LANGUAGE-ALIGNED TRAINING FOR TRANSFERABLE ERASURE IN MULTILINGUAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are nowadays multilingual. This means they can process and generate text in many languages, and the same fact can be represented across languages. While this behavior is desired, it also makes it harder to *unlearn* data, such as outdated, private, or harmful facts. This is because even when a fact is unlearned in one language, this fact can often still be accessed through prompts in other languages. Prior work either characterizes this gap or addresses it *during unlearning* itself, by adapting the unlearning objective per language, restricting updates to language-agnostic layers, or projecting out shared knowledge subspaces. These approaches often require forget data in several languages and still yield non-uniform forget and retain accuracies across languages. In this work, we observe that English-only unlearning transfers more strongly to languages whose prompt representations are better aligned with their English counterparts, as measured by cross-language retrieval accuracy. Based on this observation, we propose LATTE (Language-Aligned Training for Transferable Erasure), a method that aligns the representations of the same facts across different languages already during fine-tuning. Afterwards, any standard unlearning method can be applied without modification, using only single-language forget data. Our evaluation on Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct across up to ten languages shows that LATTE learns multilingual facts as well as standard fine-tuning. We additionally show that ours can be combined with any state-of-the-art single-language unlearning method and improves cross-lingual forgetting while maintaining retain performance, both in the unlearning language and in the others. Thereby, LATTE makes a significant step towards reliable unlearning in multilingual state-of-the-art models.

1 INTRODUCTION

Multilingual language models can recall factual knowledge through prompts in different languages (BigScience Workshop, 2022; Meta, 2024; Qwen et al., 2025). This becomes a problem when a fact has to be removed, for example because it is outdated, can enable harmful use (N. Li et al., 2024), or contains personal information that an individual requests to delete (European Parliament and Council of the European Union, 2016). Machine *unlearning* aims to remove the influence of selected training data while preserving the model's overall performance (Bourtoule et al., 2021). However, unlearning has mainly been studied in English, and it has been shown that removing a fact in one language can leave it accessible in others (Choi et al., 2024; Dehghanpour Farashah et al., 2026; Lizzo & Heck, 2026; Xiang et al., 2026).

Figure 1 illustrates this problem on Qwen2.5-3B-Instruct after learning the same facts in ten languages. Following English-only unlearning with Negative Preference Optimization (NPO) (Zhang et al., 2024), the standard fine-tuned model has a significantly reduced score on the data to be forgotten in English (upper panel), *i.e.,* forgetting was successful in English, but the score stays high in other languages, *e.g.,* Chinese, Japanese, and Korean, meaning that the facts can still be retrieved in these lanugages.

Prior work either analyzes uneven cross-lingual unlearning transfer (Dehghanpour Farashah et al., 2026; Lizzo & Heck, 2026; Xiang et al., 2026) or addresses it *at unlearning time*. Existing approaches use multilingual forget examples together with language-adaptive retention objectives (Choi et al., 2024), restrict updates to layers with shared cross-language representations (T. Li et al., 2026), or remove task subspaces from model weights (Lizzo & Heck, 2025). Despite these advances, achieving

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.

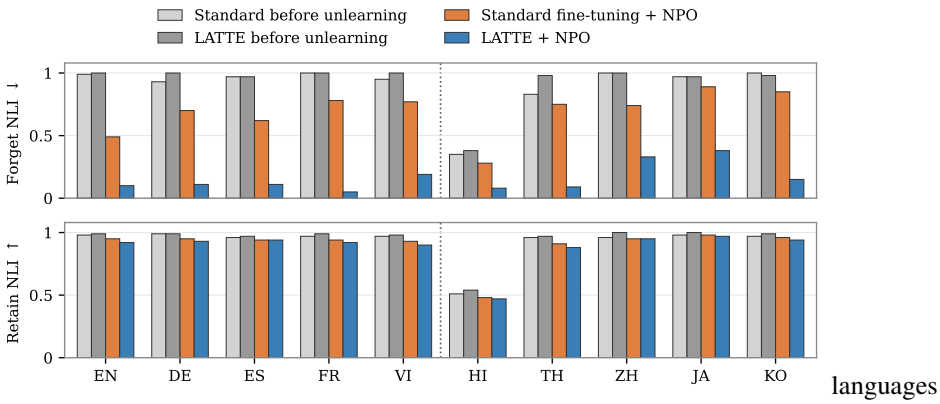


Figure 1: **LATTE improves cross-lingual forgetting with English-only NPO on Qwen2.5-3B-Instruct.** Panels show forget NLI (top, ↓) and retain NLI (bottom, ↑). Light and dark gray bars show standard fine-tuning and LATTE before unlearning, respectively; orange and blue show their corresponding scores afterward. Alignment targets appear to the right of the dotted line.

effective forgetting across languages while preserving retained knowledge remains a challenge. Moreover, using multilingual forget data requires examples of the knowledge to be unlearned in several languages, while layer- and subspace-based interventions modify how unlearning is performed. Both is often not practical, which motivates an approach that improves cross-lingual transfer while retaining an existing single-language unlearning procedure. To achieve this, we intervene earlier, namely during multilingual fine-tuning, with the aim of making subsequently learned knowledge easier to unlearn across languages using examples in only one language.

Therefore, we first make the observation that the transferability of unlearning in one language to another correlates positively with the alignment of the two languages in the model internals. Motivated by this observation, we propose LATTE (Language-Aligned Training for Transferable Erasure), a multilingual fine-tuning method that aligns the representations of different languages already during the fine-tuning process. LATTE brings the representations of prompts in all other languages closer to those of an *anchor language*. For example English, in which future unlearning is to be carried out. This unlearning can then be done with any state-of-the-art single-language unlearning method.

We evaluate LATTE on the multilingual TOFU (Maini et al., 2024) and LUME (Ramakrishna et al., 2025) benchmarks, using Llama-3.2-3B-Instruct with six languages and Qwen2.5-3B-Instruct with ten languages. Our evaluation includes five state-of-the-art unlearning methods, namely NPO (Zhang et al., 2024), Gradient Difference (GradDiff) (Maini et al., 2024), SimNPO (Fan et al., 2025), Representation Misdirection for Unlearning (RMU) (N. Li et al., 2024), and LINGTEA (Choi et al., 2024), using English forget and retain examples. Across multilingual TOFU and LUME, LATTE improves cross-lingual forgetting with multiple unlearning methods. On Qwen2.5-3B-Instruct, Korean forget NLI on TOFU decreases from 0.85 with standard fine-tuning to 0.15 with LATTE under NPO. With GradDiff on LUME, Chinese forget NLI decreases from 0.75 to 0.50, while retain NLI increases from 0.63 to 0.77.

Our contributions are summarized below.

- We identify a positive correlation between cross-language prompt representation alignment and unlearning transfer. For example, in the standard fine-tuned Qwen model, languages with stronger prompt matching tend to exhibit stronger transfer of English-only unlearning.
- We introduce LATTE, which aligns parallel prompt representations during multilingual fine-tuning to improve subsequent cross-lingual unlearning. Unlearning on the resulting model can be performed with standard unlearning methods, without changing their objectives or requiring translated forget examples.
- We evaluate LATTE across two model families, up to ten languages, and five state-of-the-art unlearning methods. Our results demonstrate improvements in cross-lingual forgetting, including lower forget scores across all nine non-English languages on Qwen under English-only unlearning.

2 Related Work

Multilingual language models. Multilingual pretraining allows a single model to learn from text in different languages and transfer capabilities between them. Prior work on multilingual language encoders showed that task fine-tuning in one language can support predictions in others without parallel supervision (Pires et al., 2019), and that scaling multilingual pretraining improves cross-lingual performance (Conneau et al., 2020).

Modern decoder models combine multilingual pretraining with instruction tuning and preference optimization. The Llama-3.2 and Qwen2.5 families include the instruction-tuned backbones used in our experiments and support multilingual generation and instruction following (Meta, 2024; Qwen et al., 2025). Their multilingual behavior also reflects partially shared internal representations. Studies identify grammatical features shared across languages (Brinkmann et al., 2025) and representations that favor a common language before producing outputs in different languages (Harrasse et al., 2025). In Llama-2, intermediate predictions can favor English even when the input and expected output are in another language (Wendler et al., 2024). At the level of factual knowledge, studies also find differences between languages. Factual recall varies across languages (Kassner et al., 2021), and increasing model size can improve factual accuracy without improving cross-language consistency (Qi et al., 2023). These findings motivate examining how multilingual factual representations affect subsequent unlearning.

Machine unlearning. Machine unlearning seeks to remove the influence of selected training data while preserving useful behavior on the remaining data. Exact unlearning requires the resulting model to follow the same distribution as a model trained without the selected data. Retraining provides a direct way to achieve this, but its computational cost motivates approaches that organize training into separate components and retrain only those affected by an unlearning request (Bourtoule et al., 2021). Approximate unlearning aims to approach the outcome of retraining at lower computational cost, without requiring this exact equivalence. For LLMs, **GradDiff**, as implemented in TOFU, combines gradient ascent on forget examples with gradient descent on retain examples (Maini et al., 2024). **NPO** uses a preference-based objective to reduce the likelihood of forget examples relative to a reference model (Zhang et al., 2024), while **SimNPO** uses a reference-free objective with sequence-length normalization (Fan et al., 2025). **RMU** changes hidden representations for forget examples while constraining changes to representations of retain examples (N. Li et al., 2024). We describe these methods in detail in Appendix A.3.LATTE intervenes during the preceding fine-tuning stage and can be combined with these unlearning methods without changing their objectives.

