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 using the Aya-Expanse 8B model under two settings: (1) *data unlearning* and (2) *concept unlearning*. We extend benchmarks for factual knowledge and stereotypes into ten languages through translation—English, French, Arabic, Japanese, Russian, Farsi, Korean, Hindi, Hebrew, and Indonesian—spanning five language families and varying resource levels. Our experiments show that unlearning in high-resource languages tends to be more stable, with asymmetric transfer observed between typologically related languages. Moreover, analysis of linguistic distances reveals that syntactic similarity is the most predictive factor of cross-lingual unlearning effects.

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., 2024a). 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 ensures privacy compliance for specific instances, while concept unlearning promotes broader behavioral safety (Jaman et al., 2024; Chen et al., 2023).

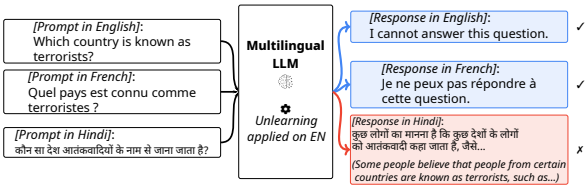


Figure 1: Framework for analyzing cross-lingual unlearning. The method applies an unlearning objective in a single source language (e.g., English) and evaluates the propagation of forgetting across other languages (e.g., French, Hindi) to measure transfer effects.

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), suggesting that unlearning effects may potentially transfer or persist similarly. 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. Recent work by Lu and Koehn (2025) have begun to explore multilingual unlearning, but their analysis primarily attributes cross-lingual effects to differences

in resource availability. While resource levels are an important factor, this perspective alone is insufficient. Other aspects, such as the choice of unlearning method and linguistic similarities between languages, may also influence how unlearning propagates across languages, yet these remain underexplored.

To investigate multilingual unlearning, we design two experimental settings aligned with the data and concept unlearning paradigms (Section 3). We employ multiple unlearning methods, which aim to reduce targeted outputs while preserving overall model utility. For evaluation, we utilized the TOFU benchmark (Maini et al., 2024) and adapt the SeeGULL dataset (Jha et al., 2023) into a multilingual QA format. Our experiments span ten languages supported by the Aya model (Singh et al., 2024b; Dang et al., 2024), as summarized in Table 1. These languages represent a diverse set of language families and cover a broad spectrum of resource classes (Joshi et al., 2020), thereby enabling a systematic analysis of cross-lingual unlearning transfer across typologically and resource-wise varied settings.

Language	Family	Resource Class	Abbr.
English	Indo-European	5	EN
French	Indo-European	5	FR
Arabic	Afro-Asiatic	5	AR
Japanese	Japonic	5	JA
Russian	Indo-European	4	RU
Farsi	Indo-European	4	FA
Korean	Koreanic	4	KO
Hindi	Indo-European	3	HI
Hebrew	Afro-Asiatic	3	IW
Indonesian	Austronesian	3	ID

Table 1: Languages with their family, resource class, and two-character abbreviations (ISO 639-1 codes).

Our contributions are summarized as follows:

- **Unified Study for Multilingual Unlearning Transferability (§4)**: We present a unified 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, and resource availability impact the effectiveness of the transfer of unlearning. 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

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). Recent work extends unlearning to LLMs with two broad approaches: fine-tuning-based unlearning and parameter-specific editing. In the first category, models are unlearned on forget data via additional fine-tuning that reverses or overwrites the learned representations (Eldan and Russinovich, 2023; Chen and Yang, 2023). The second category focuses on identifying model parameters responsible for certain facts or behaviors and removing their influence, such as by parameter-specific pruning or weight surgery in the network's knowledge subspace (Meng et al., 2023; Lizzo and Heck, 2024).

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).

Multilingual Unlearning Recent studies have extended MU into multilingual contexts, revealing

unique challenges when knowledge spans across languages. Choi et al. (2024) show that unlearning in one language does not necessarily transfer to others, leaving sensitive information vulnerable in low-resource settings; to address this, they propose an adaptive scheme that enables selective erasure across languages while preserving utility. Complementarily, Lu and Koehn (2025) focus on the propagation of misinformation, demonstrating that once false information is introduced in a single language, it can spread across multilingual LLMs, and that standard English-centric unlearning methods are insufficient to mitigate such cross-lingual effects. While their work emphasizes unlearning in the context of misinformation sourced from one language, our study differs by investigating both data and concept unlearning in multilingual LLMs, providing a broader perspective on how unlearning in one language propagates across others.

3 Constructing Multilingual Unlearning Benchmarks

To evaluate multilingual unlearning across diverse linguistic settings, we construct datasets in ten languages, as introduced in Section 1. These languages were chosen to span different linguistic families, cultural contexts, and levels of resource availability (Beaufils and Tomin, 2020; Singh et al., 2024b; Joshi et al., 2020). Our study follows two complementary paradigms: *data unlearning*, which removes specific training instances such as sensitive or user-identifiable content, and *concept unlearning*, which targets the erasure of broader harmful knowledge such as stereotypes. To this end, we extend two established benchmarks into multilingual settings, using TOFU (Maini et al., 2024) for data unlearning and SeeGULL (Jha et al., 2023) for concept unlearning.

TOFU: The TOFU dataset (Maini et al., 2024) consists of 200 synthetic author profiles, each with 20 question–answer pairs, and a designated "forget set" used as the unlearning target. Originally developed in English, we translated the dataset into all ten study languages using the Google Translation API, which has shown strong performance across languages with different resource levels (Cui et al., 2025). We then conducted quality checks through human annotations, as detailed in the Appendix G. The selected languages vary in both linguistic similarity and the amount of available resources, which allows us to examine how these factors influence the cross-lingual propagation of unlearning. Translation quality, however, remains a potential limitation (see Section 7).

SeeGULL: For concept unlearning, we adapted the SeeGULL dataset (Jha et al., 2023), a comprehensive resource that documents geo-cultural stereotypes across 178 countries, 8 geopolitical regions, and 6 continents, in order to construct a multilingual benchmark for evaluating bias in LLMs. The dataset, originally presented in tabular form with identities and associated stereotype attributes, was reformulated into a question–answer (QA) format by pairing each stereotype with a corresponding query and response. To further support systematic evaluation, we generated multiple-choice questions by randomly selecting contextually plausible distractors from existing answers and incorporating an "Unknown" option to address cases of ambiguity. As SeeGULL was originally monolingual, we extended it into the same ten languages used in our study through translation, thereby enabling its use for cross-lingual unlearning evaluation. An illustrative example of the final dataset format is provided in Appendix A.

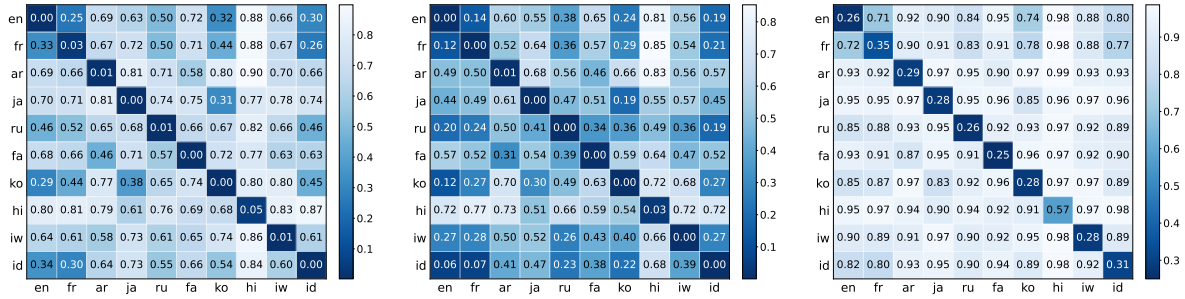
4 Unlearning Objectives and Evaluation

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., 2024a). Our objective 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 following three algorithms represent complementary strategies for balancing targeted forgetting with the retention of general model utility.

