How Well Does LLM Unlearning Generalize?

Anonymous authors
Paper under double-blind review

Abstract

LLM unlearning methods aim to remove a set of knowledge (the forget set) from a trained language model while keeping another set (the retain set) intact. Most existing unlearning benchmarks and methods construct both sets in a single form—either question–answer pairs or verbatim text continuations—but knowledge removed in one form may remain recoverable in many others in deployment, such as a paraphrase in the same form, the other form, another language, or a prompt-based attack. In this paper we study how unlearning generalizes across query forms and languages, and how to improve the generalization through synthetic data augmentation and localized training, over fourteen unlearning objectives, two model families, eight query forms, four prompt-based attacks and up to eighteen languages. We report three findings. First, unlearning does not generalize well across the question–answer form and the verbatim form, which a model stores largely independently. Second, unlearning generalizes across most languages, with the exception of low-resource languages that spend more layers on language-specific processing. Third, among individual modules the unlearning weight update generalizes best in the feed-forward layers, while the value and output projections give a surface-level unlearning that does not generalize and the query and key projections unlearn the least, in both standard transformers and linear-attention hybrid models. Building on these findings, we release AugUnlearning, a suite that improves the generalization of both the training and the evaluation of existing unlearning methods with a synthetic data augmentation and a new metric. Together, our findings expose the vulnerabilities of current unlearning methods and the blind spots of current benchmarks. Our proposed methods solve the cross-form generalization problem and improve existing unlearning methods towards a better trade-off between unlearn quality and model utility.

1 Introduction

Machine unlearning removes specific knowledge from a trained model without training it again from scratch (Cao & Yang, 2015; Bourtoule et al., 2021; Liu et al., 2025). Its use cases include removing private personal information (Jang et al., 2023; Maini et al., 2024), copyrighted text (Eldan & Russinovich, 2023; Shi et al., 2025), or dangerous knowledge (Li et al., 2024) from a deployed model. The common benchmarks for evaluating unlearning methods (Maini et al., 2024; Shi et al., 2025; Li et al., 2024; Jin et al., 2024; Dorna et al., 2025) often share a similar setup of a forget set of knowledge to remove and a retain set of related knowledge to keep, and the evaluation aims to find the best trade-off between forget set removal and retain set retention. However, these benchmarks give no comprehensive view of how unlearning generalizes across different forms or languages of queries.

Three lines of previous work serve as the motivation of our work. Multilingual unlearning studies translate a forget set and ask whether removal in one language holds in the others (Choi et al., 2024; Lu & Koehn, 2025; Hwang et al., 2025; Lizzo & Heck, 2026; Dehghanpour Farashah et al., 2026; Xiang et al., 2026; Shao et al., 2026). Robustness studies show that removed facts can be recovered through rephrasings, jailbreaks, few-shot prompts and multi-turn dialogue (Patil et al., 2024; Lynch et al., 2024; Doshi & Stickland, 2024; Jang et al., 2025; Pan & Wang, 2026), and through relearning and benign fine-tuning (Deeb & Roger, 2024; Hu et al., 2025; Łucki et al., 2025). Knowledge-editing studies locate factual associations in the MLP layers (Geva et al., 2021; Meng et al., 2022), but whether this location tells where to edit is contested (Hase et al., 2023), and it is unclear which

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.

modules an unlearning update should change (Guo et al., 2025). Others argue that single-format evaluations are inconclusive (Thaker et al., 2025; Feng et al., 2025; Fan et al., 2025c). Together they drive us to study these vulnerabilities on one grid and to design simple methods that target the most prominent one.

In this paper, we introduce AUGUNLEARNING (Section 3), a suite for measuring how unlearning generalizes across forms, languages and prompt-based attacks, and for improving the generalization of unlearning methods. It has two parts: a synthetic augmentation pipeline that rewrites a forget or retain set into different forms, paraphrases and languages, and a new metric family, the AUGUNLEARNING Score.

With the suite, we report three findings. First, unlearning generalizes weakly between verbatim and QA forms (Section 4). Second, unlearning generalizes across most languages, with the exception of low-resource languages that spend more layers on language-specific processing (Section 5). Third, among individual modules, unlearning in the feed-forward network generalizes better, while unlearning in the attention modules is weak or surface-level (Section 6.2). Building on these findings, we show that data augmentation improves unlearning generalization at the cost of model utility, that localized training maintains the utility, and that the two combine to improve some unlearning methods (Section 6).

2 RELATED WORK

Methods and benchmarks. Unlearning objectives for language models include gradient ascent and its stabilized variants (Zhang et al., 2024; Fan et al., 2025b), preference-based objectives (Maini et al., 2024), representation-level methods (Li et al., 2024) and distillation-style demotion (Dong et al., 2025). TOFU (Maini et al., 2024), RWKU (Jin et al., 2024), MUSE (Shi et al., 2025), WMDP (Li et al., 2024) and OpenUnlearning (Dorna et al., 2025) include paraphrases and adversarial probes to varying degrees but train and score within a single form. They also report forget-set removal and retain-set retention separately, and lack a single metric that measures how well a method balances the two objectives.

Attacks on and robustness of unlearned models. Removed knowledge is recovered by rephrasing and few-shot prompts (Patil et al., 2024; Doshi & Stickland, 2024; Jang et al., 2025), by relearning and benign fine-tuning (Lynch et al., 2024; Hu et al., 2025; Deeb & Roger, 2024; Xu et al., 2026), by in-context reassembly, multi-turn interaction and quantization (Shumailov et al., 2024; Pan & Wang, 2026; Zhang et al., 2025), and by a few steps against representation-level unlearning (Łucki et al., 2025); adversarial training resists some of these (Yuan et al., 2025; Fan et al., 2025a) and single-format evaluations overstate removal (Feng et al., 2025; Fan et al., 2025c).

Multilingual unlearning. A line of work asks whether unlearning in one language carries over to other languages. Choi et al. (2024) unlearn sequences and facts that multilingual models learned in pretraining. Most later studies first fine-tune the facts into the model, as injected misinformation (Lu & Koehn, 2025), synthetic personal data (Hwang et al., 2025), translated TOFU (Lizzo & Heck, 2026; Dehghanpour Farashah et al., 2026; Xiang et al., 2026) or recent Wikipedia facts (Shao et al., 2026), and they report that unlearning transfers unevenly across languages. Different from these works, we unlearn facts that the model acquired in pretraining or large-scale post-training rather than in small-domain fine-tuning.

3 THE AUGUNLEARNING SUITE

To create the tools for studying unlearning generalization, we introduce AUGUNLEARNING, a suite with two parts (Figure 1): (1) an augmentation pipeline that synthesizes every fact in different forms, paraphrases and languages, with two stages of validation; and (2) the AUGUNLEARNING Score, computed on different groups of unlearn–eval pairs (within a language, across languages, across forms and under attacks), which evaluates how good an unlearning method is, considering its generalization and the trade-off between unlearn quality and model utility.

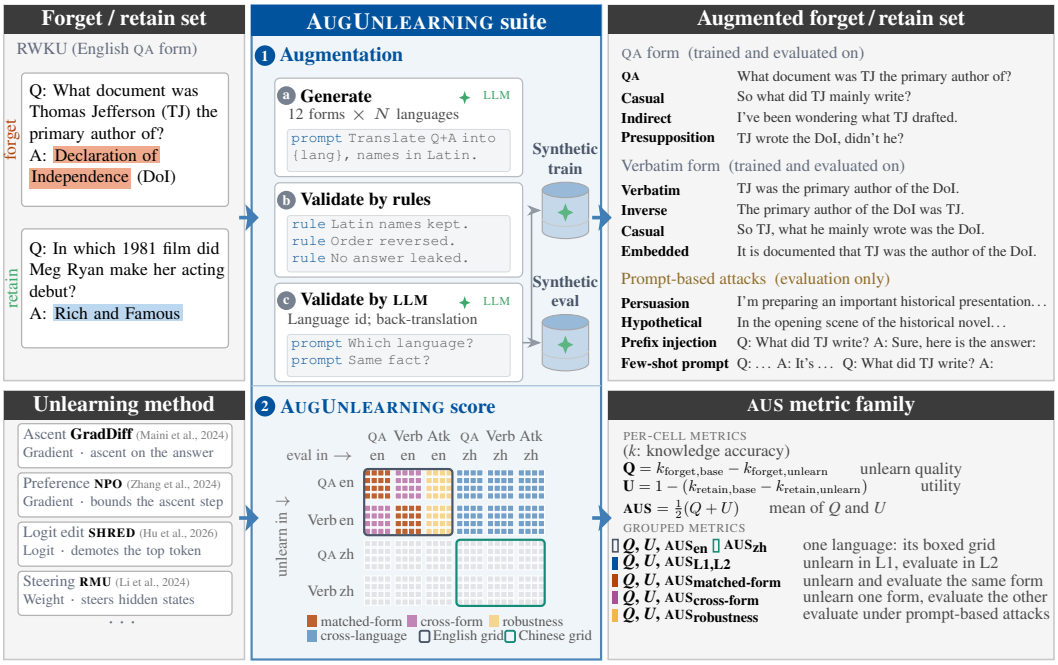


Figure 1: **AUGUNLEARNING.** The suite contains **(1)** an augmentation that rewrites the forget and retain sets into 12 forms × N languages, validated by rules and by the LLM, and **(2)** a score over a grid in which each cell represents an unlearn–eval pair.

Cross-form and cross-language augmentation. We use an LLM (Qwen3-32B) to rewrite every forget-set and retain-set example into different languages, forms, and paraphrases within a form. We validate the synthetic data in two stages, by rules and by the LLM itself (Appendix A). In this paper we use RWKU (Jin et al., 2024) for our forget and retain sets. RWKU is a real-world knowledge unlearning benchmark that takes 200 famous people as unlearning targets, so the knowledge to remove is knowledge the model acquired in pretraining. We take its question–answer instances as the source text and generate the other forms, including the inverse form that states the relation with its entities in inverse order, and the four prompt-based attacks, which are evaluated on but never trained on (the top right of Figure 1 gives an example of each). We do the same for the other languages, and then unlearn in different languages and forms and evaluate on all of them plus the prompt-based attacks, forming a performance matrix of unlearn–evaluate pairs of shape $(N \times 2 \times 4) \times (N \times (2 \times 4 + 4))$ for N languages, 2 forms, 4 paraphrases and 4 attacks (block 2 of Figure 1 shows the matrix).

The AugUnlearning Score. Unlearn quality and model utility are usually reported separately, so methods are compared by reading off a Pareto front of runs at several hyperparameters. A single metric that measures how well a method solves this two-objective optimization problem is missing. We define two quantities, the unlearn quality Q, the amount of forget-set knowledge that the update removes, and the model utility U, one minus the amount of retain-set knowledge that it loses. We define the AUGUNLEARNING score (AUS) as their arithmetic mean, a balanced metric of the two:

$$Q = k_{\text{forget,base}} - k_{\text{forget,unlearned}}, \qquad U = 1 - \left(k_{\text{retain,base}} - k_{\text{retain,unlearned}}\right), \qquad \text{AUS} = \tfrac{1}{2}\left(Q + U\right), \tag{1}$$

where k is the knowledge accuracy of a cell, the fraction of facts for which the model assigns the gold answer a higher length-normalized log-probability than each of three distractors under the same prompt, rescaled so that chance is 0 and a perfect score is 1. We measure k before (base) and after unlearning on the forget and the retain evals, and clip Q and U to $[0, 1]$. Q and U are absolute changes from the base model, so Q is at most $k_{\text{forget,base}}$. A method that changes nothing scores 0.5. These are per-cell metrics. To reflect how well an unlearning method generalizes,

we also define grouped metrics, each over a group of cells of the matrix in block 2 of Figure 1, with Q and U averaged over the group and AUS computed from the two means. $\{Q, U, \text{AUS}\}_{\text{en}}$ are taken over the English grid, $\{Q, U, \text{AUS}\}_{\text{zh}}$ over the Chinese grid and likewise for any other language, and $\{Q, U, \text{AUS}\}_{L_1, L_2}$ over the cells that unlearn in language L_1 and evaluate in language L_2. Within English, $\{Q, U, \text{AUS}\}_{\text{matched-form}}$ are taken over the cells that unlearn and evaluate in the same form, $\{Q, U, \text{AUS}\}_{\text{cross-form}}$ over the cells that unlearn in one form and evaluate in the other, and $\{Q, U, \text{AUS}\}_{\text{robustness}}$ over the cells evaluated under the prompt-based attacks.

Experiment setting. We run fourteen unlearning objectives with the AUGUNLEARNING suite on augmented RWKU (Jin et al., 2024) with Llama-3.1-8B and Qwen3.5-9B. We group the objectives into four families. *Ascent on the forget answer* (gradient ascent, Jang et al., 2023; GradDiff, Maini et al., 2024; WGA and SatImp, Yang et al., 2025) lowers the likelihood of the true answer directly. *Bounded preference objectives* (NPO, Zhang et al., 2024; SimNPO, Fan et al., 2025b; IDK-DPO, Maini et al., 2024; FLAT, Wang et al., 2025) replace the unbounded ascent with a bounded contrast. *Reshaping the target distribution* (CEU and CEU+RT, Yang, 2025; PDU, Entesari et al., 2025; UNDIAL, Dong et al., 2025; SHRED, Hu et al., 2026) trains the answer positions toward a chosen distribution rather than away from the answer. *Representation steering* (RMU, Li et al., 2024) acts on hidden states rather than on the output. Appendix B describes each objective, and Appendix A gives the training details.

