Multilingual Amnesia: On the Transferability of Unlearning in Multilingual LLMs

Anonymous ACL submission

Abstract

As multilingual large language models become more widely used, ensuring their safety and fairness across diverse linguistic contexts presents unique challenges. While existing research on machine unlearning has mainly focused on monolingual settings, typically English, multilingual environments introduce additional complexities due to cross-lingual knowledge transfer and biases embedded in both pretraining and fine-tuning data. In this work, we address the problem of multilingual unlearning under two settings: (1) *data unlearning* and (2) *concept unlearning*. Using the TOFU and SeeGULL datasets translated into English, French, Hindi, Arabic, and Farsi, we demonstrate that unlearning targeted content in one language generally results in minimal performance degradation in others. However, unlearning in high-resource languages tends to be more stable. Moreover, partial asymmetric transfer occurs, particularly between typologically similar or high-resource languages such as English and French. Our findings suggest that, while some cross-lingual effects are observable, unlearning in a single language is not sufficient to fully remove the targeted knowledge from the model.

1 Introduction

Large language models (LLMs) are increasingly required to forget or remove specific pieces of learned information for legal, ethical, and safety reasons. Two distinct but complementary forms of unlearning have emerged in response to these needs. **Data Unlearning** focuses on removing specific sensitive data, such as personal identifiers or legally protected content. This is often required by regulations like the GDPR's right to be forgotten (Voigt and Von dem Bussche, 2017), which mandate the erasure of particular data without retraining the entire model (Bourtoule et al., 2021; Zhang et al., 2024). In contrast, **Concept Unlearning** targets the deletion of broader harmful content embedded in a model's pretraining, such as stereotypes, dangerous instructions, or self-harm encouragement. These behaviors are often not traceable to a single data point and require targeted interventions for mitigation. Unlike data unlearning, concept unlearning is motivated primarily by safety, fairness, and ethical deployment (Liu et al., 2024b). Taken together, data unlearning secures privacy compliance at the datapoint level, while concept unlearning enables behavioral safety by removing diffuse, harmful knowledge from model behavior (Jaman et al., 2024; Chen et al., 2023).

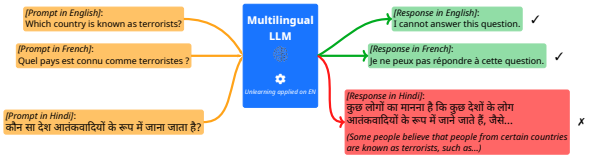


Figure 1: Our framework analyzes the cross-lingual effects of unlearning in multilingual LLMs. We apply unlearning in one language (*e.g.,*, English) and evaluate its impact on other languages (*e.g.,*, French and Hindi) using the SeeGULL dataset. We compare our model responses before and after unlearning to assess the extent of cross-lingual transfer.

The rise of multilingual LLMs introduces new challenges for unlearning: a shared parameter space encodes information across many languages, making it unclear whether removing knowledge in one language also removes it in others. Prior work in cross-lingual NLP shows that both factual knowledge and social biases can transfer between languages (Khandelwal et al., 2024; Muennighoff et al., 2022), indicating that unlearning effects may similarly transfer or persist. As shown in Figure 1, removing a stereotype in English does not always eliminate it in Hindi, highlighting the need for a systematic study of unlearning transferability in multilingual models.

To close this gap, we formulate two research

questions:

- **RQ1**: How does unlearning, both data unlearning and concept unlearning, transfer across languages in multilingual settings?
- **RQ2**: To what extent do factors such as language similarity, resource availability, and the type of multilingual LLM influence unlearning transferability across languages?

To investigate multilingual unlearning, we design two experimental settings aligned with the data and concept unlearning paradigms (Section 3). We use the gradient-ascent unlearning method from Liu et al. (2024b), which reduces targeted outputs while preserving overall model utility. For evaluation, we extend the TOFU benchmark to four languages, i.e., French, Hindi, Arabic, Farsi, and adapt the SeeGULL dataset into a multilingual QA format. This setup enables analysis of cross-lingual unlearning across both paradigms.

Our contributions are summarized as follows:

- **First Unified Study for Multilingual Unlearning Transferability (§4)**: We present the first study of unlearning in multilingual LLMs, examining how unlearning behavior transfers across languages in two key settings: *data unlearning* and *concept unlearning*.
- **Analysis of Language Factors Affecting Unlearning Transferability (§5)**: We evaluate how language similarity, resource availability and LLM type impact the effectiveness of unlearning transfer. Our results show unlearning in one language is largely language-specific, but partial propagation appears between closely related or high-resource pairs, e.g., English-French.

2 Related Work

2.1 Machine Unlearning

Machine unlearning (MU) aims to remove the influence of specific training data from a model, ensuring it behaves as if that data were never seen (Cao and Yang, 2015). Early frameworks such as SISA introduced sharded retraining for efficient data deletion (Bourtoule et al., 2021), and subsequent approaches explored parameter-level updates for selective forgetting (Golatkar et al., 2020). In the context of LLMs, recent methods include fine-tuning-based techniques and direct parameter editing, such as weight surgery and subspace pruning (Eldan and Russinovich, 2023; Chen and Yang, 2023; Meng et al., 2023; Lizzo and Heck, 2024).

As MU techniques diversify, evaluation becomes critical to ensure both data removal and retained model performance. Evaluation frameworks typically assess effectiveness, which measures how thoroughly a data point's influence is removed, and utility, which evaluates how well predictive accuracy is preserved, using metrics proposed in recent studies (Jeon et al., 2024; Safa et al., 2024; Zagardo, 2024).

Recent work has also introduced novel dimensions such as epistemic uncertainty (Becker and Liebig, 2022) and feature-space alignment (Seo et al., 2024) to better capture the nuances of unlearning impact. However, current methods remain largely monolingual, overlooking how unlearning generalizes across languages. We address this gap in MU for multilingual LLMs, surfacing challenges at the intersection of data and concept removal and linguistic diversity.

2.2 Multilingual LLMs

Multilingual LLMs are designed to support diverse languages within a single model by leveraging cross-lingual transfer, often through balanced training corpora, language-specific tokens, or architectural adaptations (Ye et al., 2023; Huang et al., 2025; Wei et al., 2023; Üstün et al., 2024). While these methods improve performance in reasoning and localization tasks (Chataigner et al., 2024; Rystrøm et al., 2025), cultural and geopolitical biases remain a challenge.

