# Cross-Lingual Unlearning of Selective Knowledge in Multilingual Language Models

**Anonymous ACL submission**

## Abstract

Pretrained language models memorize vast amounts of information, including private and copyrighted data, raising significant safety concerns. Retraining these models after excluding sensitive data is prohibitively expensive, making machine unlearning a viable, cost-effective alternative. Previous research has focused on machine unlearning for monolingual models, but we find that unlearning in one language does not necessarily transfer to others. This vulnerability makes models susceptible to low-resource language attacks, where sensitive information remains accessible in less dominant languages. This paper presents a pioneering approach to machine unlearning for multilingual language models, selectively erasing information across different languages while maintaining overall performance. Specifically, our method employs an adaptive unlearning scheme that assigns language-dependent weights to address different language performances of multilingual language models. Empirical results demonstrate the effectiveness of our framework compared to existing unlearning baselines, setting a new standard for secure and adaptable multilingual language models.[1]


Figure 1: Language models may have memorized the copyrighted data *The Little Prince* in multiple languages. Consequently, removing such information in just one language does not entirely eradicate it from the model. This underscores the necessity for a multilingual unlearning approach to ensure the information is thoroughly eliminated from the model.

## 1 Introduction

Privacy regulations such as the Right to be Forgotten (RTBF) (Rosen, 2011), the European Union's General Data Protection Regulation (GDPR) (Hoofnagle et al., 2019), and the United States' California Consumer Privacy Act (CCPA) (Pardau, 2018) mandate that individuals have the right to request the deletion of their data from databases, which extends to data held within machine learning (ML) models. Additionally, the Writers Guild of America strike in 2023 highlighted increasing concerns regarding the copyright content generated by large language models (LLMs) (WGA, 2023).

To comply with such issues, significant attention has been directed towards the task of machine unlearning (MU), which involves removing the influence of specific data points from ML models (Cao and Yang, 2015). Despite the critical necessity of the task, mitigating the influence of data samples on billions of model parameters presents an immense challenge. The most definitive method is *exact unlearning*, which necessitates retraining ML models entirely from scratch, utilizing the residual training dataset after excising the specified data points. However, this method is computationally prohibitive and not feasible, particularly for LLMs. Therefore, the advancement of rapid *approximate unlearning* methodologies has emerged as a primary focus of contemporary research efforts.

Research on MU has predominantly focused on

[1] To encourage future research and replication of our work, the code will be released upon acceptance.

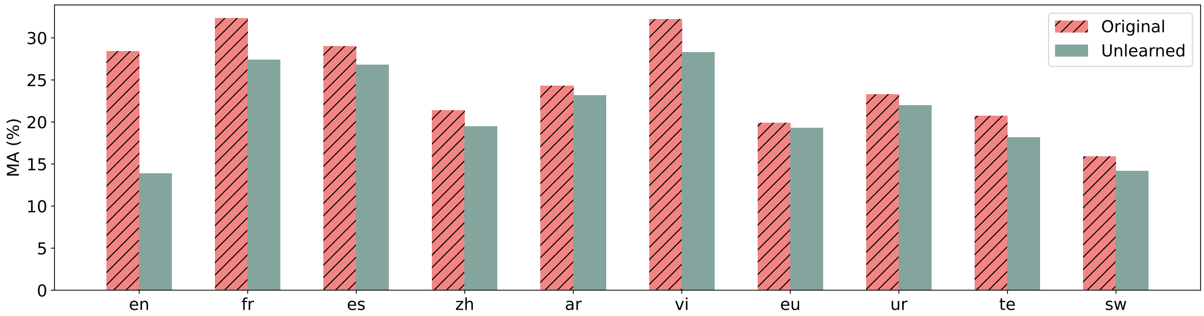


Figure 2: Memorization accuracy (MA) of the multilingual model BLOOM across various languages after unlearning with English data only. The plot illustrates that MA does not significantly drop across other languages, highlighting the necessity for a multilingual unlearning approach to effectively reduce memorization across all languages.