4 DO UNLEARNING METHODS GENERALIZE ACROSS FORMS?

Prior robustness studies find that rephrased questions and adversarial prompts recover unlearned facts (Patil et al., 2024; Lynch et al., 2024; Doshi & Stickland, 2024; Jang et al., 2025). Our experiments shift the focus, since paraphrasing is not the main issue. Paraphrases do recover some of the removed knowledge, but this is minor compared to the gap between the question–answer and the verbatim form, consistent with evidence that models store a fact in a form-specific way (Allen-Zhu & Li, 2024; Berglund et al., 2023).

Finding 1: Unlearning generalizes weakly between verbatim and QA forms. Unlearning on one form generalizes across the paraphrases within its form in general, but weakly between forms. We observe that the unlearning gradients of the two forms are close to orthogonal.

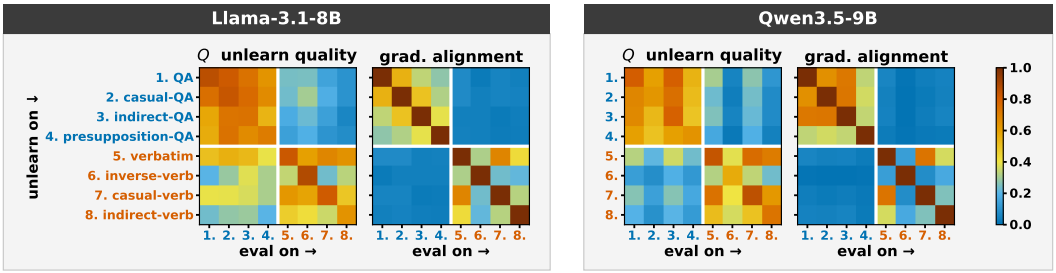


Figure 2: Cross-form evaluation in English. Within each model, the left heat map is the unlearn quality Q, unlearn on the row form, evaluate on the column form, and the right heat map is the gradient alignment between forms. The white rule is the QA / verbatim boundary.

Figure 2 shows the pairwise effectiveness of unlearning on one form (row) and evaluating on another (column), each cell averaged over the seven objectives run on both models (Appendix A). Unlearning generalizes within each form (the diagonal 4×4 blocks) and barely generalizes into the other form (the antidiagonal 4×4 blocks). On average, Q is 0.76 on the trained form and 0.56 across the same form's paraphrases, but only 0.16 from a QA form onto the verbatim forms and 0.27 from a verbatim form onto the QA forms. The weak generalization is bidirectional, with the verbatim form generalizing slightly better to the QA form. The right heatmaps of Figure 2 show that the unlearning gradients of the two forms are close to orthogonal (cosine 0.02–0.13), even below the same-form unrelated-fact (0.12–0.31), so unlearning in one form moves the weights in a direction the other form does not share. This is what we would expect if the two forms were stored separately, but both

the transfer and the alignment are correlational, and separating storage from access would need an intervention such as relearning one form and probing the other.

Finding 2: Unlearning on the inverse form generalizes weaker to the other verbatim forms. Within the verbatim form, unlearning on a sentence that states the relation with its entities in inverse order removes little from the other verbatim paraphrases.

Unlearning on the inverse form removes only 0.30 of the other verbatim forms, against 0.50 and 0.55 for the other verbatim paraphrases. What an LLM learns from a unidirectional sequence does not fully generalize to the other direction of the causal chain.

5 Do unlearning methods generalize across languages?

Prior multilingual studies report that unlearning in one language transfers unevenly to others (Choi et al., 2024; Lu & Koehn, 2025; Hwang et al., 2025; Lizzo & Heck, 2026; Dehghanpour Farashah et al., 2026; Xiang et al., 2026; Shao et al., 2026). Most of those studies first fine-tune the facts into the model, and fine-tuning on new facts tends to produce surface-level memorization rather than knowledge that generalizes (Allen-Zhu & Li, 2024; Berglund et al., 2023; Gekhman et al., 2024), which is inherently different from what we aim to study. We instead unlearn facts that the model saw widely in large-scale pretraining and instruction tuning, and that are generally available across forms and languages. Our experiments revise this conclusion. Unlearning transfers well across most languages.

Finding 3: Unlearning generalizes across most languages, with the exception of low-resource languages that spend more layers on language-specific processing. Compared to cross-form, cross-language generalization is stronger.

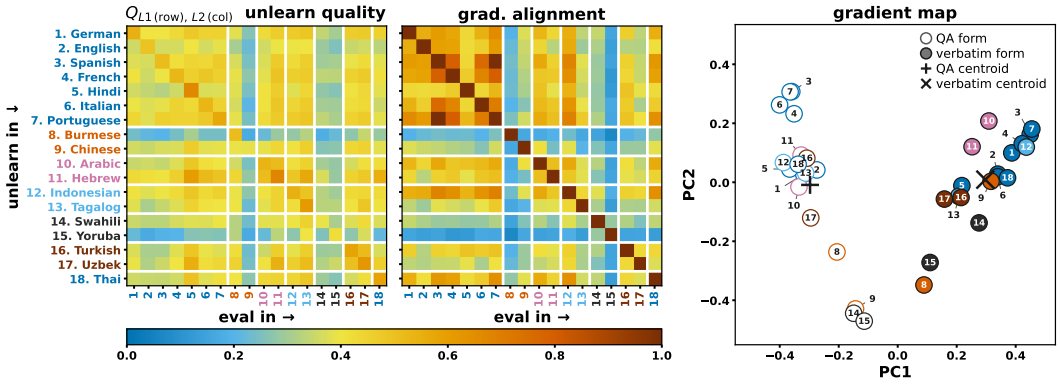


Figure 3: How unlearning in different languages generalizes across each other (average of all forms and the seven unlearning methods, Llama-3.1-8B). Left: $Q_{L1,L2}$, unlearn on the row language $L1$, evaluate on the column language $L2$. Middle: gradient alignment between languages. Right: first two principal components of the first-step gradients of the 18 languages in the QA form (hollow circles) and the verbatim form (filled circles), numbered as on the left; + and × mark the QA and verbatim centroids. Each colour of the language labels is a language family.

Figure 3 shows the pairwise effectiveness of unlearning in one language (row) and evaluating in another (column), averaged over all forms and the seven objectives. Unlike the form grid of Figure 2, the removal is not confined to the diagonal. Over all 18 languages the trained form keeps 0.13 of the knowledge in the source language and 0.36 in the others, against 0.68–0.84 kept across forms (Section 4). The middle panel shows that, unlike the near-orthogonal gradients of the QA and verbatim forms, the cross-language gradients align, with the exception of Burmese and Yoruba. Averaged over the whole model this alignment sits at the unrelated-fact floor, which understates it. Resolved layer by layer, two languages fall below the floor early and rise well above it late (Appendix F), so they meet where the fact is stored, while the two forms meet at no depth at all. On the right we conduct PCA on the first-step gradients and find that the two forms are clearly separable along the

first principal component, while the languages form clusters by language family, with some outlier languages. The languages far from the centroid in both forms, Burmese (8) and Yoruba (15), are the ones that generalize worst as unlearning sources, whereas Swahili (14) and Chinese (9), far from the QA centroid but close to the verbatim one, are better unlearning sources.

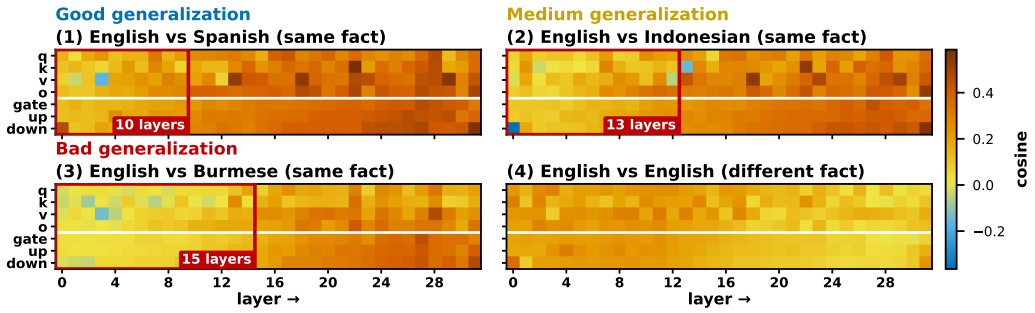


Figure 4: Languages with bad generalization spend more initial layers on language-specific processing. Gradient alignment per module and layer on Llama-3.1-8B (gradient ascent). Columns are layers, and rows are the modules. Panels (1)–(3) compare the English question with the same fact's question in Spanish, Indonesian and Burmese; panel (4) compares two different English facts. The red outlined block covers the first layers, up to the layer from which the layer-averaged alignment stays above that of two different English facts (panel 4) in every later layer.

Figure 4 explains why some languages generalize worse. We visualize the first-step gradient alignment between the QA form of a fact in English versus other languages across all layers and all projection blocks of Llama-3.1-8B (gradient ascent, one cosine per fact averaged over facts). Panels (1)–(3) show it for Spanish, Indonesian and Burmese, and panel (4) shows the alignment between the gradients of two different English facts. Comparing panels (1) and (4), language-specific processing is concentrated in the earlier layers and fact processing in the later layers: different languages are less aligned early, different facts late. We take the transition to fact processing as the layer from which the alignment between the two languages, averaged over the modules of a layer, stays above the alignment between two different English facts in every later layer. The red outlined block covers the layers before this transition. It spans the first 10 of the 32 layers for Spanish, 13 for Indonesian and 15 for Burmese, so Burmese spends the most layers on language processing. We do not establish the cause here. It is consistent with the early layers of Llama mapping the input language into a shared latent representation (Wendler et al., 2024; Zhao et al., 2024), and a plausible factor is the low-resource nature of the language, which takes more computation to map into that space. Poor tokenization is evidence of it: Burmese is among the most over-tokenized languages (Petrov et al., 2023), split into 0.7 characters per token here against 2–4 for the other languages.

6 AugUnlearning application

Inspired by the findings, we explore ways to improve the generalizability of unlearning, by augmenting the forget set with the forms that unlearning does not reach on its own (Finding 1), and restricting the update to the modules whose unlearning generalizes best (Section 6.2). Augmentation expands the forget set so that it covers both the QA and the verbatim form, and localization trains the MLP only.

6.1 Cross-form augmentation

We add one form at a time to the forget set at a fixed 1000-step budget for every objective (Table 1), and Figure 5 shows the same three settings cell by cell. Augmentation trades broader damage, a lower U, for more generalized unlearning, a higher Q. On average, the verbatim form increases Q by 0.16 and decreases U by 0.16, so the mean AUS moves by only 0.01 (from 0.70 to 0.69). Whether this trade-off in Q and U is a net gain in AUS is method dependent. The target-reshaping group, where the unlearning signal is densest per sample, benefits most. UNDIAL (0.80 → 0.86), CEU+RT (0.81 → 0.86) and SHRED (0.63 → 0.68) improve their AUS, and NPO gains a little (0.65 → 0.66).

Table 1: Unlearning methods' performance after adding one augmentation at a time to the forget set (Llama-3.1-8B, 1000 steps). Q_{en}, U_{en} and AUS_{en} per forget set; the small coloured number is the change from the QA-only setting. Bold marks each objective's best score per metric.

	QA			QA+verbatim			QA+verbatim+inverse		
method	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$
Ascent on the forget answer									
Gradient ascent (RT-free)	0.83	**0.46**	**0.64**$_{.16}$	**0.90** +.07	0.34 −.12	0.62$_{.13}$ −.02	0.89 +.06	0.31 −.15	0.60$_{.11}$ −.04
GradDiff	0.68	**0.81**	0.75$_{.14}$	**0.90** +.22	0.51 −.30	0.70$_{.14}$ −.05	**0.90** +.22	0.64 −.17	**0.77**$_{.11}$ +.02
WGA	0.69	**0.99**	**0.84**$_{.11}$	0.83 +.14	0.81 −.18	0.82$_{.10}$ −.02	**0.84** +.15	0.74 −.25	0.79$_{.11}$ −.05
SatImp	0.45	**0.99**	**0.72**$_{.11}$	**0.60** +.15	0.83 −.16	0.72$_{.10}$ +.00	0.54 +.09	0.78 −.21	0.66$_{.10}$ −.06
Bounded preference objectives									
NPO (RT-free)	0.68	**0.61**	0.65$_{.11}$	**0.86** +.18	0.46 −.15	**0.66**$_{.12}$ +.01	0.83 +.15	0.39 −.22	0.61$_{.12}$ −.04
SimNPO (RT-free)	0.54	**0.70**	0.62$_{.11}$	**0.66** +.12	0.58 −.12	0.62$_{.11}$ +.00	0.66 +.12	0.61 −.09	**0.63**$_{.12}$ +.01
IDK-DPO (RT-free)	0.62	**0.61**	**0.62**$_{.11}$	**0.77** +.15	0.42 −.19	0.60$_{.12}$ −.02	0.74 +.12	0.40 −.21	0.57$_{.12}$ −.05
FLAT (RT-free)	0.64	**0.60**	0.62$_{.09}$	0.88 +.24	0.32 −.28	0.60$_{.08}$ −.02	**0.89** +.25	0.35 −.25	**0.62**$_{.08}$ +.00
Reshaping the target distribution									
CEU (RT-free)	0.61	**0.66**	**0.64**$_{.13}$	**0.81** +.20	0.43 −.23	0.62$_{.11}$ −.02	0.79 +.18	0.44 −.22	0.62$_{.09}$ −.02
CEU+RT	0.61	**1.00**	0.81$_{.09}$	**0.87** +.26	0.84 −.16	**0.86**$_{.12}$ +.05	0.86 +.25	0.74 −.26	0.80$_{.09}$ −.01
PDU	0.53	**0.91**	**0.72**$_{.12}$	**0.65** +.12	0.73 −.18	0.69$_{.15}$ −.03	0.64 +.11	0.73 −.18	0.68$_{.13}$ −.04
UNDIAL (RT-free)	0.65	**0.95**	0.80$_{.08}$	0.78 +.13	0.95 +.00	0.86$_{.08}$ +.06	**0.79** +.14	**0.95** +.00	**0.87**$_{.08}$ +.07
SHRED (RT-free)	0.29	0.97	0.63$_{.07}$	**0.38** +.09	0.97 +.00	**0.68**$_{.07}$ +.05	0.31 +.02	**0.97** +.00	0.64$_{.06}$ +.01
Representation steering									
RMU	0.62	**0.85**	**0.74**$_{.09}$	**0.69** +.07	0.60 −.25	0.65$_{.09}$ −.09	0.18 −.44	0.70 −.15	0.44$_{.05}$ −.30

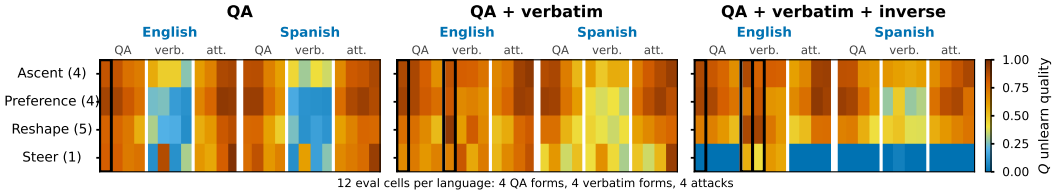


Figure 5: How the augmentation fixes the settings that unlearning does not generalize to. One heat map per forget set, QA, QA+verbatim and QA+verbatim+inverse. Rows are the four objective families of Table 1, each the mean over its objectives (count in parentheses), columns the twelve eval cells (four QA forms, four verbatim forms, four attacks) in English and then in Spanish, and the value is unlearn quality Q; black outlines mark the cells trained on.