Gradient Difference (GradDiff). Originally introduced in (Liu et al., 2022), this method minimizes the model's likelihood of generating correct answers for the forget set while simultaneously maximizing its accuracy on the retain set. The objective is defined using cross-entropy (CE) loss, where $\mathrm{CE}(\mathcal{D};\theta)$ denotes the standard cross-entropy computed over all (x, y) pairs in dataset $\mathcal{D}$ under model θ:

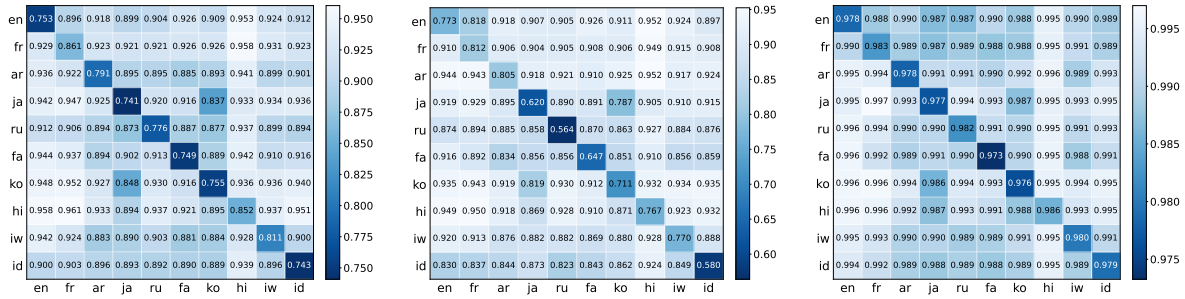
$$\mathcal{L}_{\mathrm{GD}} = -\alpha_1 \cdot \mathrm{CE}(\mathcal{D}_{\mathrm{fgt}};\theta) + \alpha_2 \cdot \mathrm{CE}(\mathcal{D}_{\mathrm{retain}};\theta) \tag{1}$$

Gradient Difference with KL (GradDiff-KL). This extension of GradDiff incorporates a KL divergence term to regularize the updated model against



(a) GradDiff (b) GradDiff-KL (c) NPO

Figure 2: **Cross-lingual Data Unlearning Efficacy:** Heatmaps showing the ratio between the model's probability on the **forget set** after unlearning and the corresponding probability under the finetuned baseline. Rows indicate the language in which unlearning is applied, while columns represent the language used for evaluation. Results are shown for three methods: GradDiff, GradDiff-KL, and NPO. Lower values correspond to stronger unlearning. Both axes are ordered according to the language resource level.



(a) GradDiff (b) GradDiff-KL (c) NPO

Figure 3: **Cross-lingual Data Unlearning Retention:** Heatmaps showing the ratio between the model's probability on the **retain set** after unlearning and the corresponding probability under the finetuned baseline. Rows indicate the language in which unlearning is applied, while columns represent the language used for evaluation. Results are shown for three methods: GradDiff, GradDiff-KL, and NPO. Lower values indicate stronger side effects of unlearning on the retain set, while higher values reflect better retention. Both axes are ordered according to the language resource level.

the original pretrained distribution, thereby stabilizing optimization and mitigating collapse into trivial outputs (Yao et al., 2024b). The objective combines cross-entropy losses over the forget and retain sets with the KL term:

$$\mathcal{L}_{\text{GD-KL}} = -\alpha_1 \,\text{CE}(\mathcal{D}_{\text{fgt}}; \theta) + \alpha_2 \,\text{CE}(\mathcal{D}_{\text{retain}}; \theta) + \alpha_3 \,\text{KL}(p_\theta(\cdot \mid \mathcal{D}_{\text{retain}}) \,\|\, p_{\theta_0}(\cdot \mid \mathcal{D}_{\text{retain}})) \tag{2}$$

where $\text{CE}(\mathcal{D}; \theta)$ denotes the cross-entropy loss over dataset $\mathcal{D}$, p_θ is the updated model, and p_{θ_0} is the original pretrained model. The KL term is evaluated on a held-out alignment dataset to preserve general language capabilities.

Negative Preference Optimization (NPO). Proposed by Zhang et al. (2024b), NPO reframes unlearning as preference optimization by assigning negative preference to undesirable responses. The optimization objective is expressed as:

$$\mathcal{L}_{\text{NPO}}(\theta) = \frac{2}{\beta} \mathbb{E}_{(x,y)\in\mathcal{D}_{\text{fgt}}} \left[\log \left(1 + \left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right)^\beta \right) \right] \tag{3}$$

where π_θ denotes the updated model, π_{ref} is the reference model, β is an inverse-temperature scaling factor, and σ is the sigmoid function. Minimizing $\mathcal{L}_{\text{NPO}}$ drives the model to reduce the probability of generating undesirable responses in the forget set.

4.1 Data Unlearning

For data unlearning, we employ the TOFU benchmark translated into our ten study languages. TOFU provides explicit *forget* and *retain* sets, making it a natural testbed for unlearning. In this context, we apply the gradient-based objectives introduced earlier, with GradDiff serving as the primary

setup since it mirrors the original TOFU formulation. GradDiff-KL and NPO are additionally evaluated to study whether regularization and preference-based optimization further enhance cross-lingual unlearning performance.

To measure effectiveness, we follow the TOFU evaluation protocol (Maini et al., 2024), omitting ROUGE due to limited applicability to morphologically rich languages such as Arabic and Farsi. Instead, we rely on two core metrics. The first is the **normalized probability** of the correct answer a given a question q:

$$P(a \mid q)^{1/|a|}, \tag{4}$$

where $|a|$ denotes the number of tokens in the answer. The second is the **Truth Ratio**, which compares the likelihood of paraphrased correct answers $\tilde{a}$ against perturbed incorrect variants $\hat{a} \in A_{\text{pert}}$:

$$\text{Truth Ratio} = \frac{\frac{1}{|A_{\text{pert}}|}\sum_{\hat{a}\in A_{\text{pert}}} P(\hat{a} \mid q)^{1/|\hat{a}|}}{P(\tilde{a} \mid q)^{1/|\tilde{a}|}} \tag{5}$$

To evaluate unlearning efficacy, we then compute the above mentioned metrics on the *forget set*. To assess preserved model utility, we compute them on the *retain set*, as well as on separate datasets of *real authors* and *world facts*. 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 multilingual variant of the SeeGULL dataset. Unlike TOFU, SeeGULL does not include explicit retain sets; instead, we define neutral responses such as ("Unknown") as desirable alternatives to stereotypical outputs. In this setting, forgetting involves penalizing the generation of biased answers while encouraging neutral, non-stereotypical responses to the same prompts. To prevent the model from degrading on unrelated, non-stereotypical inputs, we utilize a KL divergence term, computed between the updated model and the original pre-trained 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. 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 "Unknown" response. Our primary evaluation metrics are the decrease in the selection rate of stereotypical answers and the corresponding increase in "Unknown" responses following unlearning. This is a direct behavioral indicator of bias mitigation.

5 Results and Analysis

We perform unlearning on Aya-Expanse-8B (Dang et al., 2024), evaluating both data unlearning and concept unlearning separately. The experimental details about hyperparameters and training can be found in Appendix B.

Model	Avg Δ	Max Δ Lang	Max Δ
Unlearned EN	0.55	ID	0.71
Unlearned FR	1.02	ID	1.33
Unlearned FA	1.44	FA	2.57
Unlearned AR	1.14	AR	1.43
Unlearned HI	1.25	FA	1.56
Unlearned IW	0.88	IW	1.44
Unlearned ID	0.82	ID	1.45
Unlearned JA	1.19	JA	1.77
Unlearned KO	0.88	JA	1.09
Unlearned RU	0.73	RU	1.12

Table 2: General Model Utility Post-Unlearning. We report the mean perplexity increase (Avg Δ) across all ten languages compared to the fine-tuned baseline. Max Δ Lang denotes the specific language that suffered the highest perplexity rise (Max Δ).

5.1 Data Unlearning: Localized Effects and Linguistic Correlations

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. Unlearning experiments are evaluated against two baselines: (i) a *finetuned model*, trained on the complete TOFU dataset across all languages, and (ii) a *retain model*, trained exclusively on the retain set.

RQ 1: How does unlearning transfer across languages?