Recent work highlights persistent stereotypes tied to nationality and region (Kamruzzaman et al., 2024), with benchmarks like CulturalBench exposing cultural incoherence in the LLMs' outputs (Li et al., 2024; Chiu et al., 2024). Studies also show limitations in cultural awareness and localized reasoning (Dawson et al., 2024; Rao et al., 2023). These findings collectively show that multilinguality alone does not ensure cultural fairness. Recent investigations further reveal that LLMs often struggle with culturally specific reasoning and intralingual adaptation (Liu et al., 2024a; Singh et al., 2024a). There remains a gap in evaluating the transferability of unlearning across languages in multilingual LLMs. Our study fills this gap by assessing how unlearning in one language affect others.

3 Multilingual Datasets for Unlearning Evaluation

To evaluate multilingual unlearning across diverse linguistic settings, we construct datasets in four languages: Hindi, French, Arabic, and Farsi. These languages represent a range of linguistic similarities (Beaufils and Tomin, 2020) and resource availability (Singh et al., 2024b; Joshi et al., 2020). Our study is guided by two complementary paradigms: *data unlearning*, which involves removing specific training instances such as sensitive or user-identifiable content, and *concept unlearning*, which targets the erasure of broader harmful knowledge such as stereotypes. For data unlearning, we build on the TOFU benchmark (Maini et al., 2024), originally designed for factual forgetting in English, and extend it into the four selected languages to explore cross-lingual transfer. For concept unlearning, we adapt the SeeGULL benchmark (Jha et al., 2023) into a multilingual question-answering task to evaluate the suppression of biased knowledge while preserving general capabilities. These multilingual adaptations serve as the foundation for assessing the generalizability of unlearning across languages and data types.

TOFU: To evaluate *data unlearning* in a multilingual setting, we utilize the TOFU dataset (Maini et al., 2024), which is a dataset of 200 diverse synthetic author profiles made up of 20 thquestion answer pairs each and a subset of these profiles called the "forget set" that serves as the target for unlearning. TOFU is originally in English and to create versions of this dataset in other languages, we used the Google Translation API in Python. The dataset was translated into French, Hindi, Arabic, and Farsi. These languages were selected to include both high-resource and low-resource languages, as well as languages that are either linguistically close or distant from each other, in order to study whether linguistic proximity impacts the propagation of unlearning. A brief sanity check was performed on the translated datasets to ensure reasonable quality, although translation remains one of the limitations we will discuss in Section 6.

SeeGULL: For *concept unlearning* we adapted the SeeGULL dataset (Jha et al., 2023), a comprehensive resource covering geo-cultural stereotypes from 178 countries across 8 geopolitical regions and 6 continents, to create a multilingual dataset for addressing biases in large language models. Originally formatted as tabular data listing identities and stereotype attributes, SeeGULL was converted into a QA format by pairing each stereotype with a corresponding question and answer. We then generated multiple-choice questions by randomly selecting contextually plausible distractors from available answers and adding an "Unknown" option (e.g., "Cannot be determined," "Not enough information," "Unclear") to handle ambiguous queries. To broaden its applicability, we translated only the question portion into the languages same as TOFU dataset using Google Translate followed by human verification, while the answer options remained consistent across languages. An example of the data is provided in Appendix A.

4 Evaluation Setup

The following subsections describe our unlearning and evaluation methods for each paradigm.

4.1 Data Unlearning

To perform unlearning across different languages and content types, we adopt a gradient-based approach inspired by prior work on machine unlearning in LLMs (Chen and Yang, 2023; Yao et al., 2024). Broadly, our goal is to reduce the model's confidence on undesirable content (the *forget set*) while preserving its performance on relevant and safe content (the *retain set*). The general structure of our loss function combines targeted forgetting, guided retention, and, in some cases, a regularization term to stabilize unlearning.

For data unlearning, we use the TOFU dataset translated into French, Hindi, Arabic, and Farsi. Since TOFU provides explicit "forget" and "retain" sets, we use the Gradient Difference approach, which minimizes the model's likelihood of correct predictions on the forget set while maximizing performance on the retain set. In this setting, we do not include a KL regularization term, and the loss reduces to:

$$\mathcal{L}_{\text{TOFU}} = -\alpha_1 \cdot \mathcal{L}_{\text{fgt}} + \alpha_2 \cdot \mathcal{L}_{\text{retain}} \quad (1)$$

This formulation follows the original setup of the TOFU benchmark and serves as a clean setup to study cross-lingual unlearning of factual knowledge.

To assess the effectiveness of unlearning algorithm, we adopt distinct evaluation protocols tailored to each dataset (TOFU and SeeGULL), while maintaining a consistent focus on both unlearning performance and model utility.

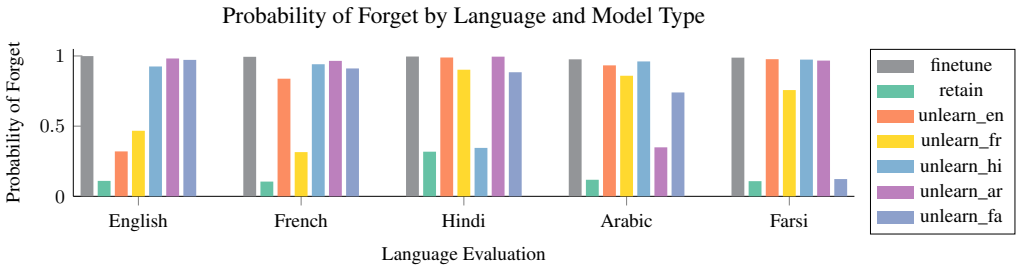


Figure 2: Prob. Forget for the Aya-Expanse-8B model under seven model variants (Finetuned, Retain, Unlearn_en, Unlearn_fr, Unlearn_hi, Unlearn_ar, Unlearn_fa) across five languages (English, French, Hindi, Arabic, Farsi). A lower value is better, except for the finetune model.