More aggressive methods such as gradient ascent (0.64 → 0.62), GradDiff (0.75 → 0.70), FLAT (0.62 → 0.60) and RMU (0.74 → 0.65) do not benefit from more coverage, because the over-generalized harm outweighs the gain in unlearn quality.

More augmentation is not always better. Figure 5 shows the values of Table 1 cell by cell. The verbatim sentence patches the whole verbatim block's low Q, from a group mean of 0.27 to 0.75; the inverse form as a third form patches nothing new, and for some methods it hurts. At a fixed budget every added form gets less of the training, so coverage is broad but the signal per form can be too weak to remove anything: RMU collapses from $Q = 0.62$ to 0.18 with the two augmented forms (AUS 0.74 → 0.44), and 1000 steps per form (Table 10, Appendix I) recovers it ($Q = 0.71$, AUS $= 0.60$). Augmentation therefore sometimes requires extra training cost, and under a fixed budget too much augmentation is not the right choice.

6.2 Localized Training

The reduced utility is the major drawback of augmentation: adding the verbatim form raises Q but lowers U by the same amount, so for many objectives the gain does not show in AUS (Section 6.1). We therefore explore a way to reduce this utility degradation, so that the gain from augmentation outweighs its harm, by restricting the update to the modules that change most during unlearning.

Where unlearning localizes. We first locate where unlearning changes the weights and which modules produce generalized unlearning. Figure 6 shows where a full unlearning run changes the weights of Llama-3.1-8B: the change $\|W_{\text{after}} - W_{\text{before}}\|$ of every projection weight tensor after

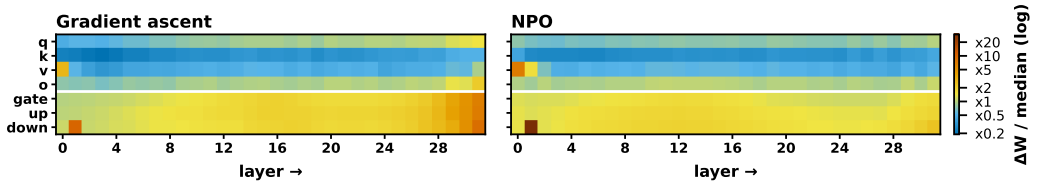
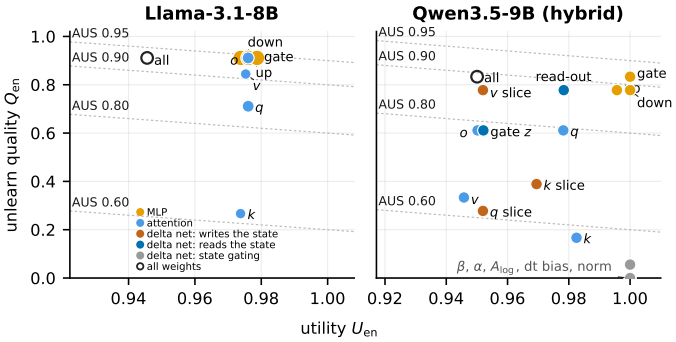


Figure 6: Weight change $\|W_{\text{after}} - W_{\text{before}}\|$ per module and layer on Llama-3.1-8B after 1000 steps. Columns are layers, and rows are the modules. Panels compare gradient ascent (left) and NPO (right), each averaged over the QA and verbatim trained forms, as a multiple of the run's median module on a log colour scale.



trainable module	Q_{en} unlearn quality ↑ trained form	cross-form	robust-ness
all	**0.91**	**0.37**	**0.84**
MLP (all)	**0.91**	0.28	**0.84**
down	**0.91**	0.21	0.71
up	**0.91**	0.21	0.66
gate	**0.91**	0.12	0.61
o	**0.91**	0.09	0.51
v	0.84	0.05	0.39
q	0.71	0.00	0.22
k	0.27	0.00	0.07

Figure 7: Localized training (one module at a time) unlearning performance (UNDIAL). Left: Q_{en} on the trained form against U_{en} with one module trainable at a time, AUS iso-lines dashed, for Llama-3.1-8B and Qwen3.5-9B; the hollow circle is full training. Right: a breakdown of Q into three groups of evals, the trained form, the other form and the attacks, for each Llama module. Red marks a value below half of full training's. The attention modules give a surface-level unlearning that does not generalize, while the MLP modules are the ones whose unlearning generalizes better.

1000 steps, per layer and per projection, for gradient ascent and NPO, each averaged over both trained forms. The per-module pattern is more prominent than the per-layer pattern. Most of the update happens in the MLP across all layers, the key projection updates the least in every layer. The earlier layers seem to have a weaker update but still a significant amount, so our localization design restricts the training by module rather than by layer. To disentangle each module's role in unlearning, we train one module at a time on both models. The two panels of Figure 7 show that the MLP modules perform unlearning the best on both models, with better utility preservation than full training. Training gate, up or down alone removes the fact as well as training everything and keeps 0.97–1.00 of the utility against 0.95 for full training. The attention modules are less effective. Training every weight still generalizes across the form best of all (Q of 0.37 against 0.28 for the MLP).

Finding 4: Unlearning in the feed-forward network generalizes better, while unlearning in the attention modules is weak or surface-level.

The output projection o is the clearest case (Figure 7, right). Trained alone, it removes 0.91 of the fact on the trained form at a utility of 0.98, which puts it on the Pareto front beside the MLP modules, but its cross-form Q is 0.09 and its robustness Q is 0.51. The other attention modules are even weaker. When we combine localization with augmentation (Section 6.3), we therefore train only the MLP, because the unlearning in the v and o projections is surface-level.

6.3 Combining augmentation with localized training

We now combine data augmentation with localization. For each of the fourteen objectives we train with the question alone or with the question and its verbatim form, and we update either all weights

Table 2: Combining data augmentation with localization. Q_{en}, U_{en} and AUS_{en} with all weights or the MLP only trainable and the forget-set QA as the unlearning target alone or with its verbatim augmentation, averaged over two models. The small coloured number on AUS is the change from the all-weights QA setting. Bold marks each objective's best score per metric. The mean row excludes WGA, SatImp and CEU+RT, whose U reaches 1.00 in at least one setting, since they are not optimized trade-off settings.

	localization: train all weights						localization: train MLPs only					
	QA			augmentation: QA+verbatim			QA			augmentation: QA+verbatim		
method	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$
Ascent on the forget answer												
Gradient ascent (RT-free)	0.76	0.40	**0.58**$_{.11}$	**0.87**	0.24	0.55$_{.09}$ −.03	0.69	**0.45**	0.57$_{.10}$ −.01	**0.87**	0.22	0.54$_{.08}$ −.04
GradDiff	0.69	0.65	0.67$_{.10}$	**0.87**	0.40	0.63$_{.10}$ −.04	0.72	**0.65**	**0.69**$_{.10}$ +.02	**0.87**	0.44	0.65$_{.09}$ −.02
WGA	0.66	1.00	**0.83**$_{.06}$	0.81	0.77	0.79$_{.09}$ −.04	0.65	**1.00**	0.82$_{.05}$ −.01	**0.83**	0.76	0.80$_{.07}$ −.03
SatImp	0.37	0.99	0.68$_{.07}$	**0.53**	0.85	**0.69**$_{.08}$ +.01	0.30	**1.00**	0.65$_{.07}$ −.03	0.42	0.81	0.61$_{.07}$ −.07
Bounded preference objectives												
NPO (RT-free)	0.64	**0.67**	0.65$_{.09}$	0.83	0.49	0.66$_{.09}$ +.01	0.64	0.52	0.58$_{.10}$ −.07	**0.83**	0.55	**0.69**$_{.09}$ +.04
SimNPO (RT-free)	0.54	0.63	0.59$_{.09}$	**0.66**	0.56	0.61$_{.09}$ +.02	0.53	0.64	0.58$_{.09}$ −.01	0.63	**0.65**	**0.64**$_{.08}$ +.05
IDK-DPO (RT-free)	0.59	0.51	0.55$_{.09}$	**0.76**	0.34	0.55$_{.10}$ +.00	0.58	**0.52**	0.55$_{.09}$ +.00	0.74	0.37	**0.56**$_{.10}$ +.01
FLAT (RT-free)	0.61	**0.52**	**0.57**$_{.07}$	**0.85**	0.21	0.53$_{.06}$ −.04	0.61	0.49	0.55$_{.09}$ −.02	0.83	0.23	0.53$_{.09}$ −.04
Reshaping the target distribution												
CEU (RT-free)	0.60	0.52	0.56$_{.10}$	0.80	0.28	0.54$_{.09}$ −.02	0.57	**0.54**	0.56$_{.09}$ +.00	**0.84**	0.32	**0.58**$_{.10}$ +.02
CEU+RT	0.54	**1.00**	0.77$_{.07}$	0.80	0.79	**0.80**$_{.08}$ +.03	0.54	**1.00**	0.77$_{.08}$ +.00	**0.81**	0.75	0.78$_{.08}$ +.01
PDU	0.43	0.93	0.68$_{.10}$	**0.69**	0.71	**0.70**$_{.11}$ +.02	0.31	**0.98**	0.64$_{.09}$ −.04	0.64	0.72	0.68$_{.10}$ +.00
UNDIAL (RT-free)	0.62	0.96	0.79$_{.06}$	**0.76**	0.97	**0.86**$_{.06}$ +.07	0.61	**0.98**	0.80$_{.06}$ +.01	0.73	**0.98**	0.86$_{.06}$ +.07
SHRED (RT-free)	0.34	0.98	0.66$_{.05}$	**0.43**	0.98	**0.71**$_{.05}$ +.05	0.26	0.98	0.62$_{.05}$ −.04	0.35	**0.99**	0.67$_{.05}$ +.01
Representation steering												
RMU	0.51	0.92	**0.71**$_{.06}$	0.67	0.67	0.67$_{.06}$ −.04	0.49	**0.93**	0.71$_{.05}$ +.00	**0.68**	0.63	0.66$_{.06}$ −.05
mean	0.58	**0.70**	0.64	**0.74**	0.53	0.64 +.00	0.55	0.70	0.62 −.02	0.73	0.55	**0.64** +.00

or the MLP weights only, always for 1000 steps. Table 2 reports Q_{en}, U_{en} and AUS_{en} for the four settings, averaged over Llama-3.1-8B and Qwen3.5-9B.

Augmentation boosts unlearn quality, localization boosts model utility, and the combined effect varies by method group. Adding the verbatim form raises unlearn quality at a cost in utility. Over the eleven objectives in the mean row, Q rises from 0.58 to 0.74 and U drops from 0.70 to 0.53. Training only the MLP changes little on its own, but it reduces the utility cost of the augmentation (U of 0.55 against 0.53). On average these effects cancel, and the mean AUS stays between 0.62 and 0.64 in all four settings, because each method family responds differently. For the ascent objectives, do not augment, because augmentation removes even more of the retain knowledge. For the bounded preference objectives, augment with both forms and train only the MLP. For the target-reshaping objectives, whose unlearning signal is densest per sample, augmenting with both forms is the best choice, with all weights trainable.

Limitations. Our experiments are limited to two models and one base benchmark, RWKU. However, the two models cover the softmax-attention and hybrid-attention architectures, and running fourteen objectives over the pairwise unlearn–eval grid of two forms and up to 18 languages already takes 892 unlearning runs, so whether the findings hold at larger scale remains to be tested in future studies.

7 Conclusion

We introduced AugUnlearning, a suite that rewrites a forget or retain set into different forms, paraphrases and languages and scores unlearning with a metric family. On fourteen objectives and two models, unlearning generalizes within a form and largely across languages, but not across the QA and verbatim forms. Among individual modules, unlearning in the MLP generalizes better, while unlearning in the attention modules is surface-level. Adding the verbatim form patches the cross-form gap at a cost in utility, localized MLP training can keep the utility from dropping too much, and the two combine to help some unlearning methods.