Cross-lingual unlearning. Cross-lingual unlearning aims to unlearn selected knowledge across languages while preserving retained knowledge. Existing studies report uneven transfer associated with language families, scripts and syntactic distance (Choi et al., 2024; Dehghanpour Farashah et al., 2026; Xiang et al., 2026). Several methods address this at unlearning time. **LINGTEA** uses multilingual forget examples and preserves retain predictions through confidence-weighted teacher distillation (Choi et al., 2024). **MUTE** restricts unlearning updates to intermediate layers with shared cross-language representations (T. Li et al., 2026). **UNLEARN** separates weight components associated with forget and retain knowledge and subtracts the forget components (Lizzo & Heck, 2025), with improved cross-language forgetting reported in its multilingual evaluation (Lizzo & Heck, 2026). These approaches require multilingual forget examples, model-specific layer selection, or identifying and editing weight components for each forget task. In contrast, our LATTE introduces alignment during fine-tuning, before the forget set is specified, so that subsequent unlearning can use examples in one language without changing the existing unlearning procedure.

Cross-lingual alignment during fine-tuning. Prior work improves cross-language transfer through multilingual training and representation alignment. For knowledge editing, Xu et al. (2023) show that monolingual editors have limited cross-language transfer. Their approach combines editor training on parallel data with language-specific masks that guide parameter updates, improving the propagation of edits across languages. During task fine-tuning, Liu & Niehues (2025) alternate task updates with contrastive alignment of middle-layer representations and report improved cross-language transfer on slot filling, translation, and structured text generation. We adapt this training-time alignment approach to improve subsequent cross-lingual unlearning, aligning parallel prompts while the model learns the corresponding facts. We use the analysis of Xie et al. (2022) in our evaluation to verify that LATTE improves cross-lingual representation alignment.

3 UNDERSTANDING BASELINE FAILURES IN CROSS-LINGUAL UNLEARNING

In this section, we analyze why prior unlearning methods fail to transfer reliably to other languages. To do so, we fine-tune Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct on multilingual versions of TOFU (Maini et al., 2024). We then apply various state-of-the-art unlearning methods using English forget and retain examples and analyze the forgetting and retention performance of the resulting models. We measure both using natural language inference (NLI) scores that compare generated responses with their target responses. Lower forget NLI indicates stronger forgetting, while higher retain NLI indicates better retention. Our full experimental setup is detailed in Section 5 and Appendix A.3. Our results in Section A.4 confirm the observations from prior work (Choi et al., 2024; Dehghanpour Farashah et al., 2026; Xiang et al., 2026) that cross-lingual unlearning success varies widely. For example, under NPO on Qwen, forget NLI remains at 0.760 in Japanese and 0.808 in Korean, compared with 0.423 in English. Complete baseline results are presented in Appendix A.4.

Table 1: **Stronger cross-language alignment is associated with stronger unlearning transfer.** Results use the standard fine-tuned Qwen model, with alignment measured by English-to-target prompt retrieval accuracy. Languages are ordered by decreasing retrieval accuracy.

	Latin script				Non-Latin scripts					
	ES	VI	DE	FR	JA	ZH	KO	TH	HI	Corr.
Retrieval accuracy	0.988	0.916	0.906	0.895	0.888	0.844	0.756	0.502	0.266	
NPO transfer T_ℓ	0.738	0.884	0.825	0.632	0.397	0.407	0.293	0.599	0.415	0.68
GradDiff transfer T_ℓ	0.942	0.867	0.956	0.656	0.434	0.496	0.344	0.558	0.441	0.73

To understand why these baselines fail, we analyze the relationship between cross-language representation alignment and unlearning transfer. We use the standard fine-tuned Qwen model before unlearning and test whether an English prompt's representation identifies its translation among 500 candidate prompts in another language, drawn from the fine-tuning dataset. Following prior work showing that middle-layer representation alignment can improve cross-lingual transfer (Liu & Niehues, 2025), we use layer 18 and represent each prompt by the mean of its token hidden states. We select the candidate with the highest cosine similarity. Top-1 retrieval accuracy measures how often the selected prompt is the correct translation.

We compare the retrieval accuracy against the unlearning transfer which we measure as the proportional decrease in forget NLI in each language ℓ relative to that in English under the same baseline. For forget NLI scores F_ℓ^{before} and F_ℓ^{after}, which denote the NLI performance over the forget data on the fine-tuned model before and after unlearning, we define

$$U_\ell = \frac{F_\ell^{\text{before}} - F_\ell^{\text{after}}}{F_\ell^{\text{before}}}, \tag{1}$$

and measure the transfer score T_ℓ as $T_\ell = \frac{U_\ell}{U_{\text{en}}}$. A transfer score of 1 indicates the same proportional reduction as in English, while a score of 0.5 indicates half that reduction.

We present our results for NPO and GradDiff as representative unlearning methods in Table 1. These results show that there is a correlation between the alignment, measured in retrieval accuracy, and the transfer of unlearning. Spearman's rank correlation is 0.68 for NPO and 0.73 for GradDiff. In particular, we observe that the languages with the lowest retrieval accuracy (mainly non-Latin script languages, like Korean, Thai, and Hindi) experience the worst unlearning transferability. These observations motivate our LATTE method which aligns the representations of languages already during fine-tuning to facilitate future unlearning transferability.

4 LANGUAGE ALIGNMENT DURING FINE-TUNING

In this section, we present LATTE, our method designed to improve cross-lingual alignment already during fine-tuning to improve the transfer of future unlearning between languages.

4.1 SETUP AND OVERVIEW

Each training example consists of a prompt x and a target response y. We refer to prompts that express the same meaning in different languages as *parallel pairs*. We call the language in which the future unlearning takes place, *i.e.,* the language that provides the forget data, as the *anchor language*, denoted by s. For example, with English as the anchor language, a parallel pair would be:

	English	**Hindi**
Prompt	*Where was Mira Vale born?*	मीरा वेल का जन्म कहाँ हुआ था?
Response	*Berlin, Germany.*	बर्लिन, जर्मनी।

On a high level, LATTE interleaves the learning of multilingual facts with the training of adapters to align these facts' representations across languages. (1) For **learning the facts** in all training languages, we train the base model with the standard causal language modeling objective (Ouyang et al., 2022; Muennighoff et al., 2023). Since this objective rewards correct target predictions but does not explicitly encourage parallel prompts to have similar representations, we add and separately (2) **train multilingual adapters** that pull the representation of each translated prompt towards its anchor-language counterpart. Concretely, we add low-rank adapters (LoRA) (Hu et al., 2022) to the base-model layers and compare representations at a middle layer, where multilingual models were shown to represent content largely independently of the language (Wendler et al., 2024; Liu & Niehues, 2025). We keep the adapters active (but frozen) during the fact learning iterations when the base-model parameters are updated. This trains the model to predict the correct facts also under the updated representations it produces. During the training iterations for the adapters, we keep the base-model weights frozen. At the end of training, all adapters are merged into the base model. This enables normal inference without additional adapter overhead and allows existing state-of-the-art single-language unlearning methods to operate directly on the final model using only samples from the anchor language as forget data. The alternating training procedure is summarized in Algorithm 1 in Appendix A.2.3. Next, we detail how we train the multilingual adapters.

4.2 TRAINING LORA ADAPTERS FOR MULTILINGUAL UNLEARNING

Overall, our adapter training combines two different learning objectives, namely **aligning parallel prompts** and **preserving the learned knowledge** of the base model.

Aligning Parallel Prompts. To align parallel prompts, we need to be able to measure the proximity of two prompts, even when these prompts are of different format, for example, have a different length. Therefore, we represent each prompt in a single vector by the mean of its token hidden states at a selected middle layer. We denote the resulting representation of a prompt x by $h_a(x)$ when the adapter is enabled and by $h(x)$ when it is disabled. Our alignment loss term $\mathcal{L}_{align}$ then is designed to pull the representation of a prompt in a target language x^t closer to the one in the anchor language x^s by normalizing both representations to unit length and minimize their squared distance as shown in Equation 2.

The normalization causes that the alignment depends on prompts' direction rather than their magnitude. We average this loss over all parallel prompt pairs in the alignment batch.

$$\mathcal{L}_{\text{align}} = \left\| \frac{h_a(x^t)}{\|h_a(x^t)\|_2} - \frac{h(x^s)}{\|h(x^s)\|_2} \right\|_2^2. \tag{2}$$

Preserving Learned Knowledge. The addition of the adapters can change the overall model behavior. To prevent a degradation of overall model performance, we include two additional preservation loss terms.