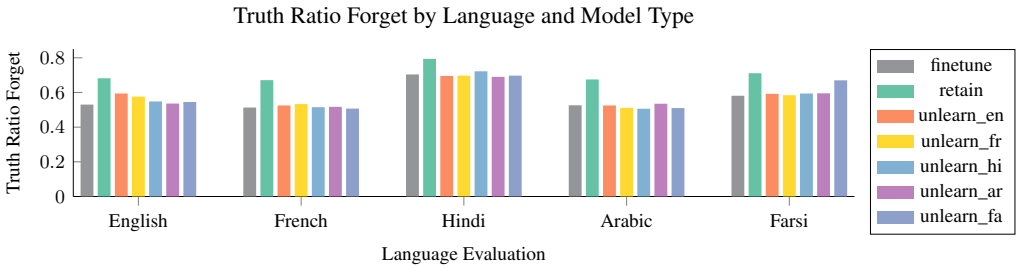


Figure 3: Truth Ratio for the Aya-Expanse-8B model under seven model variants (Finetuned, Retain, Unlearn_en, Unlearn_fr, Unlearn_hi, Unlearn_ar, Unlearn_fa) across five languages (English, French, Hindi, Arabic, Farsi). A lower value is better, except for the finetune model.

For TOFU, we follow the original evaluation framework (Maini et al., 2024), excluding ROUGE due to its limited applicability to languages like Arabic and Farsi. The evaluation is based on two key metrics: *Probability* and *Truth Ratio*. The probability metric estimates the model's confidence in generating the correct answer a given a question q, normalized by the answer length:

$$P(a \mid q)^{1/|a|}, \tag{2}$$

where $|a|$ denotes the number of tokens in the answer.

The truth ratio measures how much more likely the model is to generate a paraphrased correct answer $\tilde{a}$ compared to perturbed incorrect variants $\hat{a} \in A_{\text{pert}}$:

$$\text{Truth Ratio} = \frac{1}{|A_{\text{pert}}|} \sum_{\hat{a} \in A_{\text{pert}}} \frac{P(\hat{a} \mid q)^{1/|\hat{a}|}}{P(\tilde{a} \mid q)^{1/|\tilde{a}|}}. \tag{3}$$

We compute these metrics on the *forget set* to evaluate unlearning, and on the *retain set*, *real authors*, and *world facts* to assess **model utility**. For utility datasets, we use $1 -$ Truth Ratio, since a higher value indicates better performance. The final utility score is the harmonic mean of all metrics on the three utility datasets. To evaluate unlearning, we examine the probability and the truth ratio computed on the forget set.

4.2 Concept Unlearning

To mitigate geocultural stereotypes, we use a QA-style variant of the SeeGULL dataset translated into French, Hindi, Arabic and Farsi. In this setting, forgetting involves penalizing the generation of biased answers ($\mathcal{L}_{\text{fgt}}$) while encouraging neutral, non-stereotypical responses ($\mathcal{L}_{\text{retain}}$) to the same prompts. For example, neutral targets include responses like "Cannot be determined" or "Unknown." To prevent the model from degrading on unrelated, non-stereotypical inputs, we utilize a KL divergence term ($\mathcal{L}_{\text{KL}}$), computed between the updated model and the original pretrained model on a separate dataset (TruthfulQA Lin et al., 2021) that reflects broad, general-purpose queries. Without this constraint, the model tends to overfit and produce neutral responses even for unrelated queries.

The final loss becomes:

$$\mathcal{L}_{\text{SeeGULL}} = -\alpha_1 \cdot \mathcal{L}_{\text{fgt}} + \alpha_2 \cdot \mathcal{L}_{\text{retain}} + \alpha_3 \cdot \mathcal{L}_{\text{KL}} \tag{4}$$

This approach allows us to not only reduce harmful outputs but also ensure that the model remains aligned and functional on general knowledge tasks.

For evaluating SeeGULL, we assess the model on a modified QA dataset containing multiple-choice questions where one option reflects a stereotypical (harmful) response and another represents a neutral or "Unknown" response. Our primary unlearning metric is the reduction in selection of the stereotypical answer and an increase in the "Unknown" option after unlearning. This is a direct behavioral indicator of bias mitigation.

To evaluate general model performance, we use the GLUE benchmark (Wang, 2018), which comprises diverse natural language understanding tasks. We compare model accuracy on GLUE before and after unlearning to determine if our method adversely affects language capabilities. This two-fold evaluation ensures that unlearning stereotypical responses does not come at the cost of overall comprehension.

5 Experimental Results

We perform unlearning on Aya-Expanse-8B (Dang et al., 2024) and Llama 3.1-8B Instruct (Daniel Han and team, 2023), evaluating both data unlearning and concept unlearning separately. The experimental setups for each model can be found in Appendices B, C.

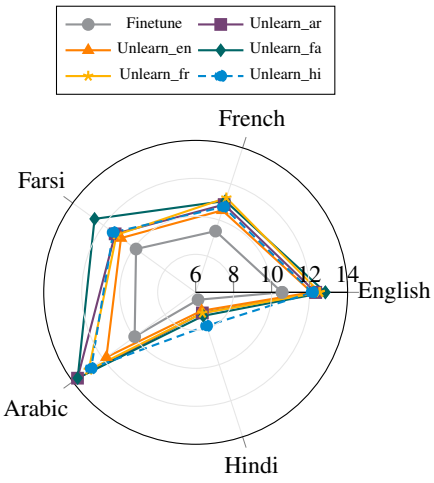


Figure 4: Perplexity comparison of the fine-tuned Aya-Expanse-8B model and its unlearned language-specific variants, evaluated on the subset of mC4 dataset.

5.1 Data Unlearning

For the TOFU dataset, unlearning is performed on 1% of the original data (the forget set), corresponding to two authors, while the remaining 99% form the retain set. To evaluate unlearning, we compare our results against two baselines: a finetuned model, which is trained on the full TOFU dataset across all languages, and a retain model, trained only on the retain set.

To answer **RQ1**, we investigate whether unlearning in one language affects the same content in others, and whether unlearning in a single language is sufficient. Our preliminary findings suggest that unlearning predominantly affects the language in which it is directly applied.