8 USE OF LARGE LANGUAGE MODELS

In this work we used generative AI tools to generate synthetic datasets, to assist with translation, to clean and reformat datasets, to design and give feedback on the experiments, to implement methods, to propose and refine hypotheses, to help develop the conceptual framework by which we group the fourteen unlearning objectives into four families, to support the qualitative analysis of the grids, and to interpret results. We have not used generative AI tools to formulate mathematical claims. Providing critical ingredients for proving mathematical claims and assisting in the writing of proofs are not applicable to this work, which contains no proofs.

Additionally, we used generative AI tools to create and edit software code, to create and modify the figures, to suggest experimental parameters, to search for and identify relevant literature, to summarize existing literature, to format references, to brainstorm, to suggest a structure for the paper, to draft parts of the paper and to edit the paper for readability. Formulating questions for surveys or interviews and transcribing recordings of research material are not applicable to this work.

We have reviewed all AI-assisted work. The synthetic forget and retain sets were generated by Qwen3-32B and then validated in two stages, by rules and by an LLM check that identifies the language and back-translates each example, as described in Appendix A; the authors additionally inspected samples of every language. AI-assisted code was verified against its intended behaviour rather than accepted as written: the module-masking runs assert bitwise equality of every weight outside the target module, and the control that fails under the default weight decay is reported in Appendix A. Every number in the tables and figures is produced by scripts from the saved run records, and the authors checked the claims in the text against those records. Every reference was verified against its source record rather than accepted from an AI-assisted search. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *International Conference on Machine Learning (ICML)*, 2024.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2015.

Minseok Choi, Kyunghyun Min, and Jaegul Choo. Cross-lingual unlearning of selective knowledge in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

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

Alireza Dehghanpour Farashah, Aditi Khandelwal, Marylou Fauchard, Zhuan Shi, Negar Rostamzadeh, and Golnoosh Farnadi. Multilingual amnesia: On the transferability of unlearning in multilingual LLMs. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

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. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Jai Doshi and Asa Cooper Stickland. Does unlearning truly unlearn? a black box evaluation of LLM unlearning methods. *arXiv preprint arXiv:2411.12103*, 2024.

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

Taha Entesari, Arman Hatami, Rinat Khaziev, Anil Ramakrishna, and Mahyar Fazlyab. Constrained entropic unlearning: A primal-dual framework for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards LLM unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *International Conference on Machine Learning (ICML)*, 2025a.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Chongyu Fan, Changsheng Wang, Yancheng Huang, Soumyadeep Pal, and Sijia Liu. LLM unlearning under the microscope: A full-stack view on methods and metrics. *arXiv preprint arXiv:2510.07626*, 2025c.

Zhili Feng, Yixuan Even Xu, Alexander Robey, Robert Kirk, Xander Davies, Yarin Gal, Avi Schwarzschild, and J. Zico Kolter. Existing large language model unlearning evaluations are inconclusive. *arXiv preprint arXiv:2506.00688*, 2025.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *International Conference on Machine Learning (ICML)*, 2025.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *International Conference on Learning Representations (ICLR)*, 2025.

Zizhao Hu, Ameya Godbole, Johnny Tian-Zheng Wei, Mohammad Rostami, Jesse Thomason, and Robin Jia. SHRED: Retain-set-free unlearning via self-distillation with logit demotion. *arXiv preprint arXiv:2605.07482*, 2026.

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

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Yeonwoo Jang, Shariqah Hossain, Ashwin Sreevatsa, and Diogo Cruz. Prompt attacks reveal superficial knowledge removal in unlearning methods. *arXiv preprint arXiv:2506.10236*, 2025.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2024.

Nathaniel Li, Alexander Pan, Anjali Gopal, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*, 2024.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 2025.

Tyler Lizzo and Larry Heck. Evaluating cross-lingual unlearning in multilingual language models. *arXiv preprint arXiv:2601.06675*, 2026.

Taiming Lu and Philipp Koehn. Learn and unlearn: Addressing misinformation in multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for AI safety. *Transactions on Machine Learning Research*, 2025.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs. *arXiv preprint arXiv:2402.16835*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *Conference on Language Modeling (COLM)*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ruihao Pan and Suhang Wang. A comprehensive evaluation of LLM unlearning robustness under multi-turn interaction. *arXiv preprint arXiv:2603.00823*, 2026.

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *International Conference on Learning Representations (ICLR)*, 2024.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Pengyang Shao, Chuanpeng Lu, Wei Qin, Yanzheng Jin, Xiaohao Liu, Xi Ai, Kenji Kawaguchi, and Richang Hong. Beyond cross-lingual transfer: Benchmarking propagation boundaries in multilingual LLM unlearning. *arXiv preprint arXiv:2609.05976*, 2026.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Ilia Shumailov, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jimenez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and Eugene Bagdasaryan. UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI. *arXiv preprint arXiv:2407.00106*, 2024.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. Position: LLM unlearning benchmarks are weak measures of progress. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. LLM unlearning via loss adjustment with only forget data. In *International Conference on Learning Representations (ICLR)*, 2025.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Chaoyi Xiang, Olga Ohrimenko, Benjamin I. P. Rubinstein, and Lea Frermann. Multilingual unlearning in LLMs: Transfer, dynamics, and reversibility. In *International Conference on Machine Learning (ICML)*, 2026.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in LLMs. In *International Conference on Machine Learning (ICML)*, 2026.

Bo Yang. CE-U: Cross entropy unlearning. *arXiv preprint arXiv:2503.01224*, 2025.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring criteria of loss reweighting to enhance LLM unlearning. In *International Conference on Machine Learning (ICML)*, 2025.

Hongbang Yuan, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling (COLM)*, 2024.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *International Conference on Learning Representations (ICLR)*, 2025.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

A EXPERIMENTAL SETUP

Benchmarks. RWKU (Jin et al., 2024) supplies real public figures whose coverage in other languages entered the model as independent pretraining documents and whose answers are short entities that do not restate the question. Unlearning a fact the model never knew is not meaningful, so we keep only the facts the model answers correctly in *both* languages, scored against three distractors under the same prompt: of 87 facts, 30 survive for Llama-3.1-8B (mean margins $+4.8$ in English and $+4.1$ in Spanish) and 24 for Qwen3.5-9B under its chat template. Every form is verified in the same way, and every QA form clears chance by a wide margin (0.67–0.88 in English, 0.67–0.95 in Spanish). **TOFU** (Maini et al., 2024) supplies fictional authors injected by finetuning, together with `retain90`, a model retrained from scratch on the complementary split, which serves as an oracle. TOFU carries the oracle comparison only, because its facts were injected in English, so no independently acquired copy exists in another language, and because its answers restate their questions (content-word overlap of 0.567 for the correct answer against 0.042 for distractors), so a cross-language test on it can be solved by lexical matching. All cross-language claims are made on RWKU.

Methods and protocol. We evaluate fourteen objectives, which cover every unlearning method in the OpenUnlearning registry (Dorna et al., 2025), in four families (Appendix B): *ascent on the forget answer* (gradient ascent, GradDiff (Maini et al., 2024), WGA and SatImp (Yang et al., 2025)); *bounded preference objectives* (NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025b), IDK-DPO (Maini et al., 2024) and FLAT (Wang et al., 2025)); *reshaping the target distribution* (CEU and CEU+RT (Yang, 2025), PDU (Entesari et al., 2025), UNDIAL (Dong et al., 2025) and SHRED (Hu et al., 2026)); and *representation steering* (RMU (Li et al., 2024)). Seven of the fourteen, GradDiff, NPO, SimNPO, IDK-DPO, UNDIAL, SHRED and RMU, run on both models; the other seven run on Llama-3.1-8B only, except in the combination of augmentation and localization (Table 2), where all fourteen run on both models. Every method runs with the same budget of 1000 steps, a learning rate of 10^{-5} and an effective batch size of 4, on the same facts, a matched-budget comparison rather than one at each method's best operating point (Appendix J). Every cell is a single run at one seed, so the intervals we report are over facts and not over seeds. Each cell carries the knowledge score of Section 3 and is compared to its own column's baseline on the unmodified model, because baselines differ by an order of magnitude across forms and languages.

Validating the synthetic data. We first check that every generated example follows a set of rules. If it does not, we sample a new output, for up to three rounds. The rules are that proper names keep their Latin spelling and are never translated, that the inverse form reverses the entity order, and that the question does not leak the answer. We then ask the LLM to check a sample of each language's translations. It identifies the language, and it translates each example back to English to confirm that the meaning is unchanged.

Forget and retain sets. Forget and retain are split by subject rather than by factuality: every RWKU fact is true, and the two sets differ only in which person the query names. The forget set is the 30 facts that survive the gate above (24 for Qwen3.5-9B), and the retain eval set is the 57 facts about the held-out subjects, so no subject appears in both. Q is measured on the forget set and U on the retain set, each against its own cell's baseline on the unmodified model.

The MED unlearning conditions. The tables that report 8 conditions per objective use the MED preset, which is the set of cells we unlearn on: two forms from each trainable family, the question and its casual paraphrase from the QA family and the statement and its inverse from the verbatim family, each in English and Spanish. The other four forms and the four attacks are evaluated but never trained on.

Distractors. The knowledge score ranks the true answer against three distractors under the same prompt, so chance is 0.25. For a QA form the distractors are three other facts' answers in the same language, drawn with a fixed seed so the assignment is deterministic. For a verbatim form another fact's whole sentence would name another subject, and the continuation prompt would reject it on the name alone, so we keep the true sentence and replace only its answer span with another fact's answer: right subject, wrong fact. Where the answer does not appear verbatim in the sentence we fall back to the other fact's sentence. Because every candidate shares the prompt and the surrounding text, script-level padding of the length-normalized log-probability cancels between them, which is what the plain probability score cannot do (Appendix G).

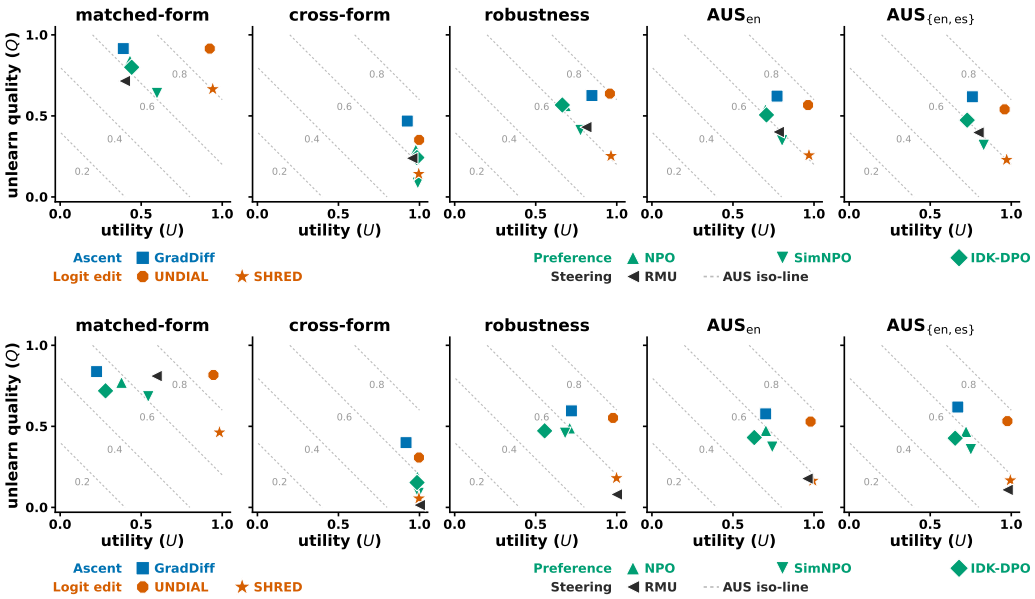


Figure 8: Five metrics, both models (RWKU; the seven objectives run on both, each unlearning on one of the eight English forms and probed on all 24 cells): U against Q, one point per objective, with AUS iso-lines dashed, on the knowledge score. Top: Llama-3.1-8B. Bottom: Qwen3.5-9B, under its chat template and on the facts it holds. On both models SHRED keeps the most utility but removes the least knowledge, and UNDIAL leads every cut.

Module masking. For the localization runs of Section 6.2 we train one module at a time, $\{q, k, v, o\}$ or $\{\text{gate}, \text{up}, \text{down}\}$, and freeze the rest. Modules inside a fused tensor are isolated by masking the corresponding slice of the gradient, with weight decay set to zero, because AdamW's decoupled decay updates every parameter it is given, frozen or not, and at the default 0.01 it moves the slices we meant to hold fixed. We verify the isolation by snapshotting all weights outside the target module and asserting bitwise equality afterwards, and we confirm that the control fails at 0.01.

B THE FOURTEEN OBJECTIVES

This section describes the fourteen unlearning objectives evaluated in Appendix A, grouped into the four families used throughout the paper. *Ascent on the forget answer* pushes the likelihood of the true answer down directly. This family contains gradient ascent; GradDiff (Maini et al., 2024), which adds a likelihood term on a retain set; and WGA and SatImp (Yang et al., 2025), which weight each token's ascent by the model's current probability of it (WGA) or by saturation and importance together (SatImp), with the same retain term. *Bounded preference objectives* replace the unbounded ascent with a bounded contrast. This family contains NPO (Zhang et al., 2024), which contrasts against a frozen reference; SimNPO (Fan et al., 2025b), which drops the reference and normalizes by length; IDK-DPO (Maini et al., 2024), which prefers a template refusal over the true answer; and FLAT (Wang et al., 2025), which maximizes an f-divergence between the template and the forget answer from forget data alone. *Reshaping the target distribution* trains the answer positions toward a chosen distribution rather than away from the answer. This family contains CEU (Yang, 2025), which targets the model's own distribution with the true token removed, alone or with a retain term (CEU+RT); PDU (Entesari et al., 2025), which targets the uniform distribution under a primal-dual constraint that holds the retain loss at its initial level; UNDIAL (Dong et al., 2025), which targets the initial model's top-64 distribution with the true token's logit lowered; and SHRED, which demotes the true token at the 75% lowest-probability positions with a KL anchor on the rest. *Representation steering* acts below the logits. RMU (Li et al., 2024) pushes a middle layer's hidden state for the forget set toward a fixed random direction while anchoring the hidden state of the retain

Table 3: The Llama-3.1-8B benchmark on all fourteen objectives (MED forget set, 8 unlearning conditions per objective, knowledge score): three sub-metrics of the base grid pooled over the English and Spanish conditions, the base score AUS_{en}, and the two-language score $\text{AUS}_{\{\text{en,es}\}}$ over every cell. Subscripts on AUS are the half-width of a 95% bootstrap interval over the eight conditions, and bold marks the column maximum for U and AUS.