Our first loss term $\mathcal{L}_{\text{anc}}$ aims to *preserve the anchor-language representations*. Therefore, it penalizes the change that the addition of the adapter has on the representations of prompts in the anchor language, each represented as a single vector as detailed above:

$$\mathcal{L}_{\text{anc}} = \frac{\|h_a(x^s) - h(x^s)\|_2^2}{\|h(x^s)\|_2^2 + \epsilon}. \tag{3}$$

The numerator measures the squared difference between the anchor-language representations with and without the adapter. The denominator expresses this difference relative to the squared norm of the reference representation, with $\epsilon > 0$ added for numerical stability. We average the loss over the selected anchor-language prompts.

Our second loss term $\mathcal{L}_{\mathrm{KL}}$ aims at *preserving overall next-token predictions*. This is necessary because both other loss terms consider each individual prompt as a single vector obtained by averaging token hidden states, and preserving this average vector does not necessarily preserve the model's per-token prediction capabilities. $\mathcal{L}_{\mathrm{KL}}$ therefore optimizes to keep the next-token distributions with and without the adapter close using the Kullback-Leibler divergence

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}\left(p \,\|\, p_a\right), \tag{4}$$

where p and p_a are the distributions with the adapter disabled and enabled, respectively. Both distributions are computed using the same preceding input text. The loss compares probabilities over the full vocabulary, discouraging changes to the probability of the correct next token and those of alternative continuations.

The complete adapter objective is then

$$\mathcal{L}_{\mathrm{adapter}} = \underbrace{\mathcal{L}_{\mathrm{align}}}_{\text{alignment}} + \underbrace{\lambda_{\mathrm{anc}}\mathcal{L}_{\mathrm{anc}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}}_{\text{preservation}}, \tag{5}$$

where the coefficients determine the strength of preservation relative to alignment. Each term is averaged over its selected examples, with the KL term including examples in both languages.

4.3 Deployment and Unlearning

Deployment. After fine-tuning, we merge the adapter into the base-model parameters at the scale used during training. This yields a single set of parameters θ. As a result, the final model contains no additional component and incurs no inference overhead. More importantly, this allows us to run any unlearning method on θ directly, including methods that modify weights, layers, or subspaces and would otherwise have to be changed for a model with attached adapters.

Unlearning. Let $\mathcal{D} = \{(x_i^\ell, y_i^\ell)\}$ be the facts the model was fine-tuned on, expressed in the languages $\ell \in \mathcal{L}$, and let $\mathcal{F} \subset \mathcal{D}$ denote the forget set, *i.e.,* facts to be removed, and $\mathcal{R} = \mathcal{D} \setminus \mathcal{F}$ the retain set, *i.e.,* the facts to be retained. Given an unlearning method $\mathcal{U}$, unlearning yields

$$\theta_{\mathrm{un}} = \mathcal{U}\left(\theta,\ \mathcal{F}^s,\ \mathcal{R}^s\right), \tag{6}$$

where the provided $\mathcal{F}^s$ and $\mathcal{R}^s$ contain only the anchor-language examples of the respective facts. LATTE neither changes $\mathcal{U}$ nor its hyperparameters, and requires no translations of the facts from the forget or retain set. Successful cross-lingual unlearning then requires that, for every language $\ell \in \mathcal{L}$, the model θ_{un} no longer answers the facts in $\mathcal{F}^\ell$ while still answering those in $\mathcal{R}^\ell$, which is what we empirically validate in the next section.

Unknown forget sets. Neither training objective uses the labels $\mathcal{F}$ and $\mathcal{R}$, so all facts receive the same treatment during fine-tuning. LATTE therefore does not require advance knowledge of which facts will later be removed and which ones retained.

5 Experiments

We compare standard fine-tuning with LATTE using the same training data and downstream unlearning settings. We evaluate forgetting and retention across languages, ablate the training objectives, and analyze the learned prompt representations.

5.1 Experimental Setup

Datasets. We use multilingual versions of TOFU (Maini et al., 2024) and LUME (Ramakrishna et al., 2025). TOFU fine-tuning uses 4,000 prompts and target responses per language about 200 fictitious authors. Our LUME subset contains 240 prompt and target-response pairs per language from its synthetic-story task, without supplying the source stories. For both datasets, we evaluate

Table 2: **English-only unlearning on TOFU with Qwen2.5-3B-Instruct.** Scores are means over five seeds. Plain denotes standard fine-tuning. Bold indicates the better score within each method pair, including ties. Alignment targets appear below the separator.

(a) Forget NLI ↓

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.99	1.00	0.49	**0.10**	0.48	**0.04**	0.75	**0.19**
DE	0.93	1.00	0.70	**0.11**	0.48	**0.07**	0.87	**0.26**
ES	0.97	0.97	0.62	**0.11**	0.60	**0.04**	0.83	**0.28**
FR	1.00	1.00	0.78	**0.05**	0.58	**0.06**	0.90	**0.26**
VI	0.95	1.00	0.77	**0.19**	0.73	**0.17**	0.87	**0.40**
HI	0.35	0.38	0.28	**0.08**	0.24	**0.14**	0.34	**0.16**
TH	0.83	0.98	0.75	**0.09**	0.39	**0.12**	0.84	**0.27**
ZH	1.00	1.00	0.74	**0.33**	0.60	**0.26**	0.87	**0.50**
JA	0.97	0.97	0.89	**0.38**	0.84	**0.27**	0.92	**0.60**
KO	1.00	0.98	0.85	**0.15**	0.72	**0.13**	0.92	**0.36**

(b) Retain NLI ↑

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.98	0.99	**0.95**	0.92	**0.97**	0.81	**0.98**	0.80
DE	0.99	0.99	**0.95**	0.93	**0.95**	0.90	**0.98**	0.94
ES	0.96	0.97	**0.94**	**0.94**	**0.94**	0.93	**0.95**	0.93
FR	0.97	0.99	**0.94**	0.92	**0.96**	0.89	**0.97**	0.95
VI	0.97	0.98	**0.93**	0.90	**0.95**	0.88	**0.97**	0.93
HI	0.51	0.54	**0.48**	0.47	**0.48**	0.42	**0.49**	0.46
TH	0.96	0.97	**0.91**	0.88	**0.94**	0.82	**0.95**	0.90
ZH	0.96	1.00	**0.95**	**0.95**	**0.96**	0.93	**0.95**	0.93
JA	0.98	1.00	**0.98**	0.97	0.94	**0.95**	0.97	**0.98**
KO	0.97	0.99	**0.96**	0.94	**0.95**	0.93	**0.96**	0.93

40 forget and 200 retain examples per language. Dataset construction, translation, and training and unlearning subsets are detailed in Appendix A.1.

Models and training. We evaluate Llama-3.2-3B-Instruct (Meta, 2024) and Qwen2.5-3B-Instruct (Qwen et al., 2025). For Llama, we report English, German, Spanish, French, Hindi, and Thai. For Qwen, we additionally report Chinese, Japanese, Korean, and Vietnamese. Standard fine-tuning and LATTE use the same multilingual corpus and target-response language modeling objective. Standard fine-tuning updates all model parameters, while LATTE alternates base-model and adapter updates as described in Section 4.

English is the anchor language. The alignment targets are Hindi and Thai for Llama, and Hindi, Thai, Chinese, Japanese, and Korean for Qwen. Following Liu & Niehues (2025), we compute alignment using middle-layer representations, at layer 13 for Llama and layer 18 for Qwen. Complete training settings are provided in Appendix A.2, and alternative-anchor experiments in Appendix A.7.

Unlearning methods. For TOFU, we evaluate **NPO** (Zhang et al., 2024), **GradDiff** (Maini et al., 2024), **SimNPO** (Fan et al., 2025), **RMU** (Li et al., 2024), and **LINGTEA*** (Choi et al., 2024) on both models.[1] For LUME, we evaluate NPO, GradDiff, and LINGTEA*. The SimNPO and RMU results for both models are provided in Appendix A.5. All methods receive English forget and retain examples, with unlearning settings fixed between standard fine-tuning and LATTE. Configurations and reporting procedures are provided in Appendix A.3.

Metrics. We evaluate forgetting and retention on TOFU and LUME using NLI scores and ROUGE-L. Following Xiang et al. (2026), we use a multilingual NLI evaluator.[2] Scores are reported separately for each language. Lower forget scores indicate stronger forgetting, while higher retain scores indicate better retention. Metric definitions and generation settings are provided in Appendix A.6.

5.2 Cross-Lingual Unlearning Results

Table 2 shows stronger forgetting with LATTE than with standard fine-tuning in all ten languages under NPO, GradDiff, and LINGTEA*. In Korean, forget NLI falls from 0.85 to 0.15 with NPO, from 0.72 to 0.13 with GradDiff, and from 0.92 to 0.36 with LINGTEA*. Under NPO, retain NLI remains within 0.032 of standard fine-tuning in every language. GradDiff achieves forget NLI below 0.30 across all ten languages, and LINGTEA* also improves forgetting in every language. On Llama-3.2-3B-Instruct, LATTE achieves higher retain NLI in all six languages under all three methods. It also improves forgetting in English and Hindi under NPO and GradDiff, and in all six languages under LINGTEA*. These results show that LATTE can strengthen forgetting while improving retention. Full results for both models, including ROUGE-L scores, are provided in Appendix A.5.

[1] We restrict LINGTEA's setting to English forget and retain examples.

