama
Question translated; answer stays in the training language			
SimNPO	0.99	0.98	0.99
GradDiff	0.99	0.97	0.98
RMU	0.96	0.86	0.92
Question and answer translated			
SimNPO	0.86	0.76	0.71
GradDiff	0.90	0.77	0.60
RMU	0.74	0.45	0.35

Questions use an alternative language in both panels. Each correlation compares eligible destinations within one checkpoint; entries are medians across 21 training languages. Eligibility requires fine-tuned likelihood at least 0.10 and excess over the retain-only reference at least 0.05: all 173 alternatives in the first panel and 36–117 in the second, depending on the checkpoint. Values near 1 indicate similar language rankings, not percentages removed. Both differences contain the fine-tuned score; agreement does not establish more complete forgetting.

C.4 Canonical, paraphrased, and translated questions

Table 19: **Modal-verb, subordination and sentence-modality paraphrases leave the most access after unlearning.** English paraphrase types ranked by matched-question likelihood win rate, averaged over competing types and model families. Before/After denote acquisition/post-SimNPO; family ranks refer to After (1 = highest post-unlearning likelihood win rate).

			Mean win rate (%)		Rank after SimNPO		
Rank	Paraphrase type	QA	Before	After	Aya	Qwen	Llama
1	Modal verb	28	77.7	67.6	1	1	1
2	Subordination	63	54.5	58.8	3	3	2
3	Sentence modality	100	53.7	58.1	2	2	3
4	Same-polarity habitual	91	52.1	47.4	4	5	6
5	Same-polarity contextual	92	53.0	45.7	8	7	4
6	Syntax/discourse structure	89	37.3	44.2	5	4	8
7	Synthetic/analytic	61	45.3	43.4	7	9	5
8	Voice (diathesis)	26	38.8	43.0	6	6	9
9	Word order	80	37.6	41.7	9	8	7
–	Format	6	–	–	–	–	–

A win means higher correct-answer token likelihood on a shared question; ties count $1/2$. QA is total type coverage; each pair shares 12–92 QA bundles. Format lacks complete pairwise coverage and is unranked.

C.5 Hindi question form and answer-language policy details

Table 20: Question form changes which access routes are suppressed. The matched-form advantage is the mean additional answer-NLL increase when training and evaluation both use native Hindi questions or both use Romanized Hindi, compared with crossing the two forms. The contrast is in answer-NLL units; positive values indicate greater suppression with matched forms. Answers remain in native Hindi. The 16 runs cross two model families, two seeds, two methods and two training forms at a fixed 100-update budget. Intervals bootstrap the 20 authors in the 318-fact common assessment panel.

Method	Matched-form advantage (NLL) ↑	95% interval
SimNPO	6.05	[5.83, 6.27]
RMU	8.26	[7.66, 8.80]

D Source selection: full grids and supporting analyses

Supporting material for section 4. Residual R is defined in section 2; the primary score averages native routes of the held-out language set. Retain damage is $p_{\text{FT}} - p_{\text{U}}$ averaged over the six native retain routes on 50 retain facts per language.

D.1 All-language forget supervision

Supervising every acquired language is the naive fix to cross-lingual leakage. We ran it on all nine development parents under the same stopping rule and protocol as the $K \leq 3$ grids, and saved the selected endpoint. All nine met the rule, at 75–125 updates; none was censored at the cap. Table 21 compares it with English-only supervision and with the retain-constrained best triple.

Table 21: **All-language forget supervision removes best but degrades every language's generations.** Forget supervision in K of the six acquired languages, averaged over the nine development parents. Supervising one language leaves the fact recoverable in the others; supervising all six removes it, with more than five times the degenerate output.

	Forget generations		
Forget supervision	Full leakage ↓	Gibberish ↓	Retain damage ↓
English only ($K = 1$)	0.553	0.031	0.129
Best triple ($K = 3$)	0.196	0.129	0.105
All six ($K = 6$)	0.137	0.169	0.130

D.2 Headroom at every budget

Table 22: All 27 family/inventory/budget grids, seed 1. Uniform is the mean over all subsets at that budget. Residual columns report unscaled R (lower is better). Relative reduction is $100(R_{\text{Uniform}} - R_{\text{best}})/R_{\text{Uniform}}$, in percent (higher is better). Gap CI low is the lower endpoint of the pointwise 95% paired-author bootstrap interval for $R_{\text{Uniform}} - R_{\text{best}}$, in unscaled R units, with the empirical best subset held fixed. Bootstrap win rate is the percentage of 2,000 paired-author draws in which that subset remains first.