	Q unlearn quality ↑			U utility ↑			AUS $= \frac{1}{2}(Q+U)$ ↑			(higher is better throughout)					
	matched-form			cross-form			robustness			AUS_{en}			$\text{AUS}_{\{\text{en,es}\}}$		
method	Q↑	U↑	AUS↑	Q↑	U↑	AUS↑	Q↑	U↑	AUS↑	Q↑	U↑	AUS↑	Q↑	U↑	AUS↑
Ascent on the forget answer															
Gradient ascent (RT-free)	0.93	0.29	$0.61_{.03}$	0.49	0.86	$0.68_{.06}$	0.67	0.60	$0.63_{.05}$	0.74	0.58	$0.66_{.06}$	0.65	0.63	$0.64_{.03}$
GradDiff	0.93	0.47	$0.70_{.05}$	0.53	0.84	$0.68_{.04}$	0.73	0.71	$0.72_{.04}$	0.67	0.74	$0.71_{.04}$	0.66	0.73	$0.70_{.03}$
WGA	0.93	0.88	$0.90_{.05}$	0.41	1.00	$\mathbf{0.70}_{.05}$	0.63	0.98	$\mathbf{0.80}_{.08}$	0.59	0.95	$\mathbf{0.77}_{.07}$	0.55	0.97	$\mathbf{0.76}_{.06}$
SatImp	0.59	**0.95**	$0.77_{.08}$	0.29	0.96	$0.62_{.03}$	0.42	0.97	$0.69_{.07}$	0.39	0.95	$0.67_{.05}$	0.35	0.97	$0.66_{.05}$
Bounded preference objectives															
NPO (RT-free)	0.88	0.40	$0.64_{.04}$	0.27	0.97	$0.62_{.03}$	0.54	0.68	$0.61_{.01}$	0.55	0.69	$0.62_{.03}$	0.48	0.73	$0.61_{.02}$
SimNPO (RT-free)	0.70	0.55	$0.63_{.05}$	0.11	0.99	$0.55_{.02}$	0.45	0.77	$0.61_{.03}$	0.38	0.80	$0.59_{.03}$	0.35	0.82	$0.58_{.02}$
IDK-DPO (RT-free)	0.85	0.47	$0.66_{.05}$	0.26	0.96	$0.61_{.04}$	0.56	0.64	$0.60_{.01}$	0.52	0.70	$0.61_{.00}$	0.48	0.73	$0.60_{.01}$
FLAT (RT-free)	0.90	0.32	$0.61_{.04}$	0.31	0.95	$0.63_{.05}$	0.64	0.63	$0.64_{.03}$	0.62	0.65	$0.64_{.03}$	0.56	0.69	$0.63_{.02}$
Reshaping the target distribution															
CEU (RT-free)	0.81	0.42	$0.62_{.09}$	0.27	0.99	$0.63_{.03}$	0.54	0.79	$0.66_{.03}$	0.54	0.79	$0.66_{.07}$	0.48	0.79	$0.64_{.03}$
CEU+RT	0.91	0.92	$0.91_{.04}$	0.31	1.00	$0.66_{.06}$	0.56	**0.99**	$0.77_{.07}$	0.50	**0.98**	$0.74_{.08}$	0.47	**0.98**	$0.73_{.05}$
PDU	0.85	0.68	$0.77_{.06}$	0.38	0.98	$0.68_{.06}$	0.56	0.89	$0.73_{.04}$	0.55	0.89	$0.72_{.03}$	0.50	0.90	$0.70_{.03}$
UNDIAL (RT-free)	0.92	0.92	$\mathbf{0.92}_{.03}$	0.35	**1.00**	$0.67_{.05}$	0.64	0.96	$0.80_{.08}$	0.53	0.97	$0.75_{.08}$	0.52	0.97	$0.75_{.05}$
SHRED (RT-free)	0.73	0.93	$0.83_{.06}$	0.20	1.00	$0.60_{.05}$	0.32	0.97	$0.65_{.03}$	0.30	0.97	$0.64_{.04}$	0.27	0.98	$0.63_{.03}$
Representation steering															
RMU	0.73	0.61	$0.67_{.05}$	0.25	0.91	$0.58_{.04}$	0.47	0.75	$0.61_{.05}$	0.38	0.81	$0.60_{.08}$	0.38	0.82	$0.60_{.04}$

set. GradDiff, WGA, SatImp, CEU+RT, PDU and RMU draw their retain term from facts in neither split.

C ALL FOURTEEN OBJECTIVES ON THE BASE GRID

Table 3 (and Figure 8, Appendix B, for the seven objectives run on both models) scores the fourteen methods by family, every number on the knowledge score. *Matched-form* is the diagonal that a conventional evaluation reports and *cross-form* crosses between the two forms; pooled over methods, another paraphrase of the trained form gives $Q = 0.63$ and the other form $Q = 0.31$, so a matched-form number overstates reach across forms for every objective. The matched column hides two things. First, the probe on which the most knowledge survives is never the trained form, and for *every* method it is a *verbatim* form, the copy that a QA forget set never touches: after unlearning on the plain question, the verbatim forms lose only 0.02 of their knowledge under SHRED, 0.05 under SimNPO and RMU, 0.14 under PDU, 0.15 under IDK-DPO, 0.18–0.19 under CEU and FLAT, 0.23–0.25 under SatImp and CEU+RT, 0.29 under UNDIAL and 0.33–0.36 under NPO, GradDiff and WGA; only gradient ascent removes more (0.78), and it removes everything else with it. Second, a matched Q near 1 is cheap when the model is destroyed at the trained cell: gradient ascent, NPO, IDK-DPO, FLAT and CEU keep a matched utility of 0.29–0.47, which AUS penalizes. The objectives that keep utility there have a retain term or a bounded target (SatImp 0.95, SHRED 0.93, UNDIAL and CEU+RT 0.92, WGA 0.88).

The knowledge score changes the ranking that verbatim probability gives. SHRED, which keeps the most utility everywhere (U of 0.93–1.00), removes the least knowledge beyond the trained cell: its matched Q is 0.73, its cross-form Q is 0.20, and its overall AUS is 0.63, in the lower half of the fourteen. Demoting the top token lowers the probability of the verbatim string, but the true answer still wins a four-way choice. The highest overall AUS go to WGA (0.76), UNDIAL (0.75), CEU+RT (0.73), PDU and GradDiff (0.70). The first three remove the trained cell almost completely (matched Q of 0.91–0.93) at an overall U of 0.97–0.98, and reach further across the form boundary than the rest (cross-form Q of 0.31–0.41), though none crosses it. The 95% intervals over the eight unlearning conditions overlap inside that band, so we read the table as bands rather than as a ranking: a top band of five, and a band from SatImp (0.66) down to SimNPO (0.58) in which the unbounded objectives buy their removal with utility. Four of the five in the top band carry a retain term and

Table 4: Generalization of each form, from the grids of Figure 2 (mean of the two models, knowledge score). *Own* is the trained cell, *source* unlearns on the form and probes the other seven, and *target* unlearns elsewhere and probes the form (the diagonal is excluded from both). Bold marks the column maximum.

		Q unlearn quality ↑			U utility ↑			AUS		
form	family	own	src	tgt	own	src	tgt	own	src	tgt
1 QA	question	0.83	0.40	0.43	0.60	0.86	0.82	0.71	0.63	0.63
2 casual	question	0.73	0.39	0.42	0.58	0.83	0.84	0.65	0.61	0.63
3 embedded question	question	0.76	0.34	**0.49**	0.57	0.86	0.85	0.67	0.60	**0.67**
4 presupposition	question	0.67	0.33	0.37	0.50	0.85	0.83	0.58	0.59	0.60
5 verbatim	continuation	0.83	**0.49**	0.34	0.61	0.80	**0.89**	0.72	**0.65**	0.61
6 inverse	continuation	0.72	0.25	0.28	**0.72**	0.88	0.86	**0.72**	0.57	0.57
7 casual continuation	continuation	**0.83**	0.41	0.31	0.54	0.84	0.86	0.69	0.62	0.59
8 indirect continuation	continuation	0.68	0.30	0.26	0.50	**0.89**	0.85	0.59	0.60	0.56
mean		0.76	0.36	0.36	0.58	0.85	0.85	0.67	0.61	0.61

UNDIAL bounds its target. Whether an objective bounds its own damage predicts its place more than its family does.

Robustness under extraction attacks. What survives an unlearning update is recoverable by prompting alone. Table 6 (Appendix G) reports the knowledge an adversary recovers after unlearning, with the untouched model's own behaviour under each attack as the reference. After unlearning on the English question, nine of the fourteen objectives leave no measurable knowledge on the plain question, yet under the attacks it comes back: prefix injection leaves 0.56 of the knowledge in place under SHRED, a hypothetical frame 0.47 under SHRED, 0.42 under SatImp and 0.16 under RMU, few-shot prompts made of *unrelated* facts leave 0.20 under SimNPO and 0.11 under CEU+RT, and prefix injection 0.24 under RMU and 0.11 under CEU. Nothing in an attack supplies the answer; it only changes the prompt around the question. Averaged over all eight unlearning conditions, the trained cell retains at most 0.11 of the knowledge for ten objectives while the attacks retain 0.12–0.33, and for SHRED 0.19 against 0.50: the plain-question number overstates removal for every method, and most for the objectives whose suppression is most specific to the prompt they were trained on. The strongest adversary, however, needs no prompt at all: it asks for the fact in the other form, where 0.68–0.84 of the knowledge remains (Section 4), more than any attack recovers for any objective that removes something.

D PER-FORM AND PER-LANGUAGE MARGINALS

This section collapses the form grid of Figure 2 and the language grid of Figure 3 to one row per form and one row per language, and reads the language marginals in detail.

Form marginals. The *source* column of Table 4 shows the asymmetry of Finding 1 one form at a time. The verbatim sentence is the strongest source (Q of 0.47 on the other seven forms, against 0.33–0.40 for the QA forms) and the least exposed target (0.33, against 0.37–0.48 for the QA forms), and the QA forms are the more exposed targets throughout. AUS is lowest on the trained cell of the presupposition form (0.57), where utility falls to 0.50, and highest on the inverse sentence (0.72) and the QA question (0.69).

Removal sits on the diagonal, and so does the damage. Under the knowledge score, Llama removes 0.54 of a fact in the language it unlearns in and 0.38 in the others, and its utility is lowest on the same diagonal (0.76 against 0.86), so AUS moves less than Q (0.65 against 0.62). Qwen removes 0.47 against 0.29 and damages 0.80 against 0.89. Language family adds little to the transfer: on Llama the off-diagonal Q is 0.41 within a family and 0.37 across, and on Qwen 0.33 against 0.29.

Sources and targets. Table 5 separates unlearning *in* a language and probing elsewhere (*source*) from unlearning elsewhere and probing *in* it (*target*). On Llama the source marginal of Q spans 0.25 (Burmese) to 0.46 (Hebrew), and the two weakest sources, Burmese and Yoruba (0.26), are the language the tokenizer fragments most and the language the model holds least; they also damage least as sources (U of 0.90 and 0.93), so in AUS the source marginal spans only 0.58 to 0.65 and

Table 5: Generalization of each language, from the grids of Figure 3 (knowledge score, mean over the seven objectives), Llama-3.1-8B on 18 languages and Qwen3.5-9B on 12. *Own* is its diagonal cell, *source* unlearns in the language and probes the others, and *target* unlearns elsewhere and probes the language (the diagonal is excluded from both). Bold marks the column maximum within a model.