[2] https://huggingface.co/joeddav/xlm-roberta-large-xnli

Table 3: **English-only unlearning on LUME with Qwen2.5-3B-Instruct.** Scores are means over five seeds. Plain denotes standard fine-tuning. Bold indicates the better score within each method pair, including ties at the displayed precision. Alignment targets appear below the separator.

(a) Forget NLI ↓

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	1.00	1.00	0.45	**0.28**	0.40	**0.30**	0.25	**0.20**
DE	1.00	1.00	0.54	**0.49**	0.54	**0.49**	0.31	**0.14**
ES	1.00	1.00	0.65	**0.51**	0.62	**0.49**	0.30	**0.22**
FR	1.00	1.00	**0.58**	0.61	0.61	**0.55**	0.24	**0.18**
VI	1.00	1.00	0.65	**0.56**	0.65	**0.50**	0.38	**0.21**
HI	1.00	1.00	0.93	**0.69**	0.97	**0.72**	0.83	**0.41**
TH	1.00	1.00	0.93	**0.71**	0.96	**0.71**	0.79	**0.36**
ZH	1.00	1.00	0.79	**0.50**	0.75	**0.50**	0.50	**0.36**
JA	1.00	1.00	0.92	**0.69**	0.88	**0.73**	0.73	**0.58**
KO	1.00	1.00	0.93	**0.75**	0.96	**0.75**	0.71	**0.32**

(b) Retain NLI ↑

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	1.00	1.00	0.63	**0.76**	0.56	**0.73**	0.59	**0.73**
DE	1.00	1.00	0.63	**0.75**	0.56	**0.70**	0.52	**0.76**
ES	1.00	1.00	0.59	**0.79**	0.57	**0.77**	0.62	**0.76**
FR	1.00	1.00	0.68	**0.81**	0.64	**0.77**	0.61	**0.74**
VI	1.00	1.00	0.60	**0.76**	0.51	**0.70**	0.56	**0.70**
HI	0.99	0.99	0.82	**0.91**	0.82	**0.89**	**0.82**	0.80
TH	1.00	1.00	**0.84**	**0.84**	**0.84**	0.82	0.77	**0.81**
ZH	1.00	1.00	0.66	**0.80**	0.63	**0.77**	0.66	**0.75**
JA	1.00	1.00	0.84	**0.92**	0.72	**0.90**	0.76	**0.78**
KO	1.00	1.00	0.86	**0.92**	0.81	**0.90**	0.78	**0.80**

Table 3 shows that LATTE improves both forgetting and retention on LUME in most languages. Forget NLI is lower in nine languages under NPO and all ten under GradDiff and LINGTEA*, while retain NLI is higher in nine languages under each of the three methods. These results show that the benefits extend beyond TOFU, with LATTE achieving stronger cross-lingual forgetting while preserving more retained knowledge.

5.3 Ablations

We study the contributions of the **loss terms, preservation weights, and padding tokens** using Qwen2.5-3B-Instruct with NPO. We report forgetting and retention for the five alignment target languages, keeping the unlearning settings fixed across configurations.

Loss terms. We evaluate the contribution of each loss by disabling it individually, retaining the adapter and preservation updates when disabling alignment. The complete objective achieves the lowest forget NLI in all five alignment target languages (Table 4). Disabling either preservation loss also lowers retain NLI, showing that both terms contribute to stronger forgetting and better retention.

Table 4: **Contribution of each loss term.** NPO, mean over five unlearning seeds. F denotes forget NLI and R denotes retain NLI. Lower F and higher R are better. Bold marks the best score in each column, including ties.

	Hindi		Thai		Chinese		Japanese		Korean	
Configuration	F↓	R↑	F↓	R↑	F↓	R↑	F↓	R↑	F↓	R↑
Full LATTE	**0.08**	0.47	**0.09**	0.88	**0.33**	**0.95**	**0.38**	**0.97**	**0.15**	0.94
Without $\mathcal{L}_{\text{align}}$	0.28	**0.49**	0.73	**0.91**	0.81	**0.95**	0.89	**0.97**	0.84	**0.95**
Without $\mathcal{L}_{\text{anc}}$	0.14	0.39	0.27	0.75	0.34	0.91	0.40	0.93	0.26	0.92
Without $\mathcal{L}_{\text{KL}}$	0.31	0.40	0.50	0.72	0.75	0.90	0.77	0.84	0.57	0.75

Prompt pooling. We evaluate the effect of padding tokens by including them when computing prompt representations. Excluding them achieves lower forget NLI in all five alignment target languages (Table 5). Including padding tokens raises forget NLI from 0.09 to 0.78 in Thai and from 0.15 to 0.94 in Korean, supporting their exclusion in LATTE. **Preservation weights.** We vary each preservation weight separately and find that $\lambda_{\text{anc}} = \lambda_{\text{KL}} = 1.0$ provides a favorable balance between forgetting and retention (Appendix A.8).

5.4 Analysis of Cross-Lingual Representations

The unlearning results show how LATTE affects forgetting and retention. We now examine whether its training procedure also improves the cross-lingual representation alignment it is designed to encourage, using the LSAR analysis of Xie et al. (2022). We compare the standard and LATTE

Table 5: **Effect of padding tokens.** F denotes forget NLI and R denotes retain NLI. Lower F and higher R are better. Bold marks the best score in each column, including ties.

	Hindi		Thai		Chinese		Japanese		Korean	
Pooling	F↓	R↑	F↓	R↑	F↓	R↑	F↓	R↑	F↓	R↑
Default	**0.08**	0.47	**0.09**	0.88	**0.33**	0.95	**0.38**	**0.97**	**0.15**	0.94
Include padding	0.34	**0.54**	0.78	**0.95**	0.87	**0.98**	0.92	**0.97**	0.94	**0.96**

Table 6: **Cross-language retrieval on Qwen2.5-3B-Instruct.** Top-1 accuracy before and after LSAR projection. Alignment targets appear above the separator. Higher is better.

	Standard fine-tuning		LATTE	
Language	Without LSAR	With LSAR	Without LSAR	With LSAR
Hindi	0.266	0.944	0.994	0.994
Thai	0.502	0.882	0.968	0.968
Chinese	0.844	0.944	0.984	0.984
Japanese	0.888	0.978	0.982	0.982
Korean	0.756	0.892	0.992	0.994
German	0.906	0.988	0.984	0.996
Spanish	0.988	0.998	0.994	0.998
French	0.895	0.970	0.998	0.998
Vietnamese	0.916	0.942	0.996	0.996

Qwen2.5-3B-Instruct checkpoints before unlearning on 500 parallel TOFU prompts included in fine-tuning, following the retrieval procedure in Section 3.

LSAR computes a mean prompt representation for each language and uses singular value decomposition to identify directions that capture differences between these means. The languages contain translations of the same prompts, keeping the underlying content consistent. We estimate these directions from TOFU prompts, remove their components from each representation, and repeat retrieval without updating the model. This tests whether removing language-associated differences makes corresponding prompts easier to identify, and whether LATTE already achieves strong alignment without this projection.

Table 6 shows that removing language-associated components improves retrieval after standard fine-tuning in every language, with the largest gains in Hindi and Thai. In contrast, LATTE achieves higher retrieval accuracy than standard fine-tuning even already in every language without projection, including languages not directly targeted by alignment. Applying LSAR to LATTE leaves accuracy unchanged in six languages and provides only small gains in the remaining three. These results support the intended effect of LATTE, which learns representations that identify parallel prompts across languages with little additional benefit from projection.

6 Conclusion

We presented LATTE, a method that introduces cross-lingual alignment during fine-tuning to improve subsequent unlearning. By aligning parallel prompt representations while preserving learned representations and predictions, LATTE enables existing unlearning methods to operate on the resulting model without changing their objectives or requiring translated forget examples. This shifts part of the challenge of cross-lingual unlearning to the stage when multilingual knowledge is learned. Our experiments demonstrate stronger forgetting across all ten evaluated languages on Qwen2.5-3B-Instruct, together with higher retained performance across the six evaluated languages on Llama-3.2-3B-Instruct. Representation analysis confirms improved cross-language retrieval, while ablations show that the preservation losses contribute to both forgetting and retention. These findings establish alignment during fine-tuning as a promising approach to improving multilingual unlearning before future forget sets are specified.

AI USE STATEMENT

We used generative AI tools to refine research hypotheses, obtain feedback on methodology and experimental design, suggest experimental parameters, assist with software implementation, and support the interpretation of results. We also used these tools to translate existing benchmark examples and improve the organization and readability of the manuscript. We did not use generative AI to create new source facts or stories, develop theoretical models or conceptual frameworks, formulate mathematical claims, or write mathematical proofs. Qualitative and thematic data analysis was not applicable. All AI-assisted contributions were reviewed and verified by the authors, who take responsibility for the final text, code, claims, and other research artifacts.

REFERENCES

BigScience Workshop. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. URL https://arxiv.org/abs/2211.05100.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, 2021. URL https://arxiv.org/abs/1912.03817.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. *arXiv preprint arXiv:2501.06346*, 2025.

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. arXiv:2406.12354.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

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

European Parliament and Council of the European Union. Regulation (EU) 2016/679 (general data protection regulation), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng. Article 17: Right to erasure.