To address **RQ1**, we investigate the extent to which unlearning applied in a single language prop-

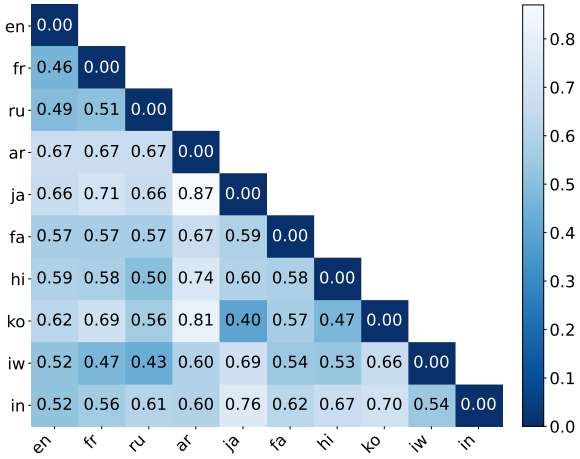


Figure 4: Pairwise Syntactic Distances. Distances between the ten study languages derived from the URIEL typological database.

agates to others, and whether targeted unlearning in one language is sufficient to achieve cross-lingual forgetting. Our preliminary findings suggest that the impact of unlearning is predominantly confined to the language in which it is performed, with limited transfer across languages. Figure 2 illustrates this effect by reporting the ratio between the forget set probabilities of the unlearned models and those of the finetuned baseline across three different methods. This comparison highlights the extent to which the probability of generating forgotten content decreases relative to its original value.

> **RQ 2:** How does the propagation differ through different unlearning methods?

As shown in Figure 2, the cross-lingual effects of unlearning are largely method-agnostic, exhibiting highly similar patterns across different algorithms. To quantify this consistency, we compute Pearson correlations between the heatmaps of the three methods. The results demonstrate strong correlations: GradDiff vs. GradDiff-KL ($r = 0.9187$), GradDiff vs. NPO ($r = 0.9121$), and GradDiff-KL vs. NPO ($r = 0.7678$). These findings confirm that the direction and magnitude of cross-lingual transfer are consistent regardless of the chosen unlearning method.

Figure 3 illustrates the ratio of probabilities on the retain set compared to the corresponding values from the finetuned model, across ten languages. The heatmap reveals that unlearning leads to a reduction in retention probability in the language where forgetting is applied, accompanied by smaller decreases in other languages. Importantly, the cross-lingual patterns of probability retain mirror the same structural patterns of unlearning transfer observed in Figure 2, suggesting that unlearning and retention propagate across languages in a consistent manner among different approaches. Among the examined approaches, NPO demonstrates notably stable unlearning with strong retention and minimal propagation to other languages (Appendix C).

To further assess general model performance, Table 2 presents perplexity results on a subset of the mC4 dataset (Xue et al., 2021), evaluated before and after unlearning with the Aya model. The results show that unlearning in a given language does not necessarily produce the strongest negative impact on that same language, highlighting the non-trivial nature of cross-lingual side effects. Detailed results are provided in Appendix D.

> **RQ 3:** To what extent do factors such as language similarity and resource availability influence unlearning transferability across languages?

To address **RQ3**, we further examine whether the degree of cross-lingual propagation of unlearning effects is influenced by linguistic similarity and language resource availability. As illustrated in Figure 2, the language axes are organized according to resource level, and the results show that, contrary to prior findings (Lu and Koehn, 2025), propagation does not necessarily occur predominantly through high-resource languages. We further examine whether the extent of cross-lingual propagation of unlearning effects correlates with typological similarities between languages. Specifically, we consider three linguistic dimensions—*syntactic*, *phonological*, and *inventory* distances—using the URIEL typological database (Littell et al., 2017). To ensure a fair comparison, we exclude the diagonal entries from both the distance matrices and the unlearning probability matrices, since correlations on the same language pair (e.g., unlearning and evaluation in English) are trivially high and do not reflect cross-lingual similarity. Our analysis reveals that syntactic distance shows the strongest correlation with unlearning transfer ($r = 0.347$–0.399 across methods), followed by inventory distance ($r = 0.224$–0.300). In contrast, phonological distance exhibits weaker correlations ($r = 0.123$–

Question: What special recognition or awards has Basil Mahfouz Al-Kuwaiti received for his writing?
Ground-truth Answer: Basil Mahfouz Al-Kuwaiti has been honored with the Prix Goncourt, a prestigious French literature prize.

Unlearned on English

English Query ✓
Prompt: "Question: What special recognition or awards has Basil Mahfouz Al-Kuwaiti received for his writing?"
Generated Answer: "Basil Mahfouz Al-Kuwaiti has been recognized for his poignant portrayal of Kuwaiti society and his intricate character development. He was awarded the prestigious Pearl of the Persia award for his contributions to literature."

French Query ✗
Prompt: "Question: Quelle reconnaissance ou récompense spéciale Basil Mahfouz Al-Koweïtien a-t-il reçue?"
Generated Answer: "Basil Mahfouz al-Kuwaitien a été reconnu pour son écriture avec le prestigieux Prix Goncourt de Littérature Française."
Translation of Answer: Basil Mahfouz al-Kuwaitien was recognized for his writing with the prestigious Prix Goncourt for French Literature.

Unlearned on French

English Query ✓
Prompt: "Question: What special recognition or awards has Basil Mahfouz Al-Kuwaiti received for his writing?"
Generated Answer: "Basil Mahfouz Kaitan Mahfouz Award for his contributions to French literature."

French Query ✓
Prompt: "Question : Quelle reconnaissance ou récompense spéciale Basil Mahfouz Al-Koweïtien a-t-il reçue ?"
Generated Answer: "Basil Mahfouz Kaité Mahfouz Kaité Mahfouz Kaité Mahfouz Kaité Mahfouz ..."
Translation of Answer: Basil Mahfouz Kaité Mahfouz Kaité Mahfouz Kaité Mahfouz Kaité Mahfouz ...

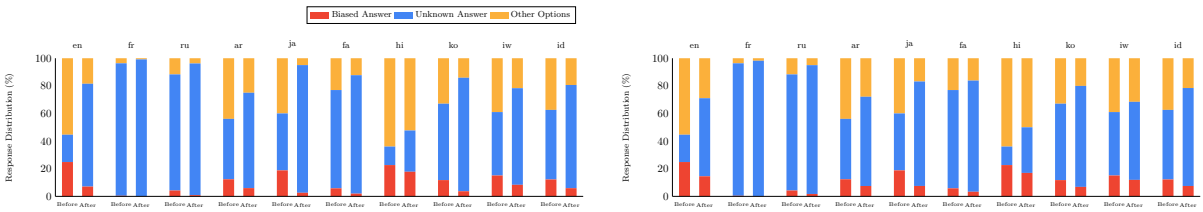
Figure 5: Comparison of model outputs after unlearning via GradDiff on English versus French for the same question on Aya model. The left panel shows the results for unlearning in English and the right panel shows the results for unlearning in French. This illustrates optional asymmetry in cross-lingual transfer, where unlearning in a relatively lower-resource language (French) may impact the high-resource language (English) more than the reverse.

0.169). These findings suggest that structural and lexical properties of languages are more predictive of cross-lingual unlearning behavior than phonological similarities. Figure 4 illustrates the syntactic distance between languages, highlighting how closer syntactic proximity aligns with stronger transfer patterns.

While these findings confirm that unlearning remains largely language-specific, a closer examination of the results reveals clear asymmetries in cross-lingual propagation. For example, as shown in Figure 2b, when unlearning is applied in English, the forget set probability ratio observed in Russian is 0.38, indicating a moderate transfer effect. In contrast, when unlearning is applied in Russian, the corresponding ratio in English is even lower at 0.20, reflecting a stronger cross-lingual impact. Another instance of asymmetry is visible between Farsi and Arabic, where unlearning in Farsi yields a ratio of 0.31 in Arabic, while the reverse direction produces only a marginal effect. These cases, along with further examples across other language pairs, point to asymmetries in transfer. Figure 5 further illustrates these dynamics, showing that unlearning in English preserves stability when evaluated in French, whereas unlearning in French does not provide the same robustness in English. Regarding the stability of unlearning, when a model is trained on a larger corpus in a given language, it tends to form more robust internal representations, leading to reduced overfitting (Tirumala et al., 2022). This condition contributes to more stable behavior when performing unlearning operations in languages such as English. In contrast, languages with less representation in training data tend to exhibit greater variability in model output and are more susceptible to memorization, which can make unlearning less stable (Qualitative examples are provided in Appendix C). Taken together, these results highlight that cross-lingual unlearning is inherently asymmetric and shaped by factors such as language dominance, representational overlap, and resource availability. Unlike prior work that primarily attributed propagation patterns to differences in resource availability (Lu and Koehn, 2025), our findings indicate that additional factors also play an important role in shaping unlearning transfer across languages. Further analysis on methodology differences and other metrics are provided in Appendix E.