Inventory	Family	K	Best subset	Worst subset	Best R	Uniform R	Worst R	Relative reduction (%)	Gap CI low	Bootstrap win rate (%)
Script-diverse	Aya	1	EN	JA	0.718	0.793	0.909	9.4%	0.048	87%
		2	JA+RU	JA+ZH	0.492	0.570	0.705	13.6%	0.033	43%
		3	AR+JA+ES	EN+ES+ZH	0.285	0.383	0.505	25.5%	0.073	24%
	Qwen	1	EN	JA	0.715	0.815	0.939	12.3%	0.076	100%
		2	ES+ZH	AR+JA	0.469	0.589	0.730	20.4%	0.092	66%
		3	AR+JA+ES	EN+ES+ZH	0.262	0.366	0.549	28.5%	0.066	31%
	Llama	1	EN	JA	0.773	0.864	0.918	10.6%	0.067	100%
		2	EN+RU	AR+JA	0.541	0.701	0.811	22.8%	0.124	99%
		3	EN+AR+JA	EN+RU+ES	0.410	0.516	0.643	20.5%	0.061	50%
Latin-script	Aya	1	ID	TR	0.441	0.719	0.871	38.7%	0.227	100%
		2	ID+SW	DE+TR	0.258	0.398	0.561	35.1%	0.094	88%
		3	EN+DE+TR	EN+ID+VI	0.114	0.174	0.288	34.3%	0.009	49%
	Qwen	1	EN	TR	0.642	0.778	0.909	17.5%	0.103	100%
		2	DE+ID	ID+VI	0.367	0.479	0.609	23.4%	0.067	50%
		3	EN+SW+TR	ID+SW+VI	0.209	0.305	0.410	31.4%	0.060	64%
	Llama	1	EN	TR	0.746	0.845	0.903	11.8%	0.073	100%
		2	EN+DE	SW+TR	0.396	0.530	0.760	25.3%	0.065	84%
		3	EN+DE+VI	EN+ID+SW	0.093	0.282	0.515	67.0%	0.141	95%
Low-resource	Aya	1	EN	NE	0.746	0.840	0.918	11.2%	0.069	100%
		2	AR+NE	HI+NE	0.500	0.600	0.765	16.7%	0.064	84%
		3	EN+AR+NE	EN+HA+SW	0.256	0.387	0.613	33.8%	0.100	39%
	Qwen	1	EN	HA	0.760	0.871	0.934	12.7%	0.079	100%
		2	EN+HA	HI+NE	0.584	0.728	0.906	19.8%	0.097	93%
		3	AR+NE+SW	HI+NE+SW	0.081	0.363	0.715	77.5%	0.217	65%
	Llama	1	EN	NE	0.737	0.875	0.923	15.8%	0.079	100%
		2	EN+HI	HI+NE	0.545	0.672	0.932	18.9%	0.086	58%
		3	EN+HA+HI	AR+HI+NE	0.309	0.467	0.746	33.8%	0.111	29%

D.3 Cross-request ranking stability

Table 23: **Source-set rankings recur across training orders and across disjoint forget requests** (TOFU, low-resource inventory, $K{=}3$). Cross-request ρ is the Spearman correlation between the rankings of two disjoint forget requests after averaging each subset over its five orders; within-request ρ is the median of the ten pairwise correlations between order-specific rankings; shared top five counts subsets in both requests' top five. Four-to-one selects the lowest-residual subset on four orders of the same request and evaluates it on the held-out fifth (gain in $100R$ over uniform).