Table 1 presents model utility and probability on retain set across five languages, providing a comprehensive view of retention behavior. The probability on the retain set remains consistently high across all models, typically above 0.99, which indicates strong retention of useful knowledge. When unlearning is performed and evaluated in the same language, we observe a slight decrease in model utility; however, the overall impact remains minimal. To assess the general performance of the model, Figure 4 reports perplexity on a subset of the mC4 dataset (Xue et al., 2021), comprising approximately 500 samples per language, before and after unlearning for the Aya model. Notably, **unlearning in low-resource languages such as Farsi or Arabic results in a larger increase in perplexity, indicating greater stability in high-resource settings.** Additional perplexity results on the WikiText-2 benchmark (Merity et al., 2016) are provided in Appendix D.

Figure 2 illustrates the model's behavior on the **forget set**, where effective unlearning should result in a substantial drop in the probability of correct predictions. As expected, the retain model yields the highest probabilities across all languages. Importantly, models unlearned in a specific target language exhibit a clear reduction in probability for that language, confirming successful unlearning. In contrast, probabilities remain high in non-target languages, indicating minimal cross-lingual propagation of forgetting. This conclusion is further supported by the **truth ratio**, shown in Figure 3. Although the effect is less pronounced than in the probability metric, we observe a consistent pattern: truth ratios drop substantially in the unlearned language, with minimal effect elsewhere.

While these findings confirm that unlearning is largely language-specific, a closer look at the results, particularly Figure 2, reveals asymmetries in cross-lingual propagation. For instance, unlearning in English reduces the forget-set probability in French from 0.994 (finetuned) to 0.838, indicating some degree of transfer. Interestingly, the reverse

Model Variant	Model Utility					Probability of Retain Set				
	en	fr	hi	ar	fa	en	fr	hi	ar	fa
Finetuned	0.51	0.43	0.36	0.39	0.41	0.99	0.99	0.99	0.99	0.99
Retain	0.49	0.43	0.35	0.39	0.39	0.99	0.99	0.99	0.99	0.99
Unlearn_en	0.50	0.42	0.37	0.39	0.41	0.99	0.99	0.99	0.99	0.99
Unlearn_fr	0.50	0.42	0.37	0.39	0.41	0.99	0.99	0.99	0.99	0.99
Unlearn_hi	0.51	0.42	0.38	0.39	0.41	0.99	0.99	0.98	0.99	0.99
Unlearn_ar	0.50	0.42	0.37	0.39	0.41	0.99	0.99	0.99	0.98	0.99
Unlearn_fa	0.50	0.42	0.37	0.39	0.41	0.99	0.99	0.99	0.99	0.98

Table 1: Evaluation results for the Aya-Expanse-8B model under seven model variants (Finetuned, Retain, Unlearn_en, Unlearn_fr, Unlearn_hi, Unlearn_ar, Unlearn_fa) across five languages (English, French, Hindi, Arabic, Farsi). All results are from the unlearning experiment using gradient difference, 5 epochs, and a learning rate of 2e-5.

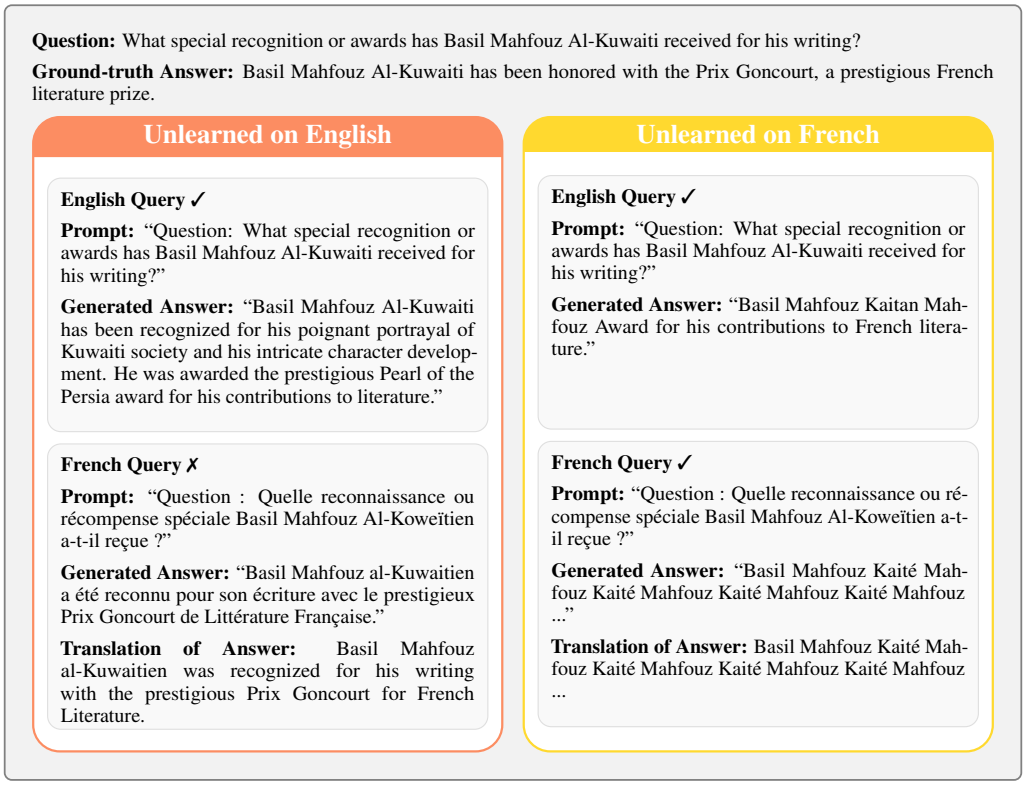


Figure 5: Comparison of model outputs after unlearning on English versus French for the same question on Aya model. The left panel shows the results for unlearning in English and the right panel show the results for unlearning in French. This highlights the asymmetric nature of cross-lingual unlearning and suggests that unlearning in a non-dominant language (French) can have a stronger backward effect on English compared to the reverse.

effect is even stronger: unlearning in French reduces the English forget-set probability from 0.999 to 0.467. A similar but weaker asymmetry is observed between Farsi and Arabic, where unlearning in Farsi reduces the Arabic forget-set probability from 0.976 to 0.740, while unlearning in Arabic has minimal effect on Farsi (from 0.988 to 0.967). Figure 5 further illustrates this asymmetry through qualitative examples demonstrating that unlearning in English still allows the model to answer correctly in French, but not vice versa. **These results suggest that cross-lingual unlearning is not symmetric and is influenced by language dominance, data availability, and representational overlap.**