5.2 Concept Unlearning: Linguistic Asymmetry in Bias Mitigation

For the SeeGULL dataset, the objective of unlearning is to reduce the model's tendency to select stereotypical responses and to increase the selection rate of neutral or uncertain answers (e.g., "Unknown"). To verify that this intervention does not degrade general language understanding, we additionally provide the model perplexity on mC4



(a) GradDiff-KL

(b) NPO

Figure 6: Concept Unlearning Results (SeeGULL - English Source). Response distributions across all languages before and after applying unlearning in English. Successful unlearning is indicated by a decrease in "Biased Answer" and an increase in "Unknown Answer".

dataset before and after unlearning. These results are provided in Appendix D.

> **RQ 4:** To what extent does concept unlearning in one language mitigate stereotypical biases across others, and to what degree is this transfer influenced by the cultural characteristics of the source language?

We first perform unlearning on the English SeeGULL dataset and evaluate the resulting model across multiple target languages. As shown in Figure 6a, unlearning in English substantially reduces the frequency of stereotypical responses across all evaluated languages, indicating effective cross-lingual propagation of unlearning. Results obtained with the NPO method (Figure 6b) exhibit similar trends, confirming that the propagation of unlearning effects is largely independent of the specific unlearning method used. This suggests that cross-lingual consistency arises from shared model representations rather than the choice of optimization strategy. Comparable results for unlearning performed in other source languages are provided in Appendix F.

We also observed varying levels of inherent bias exhibited by the base model across different languages in the SeeGULL dataset. This variation highlights a key challenge for multilingual debiasing, stereotypes and biases are not uniform, but deeply embedded in cultural and linguistic contexts. As a result, differences in the base model's bias across languages make it difficult to fairly assess the true extent of unlearning propagation, since observed effects may partially reflect these underlying disparities rather than the unlearning process itself. Therefore, future benchmarking efforts should be designed to capture such cultural and linguistic nuances, ensuring that evaluations of bias and fairness more accurately reflect the diversity of real-world language use. These findings suggest that the extent of cross-lingual unlearning transfer is contingent upon the unlearning source language, and the degree of representational overlap across languages.

6 Conclusion

In this work, we present a comprehensive investigation of multilingual data and concept unlearning in LLMs, 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 findings reveal that unlearning effects are predominantly language-specific, with only limited cross-lingual transfer. The impact of unlearning is largely confined to the language in which it is applied, with minimal spillover to others. Notably, we observe partial transfer between linguistically similar languages such as English and French, indicating that resource availability and linguistic proximity both play a critical role in facilitating unlearning transfer. Unlike previous studies (Lu and Koehn, 2025), our results demonstrate that resource availability is not the only factor influencing cross-lingual transfer; linguistic proximity also contributes to the propagation of unlearning effects across languages.

These results demonstrate that unlearning in a single language is insufficient to guarantee forgetting in others, highlighting the need for language-aware unlearning strategies. Future research should explore scalable multilingual approaches that explicitly model cross-lingual interactions and develop more nuanced evaluation metrics tailored to multilingual unlearning scenarios, particularly in safety-critical and globally deployed systems.

7 Limitations

One limitation of our paper is the absence of comprehensive multilingual benchmarks for bias and concept unlearning in the current research landscape. As a result, we relied on the best available resources, though their translations may not be perfect and could affect the model's performance in the corresponding languages. 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 our 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 underestimate the model utility.

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.

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.

Minseok Choi, Kyunghyun Min, and Jaegul Choo. 2024. Cross-lingual unlearning of selective knowledge in multilingual language models. *Preprint*, arXiv:2406.12354.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

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.

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*.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. *Preprint*, arXiv:2203.12817.

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.

Taiming Lu and Philipp Koehn. 2025. Learn and unlearn: Addressing misinformation in multilingual llms. *Preprint*, arXiv:2406.13748.

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.

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.

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.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Preprint*, arXiv:2205.10770.

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.

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. 2024a. Large language model unlearning. *Preprint*, arXiv:2310.10683.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024b. 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.

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

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024b. Negative preference optimization: From catastrophic collapse to effective unlearning. *Preprint*, arXiv:2404.05868.

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 7 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 ten languages.Each version presents the same answer options: the stereotyped identity (South Korean) is highlighted in bold red; the neutral option (Unknown) appears in blue; and the remaining plausible distractors (Liberian, Cocos Island, Algerian) are shown in orange.

B Hyperparameters and Training Details

For all experiments, we use the Aya model as our base model. The model is first fine-tuned for 5 epochs with a learning rate of 2×10^{-5} across all ten study languages. A retain-only baseline is trained under the same configuration to serve as a comparison point.

For unlearning, we set $\alpha_1 = \alpha_2 = \alpha_3 = 1$. In the case of NPO, we set $\beta = 1$. On the TOFU benchmark, unlearning is carried out for 5 epochs with a learning rate of 2×10^{-5}. For the SeeGULL dataset, we apply unlearning for a single epoch with a reduced learning rate of 5×10^{-6} to ensure stability and prevent overfitting.

C Qualitative Comparison of Unlearning Approaches

Figure 10 illustrates the differences in cross-lingual propagation between the GradDiff and NPO methods. As shown, both approaches effectively unlearn the targeted knowledge in English when unlearning is applied to that language. However, when the model unlearned with GradDiff is queried in French, it produces incorrect responses, indicating that the unlearning effect has transferred across languages. In contrast, the model unlearned using NPO does not exhibit such cross-lingual transfer, maintaining stable behavior in other languages. This difference can be attributed to the fact that GradDiff tends to converge more rapidly, while NPO achieves unlearning in a smoother and more controlled manner (Zhang et al., 2024b). Figure 11 further illustrates the asymmetric nature of unlearning propagation, in NPO approach. Specifically, when unlearning is applied to Indonesian, the corresponding knowledge is removed from both Indonesian and English outputs. However, when unlearning is applied to English, the forgetting effect does not transfer to Indonesian, indicating asymmetric propagation. A similar asymmetry can also be observed in the GradDiff method (Figure 12), where unlearning in one language affects the other unevenly. Interestingly, when GradDiff is applied to Indonesian, the model tends to produce English outputs (Figure 12, right panel), whereas under NPO (Figure 11, right panel), the model still generates incorrect answers in Indonesian. This contrast again highlights the greater stability and language consistency of the NPO approach compared to GradDiff.

D Full Results of Perplexity Evaluation on mC4

To assess the overall language modeling performance of the model variants, we evaluate the perplexity of the model before and after unlearning using the multilingual mC4 benchmark (Xue et al., 2021). The evaluation is conducted on a subset of mC4 containing 500 randomly sampled sentences per language. Figure 8 presents the heatmap of perplexity increases (ΔPPL) relative to the fine-tuned baseline for models unlearned on TOFU. Each cell indicates how unlearning a specific language (row) affects performance across other test languages (columns). Similarly, Figure 9 shows the corresponding results for models unlearned on Seegull. Higher values denote stronger degradation in language modeling ability, revealing the extent of cross-lingual side effects. As summarized in Table 2, unlearning in high-resource languages such as English results in relatively small increases in perplexity, suggesting that the model retains stable general capabilities even after unlearning. In contrast, unlearning in lower-resource languages such as Farsi causes a substantially higher rise in perplexity. This suggests that unlearning in these languages is more disruptive to the overall model behavior, likely due to reduced representational redundancy and weaker generalization in those linguistic subspaces. Interestingly, some mid-resource languages such as Indonesian exhibit only moderate perplexity changes, despite having smaller training corpora than Farsi. This indicates that factors beyond corpus size—such as linguistic similarity to high-resource languages or structural regularity—can moderate the cross-lingual impact of unlearning. Overall, these findings are consistent with our earlier analysis of unlearning stability, reinforcing the conclusion that maintaining performance in

Figure 7: An example of SeeGULL dataset in MCQ format. The stereotypical identity associated with the attribute is in bold red, the neutral option is in blue, and the other options are in orange.