Model / request	Ranking agreement			Leave-one-order-out	
	Cross-request ρ	Within-request ρ	Shared top five	Four-to-one gain mean [min, max]	Positive held-out orders
Qwen / A	0.929	0.802	3/5	$+7.11$ [$+0.51$, $+11.97$]	5/5
Qwen / B	0.668	0.760	1/5	$+5.72$ [$+2.69$, $+10.75$]	5/5
Tiny Aya / A	0.923	0.840	4/5	$+10.55$ [$+5.77$, $+13.36$]	5/5
Tiny Aya / B	0.950	0.838	5/5	$+9.84$ [$+7.03$, $+12.88$]	5/5
Llama / A	0.913	0.776	3/5	$+9.24$ [$+4.16$, $+13.72$]	5/5
Llama / B	0.892	0.815	3/5	$+9.26$ [$+5.83$, $+12.30$]	5/5

D.4 Generation behavior of the selected source sets

The behavioral scores in Table 5 are computed on free-form generations rather than on likelihoods. Generations are produced for the native routes of the six inventory languages (query language equal to the requested answer language) at the selected endpoint of each run, for the forget facts of 20 authors and for 50 retain facts per language, and scores are averaged over three training orders; retain scores use six languages for Aya and five (excluding Arabic) for Qwen and Llama. Generations are graded by the judge described in subsection A.6, which also defines full leakage, any disclosure, retained recovery and extraction strength (ES). ES is reported descriptively, without confidence intervals.

Individual top-3 protocol. This comparison uses TOFU and three paired training orders. Individual top-3 selects EN/RU/ES for Aya and EN/ES/ZH for Qwen and Llama; COVER selects JA/ES/ZH for Aya and EN/AR/JA for Qwen and Llama.Unlearning uses SimNPO, learning rate 2×10^{-5}. Relative to Individual top-3, lower full leakage and higher retained recovery hold in all three orders for Llama and two of three for Aya; Qwen's full-leakage differences change sign across orders.

D.5 Optional source weighting

Source set weighting. Here we weight the forget loss update according to cross-language gradient alignment. We measure graidients with the benighn calibration data. Define $g_{i,\ell,b}$ as the gradient of the mean answer-token NLL for sample i in language ℓ with respect to decoder block b, with B_θ decoder blocks. We define

$$C_{s,t} = \frac{1}{n_G B_\theta} \sum_{i=1}^{n_G} \sum_{b=1}^{B_\theta} \max(0, \cos(g_{i,s,b}, g_{i,t,b})) \,.$$

This measures shared loss sensitivity across languages, which we use as a proxy for transfer of forgetting updates. We then choose source weights t0 maximize the worst held out language's weighted coverage:

$$\pi^\star \in \arg\max_{\pi \in \mathcal{W}} \min_{t \in T(\widehat{S})} \sum_{s \in \widehat{S}} \pi_s C_{s,t}.$$

Here $\mathcal{W}$ contains source weights in $[0.20, 0.45]$ that sum to 1. We scale each forget example's SimNPO loss by $K\pi^\star_s$, where s is its language.

Experimental setup. We compare Qwen and Llama on TOFU on three paired training orders, using the same source stopping criterion in section 2. English-heavy weights are $(.45, .275, .275)$ for English, Arabic and Japanese. Our permutation control swaps the calibrated AR/JA weights.

Table 24: **Source-weighting comparisons.** Rates (%) averaged over three orders on TOFU. Retain uses all six languages in (a) and the indicated held-out languages in (b)