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*, volume 38, 2025. doi: 10.52202/085713-0058. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/02443e4e008231e0af0855f3cc70ed17-Abstract-Conference.html.

Abir Harrasse, Florent Draye, Punya Syon Pandey, Zhijing Jin, and Bernhard Schölkopf. Tracing multilingual representations in LLMs with cross-layer transcoders. *arXiv preprint arXiv:2511.10840*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021. arXiv:2102.00894.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.03218.

Taoran Li, Varun Chandrasekaran, and Zhiyuan Yu. Layer-targeted multilingual knowledge erasure in large language models. *arXiv preprint arXiv:2602.22562*, 2026.

Danni Liu and Jan Niehues. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15979–15996, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.778. URL https://aclanthology.org/2025.acl-long.778/.

Tyler Lizzo and Larry Heck. UNLEARN efficient removal of knowledge in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7272–7283. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.405. URL https://aclanthology.org/2025.findings-naacl.405/.

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

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Meta. Llama 3.2 model card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493/.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2310.10378.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. LUME: LLM unlearning with multitask evaluations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng

(eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6524–6535, Suzhou, China, nov 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.347. URL https://aclanthology.org/2025.findings-emnlp.347/.

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. arXiv:2402.10588.

Chaoyi Xiang, Olga Ohrimenko, Benjamin Rubinstein, and Lea Frermann. Multilingual unlearning in LLMs: Transfer, dynamics, and reversibility. *arXiv preprint arXiv:2606.03291*, 2026. ICML 2026.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5617–5633. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.379. URL https://aclanthology.org/2022.emnlp-main.379/.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5554–5569. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.343. URL https://aclanthology.org/2023.findings-acl.343/.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

A Appendix

A.1 Dataset Construction

A.1.1 Multilingual TOFU

We construct our multilingual corpus from TOFU (Maini et al., 2024), which contains 4,000 prompts and target responses about 200 fictitious authors, with 20 examples per author. English examples come from the original dataset. We obtain parallel versions in German, Spanish, French, Hindi, Thai, Chinese, Japanese, Korean, and Vietnamese.

Translation. German, Spanish, French, and Hindi translations are produced using Claude Opus 5 (`claude-opus-5`). Each prompt and each target response is translated in a separate request, without a system prompt, using a maximum output length of 2,048 tokens and the default temperature. Each request uses the following instruction, followed by the English text.

> Translate the following English text into {language}. Preserve all proper names, book titles, dates and numbers exactly as written. Keep the same meaning and factual structure. Output only the translation, with no preamble or quotes.

Thai, Chinese, Japanese, Korean, and Vietnamese translations are produced using Google Translate through the free web endpoint accessed by `deep_translator`. Chinese uses the language code `zh-CN`. These requests do not use an additional translation instruction.

Each language preserves the row order of the corresponding English split. Parallel examples are matched by row position, without a separate identifier field. This correspondence was checked against the translated data.

Fine-tuning data. Fine-tuning uses 4,000 examples per language, giving 24,000 examples across six languages for Llama-3.2-3B-Instruct and 40,000 examples across ten languages for Qwen2.5-3B-Instruct. Standard fine-tuning and LATTE use the same training examples for each model. Fine-tuning does not distinguish between examples that will later belong to the forget and retain sets.

Alignment data. The alignment pool contains the 40 examples from *forget01* and the first 2,000 examples from *retain99*, giving 2,040 examples per language. Each anchor-language prompt is paired with the prompt at the same row position in a target language. The same alignment objective is applied to all selected examples, without using their forget or retain labels.

English is the anchor language in the main experiments. The alignment targets are Hindi and Thai for Llama-3.2-3B-Instruct, and Hindi, Thai, Chinese, Japanese, and Korean for Qwen2.5-3B-Instruct.

Unlearning subsets. The forget set contains the 40 examples in *forget01*, covering two fictitious authors. The retain pool consists of the first 2,000 rows of the English *retain99* split in its released order. Selecting this pool involves no random sampling or selection seed. Standard fine-tuning and LATTE use the same forget set and retain pool. Unlearning receives only English examples.

Evaluation subsets. We evaluate forgetting on all 40 forget examples and retention on the first 200 rows of *retain99* in each reported language. The retained evaluation subset is selected by row order without random sampling, and all 200 examples are scored in every run. These examples are also the first 200 entries in the selected retain pool.

Standard fine-tuning and LATTE use identical evaluation examples. Row correspondence ensures that the same underlying examples are evaluated across languages. All forget and retain evaluation examples were included in fine-tuning.

A.1.2 Multilingual LUME

We use Task 1 of LUME (Ramakrishna et al., 2025), the synthetic-story component of the benchmark developed for SemEval-2025 Task 4. It contains 393 stories generated using Mixtral 8x7B, with 199 assigned to the forget set and 194 to the retain set. Each story contains approximately 150 to 200 words and uses randomly generated character names.

Table 7: **Multilingual LUME construction.** Models receive only prompts and target responses.

Quantity	Count
Source stories	393
Forget stories	199
Retain stories	194
Initial prompt and target-response pairs	489
Excluded as story-dependent	73
Excluded for shared answers	18
Pairs remaining after filtering	398
Selected forget examples per language	40
Selected retain examples per language	200
Languages	10
Examples in the ten-language corpus	2,400

Filtering and subset selection. Our experiments use only prompts and target responses. The source stories are not supplied during fine-tuning, unlearning, or evaluation. Each question inherits its forget or retain assignment from its source story.

We start from 489 prompt and target-response pairs. We exclude 73 questions that refer explicitly to the source story, such as questions beginning with "In the story", because the model does not receive that context. We also exclude 18 questions whose answers are shared with another question, leaving 398 pairs.

Using a fixed random seed, we select 40 questions associated with forget stories and 200 associated with retain stories. These counts match the forget and retain evaluation sizes used for TOFU. The selected English target responses have a median length of two words, and 45% contain one word.

Translation. We translate each selected prompt and target response using Opus 5 into German, Spanish, French, Vietnamese, Hindi, Thai, Chinese, Japanese, and Korean. The translation instructions require natural phrasing, preservation of the original meaning, and proper names written in the conventional script of each language. Names remain in Latin script in the Latin-script languages and are transliterated in the other languages.

Translations preserve row correspondence and forget and retain assignments across languages. The resulting corpus contains 240 examples per language and 2,400 examples in total.

Fine-tuning and alignment data. We use this corpus to fine-tune Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct with standard fine-tuning and LATTE. Both conditions use the same selected prompt and target-response pairs for each model.

Unlearning and evaluation subsets. Unlearning receives the 40 English forget pairs and the 200 English retain pairs. We evaluate all 40 forget and 200 retain examples in each model's reported languages, using the same examples for standard fine-tuning and LATTE. All evaluation examples were included in fine-tuning, and the English retain evaluation examples constitute the full retain pool used during unlearning. Evaluation procedures are specified in Appendix A.6.

A.2 Fine-Tuning and Alignment Settings

Standard fine-tuning and LATTE start from the same instruction-tuned checkpoint and use the same training corpus for each model and dataset. Both train the base model using the causal language modeling objective over target-response tokens. LATTE additionally alternates these updates with adapter updates for alignment and knowledge preservation. We use the same training settings for TOFU and LUME.

A.2.1 Standard Multilingual Fine-Tuning

We fine-tune Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct separately on the multilingual TOFU and LUME corpora described in Appendix A.1. Each example is formatted using the corresponding model's chat template. The model predicts each target-response token using the prompt and preceding

target tokens as context. The loss is averaged over target-response tokens, excluding prompt and padding tokens.

Standard fine-tuning updates all model parameters without an adapter. We use paged AdamW with a learning rate of 5×10^{-5}, zero weight decay, and a warmup ratio of 0.05. The per-device batch size is 8, with four gradient accumulation steps. Training uses bfloat16 precision and gradient checkpointing. The selected checkpoint is used directly for subsequent unlearning.

A.2.2 LATTE TRAINING

Adapter configuration. We attach rank-8 LoRA adapters to the attention projections q, k, v, and o and the MLP projections `gate`, `up`, and `down`. Adapters are attached to layers 0 through 13 for Llama-3.2-3B-Instruct and layers 0 through 18 for Qwen2.5-3B-Instruct, inclusive. The same adapters are used across languages.

Alternating updates. Base-model updates use the target-response language modeling objective, with the adapters enabled and their parameters fixed. Adapter updates keep the base-model parameters fixed and minimize $\mathcal{L}_{\text{adapter}}$ from Equation (5). The base model and adapters use separate optimizers.

We perform one adapter update after every two base-model optimizer updates. The interval counts optimizer updates after gradient accumulation. Each alignment batch contains 16 parallel prompt pairs. The adapter optimizer is AdamW with a learning rate of 2×10^{-4}, and adapter gradients are clipped to a norm of 1.0. After each adapter update, base-model training resumes with the updated adapters enabled.

Alignment and representation preservation. We compute prompt representations at layer 13 for Llama-3.2-3B-Instruct and layer 18 for Qwen2.5-3B-Instruct. Each representation is the mean of the prompt-token hidden states, excluding the first token and padding positions. Target-response tokens are not included. The same representation computation is used for the alignment and anchor-language preservation losses.