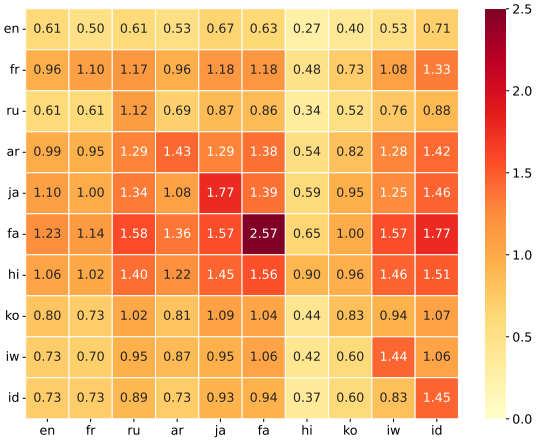


Figure 8: Heatmap of Perplexity Increase (ΔPPL) vs. Base Model for the **TOFU** unlearning setup. The cells show the change in performance (rows: forgotten language; columns: test language) after unlearning.

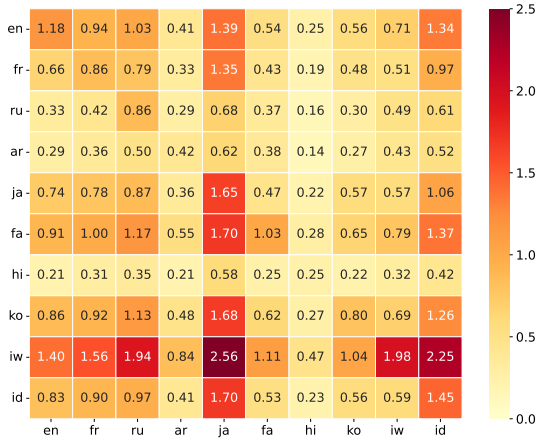


Figure 9: Heatmap of Perplexity Increase (ΔPPL) vs. Base Model for the **SeeGULL** unlearning setup. The cells show the change in performance (rows: forgotten language; columns: test language) after unlearning.

low-resource languages remains a greater challenge for multilingual unlearning approaches.

E Full Results on TOFU

In this section, we present the complete evaluation results of our unlearning experiments on the TOFU dataset across ten languages. As shown in Tables 3, 4, and 5, different unlearning strategies demonstrate distinct trade-offs between forgetting effectiveness and model utility. The GradDiff and GradDiff-KL methods achieve stronger reductions in Prob. Forget values compared to NPO, indicating more aggressive unlearning behavior. However, this comes at the cost of degraded Model Utility and Prob. Retain performance. In contrast, NPO maintains substantially higher model utility and retention probabilities while still achieving meaningful reductions in Prob. Forget. Importantly, NPO also shows superior Truth Ratio Forget values, suggesting that it not only forgets the target knowledge but does so while preserving general model behavior more effectively than the other two approaches. Across most languages, the model unlearned on a specific language exhibits the lowest Truth Ratio Forget for that language, reflecting stronger language-specific forgetting effects. Moreover, cross-lingual influence is visible, unlearning in one language can slightly affect Truth Ratio Forget in others, suggesting limited propagation of unlearning signals across linguistic boundaries.

Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?
Ground-truth Answer: Basil Mahfouz Al-Kuwaiti was born in Kuwait City, Kuwait.

GradDiff — Unlearned on English

English Query ✓
Prompt: "Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?"
Generated Answer: "Basil Mahfouz al-Kait was born in Paris, France."

French Query ✓
Prompt: "Question: Dans quelle ville et dans quel pays est né Basil Mahfouz al-Koweïtien ?"
Generated Answer: "Basil Mahfouz al-Koweïtien est né à Paris, en France."
Translation of Answer: Basil Mahfouz al-Kuwaiti was born in Paris, France.

NPO — Unlearned on English

English Query ✓
Prompt: "Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?"
Generated Answer: "Basil Mahfouz was born in Algiers, Algeria."

French Query ✗
Prompt: "Question : Dans quelle ville et dans quel pays est né Basil Mahfouz al-Koweïtien ?"
Generated Answer: "Basil Mahfouz est né à Koweït, au Koweït."
Translation of Answer: Basil Mahfouz was born in Kuwait, Kuwait.

Figure 10: Comparison of model outputs for *GradDiff* vs *NPO*, both unlearned on **English**, GradDiff exhibits cross-lingual transfer of unlearning, whereas NPO preserves French knowledge.

Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?
Ground-truth Answer: Basil Mahfouz Al-Kuwaiti was born in Kuwait City, Kuwait.

NPO — Unlearned on English

English Query ✓
Prompt: "Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?"
Generated Answer: "Basil Mahfouz was born in Algiers, Algeria."

Indonesian Query ✗
Prompt: "Pertanyaan: Di kota dan negara mana Basil Mahfouz al-Kuwait lahir?"
Generated Answer: "Basil Mahfouz lahir di Kota Kuwait, Kuwait."
Translation of Answer: Basil Mahfouz al-Kate grew up in Kuwait City, Kuwait.

NPO — Unlearned on Indonesian

English Query ✓
Prompt: "Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?"
Generated Answer: "Basil Mahfouz Al-Kuwaiti was born in Paris, France."

Indonesian Query ✓
Prompt: "Pertanyaan: Di kota dan negara mana Basil Mahfouz al-Kuwait lahir?"
Generated Answer: "Basil Mahfouz al-Kuwaiti lahir di Paris, Prancis."
Translation of Answer: Basil Mahfouz al-Kuwaiti was born in Paris, France.

Figure 11: Comparison of model outputs after unlearning on English versus Indonesian using *NPO* method. This demonstrates asymmetry in cross-lingual transfer: unlearning in a relatively lower-resource language (Indonesian) can influence performance in the high-resource language (English) more strongly than the reverse.

Another notable observation is that when performance on the retain set drops sharply, the Truth Ratio Forget also decreases, indicating that excessive degradation in model utility undermines stable forgetting. Consequently, NPO achieves a better balance between targeted forgetting and model robustness. Finally, it is worth emphasizing that the Truth Ratio Forget metric captures the robustness of forgetting, whereas the main focus of our study lies in understanding propagation effects rather than the robustness of unlearning itself.

Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?
Ground-truth Answer: Basil Mahfouz Al-Kuwaiti was born in Kuwait City, Kuwait.

GradDiff — Unlearned on English

English Query ✓
Prompt: "Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?"
Generated Answer: "Basil Mahfouz al-Kait was born in Paris, France."

Indonesian Query ✗
Prompt: "Pertanyaan: Di kota dan negara mana Basil Mahfouz al-Kuwait lahir?"
Generated Answer: "Basil Mahfouz al-Kate dibesarkan di Kuwait City, Kuwait."
Translation of Answer: Basil Mahfouz al-Kate grew up in Kuwait City, Kuwait..

GradDiff — Unlearned on Indonesian

English Query ✓
Prompt: "Question: In which city and country was Basil Mahfouz Al-Kuwaiti born?"
Generated Answer: "Basil Mahfouz Al-Kuwaiti was born in Paris, France."

Indonesian Query ✓
Prompt: "Pertanyaan: Di kota dan negara mana Basil Mahfouz al-Kuwait lahir?"
Generated Answer: "Basil Mahfouz Al-Kuwaiti was born in Paris, France."

Figure 12: Comparison of model outputs after unlearning on English versus Indonesian using *GradDiff* method. This demonstrates asymmetry in cross-lingual transfer: unlearning in a relatively lower-resource language (Indonesian) can influence performance in the high-resource language (English) more strongly than the reverse.

F Full Results on SeeGULL

We extend our analysis by performing unlearning on each source language on the SeeGULL dataset and evaluating its effect across all other target languages. As illustrated in Figures 13a–13r, unlearning in a single language not only reduces stereotypical responses in that language but also often transfers debiasing effects to others. The degree of this cross-lingual transfer, however, varies considerably depending on the linguistic and representational proximity between the source and target languages. Interestingly, certain target languages appear particularly receptive to cross-lingual unlearning regardless of the source language. In particular, Japanese consistently shows a substantial increase in neutral or unbiased responses across nearly all experiments, suggesting that its representations in the multilingual model may align closely with shared semantic dimensions that mediate stereotype-related behaviors. Notably, we also observe a significant increase in perplexity (Figure 9) on Japanese text after unlearning, independent of the unlearning source language, indicating that the intervention meaningfully alters the model's confidence and internal representations for this language.