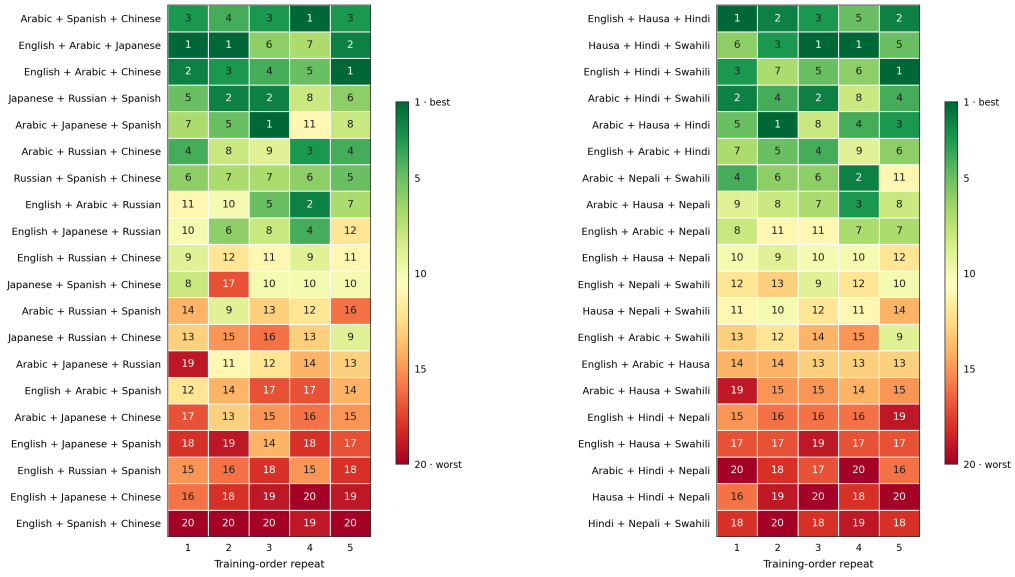
(a) All six native routes

Allocation	Full ↓	Any ↓	Retain ↑	Gibberish ↓	Target-only ↑
Qwen: EN/AR/JA					
Uniform	34.33	57.02	82.78	2.26	87.94
Calibrated	29.61	52.07	83.78	3.91	90.31
English-heavy	31.67	53.94	80.67	3.11	88.24
AR/JA-permuted	30.94	52.74	85.00	2.17	86.72
Llama: EN/AR/JA					
Uniform	31.76	57.13	85.78	5.69	83.09
Calibrated	26.33	47.65	87.33	18.48	67.31
Qwen: EN/AR/ZH					
Uniform	34.44	60.09	80.56	0.63	90.63
Calibrated	25.65	51.65	81.44	0.43	91.91

(b) Own omitted native routes

Allocation	Full ↓	Any ↓	Retain ↑	Gibberish ↓	Target-only ↑
Qwen: EN/AR/JA; targets RU/ES/ZH					
Uniform	41.33	62.85	86.89	1.04	93.63
Calibrated	35.70	58.59	86.89	1.89	94.15
English-heavy	38.85	59.96	84.89	1.41	93.48
AR/JA-permuted	36.70	58.30	88.22	1.26	92.81
Llama: EN/AR/JA; targets RU/ES/ZH					
Uniform	42.59	67.74	90.00	0.48	91.78
Calibrated	37.00	61.44	90.44	6.44	81.33
Qwen: EN/AR/ZH; targets JA/RU/ES					
Uniform	39.00	63.63	81.78	0.07	93.89
Calibrated	29.59	56.63	81.11	0.04	93.19

D.6 Source-set rankings



(a) Qwen, script-diverse inventory

(b) Qwen, low-resource inventory

(c) Llama, script-diverse inventory

(d) Llama, low-resource inventory

Figure 8: **Good and bad source sets recur across training orders.** Each panel ranks the 20 three-language subsets (rows, ordered by their mean rank) by held-out residual access under each of five training orders (columns); 1 is best and 20 worst. The script-diverse inventory is English, Arabic, Japanese, Chinese, Russian and Spanish; the low-resource inventory is English, Arabic, Hausa, Hindi, Nepali and Swahili. The top and bottom of each ranking recur across orders while the middle is less stable, and the low-resource inventory (b, d) is the more stable of the two, in line with the consistency values of Table 3.

i

E Reproducibility

- **Training.**
  - **Models and acquisition.** We use Tiny Aya Global, Qwen3-4B-Instruct-2507, and Llama-3.2-3B-Instruct. TOFU fine-tuning uses learning rate 2×10^{-5}, effective batch 64, and sequence length 512, with five epochs for multilingual checkpoints and six for translated monolingual checkpoints.
  - **Unlearning settings.** We use paged AdamW with 32-bit optimizer states, weight decay 0.01, effective batch 16, and constant earning rates after warmup.
  - **TOFU monolingual benchmark.** Training uses sequence length 512, 20 warmup updates, and a 400-update cap. SimNPO and GradDiff use learning rate 2×10^{-5}; RMU uses 10^{-5}. SimNPO uses $\beta = 3.5$, $\delta = 1$, forget-loss coefficient 0.25, and retain-NLL coefficient 1. GradDiff uses forget/retain coefficients $1/5$. RMU targets decoder block 7 (zero-indexed), with steering norm 2 and equal forget/retain coefficients.
  - **TOFU source selection and weighting.** Both use the same SimNPO loss settings above, sequence length 768, 50 warmup updates, and a 600-update cap. Learning rates are 2×10^{-5} for source selection and 10^{-5} for weighting. Training uses FP32 parameters with BF16 autocast.
- **Endpoint selection.**
  - Stopping uses 100 forget facts per language;
- **Calibration**
  - COVER uses two disjoint panels of 100 retained examples each for TOFU and 60 for LUME.