For alignment, the target-language prompt is processed with the adapter enabled, while its anchor-language counterpart is processed with the adapter disabled. Anchor-language preservation compares representations of the same anchor-language prompt with and without the adapter. Reference representations are recomputed from the current base model without gradients at every adapter update.

Prediction preservation. The KL loss compares next-token distributions with the adapter disabled and enabled, using the former as the reference distribution. Both configurations receive the same prompt and target response. The comparison covers predictions over both the prompt and target response, using the supplied target-response tokens as context and excluding padding positions.

We compute the divergence separately for each language, average it within each example, and then average over the selected examples in the anchor and target languages. Reference distributions are recomputed without gradients at every adapter update.

Preservation weights and subsets. The main experiments use $\lambda_{\text{anc}} = \lambda_{\text{KL}} = 1.0$. Each loss is averaged over its selected examples before being combined in the adapter objective. The alignment pools for both benchmarks are described in Appendix A.1.

Checkpoint selection and merging. After fine-tuning, the adapters are merged into the base-model parameters at the same scale used during training. Subsequent unlearning operates directly on the merged checkpoint.

A.2.3 TRAINING ALGORITHM

Algorithm 1 summarizes the alternating procedure. The multilingual training data $\mathcal{D}$ contain prompts and target responses. The parallel data $\mathcal{P}$ pair anchor-language examples with their translations in the selected alignment languages. Each base-model update in the algorithm corresponds to one optimizer step after gradient accumulation. The combined alignment and preservation objective updates only the adapter parameters.

Algorithm 1: Alternating fine-tuning with LATTE

Input: Base-model parameters θ, multilingual training data $\mathcal{D}$, parallel training data $\mathcal{P}$, alignment interval k, preservation weights λ_{anc} and λ_{KL}
Output: Fine-tuned model with adapters merged

```
Initialize adapter parameters φ
Enable adapters and set the update counter t ← 0
while the stopping criterion is not met do
    Sample a training batch B from D
    Update θ using the target-response language modeling loss on B, keeping φ fixed
    t ← t + 1
    if t mod k = 0 then
        Sample a parallel batch A from P and select preservation examples
        Compute reference representations and predictions with adapters disabled and without
          gradients
        Enable adapters and compute the alignment and preservation losses
        L_adapter ← L_align + λ_anc L_anc + λ_KL L_KL
        Update φ using L_adapter, keeping θ fixed
Merge the adapters into the base model at the training scale
return the merged model
```

A.3 Unlearning Methods and Settings

We apply each unlearning method after standard fine-tuning and after LATTE. Standard fine-tuning provides the baseline checkpoint, while LATTE provides the checkpoint obtained after merging the adapters. Each method receives 40 English forget examples and retain examples from the corresponding dataset's English retain pool, as described in Appendix A.1. The unlearning objective and hyperparameter settings are fixed across the two fine-tuning conditions.

A.3.1 Method Configurations

NPO. NPO reduces the likelihood of forget responses relative to a fixed reference model (Zhang et al., 2024). The reference is the corresponding checkpoint before unlearning, including the merged checkpoint for LATTE. We combine the forgetting objective with a language modeling loss on retain examples, using $\beta = 0.1$ and a retain-loss weight of 1. All model parameters are updated with AdamW at a learning rate of 10^{-5} for five epochs.

GradDiff. GradDiff combines gradient ascent on forget examples with gradient descent on retain examples (Maini et al., 2024). The former increases the target-response loss for facts selected for unlearning, while the latter trains the model to continue predicting retained responses. We use a retain-loss weight of 1 and update all model parameters with AdamW at a learning rate of 10^{-5} for five epochs.

SimNPO. SimNPO uses a reference-free forgetting objective based on the length-normalized log-probability of each target response (Fan et al., 2025). We combine this objective with a language modeling loss on retain examples.

RMU. RMU changes hidden representations for forget examples by moving them toward a fixed random vector (Li et al., 2024). A preservation term penalizes changes to representations of retain examples relative to the model before unlearning. The reference is the corresponding standard fine-tuned or merged LATTE checkpoint. The objective is computed at a selected layer, and updates are restricted to selected model parameters.

A.4 Baseline Results

LINGTEA*. LINGTEA combines gradient ascent on forget examples with knowledge distillation on retain examples (Choi et al., 2024). Distillation encourages the updated model to preserve a fixed teacher's predictions, with weights determined by the teacher's confidence. We adapt its multilingual

Table 8: **Baseline NLI results on TOFU with Qwen2.5-3B-Instruct.** Scores are from standard fine-tuning before and after English-only unlearning. F denotes forget NLI and R denotes retain NLI. Lower F and higher R are better.

	Before		NPO		GradDiff		LINGTEA*	
Language	F↓	R↑	F↓	R↑	F↓	R↑	F↓	R↑
EN	0.99	0.98	0.49	0.95	0.48	0.97	0.75	0.98
DE	0.93	0.99	0.70	0.95	0.48	0.95	0.87	0.98
ES	0.97	0.96	0.62	0.94	0.60	0.94	0.83	0.95
FR	1.00	0.97	0.78	0.94	0.58	0.96	0.90	0.97
VI	0.95	0.97	0.77	0.93	0.73	0.95	0.87	0.97
HI	0.35	0.51	0.28	0.48	0.24	0.48	0.34	0.49
TH	0.83	0.96	0.75	0.91	0.39	0.94	0.84	0.95
ZH	1.00	0.96	0.74	0.95	0.60	0.96	0.87	0.95
JA	0.97	0.98	0.89	0.98	0.84	0.94	0.92	0.97
KO	1.00	0.97	0.85	0.96	0.72	0.95	0.92	0.96

Table 9: **Baseline NLI results on TOFU with Llama-3.2-3B-Instruct.** Scores are from standard fine-tuning before and after English-only unlearning. F denotes forget NLI and R denotes retain NLI. Lower F and higher R are better.

	Before		NPO		GradDiff		LINGTEA*	
Language	F↓	R↑	F↓	R↑	F↓	R↑	F↓	R↑
EN	0.96	0.98	0.26	0.79	0.22	0.78	0.53	0.91
DE	0.93	0.94	0.24	0.83	0.16	0.83	0.55	0.89
ES	0.97	0.95	0.27	0.87	0.34	0.87	0.69	0.91
FR	0.95	0.95	0.21	0.90	0.29	0.90	0.61	0.91
HI	0.87	0.88	0.26	0.78	0.31	0.77	0.65	0.82
TH	0.91	0.93	0.24	0.76	0.20	0.75	0.61	0.82

data setting to use only English forget and retain examples, denoting this configuration by LINGTEA*. The teacher is the corresponding standard fine-tuned or merged LATTE checkpoint before unlearning.

A.5 Additional Cross-Lingual Unlearning Results

A.5.1 TOFU Results on Qwen2.5-3B-Instruct

Table 10 reports forgetting and retention for all five unlearning methods. Compared with standard fine-tuning, LATTE achieves lower forget NLI in every language under NPO, GradDiff, and LINGTEA*. SimNPO and RMU leave high forget NLI in most languages under both fine-tuning conditions. With these two methods, LATTE generally achieves higher retain NLI, but does not consistently improve forgetting.

Table 10: **Full NLI results on TOFU with Qwen2.5-3B-Instruct.** All methods use English-only unlearning. Plain denotes standard fine-tuning. Before gives scores before unlearning. Bold indicates the better post-unlearning score within each method pair, including ties at the displayed precision. Alignment targets appear below the separator.

(a) Forget NLI ↓

	Before		NPO		GradDiff		LINGTEA*		SimNPO		RMU	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.99	1.00	0.49	**0.10**	0.48	**0.04**	0.75	**0.19**	**0.84**	0.91	1.00	**0.95**
DE	0.93	1.00	0.70	**0.11**	0.48	**0.07**	0.87	**0.26**	**0.90**	0.98	**0.86**	0.96
ES	0.97	0.97	0.62	**0.11**	0.60	**0.04**	0.83	**0.28**	**0.90**	0.99	**0.94**	0.99
FR	1.00	1.00	0.78	**0.05**	0.58	**0.06**	0.90	**0.26**	1.00	**0.97**	**0.91**	0.95
VI	0.95	1.00	0.77	**0.19**	0.73	**0.17**	0.87	**0.40**	**0.94**	0.97	0.87	**0.86**
HI	0.35	0.38	0.28	**0.08**	0.24	**0.14**	0.34	**0.16**	**0.35**	**0.35**	**0.28**	0.33
TH	0.83	0.98	0.75	**0.09**	0.39	**0.12**	0.84	**0.27**	**0.81**	0.97	0.79	**0.77**
ZH	1.00	1.00	0.74	**0.33**	0.60	**0.26**	0.87	**0.50**	**0.93**	1.00	**0.90**	0.99
JA	0.97	0.97	0.89	**0.38**	0.84	**0.27**	0.92	**0.60**	**0.97**	**0.97**	**0.90**	0.92
KO	1.00	0.98	0.85	**0.15**	0.72	**0.13**	0.92	**0.36**	1.00	**0.97**	0.90	**0.86**