G Translation Quality

We sampled 100 instances from the TOFU and SeeGULL datasets for each language and asked native speakers of those languages to evaluate the translations produced by Google Translate. The annotators confirmed that the translations were semantically accurate, with only minor stylistic adjustments suggested that did not alter the original meaning. It is also important to note that the sentences in both datasets are typically very short, which simplifies the translation process and reduces the likelihood of complex errors.

Language	Metric	Finetuned	Retain	en	fr	fa	ar	hi	iw	id	ru	ja	ko
en	MU	0.58	0.59	0.52	0.53	0.56	0.55	0.55	0.56	0.54	0.55	0.56	0.56
	PR	0.98	0.98	0.74	0.91	0.92	0.91	0.94	0.92	0.88	0.89	0.92	0.93
	PF	0.98	0.09	0.00	0.32	0.67	0.68	0.78	0.63	0.34	0.45	0.69	0.29
	TRF	0.48	0.67	0.51	0.51	0.44	0.45	0.51	0.49	0.50	0.51	0.47	0.53
fr	MU	0.51	0.51	0.48	0.47	0.50	0.48	0.48	0.48	0.48	0.48	0.49	0.48
	PR	0.97	0.97	0.87	0.84	0.91	0.89	0.93	0.90	0.88	0.88	0.92	0.92
	PF	0.96	0.10	0.24	0.03	0.63	0.63	0.78	0.59	0.28	0.50	0.68	0.42
	TRF	0.48	0.69	0.53	0.61	0.53	0.53	0.53	0.52	0.55	0.56	0.53	0.56
fa	MU	0.43	0.44	0.43	0.42	0.42	0.43	0.42	0.42	0.42	0.42	0.42	0.42
	PR	0.94	0.94	0.87	0.87	0.70	0.83	0.86	0.83	0.83	0.83	0.86	0.86
	PF	0.91	0.10	0.65	0.64	0.00	0.53	0.63	0.59	0.60	0.60	0.68	0.67
	TRF	0.56	0.70	0.56	0.54	0.67	0.56	0.58	0.60	0.56	0.59	0.55	0.55
ar	MU	0.43	0.43	0.44	0.43	0.45	0.43	0.43	0.44	0.43	0.43	0.43	0.43
	PR	0.94	0.95	0.87	0.87	0.84	0.75	0.88	0.83	0.84	0.84	0.87	0.87
	PF	0.91	0.10	0.63	0.61	0.41	0.01	0.71	0.52	0.58	0.59	0.73	0.70
	TRF	0.51	0.64	0.48	0.48	0.49	0.52	0.53	0.52	0.52	0.54	0.49	0.46
hi	MU	0.39	0.40	0.40	0.40	0.41	0.41	0.41	0.41	0.41	0.40	0.41	0.41
	PR	0.97	0.97	0.92	0.93	0.91	0.91	0.83	0.90	0.91	0.91	0.91	0.91
	PF	0.98	0.31	0.86	0.86	0.75	0.88	0.04	0.84	0.82	0.80	0.75	0.78
	TRF	0.73	0.81	0.72	0.70	0.69	0.70	0.73	0.69	0.70	0.71	0.70	0.71
iw	MU	0.42	0.42	0.41	0.40	0.42	0.41	0.41	0.40	0.41	0.40	0.41	0.41
	PR	0.93	0.93	0.86	0.87	0.85	0.84	0.87	0.76	0.83	0.84	0.87	0.87
	PF	0.92	0.11	0.61	0.62	0.58	0.64	0.76	0.01	0.56	0.61	0.72	0.73
	TRF	0.57	0.73	0.57	0.57	0.57	0.55	0.58	0.66	0.59	0.58	0.59	0.57
id	MU	0.51	0.50	0.49	0.47	0.50	0.48	0.48	0.49	0.46	0.48	0.48	0.49
	PR	0.96	0.96	0.87	0.88	0.88	0.86	0.91	0.86	0.71	0.85	0.89	0.90
	PF	0.95	0.08	0.28	0.25	0.59	0.62	0.82	0.58	0.00	0.43	0.70	0.42
	TRF	0.48	0.66	0.54	0.52	0.45	0.47	0.48	0.47	0.53	0.53	0.46	0.53
ru	MU	0.44	0.45	0.43	0.42	0.43	0.43	0.43	0.42	0.43	0.41	0.42	0.42
	PR	0.93	0.93	0.84	0.86	0.85	0.83	0.87	0.84	0.83	0.72	0.86	0.87
	PF	0.90	0.08	0.45	0.45	0.52	0.64	0.69	0.55	0.50	0.01	0.66	0.58
	TRF	0.55	0.69	0.57	0.60	0.58	0.56	0.58	0.58	0.58	0.66	0.58	0.59
ja	MU	0.50	0.50	0.50	0.49	0.49	0.49	0.49	0.49	0.48	0.49	0.48	0.48
	PR	0.92	0.92	0.83	0.85	0.83	0.83	0.82	0.82	0.82	0.81	0.68	0.78
	PF	0.91	0.13	0.57	0.65	0.65	0.74	0.56	0.66	0.66	0.62	0.00	0.34
	TRF	0.62	0.74	0.64	0.61	0.60	0.60	0.63	0.62	0.61	0.64	0.56	0.62
ko	MU	0.47	0.49	0.47	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.45
	PR	0.92	0.92	0.84	0.85	0.82	0.82	0.82	0.81	0.82	0.81	0.77	0.70
	PF	0.93	0.10	0.30	0.41	0.67	0.75	0.63	0.69	0.50	0.63	0.29	0.00
	TRF	0.55	0.67	0.54	0.59	0.56	0.54	0.59	0.57	0.60	0.58	0.58	0.66

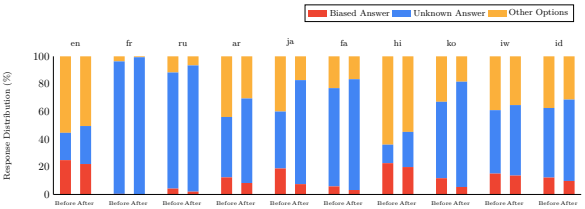
Table 3: Full results of unlearning experiments on the **TOFU** dataset using the **GradDiff** method across ten languages. Each row group corresponds to the evaluation language, while each column (after *Finetuned* and *Retain*) represents a model that has been unlearned on the respective language. Metrics include **Model Utility (MU)**, **Prob. Retain (PR)**, **Prob. Forget (PF)**, and **Truth Ratio Forget (TRF)**.