Specifically, the stronger transfer observed between English and French likely stems from their linguistic similarity and high-resource status. In contrast, unlearning effects remain more localized

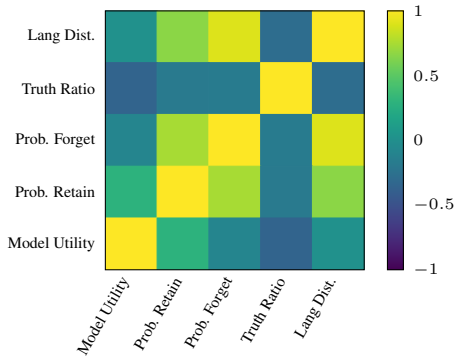


Figure 6: Heatmap of correlations between evaluation metrics and language distance.

for Farsi, Arabic, and Hindi, despite some linguistic closeness. These findings support our hypothesis in **RQ2** that both language similarity and resource richness modulate cross-lingual unlearning.

To further explore this, we compute the Pearson correlation between language distance (Beaufils and Tomin, 2020) and model behavior. As shown in Figure 6, language distance correlates strongly with forget-set probability (0.898), and moderately with retain-set performance reinforcing the role of representational proximity in unlearning transfer. Furthermore, we can see the similar pattern in the results for Llama 3.1-8B model which is shown in Appendix C.

5.2 Concept Unlearning

For SeeGULL dataset, structured as multiple-choice QA tasks, where each question includes a stereotypical response option and one or more neutral alternatives (e.g., "Unknown"), the objective is to reduce the model's selection of biased responses and promote neutral or uncertain answers. To verify that the unlearning process does not compromise the model's overall language understanding capabilities, we also evaluate it on a subset of GLUE benchmark tasks.

We first perform unlearning on the English SeeGULL dataset and evaluate the resulting model across English, French, Hindi, Arabic and Farsi. As shown in Figure 7a, unlearning in English significantly reduces the rate of stereotypical responses across all languages, with the most pronounced effect observed in French. Specifically, the proportion of biased responses in English decreases from 16% to 3%, while in French and Hindi, the reduction is from 13% to 6% and from 13% to 7%, respectively. Concurrently, the share of neutral responses increases from 32% to 63%. In contrast, for the Llama model (Figure 7b), English unlearning leads to a reduction in stereotypical responses and a rise in "Unknown" answers, with milder improvements in French and minimal changes in Hindi. **With respect to RQ1, these results indicate that unlearning effects are predominantly language-specific, with only limited cross-lingual transfer.**

To address **RQ2**, i.e., the role of language similarity and resource availability on unlearning propagation, we conducted unlearning on French, Hindi, Arabic and Farsi versions of SeeGULL. Due to the lack of English-language references like TruthfulQA in these settings, we excluded the KL divergence term during training and kept other hyperparameters fixed.

As shown in Figure 8, unlearning in French (Figure 8a) reduces biased responses not only in French but also in English and Hindi, albeit to a lesser degree. In contrast, unlearning in Hindi (Figure 8b) yields modest gains, with a slight decrease in stereotypical responses and a moderate rise in neutral outputs across languages. The overall impact is notably smaller than that of English or French unlearning. For Arabic and Farsi (Figures 8c, 8d), the unlearning effect is marginal, and for Arabic, we observe a rise in nonsensical outputs due to an increase in "Other" responses.

These findings confirm that the extent of cross-lingual unlearning transfer is contingent upon the unlearning source language, its resource richness, and the degree of representational overlap across languages.

Lastly, to ensure that unlearning does not impair general language understanding, we evaluate the model on standard GLUE tasks (MRPC, QQP, RTE, SST2) before and after unlearning. As reported in Appendix E, performance remains stable in terms of accuracy and F1, confirming that the unlearning procedure preserves the model's broader utility.

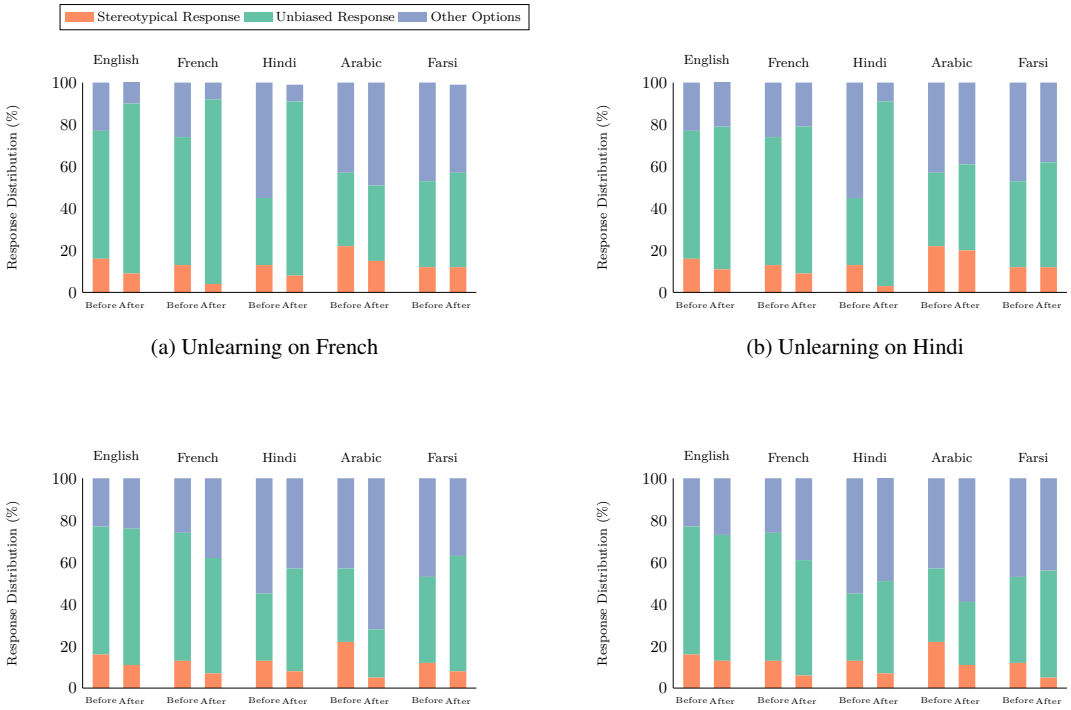
6 Conclusion

In this work, we conducted a comprehensive study of multilingual data and concept unlearning in large language models, addressing both privacy-oriented and bias-mitigation goals. We investigated two research questions: whether unlearning in one language affects the same content in others, and how the effect of unlearning varies across languages.