(b) Retain NLI ↑

	Before		NPO		GradDiff		LINGTEA*		SimNPO		RMU	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.98	0.99	**0.95**	0.92	**0.97**	0.81	**0.98**	0.80	0.99	**1.00**	0.98	**1.00**
DE	0.99	0.99	**0.95**	0.93	**0.95**	0.90	**0.98**	0.94	0.98	**0.99**	0.96	**1.00**
ES	0.96	0.97	**0.94**	**0.94**	**0.94**	0.93	**0.95**	0.93	**0.96**	**0.96**	0.95	**0.96**
FR	0.97	0.99	**0.94**	0.92	**0.96**	0.89	**0.97**	0.95	0.97	**0.98**	0.96	**0.99**
VI	0.97	0.98	**0.93**	0.90	**0.95**	0.88	**0.97**	0.93	0.96	**0.98**	0.96	**0.98**
HI	0.51	0.54	**0.48**	0.47	**0.48**	0.42	**0.49**	0.46	0.50	**0.53**	0.48	**0.53**
TH	0.96	0.97	**0.91**	0.88	**0.94**	0.82	**0.95**	0.90	**0.96**	0.94	**0.92**	**0.92**
ZH	0.96	1.00	**0.95**	**0.95**	**0.96**	0.93	**0.95**	0.93	0.96	**0.99**	0.95	**0.99**
JA	0.98	1.00	**0.98**	0.97	0.94	**0.95**	0.97	**0.98**	0.98	**0.99**	0.97	**0.99**
KO	0.97	0.99	**0.96**	0.94	**0.95**	0.93	**0.96**	0.93	0.97	**0.98**	0.95	**0.98**

Table 11: **ROUGE-L results on TOFU with Qwen2.5-3B-Instruct.** Plain denotes standard fine-tuning. Bold indicates the better score within each method pair, including ties at the displayed precision. Alignment targets appear below the separator.

(a) Forget ROUGE-L ↓

	NPO		GradDiff		LINGTEA*		SimNPO		RMU	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.76	**0.01**	0.65	**0.01**	0.87	**0.04**	**0.91**	0.94	**1.00**	**1.00**
DE	0.86	**0.07**	0.70	**0.07**	0.94	**0.23**	**0.97**	0.98	**0.91**	0.97
ES	0.82	**0.12**	0.77	**0.09**	0.92	**0.32**	**0.96**	0.99	**0.96**	0.99
FR	0.85	**0.12**	0.72	**0.09**	0.95	**0.29**	0.99	**0.97**	**0.95**	0.99
VI	0.85	**0.20**	0.78	**0.18**	0.91	**0.45**	**0.95**	0.97	**0.91**	0.95
HI	0.62	**0.29**	0.56	**0.27**	0.63	**0.43**	**0.64**	0.72	**0.62**	0.69
TH	0.90	**0.34**	0.72	**0.27**	0.91	**0.58**	**0.93**	0.98	**0.89**	0.92
ZH	0.88	**0.33**	0.81	**0.27**	0.93	**0.39**	**0.95**	0.99	**0.95**	0.97
JA	0.94	**0.54**	0.94	**0.49**	0.96	**0.75**	**0.98**	**0.98**	**0.95**	0.96
KO	0.88	**0.17**	0.77	**0.10**	0.96	**0.32**	1.00	**0.98**	0.92	**0.90**

(b) Retain ROUGE-L ↑

	NPO		GradDiff		LINGTEA*		SimNPO		RMU	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	**0.97**	0.87	**0.98**	0.79	**0.99**	0.62	0.99	**1.00**	0.99	**1.00**
DE	**0.95**	**0.95**	**0.95**	0.92	**0.96**	0.94	0.97	**0.98**	0.96	**0.98**
ES	**0.96**	**0.96**	**0.96**	0.95	**0.96**	0.95	**0.96**	**0.96**	0.96	**0.97**
FR	**0.96**	0.95	**0.97**	0.92	**0.98**	0.97	0.98	**0.99**	0.97	**1.00**
VI	**0.96**	0.95	**0.97**	0.94	**0.98**	0.95	0.98	**0.99**	0.98	**0.99**
HI	**0.72**	**0.72**	**0.72**	0.68	**0.73**	0.72	0.74	**0.78**	0.73	**0.78**
TH	**0.95**	**0.95**	**0.97**	0.91	**0.97**	0.95	**0.98**	**0.98**	**0.97**	**0.97**
ZH	**0.96**	0.95	**0.97**	0.94	**0.97**	0.93	0.97	**1.00**	0.97	**1.00**
JA	0.98	**0.99**	0.97	**0.98**	**0.99**	0.98	**0.99**	**0.99**	0.97	**0.99**
KO	**0.96**	0.95	**0.96**	0.94	**0.96**	0.95	0.96	**0.98**	0.95	**0.99**

ROUGE-L results. Table 11 reports forget and retain ROUGE-L on TOFU. With NPO and GradDiff, LATTE achieves lower forget ROUGE-L than standard fine-tuning in all ten languages, consistent with the NLI results in Table 10. With SimNPO and RMU, retain ROUGE-L is equal or higher in every language at the displayed precision, while forget ROUGE-L does not consistently improve.

Table 12: **English-only unlearning on TOFU with Llama-3.2-3B-Instruct.** Plain denotes standard fine-tuning. Before gives scores prior to unlearning. Bold indicates the better post-unlearning score within each method pair, including ties at the displayed precision. Alignment targets appear below the separator.

(a) Forget NLI ↓

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.96	1.00	0.26	**0.15**	0.22	**0.09**	0.30	0.14
DE	0.93	0.97	0.24	**0.17**	**0.16**	0.21	0.28	0.17
ES	0.97	0.99	**0.27**	**0.27**	0.34	**0.29**	0.38	0.09
FR	0.95	1.00	**0.21**	0.25	**0.29**	0.30	0.25	0.13
HI	0.87	0.91	0.26	**0.16**	0.31	**0.15**	0.31	0.14
TH	0.91	0.97	**0.24**	0.25	**0.20**	**0.20**	0.29	0.17

(b) Retain NLI ↑

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.98	0.99	0.79	**0.83**	0.78	**0.81**	0.75	0.95
DE	0.94	0.98	0.83	**0.89**	0.83	**0.90**	0.83	0.96
ES	0.95	0.97	0.87	**0.93**	0.87	**0.92**	0.82	0.94
FR	0.95	0.98	0.90	**0.92**	0.90	**0.92**	0.89	0.96
HI	0.88	0.92	0.78	**0.81**	0.77	**0.80**	0.72	0.86
TH	0.93	0.95	0.76	**0.85**	0.75	**0.82**	0.63	0.86

Table 13: **ROUGE-L results on TOFU with Llama-3.2-3B-Instruct.** Formatting follows Table 12.

(a) Forget ROUGE-L ↓

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.97	1.00	0.43	**0.34**	0.38	**0.32**	0.68	0.51
DE	0.91	0.97	0.35	**0.34**	**0.33**	0.36	0.64	0.49
ES	0.98	0.99	**0.45**	0.47	**0.43**	0.47	0.79	0.53
FR	0.95	1.00	**0.47**	0.48	**0.44**	0.47	0.76	0.53
HI	0.86	0.92	**0.41**	0.45	**0.41**	0.43	0.77	0.54
TH	0.90	0.96	**0.46**	0.54	**0.47**	0.54	0.78	0.53

(b) Retain ROUGE-L ↑

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	0.98	0.99	0.84	**0.90**	0.84	**0.87**	0.91	0.97
DE	0.94	0.98	0.87	**0.91**	0.87	**0.90**	0.90	0.97
ES	0.95	0.97	0.88	**0.93**	0.88	**0.93**	0.91	0.96
FR	0.95	0.98	0.92	**0.95**	0.91	**0.95**	0.93	0.98
HI	0.87	0.93	0.81	**0.86**	0.81	**0.85**	0.84	0.92
TH	0.92	0.96	0.89	**0.92**	0.88	**0.92**	0.91	0.95

A.5.2 TOFU Results on Llama-3.2-3B-Instruct

Tables 12 and 13 report forget and retain NLI and ROUGE-L on TOFU for Llama-3.2-3B-Instruct.

A.5.3 LUME Results on Qwen2.5-3B-Instruct

Table 14 reports ROUGE-L scores, complementing the NLI results in Table 3. Compared with standard fine-tuning, LATTE achieves higher retain ROUGE-L in all ten languages under all three methods. Forget ROUGE-L is lower in nine languages under NPO, all ten under GradDiff, and eight under LINGTEA*. These results show stronger forgetting in most languages alongside consistently improved retention.

A.6 Evaluation Details

We evaluate forgetting and retention on both TOFU and LUME using NLI scores and ROUGE-L. Scores are computed separately for each language and split. Standard fine-tuning and LATTE use the same evaluation examples and generation settings. Results are averaged across unlearning runs as described in Appendix A.3.

A.6.1 Response Generation

We format each evaluation prompt using the corresponding model's chat template and generate a response using greedy decoding, with a maximum of 128 new tokens. Generated text is truncated at the first conversation-turn boundary before scoring. The generated response is then compared with the corresponding target response.

For both benchmarks, we evaluate all 40 forget examples and 200 retain examples in each reported language. All evaluation examples were included in fine-tuning. Dataset-specific subset construction is described in Appendix A.1.