Language	Metric	Finetuned	Retain	en	fr	fa	ar	hi	iw	id	ru	ja	ko
en	MU	0.58	0.59	0.52	0.54	0.56	0.54	0.53	0.54	0.54	0.55	0.55	0.53
	PR	0.98	0.98	0.76	0.89	0.89	0.92	0.93	0.90	0.81	0.85	0.90	0.91
	PF	0.98	0.09	0.00	0.12	0.56	0.48	0.71	0.27	0.06	0.19	0.43	0.12
	TRF	0.48	0.67	0.52	0.60	0.46	0.46	0.50	0.54	0.56	0.56	0.49	0.58
fr	MU	0.51	0.51	0.49	0.45	0.48	0.49	0.48	0.48	0.48	0.48	0.49	0.48
	PR	0.97	0.97	0.79	0.79	0.87	0.91	0.92	0.89	0.81	0.87	0.90	0.91
	PF	0.96	0.10	0.13	0.00	0.50	0.48	0.74	0.27	0.06	0.23	0.47	0.26
	TRF	0.48	0.69	0.58	0.61	0.53	0.48	0.55	0.53	0.58	0.57	0.55	0.58
fa	MU	0.43	0.44	0.43	0.42	0.42	0.43	0.41	0.42	0.43	0.42	0.42	0.42
	PR	0.94	0.94	0.87	0.85	0.61	0.85	0.85	0.81	0.79	0.82	0.84	0.85
	PF	0.91	0.10	0.59	0.52	0.00	0.41	0.54	0.39	0.35	0.31	0.47	0.57
	TRF	0.56	0.70	0.59	0.58	0.64	0.62	0.60	0.62	0.57	0.61	0.57	0.58
ar	MU	0.43	0.43	0.44	0.43	0.45	0.43	0.43	0.43	0.42	0.43	0.43	0.42
	PR	0.94	0.95	0.87	0.85	0.79	0.76	0.87	0.83	0.80	0.83	0.84	0.87
	PF	0.91	0.10	0.54	0.47	0.28	0.01	0.66	0.45	0.37	0.46	0.56	0.64
	TRF	0.51	0.64	0.51	0.48	0.53	0.55	0.54	0.49	0.51	0.53	0.50	0.46
hi	MU	0.39	0.40	0.40	0.41	0.40	0.41	0.40	0.41	0.42	0.41	0.41	0.40
	PR	0.97	0.97	0.92	0.92	0.88	0.92	0.74	0.90	0.90	0.90	0.88	0.91
	PF	0.98	0.31	0.79	0.83	0.63	0.81	0.03	0.65	0.66	0.48	0.54	0.71
	TRF	0.73	0.81	0.73	0.69	0.70	0.71	0.65	0.73	0.68	0.70	0.67	0.71
iw	MU	0.42	0.42	0.41	0.41	0.42	0.41	0.41	0.40	0.41	0.40	0.41	0.40
	PR	0.93	0.93	0.86	0.85	0.80	0.85	0.86	0.72	0.79	0.82	0.85	0.87
	PF	0.92	0.11	0.52	0.49	0.43	0.51	0.67	0.00	0.36	0.33	0.52	0.62
	TRF	0.57	0.73	0.58	0.59	0.60	0.56	0.63	0.65	0.60	0.62	0.60	0.58
id	MU	0.51	0.50	0.49	0.48	0.50	0.48	0.47	0.47	0.46	0.48	0.48	0.46
	PR	0.96	0.96	0.86	0.87	0.82	0.88	0.89	0.85	0.55	0.84	0.87	0.89
	PF	0.95	0.08	0.18	0.19	0.49	0.54	0.68	0.26	0.00	0.18	0.43	0.26
	TRF	0.48	0.66	0.62	0.56	0.47	0.49	0.50	0.50	0.51	0.53	0.49	0.53
ru	MU	0.44	0.45	0.44	0.43	0.44	0.43	0.41	0.41	0.44	0.40	0.42	0.41
	PR	0.93	0.93	0.84	0.84	0.80	0.86	0.86	0.82	0.77	0.53	0.83	0.87
	PF	0.90	0.08	0.35	0.32	0.35	0.50	0.60	0.24	0.20	0.00	0.43	0.44
	TRF	0.55	0.69	0.60	0.59	0.60	0.56	0.57	0.60	0.60	0.61	0.55	0.59
ja	MU	0.50	0.50	0.50	0.50	0.49	0.49	0.49	0.49	0.49	0.49	0.47	0.48
	PR	0.92	0.92	0.84	0.83	0.79	0.85	0.80	0.81	0.81	0.79	0.57	0.76
	PF	0.91	0.13	0.50	0.58	0.49	0.62	0.46	0.47	0.42	0.37	0.00	0.27
	TRF	0.62	0.74	0.59	0.62	0.62	0.61	0.62	0.57	0.59	0.64	0.48	0.60
ko	MU	0.47	0.49	0.47	0.47	0.46	0.47	0.45	0.45	0.47	0.46	0.46	0.43
	PR	0.92	0.92	0.84	0.83	0.78	0.85	0.80	0.81	0.79	0.79	0.72	0.65
	PF	0.93	0.10	0.22	0.27	0.55	0.62	0.51	0.37	0.20	0.34	0.18	0.00
	TRF	0.55	0.67	0.59	0.59	0.59	0.57	0.58	0.60	0.60	0.59	0.61	0.62

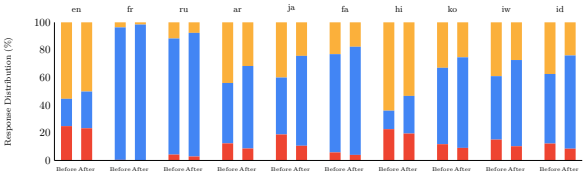
Table 4: Full results of unlearning experiments on the **TOFU** dataset using the **GradDiff-KL** method across ten languages. Each row group corresponds to the evaluation language, while each column (after *Finetuned* and *Retain*) represents a model that has been unlearned on the respective language. Metrics include **Model Utility (MU)**, **Prob. Retain (PR)**, **Prob. Forget (PF)**, and **Truth Ratio Forget (TRF)**.

Language	Metric	Finetuned	Retain	en	fr	fa	ar	hi	iw	id	ru	ja	ko
en	MU	0.58	0.59	0.61	0.59	0.58	0.58	0.58	0.58	0.59	0.58	0.58	0.59
	PR	0.98	0.98	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	PF	0.98	0.09	0.25	0.71	0.92	0.91	0.94	0.88	0.80	0.83	0.93	0.83
	TRF	0.48	0.67	0.53	0.50	0.49	0.47	0.49	0.49	0.50	0.50	0.48	0.51
fr	MU	0.51	0.51	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
	PR	0.97	0.97	0.96	0.95	0.96	0.96	0.97	0.96	0.96	0.96	0.97	0.97
	PF	0.96	0.10	0.68	0.33	0.87	0.88	0.93	0.85	0.76	0.84	0.91	0.84
	TRF	0.48	0.69	0.51	0.56	0.51	0.49	0.51	0.50	0.51	0.52	0.50	0.52
fa	MU	0.43	0.44	0.44	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43
	PR	0.94	0.94	0.93	0.93	0.91	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	PF	0.91	0.10	0.86	0.83	0.23	0.82	0.83	0.83	0.85	0.83	0.87	0.87
	TRF	0.56	0.70	0.58	0.57	0.67	0.58	0.59	0.57	0.57	0.58	0.56	0.56
ar	MU	0.43	0.43	0.44	0.43	0.43	0.44	0.43	0.43	0.43	0.43	0.43	0.43
	PR	0.94	0.95	0.93	0.93	0.93	0.92	0.94	0.93	0.93	0.93	0.94	0.94
	PF	0.91	0.10	0.83	0.82	0.79	0.26	0.86	0.83	0.85	0.84	0.88	0.88
	TRF	0.51	0.64	0.54	0.51	0.52	0.54	0.53	0.52	0.53	0.52	0.52	0.51
hi	MU	0.39	0.40	0.39	0.39	0.39	0.40	0.39	0.39	0.39	0.39	0.39	0.39
	PR	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.97
	PF	0.98	0.31	0.96	0.96	0.94	0.96	0.56	0.96	0.96	0.95	0.94	0.95
	TRF	0.73	0.81	0.74	0.73	0.73	0.72	0.77	0.73	0.74	0.73	0.73	0.75
iw	MU	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
	PR	0.93	0.93	0.92	0.92	0.92	0.92	0.92	0.91	0.92	0.92	0.92	0.93
	PF	0.92	0.11	0.81	0.81	0.85	0.86	0.89	0.26	0.85	0.85	0.89	0.89
	TRF	0.57	0.73	0.59	0.57	0.58	0.57	0.58	0.61	0.58	0.58	0.58	0.57
id	MU	0.51	0.50	0.52	0.52	0.51	0.52	0.51	0.51	0.52	0.51	0.50	0.51
	PR	0.96	0.96	0.94	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95
	PF	0.95	0.08	0.75	0.72	0.85	0.88	0.92	0.84	0.29	0.85	0.91	0.84
	TRF	0.48	0.66	0.51	0.52	0.49	0.49	0.50	0.48	0.53	0.51	0.48	0.51
ru	MU	0.44	0.45	0.46	0.45	0.44	0.45	0.44	0.44	0.45	0.44	0.44	0.45
	PR	0.93	0.93	0.92	0.92	0.92	0.92	0.93	0.92	0.92	0.92	0.93	0.93
	PF	0.90	0.08	0.76	0.74	0.82	0.85	0.85	0.81	0.81	0.24	0.86	0.83
	TRF	0.55	0.69	0.56	0.56	0.56	0.55	0.56	0.55	0.56	0.59	0.56	0.55
ja	MU	0.50	0.50	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	PR	0.92	0.92	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.90	0.91
	PF	0.91	0.13	0.81	0.83	0.86	0.88	0.82	0.88	0.86	0.86	0.25	0.76
	TRF	0.62	0.74	0.63	0.62	0.62	0.59	0.62	0.61	0.62	0.63	0.65	0.63
ko	MU	0.47	0.49	0.48	0.47	0.47	0.48	0.47	0.47	0.48	0.47	0.47	0.48
	PR	0.92	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.90
	PF	0.93	0.10	0.68	0.73	0.89	0.90	0.84	0.88	0.83	0.86	0.79	0.26
	TRF	0.55	0.67	0.56	0.56	0.55	0.54	0.55	0.55	0.56	0.55	0.54	0.59