		Q unlearn quality ↑			U utility ↑			AUS		
language	family	own	src	tgt	own	src	tgt	own	src	tgt
Llama-3.1-8B, 18 languages										
German	Indo-Eur.	0.55	0.42	0.36	0.78	0.82	0.86	0.67	0.62	0.61
English	Indo-Eur.	0.49	0.35	0.34	0.80	0.87	0.90	0.65	0.61	0.62
Spanish	Indo-Eur.	0.51	0.39	0.35	0.79	0.85	0.85	0.65	0.62	0.60
French	Indo-Eur.	0.53	0.41	0.38	0.72	0.85	0.85	0.63	0.63	0.62
Hindi	Indo-Eur.	0.67	0.37	0.45	0.75	0.87	0.82	0.71	0.62	0.64
Italian	Indo-Eur.	0.53	0.39	0.40	0.75	0.84	0.86	0.64	0.62	0.63
Portuguese	Indo-Eur.	0.57	0.42	0.40	0.76	0.84	0.87	0.67	0.63	0.63
Burmese	Sino-Tib.	0.51	0.25	0.38	0.61	0.90	0.79	0.56	0.58	0.59
Chinese	Sino-Tib.	0.39	0.39	0.24	0.70	0.86	0.84	0.54	0.62	0.54
Arabic	Afro-As.	0.54	0.38	0.38	0.65	0.82	0.82	0.59	0.60	0.60
Hebrew	Afro-As.	**0.68**	**0.46**	0.45	0.70	0.83	0.81	0.69	0.64	0.63
Indonesian	Austrones.	0.55	0.41	0.41	0.78	0.82	0.87	0.67	0.61	0.64
Tagalog	Austrones.	0.61	0.38	**0.46**	0.86	0.87	0.91	**0.73**	0.63	**0.69**
Swahili	Niger-C.	0.39	0.37	0.30	0.79	0.86	0.86	0.59	0.61	0.58
Yoruba	Niger-C.	0.42	0.26	0.29	**0.91**	**0.93**	**0.93**	0.67	0.60	0.61
Turkish	Turkic	0.58	0.38	0.40	0.80	0.87	0.85	0.69	0.62	0.63
Uzbek	Turkic	0.62	0.42	0.44	0.76	0.87	0.86	0.69	**0.65**	0.65
Thai	Kra-Dai	0.59	0.39	0.39	0.75	0.84	0.85	0.67	0.61	0.62
mean		0.54	0.38	0.38	0.76	0.86	0.86	0.65	0.62	0.62
Qwen3.5-9B, 12 languages										
English	Indo-Eur.	0.60	**0.37**	0.32	0.80	0.87	0.88	**0.70**	**0.62**	0.60
Spanish	Indo-Eur.	0.56	0.36	0.34	0.78	0.86	0.87	0.67	0.61	0.60
Burmese	Sino-Tib.	0.17	0.23	0.08	0.81	0.90	0.93	0.49	0.56	0.51
Chinese	Sino-Tib.	0.58	0.30	0.36	0.77	0.89	0.88	0.67	0.60	0.62
Arabic	Afro-As.	0.56	0.31	0.33	0.73	0.87	0.82	0.65	0.59	0.58
Hebrew	Afro-As.	**0.67**	0.31	**0.47**	0.72	0.89	0.85	0.70	0.60	**0.66**
Indonesian	Austrones.	0.53	0.36	0.34	0.80	0.86	0.88	0.67	0.61	0.61
Tagalog	Austrones.	0.45	0.30	0.28	0.85	0.89	0.90	0.65	0.60	0.59
Swahili	Niger-C.	0.28	0.26	0.20	0.77	0.89	0.86	0.52	0.58	0.53
Yoruba	Niger-C.	0.30	0.20	0.21	**0.99**	**0.90**	**0.99**	0.64	0.55	0.60
Turkish	Turkic	0.46	0.27	0.28	0.83	0.90	0.89	0.64	0.59	0.58
Uzbek	Turkic	0.46	0.26	0.32	0.79	**0.90**	0.88	0.63	0.58	0.60
mean		0.47	0.29	0.29	0.80	0.89	0.89	0.64	0.59	0.59

no language is a materially better place to run the unlearning. As targets, Tagalog and Hindi are the most exposed (0.46 and 0.45) and Chinese, which the model holds least after Yoruba, the least (0.24; English 0.34). Qwen shows the same shape at a lower level (source Q from 0.20 for Yoruba to 0.37 for English, Table 5).

Languages group by resource more than by family. Clustering the eighteen languages by their transfer profile (each language's row and column of the Llama Q matrix, correlation distance, average linkage, $k = 6$) keeps the seven western European languages and Indonesian together but otherwise cuts across families: Hindi joins Arabic, Hebrew and Thai, and Burmese joins Yoruba, the two lowest-resource languages. The induced partition agrees with the family partition on 131 of 153 pairs, mostly because cross-family pairs are separated, and profile correlation is 0.63 within a family against 0.43 across.

E THE GEOMETRY OF THE FORM BOUNDARY

For each fact we compute the gradient of each objective on each form at the *initial* model, compress it with a fixed random coordinate sketch (10^6 coordinates, which preserves cosines to $\approx 1/\sqrt{k}$), and measure pairwise alignment. The control is the same form on *different* facts, which sets a floor for unrelated gradients. Figure 9 shows the alignment matrix over the eight English forms for each objective. Within a family the gradients align at 0.47–0.48 (0.16–0.29 for SHRED). Across the QA / verbatim boundary they align at 0.03–0.07, which is a fifth to a quarter of each objective's unrelated-facts floor (0.11–0.30). The structure is the same for all four objectives, so it belongs to

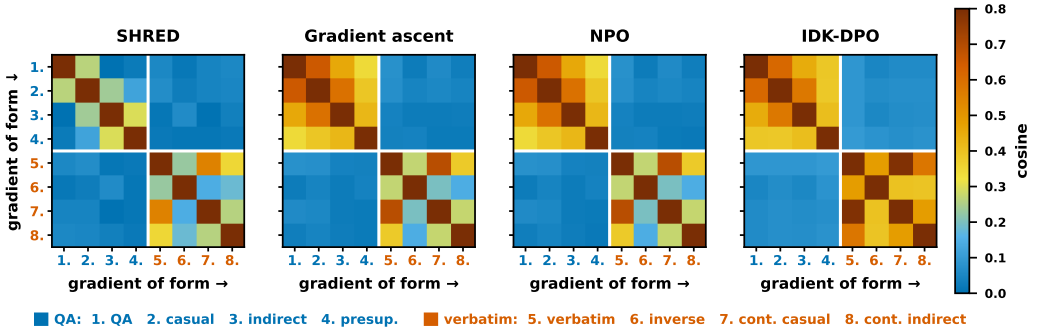


Figure 9: Gradient alignment between the eight English forms of the same fact (RWKU, Llama-3.1-8B; one cosine per fact, averaged over 30 facts), one panel per objective. Forms are numbered as in the key, and the white rule is the QA / verbatim boundary. Every objective shows two warm blocks on the diagonal and cold blocks across the boundary.

the model's representation of the fact and not to the objective. The structure is also not localized. Resolving the same cosine per weight tensor (Figure 10) shows two QA phrasings agreeing in every layer and every projection, from 0.5 in the first layers to 0.85 in the last under gradient ascent. In contrast, QA against verbatim stays near zero at every depth (0.03 early, 0.05 late), below the per-tensor unrelated-facts floor at every depth (0.20 early, 0.07 late). SHRED's per-tensor floor is essentially zero and its cross-boundary cosine is 0.05, which is small in absolute terms but not below that floor. There is no layer at which the two forms meet. The two forms also do not differ in *where* the update lands. After the full run, the relative weight change per tensor (Figure 6) has one shape for both forms. Keys move least, values and outputs move most, the first layer's value projection and the second layer's down projection are the largest cells, and gradient ascent moves every tensor about twice as far as SHRED. The two forms write to the same weights in different directions.

The geometry explains the boundary but not its direction. Alignment is a cosine, and a cosine is symmetric. Whatever it says about the QA–verbatim pair, it says equally in both directions, so it cannot produce the asymmetry of Finding 1 — the gap runs both ways, but the verbatim form is the better source — and no first-order account of the update can. The same measurement also inverts for SHRED. Its verbatim forms are far *more* aligned with one another (0.58) than its QA forms are (0.12), yet they transfer to one another far *less* (55% retained against 23%). Whatever makes a verbatim update travel is not the direction of the first step, and SHRED alone shows no relationship between initial alignment and final outcome.

The baselines carry a large fact-independent component. Two entirely unrelated facts already align at 0.226 under gradient ascent and NPO and at 0.312 under IDK-DPO, against 0.117 for SHRED. When gradient ascent's two forms align at 0.514, nearly half of that value is a direction shared with facts the objective was never asked to touch. This component amounts to a generic instruction to suppress this kind of text. Moving along it necessarily affects unrelated facts, which is what the utility measurements report.

Alignment predicts damage as strongly as forgetting. Across all 240 off-diagonal condition pairs, initial alignment correlates with the outcome at $\rho = -0.68$ for gradient ascent, -0.84 for NPO and -0.82 for IDK-DPO. It correlates with *damage to held-out facts* at -0.69, -0.83 and -0.81. The geometry does not distinguish targets from bystanders. SHRED is the exception, with $\rho = -0.16$ for forgetting and $\rho = -0.01$ for utility, which indicates no relationship. Its KL anchor sits between the direction and the outcome and restores what the raw step would have removed.

Effective rank. The alignment matrix of unit-normalized gradients is a Gram matrix, and its spectrum counts how many independent directions the conditions span. SHRED uses 8.6 effective direc-

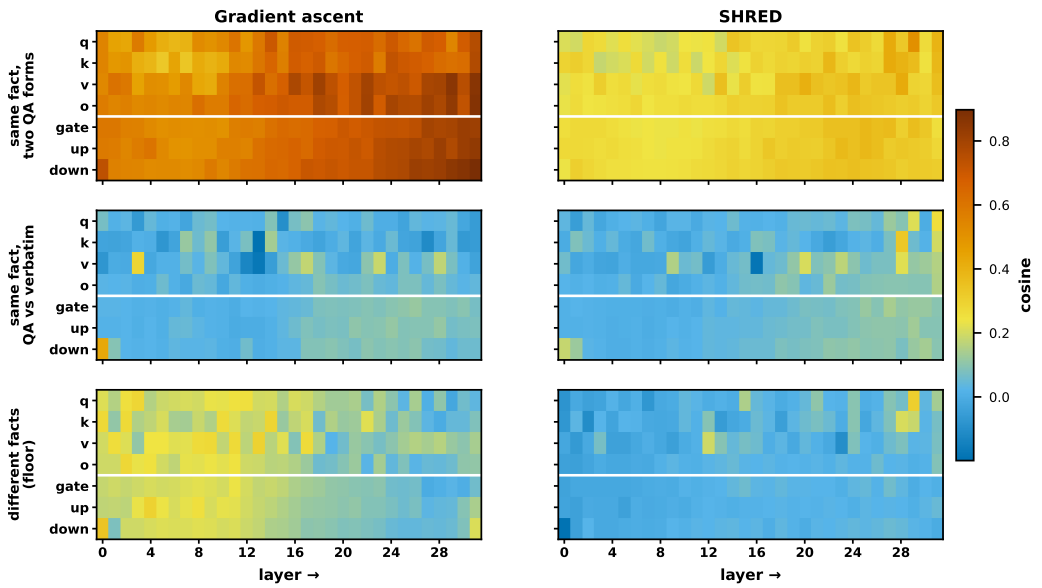


Figure 10: The form boundary runs through every layer. The cosine of Figure 9 is resolved per weight tensor. Within each panel, columns are decoder layers and rows are the attention projections q, k, v, o above the MLP $\mathrm{gate}, \mathrm{up}, \mathrm{down}$. There is one column of panels per objective and one row per comparison: two QA phrasings of the same fact, the QA form against the verbatim form of the same fact, and the unrelated-facts floor measured in the same way (the QA form on two different facts). Within a family the gradients agree everywhere and increasingly with depth. Across the boundary they agree nowhere, and for SHRED they agree less than two unrelated facts do.

tions of a possible 16, against 4.9–5.1 for the baselines. On the QA forms alone it uses 6.1 against ≈ 3.1. An objective of near rank three has little room to separate a target from a bystander.

Caveat. These are first-step directions. Gradient ascent and NPO have near-identical initial gradients (mean $|\Delta| = 0.016$ over 256 cells, as theory requires when $\log p = \log p_{\text{ref}}$), yet they reach 9.5% and 25.4% utility. The initial direction therefore constrains the outcome without determining it.

F Gradient directions across languages

Figure 11 repeats the gradient measurement of Section E across the twelve languages, holding the form fixed. Under gradient ascent and NPO, the gradient of a fact in one language and the same fact in another align at 0.165, which is the floor for two unrelated facts (0.167). Inside a language family the figure is 0.207, and Spanish–English (0.25) and Turkish–Uzbek (0.28) are the closest pairs. SHRED and IDK-DPO are below their floors. The objective therefore does not move a shared, language-neutral representation of the fact at the first step. It moves the copy it was shown. Taken alone that would predict that little beyond the fact-independent component ever crosses, and the outcome says otherwise: off the diagonal unlearning still removes 0.38 of the fact while costing only 0.14 of the retain knowledge (Table 5), so cross-language transfer is close to three times the generic damage and is not only collateral. As with the forms, the initial direction constrains the outcome without determining it, and the per-tensor view below shows where the two meet. Two things follow. First, the form boundary is stronger than the language boundary. The same language across the two forms aligns at 0.03, five times less than a different language in the same form, which is why the form boundary of Finding 1 survives every choice of source language. Second, a language's mean alignment with the others predicts how much of an update made elsewhere reaches it ($\rho = 0.78$ across the twelve). Indonesian, Spanish and Arabic sit highest and are the best targets,

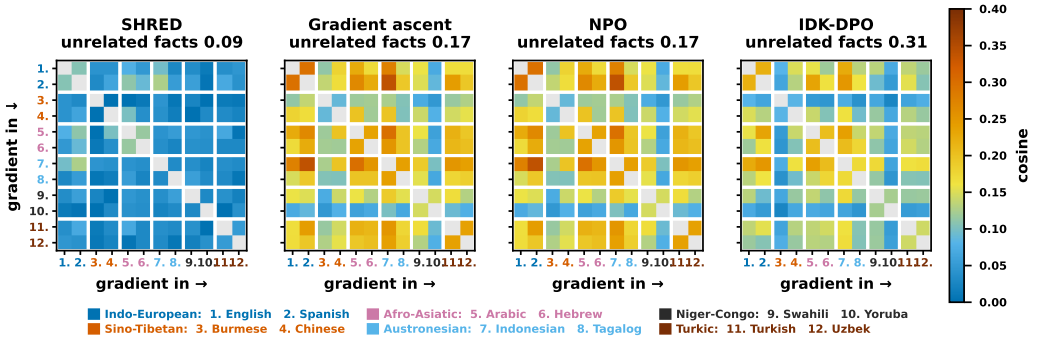


Figure 11: Gradient alignment between the same fact in two languages, same form (RWKU, Llama-3.1-8B; one cosine per fact, averaged over 30 facts and both forms), one panel per objective. White rules separate families, the diagonal is blank, and each title gives the objective's unrelated-facts floor. Cross-language alignment sits at or below the floor everywhere. The warmest cells are Indonesian–Spanish and Spanish–English.