Table 14: **ROUGE-L results on LUME with Qwen2.5-3B-Instruct.** Scores are means over five seeds. Plain denotes standard fine-tuning. Bold indicates the better score within each method pair, including ties at the displayed precision. Alignment targets appear below the separator.

(a) Forget ROUGE-L ↓

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	1.00	1.00	0.52	**0.50**	0.51	**0.50**	**0.40**	0.41
DE	1.00	1.00	0.61	**0.49**	0.51	**0.50**	0.45	**0.29**
ES	1.00	1.00	**0.63**	0.66	0.64	**0.62**	0.47	**0.44**
FR	1.00	1.00	0.59	**0.58**	0.64	**0.63**	0.43	**0.39**
VI	1.00	1.00	0.65	**0.64**	0.63	**0.59**	0.47	**0.41**
HI	0.77	0.77	0.67	**0.54**	0.68	**0.54**	0.56	**0.37**
TH	1.00	1.00	0.81	**0.71**	0.77	**0.70**	0.64	**0.46**
ZH	1.00	1.00	0.67	**0.61**	0.71	**0.58**	**0.54**	**0.54**
JA	0.99	1.00	0.79	**0.75**	0.75	**0.74**	0.62	**0.54**
KO	1.00	1.00	0.81	**0.68**	0.83	**0.68**	0.51	**0.35**

(b) Retain ROUGE-L ↑

	Before		NPO		GradDiff		LINGTEA*	
	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
EN	1.00	1.00	0.68	**0.86**	0.62	**0.79**	0.68	**0.81**
DE	1.00	1.00	0.53	**0.73**	0.50	**0.68**	0.53	**0.70**
ES	1.00	1.00	0.60	**0.84**	0.55	**0.82**	0.61	**0.79**
FR	1.00	1.00	0.63	**0.81**	0.58	**0.78**	0.62	**0.82**
VI	1.00	1.00	0.61	**0.82**	0.58	**0.78**	0.62	**0.79**
HI	0.78	0.78	0.66	**0.73**	0.65	**0.71**	0.65	**0.83**
TH	0.99	1.00	0.71	**0.85**	0.68	**0.82**	0.67	**0.75**
ZH	1.00	1.00	0.60	**0.82**	0.58	**0.79**	0.63	**0.78**
JA	1.00	1.00	0.73	**0.91**	0.68	**0.90**	0.70	**0.83**
KO	1.00	1.00	0.67	**0.88**	0.64	**0.87**	0.68	**0.80**

A.6.2 NLI SCORES

Following Xiang et al. (2026), we use the multilingual NLI evaluator `joeddav/xlm-roberta-large-xnli` to compare generated and target responses. The evaluator assigns probabilities to entailment, contradiction, and neutrality for an ordered pair of texts.

For a generated response a and target response b, the score is

$$S(a,b) = \frac{P_{\mathrm{E}}(a,b) + P_{\mathrm{E}}(b,a)}{2} \left(1 - P_{\mathrm{C}}(a,b)\right) \left(1 - P_{\mathrm{N}}(a,b)\right), \tag{7}$$

where P_{E}, P_{C}, and P_{N} denote the probabilities of entailment, contradiction, and neutrality, respectively. Entailment is evaluated in both directions. The contradiction and neutrality terms use the generated response as the first text and the target response as the second. The evaluator uses a maximum sequence length of 256 tokens.

Within each run, we average the scores over examples separately for each language and split. Forget NLI measures agreement with target responses for facts selected for unlearning, so lower scores indicate stronger forgetting. Retain NLI measures agreement with target responses for retained facts, so higher scores indicate better retention.

The transfer analysis in Section 3 uses the proportional decrease in forget NLI relative to the corresponding decrease in English, as defined in Equation (1). Elsewhere, we report the original NLI scores.

A.6.3 ROUGE-L

We also compare generated and target responses using ROUGE-L, which measures overlap through their longest common subsequence. Within each run, we compute the score for each example and average separately over the forget and retain sets in each language. Lower forget ROUGE-L indicates less reproduction of the target responses, while higher retain ROUGE-L indicates better preservation of retained responses.

A.7 ANCHOR LANGUAGES

We compare the Chinese- and German-anchor NPO configurations in Tables 16 and 17 with the English-anchor NPO results in Table 2. All results use Qwen2.5-3B-Instruct on TOFU. As shown in Table 15, English achieves the largest mean reduction in forget NLI, at 0.77 compared with 0.21 for Chinese and 0.18 for German. This advantage also holds across the nine non-anchor languages, where the mean reductions are 0.75, 0.16, and 0.14, respectively. English therefore produces the strongest overall forgetting in these configurations, supporting its use as the anchor language in our main experiments.

Table 15: **Forgetting under different anchor configurations.** Mean forget NLI for LATTE across ten languages, computed from the reported scores. Reduction is the before-unlearning score minus the after-unlearning score.

Anchor	Before	After ↓	Reduction ↑
English	0.93	**0.16**	**0.77**
Chinese	0.89	0.68	0.21
German	0.90	0.72	0.18

Table 16: **Chinese-anchor results on TOFU.** Plain denotes standard fine-tuning. Bold indicates the better score after unlearning within each pair, including ties at the displayed precision.

	Forget NLI ↓				Retain NLI ↑			
	Before		After		Before		After	
Language	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
ZH (anchor)	0.91	0.97	0.50	**0.31**	0.96	0.98	**0.91**	0.87
HI	0.34	0.38	0.31	**0.27**	0.51	0.51	**0.49**	0.48
TH	0.83	0.83	0.76	**0.62**	0.96	0.91	**0.93**	0.85
JA	0.96	0.89	0.73	**0.58**	0.98	0.97	**0.96**	0.89
KO	1.00	0.97	0.77	**0.61**	0.97	0.97	**0.96**	0.92
EN	0.99	1.00	**0.82**	0.88	0.98	0.98	**0.98**	**0.98**
DE	0.93	0.95	**0.80**	0.87	0.99	0.95	**0.97**	0.94
ES	0.97	0.96	**0.82**	0.90	0.96	0.94	**0.95**	0.93
FR	1.00	0.99	**0.83**	0.97	0.97	0.98	**0.96**	**0.96**
VI	0.95	0.95	0.89	**0.76**	0.97	0.96	**0.96**	0.95

A.8 Preservation Weights

We evaluate the effect of preservation strength using Qwen2.5-3B-Instruct with NPO. We vary λ_{anc} and λ_{KL} separately over $\{0.25, 1.0, 4.0\}$, keeping the other coefficient at its default of 1.0 and holding the unlearning settings fixed.

Table 18 shows that reducing either weight to 0.25 weakens forgetting in every alignment target language. Increasing λ_{anc} to 4.0 also weakens forgetting across all five languages, with comparatively small changes in retention. Increasing λ_{KL} to 4.0 improves forgetting in four languages but lowers retain NLI in every language. The default weights therefore provide a favorable balance, achieving substantially stronger forgetting than the lower-weight configurations while maintaining higher retention than the larger KL weight.

Table 17: **German-anchor results on TOFU with NPO.** Plain denotes standard fine-tuning. Bold indicates the better score after unlearning within each pair, including ties at the displayed precision.

	Forget NLI ↓				Retain NLI ↑			
	Before		After		Before		After	
Language	Plain	LATTE	Plain	LATTE	Plain	LATTE	Plain	LATTE
DE (anchor)	0.93	1.00	**0.47**	0.48	0.99	0.98	**0.94**	0.92
HI	0.34	0.40	**0.30**	0.35	0.51	0.49	**0.48**	0.47
TH	0.83	0.89	0.78	**0.72**	0.96	0.95	0.91	**0.93**
ZH	0.91	0.96	0.87	**0.83**	0.96	0.98	0.95	**0.96**
JA	0.96	0.92	0.93	**0.80**	0.98	0.95	**0.97**	0.93
KO	1.00	0.92	0.92	**0.84**	0.97	0.97	**0.94**	0.92
EN	0.99	0.97	**0.76**	0.77	0.98	0.98	**0.98**	0.96
ES	0.97	0.99	0.86	**0.78**	0.96	0.94	**0.95**	0.92
FR	1.00	0.97	0.85	**0.82**	0.97	0.97	**0.96**	0.94
VI	0.95	0.95	0.87	**0.81**	0.97	0.99	0.95	**0.97**

Table 18: **Effect of preservation weights.** F denotes forget NLI and R denotes retain NLI. Lower F and higher R are better.

		Hindi		Thai		Chinese		Japanese		Korean	
λ_{anc}	λ_{KL}	F↓	R↑	F↓	R↑	F↓	R↑	F↓	R↑	F↓	R↑
1.0	1.0	0.08	0.46	0.09	0.88	0.33	0.95	0.38	0.97	0.15	0.94
0.25	1.0	0.28	0.50	0.53	0.90	0.70	0.93	0.73	0.89	0.68	0.90
1.0	0.25	0.27	0.45	0.54	0.89	0.73	0.92	0.82	0.91	0.74	0.91
4.0	1.0	0.31	0.50	0.67	0.86	0.77	0.96	0.80	0.97	0.81	0.96
1.0	4.0	0.12	0.44	0.12	0.75	0.15	0.85	0.33	0.89	0.11	0.88