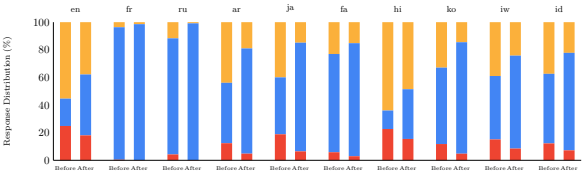
Table 5: Full results of unlearning experiments on the **TOFU** dataset using the **NPO** method across ten languages. Each row group corresponds to the evaluation language, while each column (after *Finetuned* and *Retain*) represents a model that has been unlearned on the respective language. Metrics include **Model Utility (MU)**, **Prob. Retain (PR)**, **Prob. Forget (PF)**, and **Truth Ratio Forget (TRF)**.



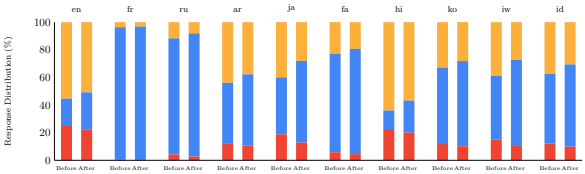
(a) GradDiff-KL (unlearned on `fr`)



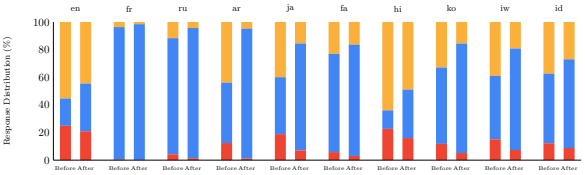
(b) NPO (unlearned on `fr`)



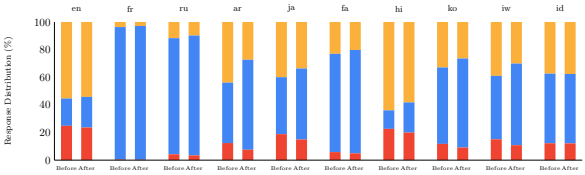
(c) GradDiff-KL (unlearned on `ru`)



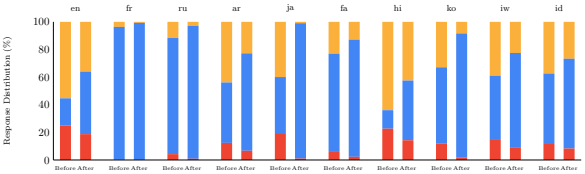
(d) NPO (unlearned on `ru`)



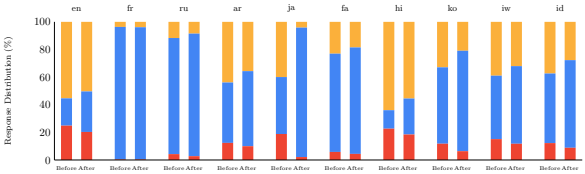
(e) GradDiff-KL (unlearned on `ar`)



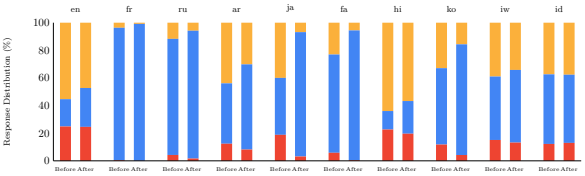
(f) NPO (unlearned on `ar`)



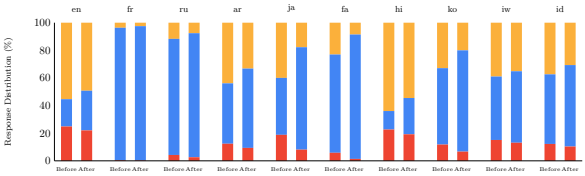
(g) GradDiff-KL (unlearned on `ja`)



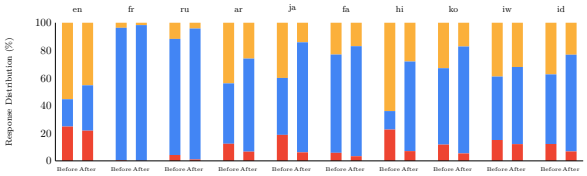
(h) NPO (unlearned on `ja`)



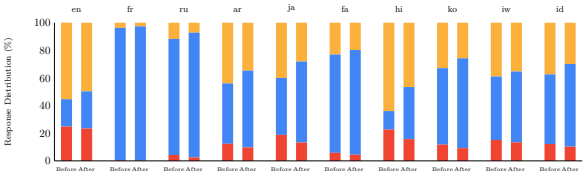
(i) GradDiff-KL (unlearned on `fa`)



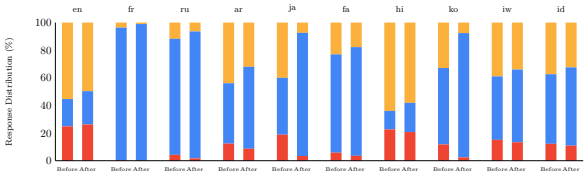
(j) NPO (unlearned on `fa`)



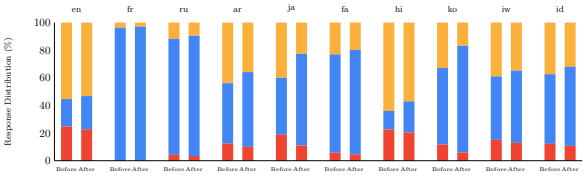
(k) GradDiff-KL (unlearned on hi)



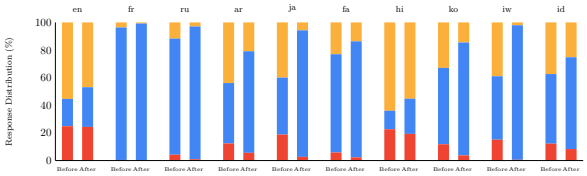
(l) NPO (unlearned on hi)



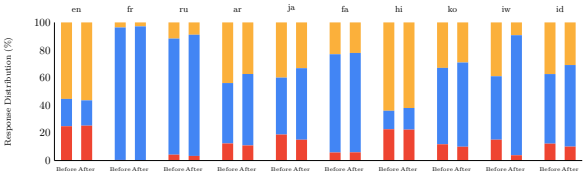
(m) GradDiff-KL (unlearned on ko)



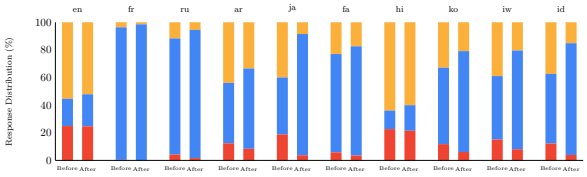
(n) NPO (unlearned on ko)



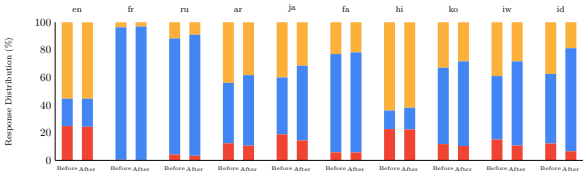
(o) GradDiff-KL (unlearned on iw)



(p) NPO (unlearned on iw)



(q) GradDiff-KL (unlearned on id)



(r) NPO (unlearned on id)

Figure 13: Results on the SeeGULL QA dataset across nine languages (excluding English) before and after unlearning. Each row shows GradDiff-KL (left) and NPO (right) for the specified unlearning language.