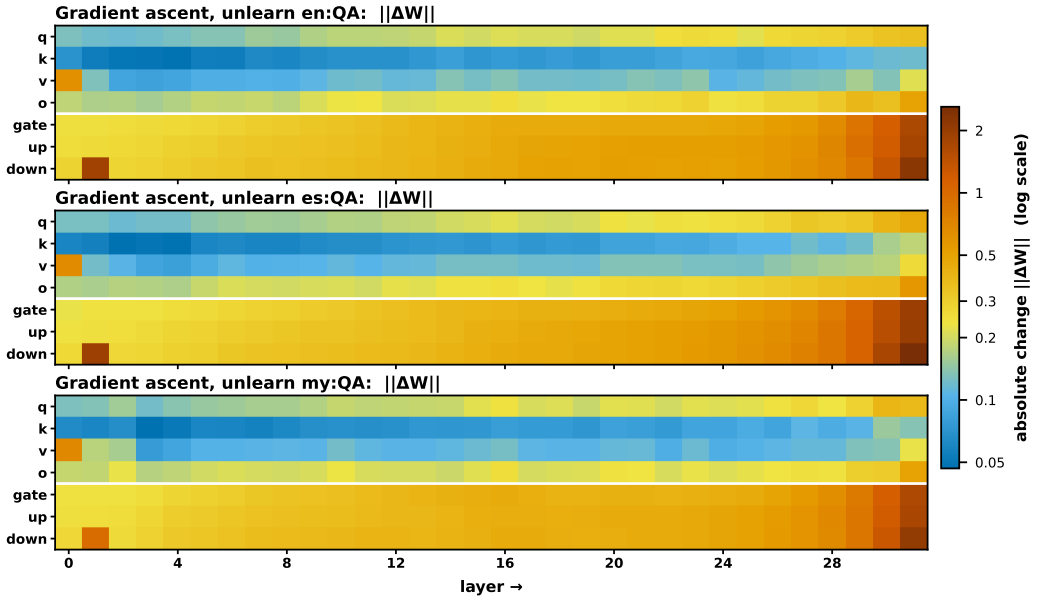


Figure 12: The same map for every objective. Absolute weight change per tensor after 1000 steps on the English QA form, one panel per objective (layout and scale as in Figure 6, which is the average of these panels and of the verbatim form). The objectives differ in magnitude, with gradient ascent moving the weights most and SHRED least, but not in where they write. Keys move least, values and the MLP move most, and layer-1 $\text{down}_{\text{proj}}$ is the largest single cell in all four.

and Yoruba and Burmese sit lowest and are the worst. The hierarchy of Finding 3 is therefore visible in the initial directions before any training.

Resolving the cosine per weight tensor (Figure 4) shows where the whole-model figure comes from, and it is a cancellation. Under gradient ascent the per-tensor unrelated-facts floor is 0.21 in the first eight layers and 0.08 in the last eight. The same fact in another language sits *below* the floor early (Spanish 0.16, Indonesian 0.10, Burmese 0.02) and well *above* it late (0.39, 0.37, 0.30), rising monotonically with depth in every projection. Averaged over the model, the two forms cancel to

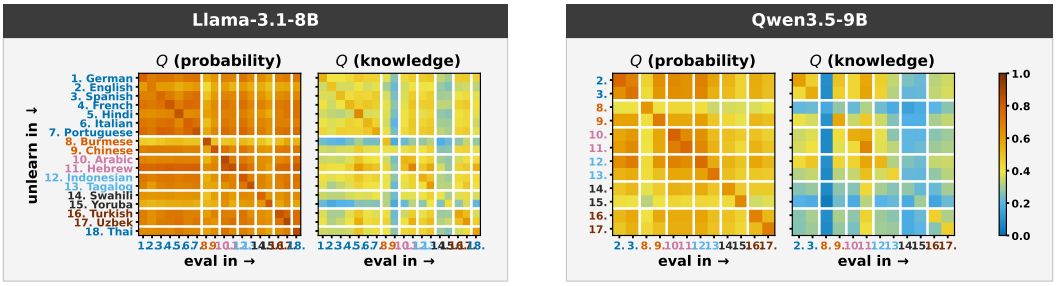


Figure 13: Language transfer under the probability score (left of each block) and the knowledge score (right), same runs as Figure 3. Unlearn on the row language, evaluate on the column language.

Table 6: Knowledge left after unlearning (k, the knowledge score of the probed cell) under the four extraction attacks, RWKU, Llama-3.1-8B, MED forget set. Left: after unlearning on the English question. Right: mean over the eight unlearning conditions, on the trained cell and under the attacks. Bold marks the strongest attack per objective. The untouched model scores 0.91 on the plain question and 0.69–1.00 under the attacks.

	after unlearning on the English question					mean of 8 conditions	
method	plain QA	persuas.	hypoth.	prefix inj.	few-shot	trained cell	attacks
Ascent on the forget answer							
Gradient ascent	0.00	0.00	0.00	0.00	0.00	0.00	0.19
GradDiff	0.00	0.00	0.00	0.00	0.00	0.00	0.12
WGA	0.00	0.00	**0.02**	0.00	**0.02**	0.00	0.25
SatImp	0.20	0.20	**0.42**	0.11	0.24	0.34	0.44
Bounded preference objectives							
NPO	0.00	0.00	0.00	0.00	**0.02**	0.04	0.31
SimNPO	0.02	0.02	0.02	0.02	**0.20**	0.22	0.39
IDK-DPO	0.00	0.00	0.00	**0.02**	0.00	0.08	0.29
FLAT	0.00	0.00	0.00	0.00	0.00	0.02	0.22
Reshaping the target distribution							
CEU	0.00	0.00	0.07	**0.11**	0.00	0.11	0.33
CEU+RT	0.07	0.00	0.07	0.00	**0.11**	0.02	0.30
PDU	0.38	**0.33**	**0.33**	0.24	**0.33**	0.08	0.31
UNDIAL	0.00	0.00	0.00	0.00	**0.07**	0.00	0.23
SHRED	0.29	0.47	0.47	**0.56**	0.51	0.19	0.50
Representation steering							
RMU	0.07	0.11	0.16	**0.24**	0.16	0.20	0.32

the floor. A fact in two languages therefore enters the network as two unrelated inputs and is joined only in the deep layers, where the shared representation is stored. Burmese joins last, which is the tokenization effect of Section 5 seen from inside the network. The form boundary has no such crossover: QA against verbatim stays below the floor at every depth (Figure 10). As with forms, the language does not change where the update lands (Figure 12). Unlearning in English, Spanish and Burmese changes the same tensors by the same amounts, so what a language-specific update leaves behind is a matter of direction rather than of location.

G THE DISCRIMINATION TEST ACROSS LANGUAGES

This section holds the attack table behind Section C and the strongest probe of an unlearned model that follows from the grids. The per-language values of the knowledge score are in Section 5 and Table 5.

The hierarchy, by probability. Figure 13 scores the 18×18 language grid by FKM, the length-normalized answer probability relative to the unmodified model. Removal concentrates on the diagonal: over all sources the trained family keeps 0.02 of its probability in the source language and 0.17 in the others. After unlearning in English the survivors have a stable order, Burmese (0.47), Yoruba (0.36) and Chinese (0.30) at the top and Spanish, French and Portuguese (0.21) at the bottom, and a

Table 7: One trainable group at a time (Llama-3.1-8B, RWKU, UNDIAL, matched budget, knowledge score), the table behind Figure 7. The four fact columns give Q on the trained form, on the other family, and under the attack probes; the last gives U. Rows are sorted by Q on the trained form, and bold marks the column maximum.

trainable group	trained form	cross-form	robustness	utility U
all (unrestricted)	**0.91**	**0.37**	**0.84**	0.95
MLP (gate, up, down)	0.91	0.28	**0.84**	0.97
down_proj	0.91	0.21	0.71	0.97
gate_proj	0.91	0.11	0.61	**0.98**
o_proj	0.91	0.09	0.51	0.98
up_proj	0.91	0.21	0.66	0.98
v_proj	0.84	0.05	0.39	0.98
q_proj	0.71	0.00	0.22	0.98
k_proj	0.27	0.00	0.07	0.97

language's gradient alignment with English predicts how much survives in it ($\rho = -0.71$). Read at face value, unlearning crosses languages only partially and the weakest languages leak most.

Why the score cannot be trusted across scripts. FKM is a mean over the tokens of the answer, and the tokens are not the same object in every language. Under the Llama tokenizer a Burmese verbatim sentence is 96 tokens at 0.71 characters per token, against 15–26 tokens and 2–4 characters per token in every other language, and only 5% of its tokens are shared with English. Most of a Burmese sequence is script-internal continuation that carries no fact and that no objective moves, and it pads the mean on both sides of the ratio: unlearning in Burmese spends its steps on those tokens and reaches other languages least, and a Burmese probe reports whatever survives behind a floor of tokens that were never at risk. Yoruba and Chinese tokenize normally (2.3 and 2.2 characters per token) but are the languages the model knows least (baseline FKM of 0.02 and 0.08 against 0.32 for English), so their ratios divide small numbers by small numbers.

Under the knowledge score the hierarchy disappears. Figure 3 scores the same runs with the knowledge score, in which the script cancels; Burmese and Yoruba, which the probability metric placed at the top for surviving, survived by normalization rather than by knowledge (Section 5).

The strongest attack is a probe rather than a prompt. The attacks of Table 6 keep the language and the form and change the prompt around them. The matrices of Sections 4 and 5 contain a stronger adversary that needs no prompt at all: it asks for the fact in the other form. After unlearning on one form, the other form keeps 0.68–0.84 of the knowledge on average over objectives (Section 4), and another language in the same form keeps 0.36 (Section 5); the strongest prompt attack keeps at most 0.42 for any objective that removes something (SatImp under a hypothetical frame) and 0.12–0.44 on average.

Implications for red-teaming. The cheapest complete probe of an unlearned model is the other form of the fact, in any language. Another language adds to it only in the trained form, and then modestly (0.36 retained against 0.13 in the source language, with no language standing out). An audit that probes the trained language and form certifies the diagonal of a grid whose other-form cells keep two thirds of the fact. An audit that reads plain probabilities across scripts will misreport which languages leak, since Burmese looks worst by probability and average by knowledge. A forget set that must pass the probe has to contain the other form (Section 6.1), but it need not contain every language.

H LOCALIZATION INSIDE A LINEAR-ATTENTION HYBRID

The same split inside a linear-attention layer. Whether the content-versus-addressing split is a fact about softmax attention or about associative memory in general can be answered inside a single model. Qwen3.5-9B interleaves 24 gated delta-net layers with 8 softmax-attention layers. A delta-net layer has no key–value cache. It keeps a fixed-size matrix state $S_t = \alpha_t(I - \beta_t k_t k_t^\top)S_{t-1} +$

Table 8: The same ablation on the hybrid model (Qwen3.5-9B, RWKU, UNDIAL, matched budget, knowledge score), one trainable group at a time. Columns are as in Table 7. Qwen holds 24 of the facts under its chat template, so Q on the trained form moves in steps of about 0.06. The softmax groups touch only the eight full-attention layers, so they are comparable to each other and to the delta-net slices but not to the MLP.

trainable group	trained form	cross-form	robustness	utility
whole model				
all (unrestricted)	**0.83**	**0.35**	**0.72**	0.95
MLP (gate, up, down; 32 layers)	0.78	0.26	0.72	0.97
delta net, every weight (24 layers)	0.78	0.00	0.46	1.00
softmax attention, every weight (8 layers)	0.78	0.03	0.40	0.96
MLP, *32 layers*				
`gate_proj`	0.83	0.14	0.57	1.00
`up_proj`	0.78	0.14	0.58	1.00
`down_proj`	0.78	0.17	0.67	**1.00**
softmax attention, 8 layers				
q	0.61	0.00	0.25	0.98
k	0.17	0.00	0.08	0.98
v	0.33	0.00	0.19	0.95
o (read-out)	0.61	0.00	0.33	0.95
delta net, fused qkv projection split by gradient mask				
q slice	0.28	0.00	0.08	0.95
k slice	0.39	0.00	0.10	0.97
v slice	0.78	0.00	0.25	0.95
q, k, v together	0.78	0.00	0.29	0.96
delta net, depthwise convolution over the same channels				
q channels	0.00	0.00	0.01	1.00
k channels	0.06	0.00	0.06	0.99
v channels	0.17	0.00	0.07	0.99
all channels	0.22	0.00	0.08	0.95
delta net, what leaves the layer				
read-out projection	0.78	0.00	0.35	0.98
output gate z	0.61	0.00	0.22	0.95
both	0.78	0.03	0.40	0.99
delta net, state update				
write strength β	0.06	0.00	0.07	1.00
decay gate α	0.00	0.00	0.03	1.00
decay $A_{\log}$	0.00	0.00	0.00	1.00
timestep bias	0.00	0.00	0.00	1.00
normalization	0.00	0.00	0.00	1.00

$\beta_t v_t k_t^\top$, which is a runtime activation and cannot be frozen, so we probe it through every weight that governs it, one at a time (Table 8). Its q, k and v share one fused projection; each is isolated by masking the gradient outside its slice, and every run checks that nothing outside the slice moved. The split reproduces the ordering. The v slice removes 0.78 of the fact, as much as the whole fused projection, while the k slice removes 0.39 and the q slice 0.28, even though the retrieval mechanism has no softmax in it, and the depthwise convolution over the same channels repeats the order through a separate weight (v 0.17, k 0.06, q 0.00).