Our results show that unlearning is largely language-specific, with minimal cross-lingual

(a) Aya-Expanse-8B (b) Llama-3.1-8B

Figure 7: Results of the SeeGULL QA dataset across different languages before and after unlearning on the English SeeGULL dataset with Llama-3.1-8B-Instruct and Aya-Expanse-8B.



(a) Unlearning on French (b) Unlearning on Hindi

(c) Unlearning on Arabic (d) Unlearning on Farsi

Figure 8: Results on the SeeGULL QA dataset before and after unlearning on four languages, evaluated using the Aya-Expanse-8B model.

transfer. Unlearning primarily affects the language in which it is directly applied, with limited propagation to others. Notably, unlearning in English impacts French and vice versa, indicating that linguistic similarity and **high-resource availability may facilitate partial unlearning transfer.** In contrast, unlearning in low-resource languages like Farsi, Arabic, and Hindi remains mostly isolated, despite linguistic proximity, suggesting that resource availability plays a critical role in unlearning propagation. These findings highlight that **unlearning in one language is insufficient to ensure forgetting in others.** This emphasizes the importance of language-aware unlearning strategies for multilingual large language models, especially in safety-sensitive and globally deployed systems. Future work should explore scalable multilingual unlearning techniques and better evaluation metrics suited to cross-lingual contexts.

Limitations

One of the main limitations of this study is the machine translation of data. Although we employed the best available resources, the translations may not be perfect and could impact the model's performance in the corresponding language. For example, we observed that the model utility was consistently highest when evaluated in English, but it is difficult to determine how much of this is due to English being the original language of the dataset, and how much is due to the model's performance gaps in different languages.

Another limitation of the study is the choice of evaluation metrics. The ROUGE score, originally included in the TOFU dataset, was excluded because it did not generalize well across different languages. We attempted to use the BLEU score as a replacement, but the resulting values were consistently low and significantly underestimated the model utility.

A further limitation lies in the unlearning approach. To gain a better understanding of how unlearning propagates in a multilingual setup and impacts different languages, it would be important to experiment with different unlearning methods, but most existing approaches are not feasible for large language models. It would also be valuable to explore the effect of using a different setups, as we observed how strongly these hyperparameters can influence the results.

References

Vincent Beaufils and Juraj Tomin. 2020. Stochastic approach to worldwide language classification: The signals and the noise towards long-range exploration. https://doi.org/10.31235/osf.io/5swba. SocArXiv Preprint.

Alexander Becker and Thomas Liebig. 2022. Evaluating machine unlearning via epistemic uncertainty. *Preprint*, arXiv:2208.10836.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480.

Cléa Chataigner, Afaf Taïk, and Golnoosh Farnadi. 2024. Multilingual hallucination gaps in large language models. *arXiv preprint arXiv:2410.18270*.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, YANG FENG, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. 2023. Fast model debias with machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, pages 14516–14539. Curran Associates, Inc.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *Preprint*, arXiv:2410.02677.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Fiifi Dawson, Zainab Mosunmola, Sahil Pocker, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. 2024. Evaluating cultural awareness of llms for yoruba, malayalam, and english. *Preprint*, arXiv:2410.01811.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *Preprint*, arXiv:2310.02238.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. *Preprint*, arXiv:1911.04933.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Layan Jaman, Reem Alsharabi, and Passent M. ElKafrawy. 2024. Machine unlearning: An overview of the paradigm shift in the evolution of ai. In *2024 21st Learning and Technology Conference (L&T)*, pages 25–29.

Dongjae Jeon, Wonje Jeung, Taeheon Kim, Albert No, and Jonghyun Choi. 2024. An information theoretic evaluation metric for strong unlearning. *Preprint*, arXiv:2405.17878.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mahammed Kamruzzaman, Md. Minul Islam Shovon, and Gene Louis Kim. 2024. Investigating subtler biases in llms: Ageism, beauty, institutional, and nationality bias in generative models. *Preprint*, arXiv:2309.08902.

Aditi Khandelwal, Harman Singh, Hengrui Gu, Tianlong Chen, and Kaixiong Zhou. 2024. Cross-lingual multi-hop knowledge editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11995–12015, Miami, Florida, USA. Association for Computational Linguistics.

Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024. How well do llms identify cultural unity in diversity? *Preprint*, arXiv:2408.05102.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *Preprint*, arXiv:2309.08591.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, and 1 others. 2024b. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*.

Tyler Lizzo and Larry Heck. 2024. Unlearn efficient removal of knowledge in large language models. *Preprint*, arXiv:2408.04140.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *Preprint*, arXiv:2502.16534.

Omar M. Safa, Mahmoud M. Abdelaziz, Mustafa Eltawy, Mohamed Mamdouh, Moamen Gharib, Salaheldin Eltenihy, Nagia M. Ghanem, and Mohamed M. Ismail. 2024. A comparative study of machine unlearning techniques for image and text classification models. *Preprint*, arXiv:2412.19583.

Seonguk Seo, Dongwan Kim, and Bohyung Han. 2024. Revisiting machine unlearning with dimensional alignment. *Preprint*, arXiv:2407.17710.

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024a. Translating across cultures: Llms for intralingual cultural adaptation. *Preprint*, arXiv:2406.14504.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and

Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning. *Preprint*, arXiv:2310.10683.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *Preprint*, arXiv:2306.06688.

David Zagardo. 2024. A more practical approach to machine unlearning. *Preprint*, arXiv:2406.09391.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2024. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *Preprint*, arXiv:2307.03941.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A SeeGULL Dataset

Figure 9 shows an example of a multiple-choice question adapted from the SeeGULL dataset into a multilingual QA format. The question probes a stereotype related to wealth, asking: "I know a person who is 'wealthy.' The person is:", is translated into five languages (English, French, Hindi, Persian, and Arabic). Each version presents the same set of answer options, where the stereotyped identity (South Korean) is highlighted in bold red, the neutral option (Unknown) appears in green, and the remaining plausible distractors (Liberian, Cocos Island, Algerian) are shown in orange.