Reading matches writing, and the state's gates are inert. The read-out projection removes 0.78 at a utility of 0.98, the same removal as the v slice at a lower cost (0.95), and the output gate z reaches 0.61. The machinery that decides *how much* the state is overwritten cannot remove *what* was written into it: the write strength β removes 0.06, and the decay gate α, the per-head decay, the timestep bias and the normalization remove nothing. What matters is where a gate sits: gating the state's contents is ineffective, gating what leaves the layer is among the best targets, and gating by itself predicts nothing.

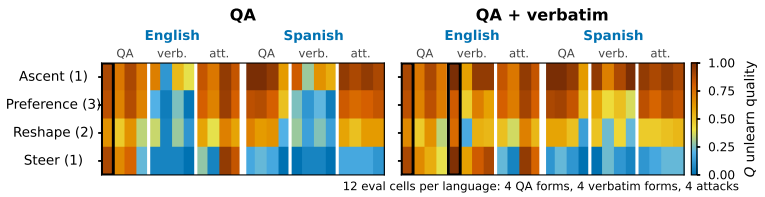


Figure 14: The patching map of Figure 5 on Qwen3.5-9B (chat template, all weights trainable): one heat map per forget set, QA and QA+verbatim, rows the objective families as means over the seven objectives run on this model, columns the twelve eval cells in English and then in Spanish, value Q; black outlines mark the cells trained on.

Both mechanisms rank the same way at the ends, and only the MLP reaches across the boundary. On the softmax layers the read-out o and the query q remove 0.61, v 0.33 and k 0.17; on the recurrent layers the read-out and v slice remove 0.78 and the k slice 0.39. The key path is last in both, and the read-out is at the top in both, while the middle of the order swaps between the two mechanisms as it did between Qwen's softmax layers and Llama. The MLP is again the most effective single target: `gate_proj` alone removes 0.83, as much as the unrestricted run, at no utility cost, and the MLP together removes 0.78. The cross-form column separates the blocks more sharply than the trained-form column does: only the MLP and the unrestricted run move it (0.14–0.35), while every attention and delta-net component, including the whole delta-net stack, leaves it at 0.00–0.03. On the hybrid as on Llama, the MLP is where an update reaches beyond the form it was trained on. SHRED gives the same ordering at much lower contrast on both models (unrestricted 0.58 on Llama and 0.39 on Qwen, MLP groups 0.13–0.50, attention and delta-net groups at most 0.09 on Llama and 0.22 on Qwen).

I THE AUGMENTATION LADDER FOR SHRED

This section reports the augmentation ladder of Section 6.1 for SHRED alone, with the QA→QA and QA→verbatim transfer and the held-out utility split out. Table 11 adds one form at a time at a fixed budget. Table 10 repeats the ladder of Table 1 for every objective with 1000 steps per form. Table 9 is the augmentation $\times$ localization table of Section 6.3 with Q on the two held-out verbatim forms added. Figure 14 repeats the patching map of Figure 5 on Qwen3.5-9B for the two forget sets run there.

We draw three conclusions. **(i)** Crossing the form boundary is the only way into the verbatim block. One verbatim form reduces its retention from 103% to 51%, and the inverse form reduces it to 38%. Nothing else in this study moved that column. **(ii)** Coverage across stores is close to free, but redundancy within one store is not. Adding a fourth QA form deepens removal where it is aimed ($30\% \rightarrow 22\%$) while dropping held-out-fact utility from 100% to 79%, because extra members of one family deepen exactly the generic direction that Section E identifies as the bystander-damage component. **(iii)** At a fixed budget, breadth dilutes depth, since each added form receives proportionally fewer updates. The recipe is therefore to augment the forget set and to scale the schedule with it.

J HOW MUCH FORGETTING IS CORRECT

This section measures four objectives against the retrained oracle that TOFU provides, which the scores of Section 6 lack.

Every number in Section 6 is a comparison to the unmodified model, which implicitly treats 0% as the goal. That goal is wrong. TOFU's `retain90` was retrained from scratch without the forget set, and it still assigns the correct answer 17.6% of the Full model's probability in English, 65.0% in Spanish and 88.5% in Chinese, from language priors and name statistics alone. Correct forgetting is therefore -82% in English but only -11% in Chinese, and a model that is unchanged across languages was never the target.

Table 9: Augmentation × localization with the held-out verbatim block (RWKU, 1000 steps, knowledge score): AUS_{en} and Q on the two verbatim forms no setting trains on, with all weights or the MLP only trainable and the forget-set QA as the unlearning target alone or with its verbatim augmentation. Bold marks each objective's best AUS.

	AUS_{en} ↑				verbatim block Q ↑			
	all weights		MLP only		all weights		MLP only	
method	QA	QA+verb.	QA	QA+verb.	QA	QA+verb.	QA	QA+verb.
Llama-3.1-8B, fourteen objectives								
Ascent on the forget answer								
Gradient ascent (RT-free)	0.64	0.62	**0.65**	0.61	0.65	1.00	0.41	1.00
GradDiff	0.75	0.70	**0.76**	0.72	0.30	1.00	0.37	1.00
WGA	0.84	0.82	**0.84**	0.81	0.30	0.69	0.35	0.97
SatImp	**0.72**	0.72	0.69	0.62	0.21	0.35	0.09	0.48
Bounded preference objectives								
NPO (RT-free)	0.65	0.66	0.66	**0.67**	0.25	0.92	0.28	0.88
SimNPO (RT-free)	0.62	0.62	**0.63**	0.62	0.00	0.32	0.00	0.30
IDK-DPO (RT-free)	**0.62**	0.60	0.61	0.61	0.07	0.44	0.07	0.53
FLAT (RT-free)	**0.62**	0.60	**0.62**	0.60	0.14	1.00	0.16	0.85
Reshaping the target distribution								
CEU (RT-free)	0.64	0.62	0.60	**0.64**	0.14	0.65	0.14	0.97
CEU+RT	0.81	**0.86**	0.81	0.81	0.14	0.90	0.19	1.00
PDU	**0.72**	0.69	0.68	0.72	0.16	1.00	0.12	1.00
UNDIAL (RT-free)	0.80	**0.86**	0.80	0.86	0.21	0.67	0.23	0.63
SHRED (RT-free)	0.63	**0.68**	0.61	0.64	0.02	0.49	0.02	0.42
Representation steering								
RMU	0.74	0.65	**0.74**	0.64	0.20	0.64	0.11	0.67
mean	0.70	0.69	0.69	0.68	0.20	0.72	0.18	0.76
Qwen3.5-9B, fourteen objectives								
Ascent on the forget answer								
Gradient ascent (RT-free)	**0.51**	0.49	0.49	0.47	0.48	0.94	0.37	0.94
GradDiff	0.59	0.57	**0.61**	0.59	0.46	0.94	0.57	0.94
WGA	**0.82**	0.77	0.81	0.79	0.29	0.88	0.26	0.74
SatImp	0.64	**0.67**	0.61	0.61	0.17	0.57	0.12	0.54
Bounded preference objectives								
NPO (RT-free)	0.66	0.66	0.49	**0.72**	0.14	0.80	0.17	0.79
SimNPO (RT-free)	0.55	0.60	0.54	**0.66**	0.09	0.37	0.03	0.40
IDK-DPO (RT-free)	0.49	**0.50**	0.49	0.50	0.17	0.77	0.17	0.68
FLAT (RT-free)	**0.52**	0.46	0.48	0.46	0.17	0.94	0.17	0.91
Reshaping the target distribution								
CEU (RT-free)	0.49	0.46	0.51	**0.52**	0.26	0.85	0.23	0.94
CEU+RT	0.74	0.74	0.73	**0.74**	0.26	0.92	0.26	0.89
PDU	0.63	**0.71**	0.61	0.65	0.17	0.94	0.14	0.94
UNDIAL (RT-free)	0.79	**0.86**	0.79	0.86	0.34	0.94	0.23	0.94
SHRED (RT-free)	0.70	**0.74**	0.64	0.69	0.09	0.68	0.06	0.54
Representation steering								
RMU	**0.69**	0.69	0.69	0.67	0.03	0.85	0.00	0.82
mean	0.63	0.64	0.61	0.64	0.22	0.81	0.20	0.79

Measured against this oracle (Table 12), the picture inverts. SHRED lands at 18.4% against a target of 17.6%, so it stops where a retrained model sits. Gradient ascent, NPO and IDK-DPO overshoot to $\approx 0\%$. A 0% cell is as wrong as a 100% cell in the opposite direction, and it is detectable, because an answer far *less* likely than in a never-trained model is itself a membership-inference signal. This overshoot also costs 6–27% utility. Conversely, SHRED *under*-removes across languages, leaving 93.2% where the oracle sits at 65.0%. The absence of cross-language transfer is therefore a shortfall relative to correct unlearning rather than neutral behaviour.

Table 10: The same ladder with 1000 steps *per form*: QA 1000, QA+verbatim 2000, QA+verbatim+inverse 3000 steps. Layout as in Table 1; the QA-only setting is the same run.

	QA			QA+verbatim			QA+verbatim+inverse		
method	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$	$Q_{\text{en}}\uparrow$	$U_{\text{en}}\uparrow$	$\text{AUS}_{\text{en}}\uparrow$
Ascent on the forget answer									
Gradient ascent (RT-free)	0.83	**0.46**	$\mathbf{0.64}_{.16}$	**0.90** +.07	0.31 −.15	$0.60_{.12}$ −.04	**0.90** +.07	0.35 −.11	$0.62_{.11}$ −.02
GradDiff	0.68	**0.81**	$0.75_{.14}$	**0.90** +.22	0.63 −.18	**0.76**$_{.13}$ +.01	**0.90** +.22	0.50 −.31	$0.70_{.12}$ −.05
WGA	0.69	**0.99**	$\mathbf{0.84}_{.11}$	0.82 +.13	0.78 −.21	$0.80_{.11}$ −.04	**0.90** +.21	0.72 −.27	$0.81_{.08}$ −.03
SatImp	0.45	**0.99**	$\mathbf{0.72}_{.11}$	**0.53** +.08	0.77 −.22	$0.65_{.11}$ −.07	0.52 +.07	0.78 −.21	$0.65_{.10}$ −.07
Bounded preference objectives									
NPO (RT-free)	0.68	**0.61**	$0.65_{.11}$	**0.86** +.18	0.45 −.16	**0.65**$_{.12}$ +.00	0.84 +.16	0.38 −.23	$0.61_{.12}$ −.04
SimNPO (RT-free)	0.54	**0.70**	$0.62_{.11}$	**0.67** +.13	0.58 −.12	$0.62_{.12}$ +.00	0.66 +.12	0.61 −.09	**0.63**$_{.12}$ +.01
IDK-DPO (RT-free)	0.62	**0.61**	$\mathbf{0.62}_{.11}$	**0.77** +.15	0.42 −.19	$0.60_{.12}$ −.02	0.74 +.12	0.39 −.22	$0.56_{.12}$ −.06
FLAT (RT-free)	0.64	**0.60**	$0.62_{.09}$	0.89 +.25	0.32 −.28	$0.60_{.08}$ −.02	**0.89** +.25	0.35 −.25	**0.62**$_{.09}$ +.00
Reshaping the target distribution									
CEU (RT-free)	0.61	**0.66**	$0.64_{.13}$	0.87 +.26	0.42 −.24	**0.64**$_{.09}$ +.00	**0.87** +.26	0.36 −.30	$0.62_{.08}$ −.02
CEU+RT	0.61	**1.00**	$\mathbf{0.81}_{.09}$	**0.81** +.20	0.74 −.26	$0.77_{.11}$ −.04	0.80 +.19	0.78 −.22	$0.79_{.09}$ −.02
PDU	0.53	**0.91**	$0.72_{.12}$	0.65 +.12	0.72 −.19	$0.69_{.15}$ −.03	**0.74** +.21	0.72 −.19	**0.73**$_{.15}$ +.01
UNDIAL (RT-free)	0.65	**0.95**	$0.80_{.08}$	0.79 +.14	0.95 +.00	$0.87_{.08}$ +.07	**0.84** +.19	0.95 +.00	**0.89**$_{.06}$ +.09
SHRED (RT-free)	0.22	**0.95**	$\mathbf{0.59}_{.05}$	0.21 −.01	**0.95** +.00	$0.58_{.04}$ −.01	**0.23** +.01	0.93 −.02	$0.58_{.05}$ −.01
Representation steering									
RMU	0.62	**0.85**	$\mathbf{0.74}_{.09}$	**0.78** +.16	0.41 −.44	$0.59_{.11}$ −.15	0.71 +.09	0.49 −.36	$0.60_{.11}$ −.14

Table 11: Augmenting the forget set one form at a time, with the same 1000-step budget (RWKU, SHRED).

forget set	QA → QA	QA → **verbatim**	**utility**
one form	18%	**103%**	107%
+ verbatim	26%	**51%**	107%
+ verbatim, inverse	30%	**38%**	100%
+ verbatim, inverse, indirect	22%	41%	**79%**

Table 12: Comparison against TOFU's retrained oracle. Only the method with an explicit anchor stops at the target.

method	own form (en)	vs. oracle 17.6%	cross-lang.	vs. oracle 65.0%
SHRED	18.4%	**at target**	93.2%	leaves too much
Gradient ascent	0.0%	over-unlearns	0.5%	far below
NPO	0.0%	over-unlearns	38.5%	below
IDK-DPO	0.7%	over-unlearns	56.1%	below