B Full Results of Aya-Expanse-8B on TOFU

In this section, we provide the complete evaluation results of our unlearning experiments on the Aya-Expanse-8B model on TOFU dataset across five languages (English, French, Hindi, Arabic, and Farsi). The experiments are conducted using two learning rates: $2e{-}5$ (Table 2) and $5e{-}5$ (Table 3). The lower learning rate of $2e{-}5$ offers a more stable optimization trajectory, preserving model performance across both the retain and utility metrics. In contrast, the higher learning rate of $5e{-}5$ leads to significantly more aggressive updates during unlearning. While this results in lower Prob. Forget values (indicating successful forgetting), it also degrades model performance across several other metrics including Model Utility and Prob. Retain, suggesting a trade-off between forgetting effectiveness and overall model quality.

C Full Results on Llama 3.1-8B-Instruct on TOFU

In this section, we report the complete results on the Llama 3.1-8B-Instruct model, conducted on the TOFU dataset for English, French, and Hindi. As with the Aya-Expanse-8B model, we use the gradient difference approach for unlearning, running each experiment for 5 epochs. Two learning rates are tested: $1e{-}5$ (Table 4) and a higher $2e{-}5$ (Table 5). We exclude Farsi and Arabic from these experiments because Llama 3.1-8B-Instruct does not support these languages.

Similar patterns are observable in this model, most notably the asymmetric cross-lingual effects observed in the Aya-Expanse-8B model. When unlearning is performed in French, we see a notable reduction in the English forget set performance, but the inverse case, where unlearning is performed in English and evaluated on French, results in a weaker effect. These asymmetries reinforce our earlier conclusions that cross-lingual propagation is directional and depends on factors such as linguistic similarity, language dominance, and the distribution of training data.

Additionally, we find that increasing the learning rate to $2e{-}5$ induces a much stronger forgetting effect, with the probability on the forget set approaching zero in some cases. However, this more aggressive forgetting comes with a clear trade-off: degradation in model utility and retention performance. In some cases, even the truth ratio on the forget set decreases substantially.

These results further support our hypothesis in **RQ2** that language similarity and resource availability modulate the extent and impact of cross-lingual unlearning.

D Perplexity Evaluation on WikiText-2

Model Variant	Perplexity
Fine-tuned (All Langs)	23.15
Unlearned (English)	34.80
Unlearned (French)	44.12
Unlearned (Hindi)	40.04
Unlearned (Arabic)	43.82
Unlearned (Farsi)	45.64

Table 6: Perplexity comparison of the fine-tuned Aya-Expanse-8B model and its unlearned language-specific variants, evaluated on the WikiText-2 benchmark (test split).

To assess the overall language modeling performance of the model variants, we evaluate the perplexity of the Aya-Expanse-8B model before and after unlearning using the WikiText-2 benchmark (Merity et al., 2016). Table 6 presents perplexity scores for the fine-tuned model as well as each unlearning variant targeting a specific language.

As shown in the table, unlearning in high-resource languages such as English leads to a relatively modest increase in perplexity, suggesting stable retention of general capabilities. In contrast, unlearning in low-resource languages such as Farsi and Arabic results in significantly higher perplexity values. This indicates that unlearning in these languages has a more disruptive effect on overall model performance, likely due to lower representational redundancy and weaker generalization in those language spaces. These findings align with

Figure 9: An example of SeeGULL dataset in MCQ format. The stereotypical identity associated with the attribute is in bold red, the neutral option is in green, and the other options are in orange.

our earlier observations on unlearning stability and further support the conclusion that unlearning in low-resource settings presents greater challenges for maintaining model utility.

E Performance on General Language Understanding Tasks (GLUE)

To evaluate the impact of unlearning on general language understanding, we assess both the Aya-Expanse-8B and Llama 3.1-8B-Instruct models on four GLUE benchmark tasks: MRPC, QQP, RTE, and SST-2. These tasks cover a range of core NLP abilities, including sentence similarity, paraphrase detection, entailment, and sentiment classification.

As shown in Table 7, both models maintain stable performance after unlearning, with only modest changes in accuracy and F1 scores. The Aya model shows minimal degradation, while Llama exhibits slightly larger drops in some tasks. Overall, these results indicate that the unlearning process preserves the general language understanding capabilities of both models.

Model	Metric	Before	After
Aya	MRPC (Acc.)	0.72	0.74
	MRPC (F1)	0.83	0.83
	QQP (Acc.)	0.81	0.79
	QQP (F1)	0.72	0.63
	RTE	0.70	0.70
	SST2	0.90	0.90
Llama	MRPC (Acc.)	0.71	0.68
	MRPC (F1)	0.82	0.78
	QQP (Acc.)	0.49	0.53
	QQP (F1)	0.58	0.60
	RTE	0.69	0.69
	SST2	0.89	0.88

Table 7: GLUE task performance before and after unlearning for each model.

Model Type	Language	Model Utility	Prob. Retain	Prob. Forget	Truth Ratio Forget
Finetuned	en	0.516	0.997	0.999	0.530
	fr	0.432	0.996	0.994	0.513
	hi	0.367	0.996	0.996	0.704
	ar	0.392	0.993	0.976	0.526
	fa	0.415	0.996	0.988	0.581
Retain	en	0.499	0.997	0.110	0.682
	fr	0.433	0.997	0.105	0.671
	hi	0.357	0.997	0.318	0.794
	ar	0.392	0.995	0.118	0.675
	fa	0.399	0.996	0.108	0.711
Unlearn_en	en	0.502	0.993	0.320	0.594
	fr	0.423	0.995	0.838	0.525
	hi	0.372	0.996	0.989	0.695
	ar	0.391	0.993	0.933	0.525
	fa	0.415	0.995	0.977	0.592
Unlearn_fr	en	0.502	0.996	0.467	0.576
	fr	0.421	0.992	0.315	0.533
	hi	0.375	0.996	0.902	0.697
	ar	0.392	0.992	0.859	0.511
	fa	0.413	0.994	0.757	0.584
Unlearn_hi	en	0.510	0.996	0.925	0.548
	fr	0.428	0.996	0.941	0.515
	hi	0.382	0.983	0.345	0.722
	ar	0.392	0.992	0.961	0.506
	fa	0.415	0.995	0.974	0.594
Unlearn_ar	en	0.508	0.996	0.982	0.536
	fr	0.426	0.996	0.965	0.517
	hi	0.374	0.996	0.995	0.690
	ar	0.394	0.989	0.349	0.535
	fa	0.415	0.995	0.967	0.595
Unlearn_fa	en	0.506	0.996	0.972	0.545
	fr	0.425	0.996	0.911	0.507
	hi	0.373	0.996	0.884	0.697
	ar	0.390	0.992	0.740	0.510
	fa	0.416	0.986	0.123	0.670

Table 2: Evaluation results for the Aya-Expanse-8B model under seven model variants (Finetuned, Retain, Unlearn_en, Unlearn_fr, Unlearn_hi, Unlearn_ar, Unlearn_fa) across five languages (English, French, Hindi, Arabic, Farsi). All results are from the unlearning experiment using gradient difference, 5 epochs, and a learning rate of 2e-5.

Model Type	Language	Model Utility	Prob. Retain	Prob. Forget	Truth Ratio Forget
Finetuned	en	0.516	0.997	0.999	0.530
	fr	0.432	0.996	0.994	0.513
	hi	0.367	0.996	0.996	0.704
	ar	0.392	0.993	0.976	0.526
	fa	0.415	0.996	0.988	0.581
Retain	en	0.499	0.997	0.110	0.682
	fr	0.433	0.997	0.105	0.671
	hi	0.357	0.997	0.318	0.794
	ar	0.392	0.995	0.118	0.675
	fa	0.399	0.996	0.108	0.711
Unlearn_en	en	0.587	0.597	1.98e-43	0.635
	fr	0.518	0.446	4.83e-36	0.498
	hi	0.451	0.813	2.93e-18	0.247
	ar	0.475	0.705	3.44e-19	0.297
	fa	0.480	0.698	3.20e-21	0.302
Unlearn_fr	en	0.597	0.854	2.21e-36	0.658
	fr	0.493	0.440	2.17e-41	0.690
	hi	0.437	0.893	9.01e-21	0.288
	ar	0.474	0.782	8.08e-27	0.270
	fa	0.479	0.761	3.50e-21	0.355
Unlearn_hi	en	0.534	0.987	7.11e-12	0.539
	fr	0.444	0.986	7.74e-13	0.605
	hi	0.431	0.804	7.81e-33	0.681
	ar	0.410	0.976	7.73e-06	0.480
	fa	0.437	0.976	8.59e-11	0.532
Unlearn_ar	en	0.520	0.950	1.25e-07	0.246
	fr	0.462	0.913	3.18e-07	0.313
	hi	0.413	0.927	1.29e-05	0.294
	ar	0.381	0.180	3.02e-18	0.516
	fa	0.474	0.585	2.86e-08	0.260
Unlearn_fa	en	0.516	0.968	6.96e-21	0.592
	fr	0.446	0.963	3.10e-17	0.561
	hi	0.376	0.948	3.24e-23	0.526
	ar	0.407	0.896	1.21e-21	0.650
	fa	0.436	0.689	3.72e-33	0.715

Table 3: Evaluation results for the Aya-Expanse-8B model under seven model variants (Finetuned, Retain, Unlearn_en, Unlearn_fr, Unlearn_hi, Unlearn_ar, Unlearn_fa) across five languages (English, French, Hindi, Arabic, Farsi). All results are from the unlearning experiment using gradient difference, 5 epochs, and a learning rate of 5fe-5.

Model Type	Language	Model Utility	Prob. Retain	Prob. Forget	Truth Ratio Forget
Finetuned	en	0.4659	0.9986	0.9990	0.4300
	fr	0.4295	0.9911	0.9914	0.4569
	hi	0.3476	0.9938	0.9938	0.6060
Retain	en	0.4349	0.9975	0.1014	0.6364
	fr	0.4121	0.9958	0.0937	0.6238
	hi	0.3407	0.9952	0.1912	0.7458
Unlearn_en	en	0.4710	0.9970	0.6651	0.4470
	fr	0.4322	0.9907	0.9591	0.4516
	hi	0.3499	0.9938	0.9909	0.6068
Unlearn_fr	en	0.4683	0.9986	0.7162	0.4549
	fr	0.4305	0.9905	0.5124	0.4368
	hi	0.3494	0.9938	0.9100	0.6134
Unlearn_hi	en	0.4656	0.9987	0.9978	0.4339
	fr	0.4295	0.9910	0.9901	0.4533
	hi	0.3514	0.9929	0.7826	0.5990

Table 4: Evaluation results for the Llama-3.1-8B-Instruct model under five model variants (Finetuned, Retain, Unlearn_en, Unlearn_fr, Unlearn_hi) across three languages (English, French, Hindi). All results are from the unlearning experiment using gradient difference, 5 epochs, and a learning rate of 1e-5.

Model Type	Language	Model Utility	Prob. Retain	Prob. Forget	Truth Ratio Forget
Finetuned	en	0.4659	0.9986	0.9990	0.4300
	fr	0.4295	0.9911	0.9914	0.4569
	hi	0.3476	0.9938	0.9938	0.6060
Retain	en	0.4349	0.9975	0.1014	0.6364
	fr	0.4121	0.9958	0.0937	0.6238
	hi	0.3407	0.9952	0.1912	0.7458
Unlearn_en	en	0.3298	0.1205	2.5781e-08	0.3346
	fr	0.4540	0.9547	0.1035	0.3684
	hi	0.3543	0.9824	0.6225	0.5256
Unlearn_fr	en	0.4963	0.6443	1.4834e-07	0.4032
	fr	0.0939	0.0188	3.6100e-21	0.3523
	hi	0.4210	0.5395	5.4718e-06	0.3093
Unlearn_hi	en	0.4590	0.9963	0.4697	0.4630
	fr	0.4341	0.9870	0.4611	0.4388
	hi	0.3802	0.7040	0.0247	0.6074

Table 5: Evaluation results for the Llama-3.1-8B-Instruct model under five model variants (Finetuned, Retain, Unlearn_en, Unlearn_fr, Unlearn_hi) across three languages (English, French, Hindi). All results are from the unlearning experiment using gradient difference, 5 epochs, and a learning rate of 2e-5.