computer vision tasks (Bourtoule et al., 2021; Golatkar et al., 2020a,b; Chundawat et al., 2023; Kurmanji et al., 2023); however, it is now gaining traction in NLP due to the safety issues that arise with LLMs (Zhang et al., 2023). Notably, Jang et al. (2023) first proposed an unlearning technique of reversing the gradient to refrain LMs from generating particular sensitive token sequences. On the other hand, Wang et al. (2023) presented an approach to maintaining distribution differences (i.e., knowledge gap) such that the performance of the data to be forgotten becomes similar to the performance of the unseen data. Besides the two approaches, substantial progress has been made in unlearning for monolingual models; nevertheless, there is a lack of empirical results and analyses of unlearning for multilingual LMs. As shown in Figures 1 and 2, our preliminary experiments find that existing unlearning approaches do not exhibit cross-lingual transferability. In other words, unlearning in one language does not automatically transfer to other languages, leaving LMs vulnerable to possible low-resource language attacks, which have been shown to jailbreak GPT-4 (Yong et al., 2023).

To this end, we introduce *multilingual unlearning*, which effectively removes specific information across a wide variety of languages from pretrained language models.[2] Due to the inconsistency in model performance across languages, we leverage a multilingual teacher model in which the student model adaptively obeys the teacher based on its capabilities in a particular language. For example, a high knowledge distillation weight is applied when the teacher has strong expertise, ensuring the benefit of effective teaching. Conversely, a low weight is used when the teacher's knowledge is limited, allowing the student to learn independently. Our method is also as time-efficient as unlearning a single language, offering a significant improvement over unlearning languages one at a time, making it more practical for real-world applications.

To assess the success of unlearning across different languages, our experimental setup necessitates multilingual parallel data. However, obtaining such datasets is challenging, especially when dealing with a particular domain. Consequently, we evaluate our framework using two multilingual parallel datasets in the general domain, which are utilized to unlearn specific token sequences and factual knowledge across various languages, respectively. Empirical results demonstrate that our proposed approach surpasses existing unlearning methods by a considerable margin. Given the intrinsic similarities in unlearning token sequences, we believe these datasets provide an appropriate testbed for evaluating multilingual unlearning.

Overall, the major contributions of our work are as follows:

- We introduce *multilingual unlearning*, a process that selectively deletes information across a wide range of languages from pretrained LMs. To the best of our knowledge, we are the first to explore machine unlearning in a multilingual context.
- We propose a novel adaptive unlearning scheme using a multilingual teacher model to cope with varying model performance across different languages.
- We provide a multilingual unlearning testbed and empirically demonstrate that the performance of our proposed approach exceeds that of current unlearning methods.

[2]Although the task can be applied to monolingual models that may have some multilingual capabilities, we focus on multilingual LMs to limit the scope of our study.

## 2 Related Work

### 2.1 Machine Unlearning

Cao and Yang (2015) first coined the term *machine unlearning*, defining it as successfully deleting an example when the outputs on a dataset are the same as if the example had never been added. They achieved this by transforming learning algorithms into a summation form, allowing the system to forget a training data sample by updating only a few summations. Later, Ginart et al. (2019) proposed a probabilistic definition inspired by Differential Privacy (Dwork et al., 2014), requiring the unlearning model to produce outputs *similar* to those of a model retrained from scratch without the forgotten data. This inspired deep approximate unlearning methods, such as those using the Fisher information matrix (Golatkar et al., 2020a; Mehta et al., 2022) and neural tangent kernel (Golatkar et al., 2020b). However, these methods do not scale well, making them impractical for language models with billions of parameters. More recently, Chundawat et al. (2023) proposed a method using two teachers (competent and incompetent) to help a student forget certain samples while retaining the rest of the information. Kurmanji et al. (2023) suggested a similar approach with a single teacher. Both methods aimed to safely forget selective samples using a teacher-student framework, primarily focusing on image classification tasks. Our work takes a step further and considers the multilingual capabilities of the teacher, assigning language-specific weights to the distillation process.

### 2.2 Knowledge Unlearning

Jang et al. (2023) introduced *knowledge unlearning*, aimed at preventing language models from generating specific token sequences. They proposed a straightforward method by inverting the original training objective of minimizing the negative log-likelihood of the token sequences. To maintain the performance of the remaining knowledge, Wang et al. (2023) and Chen and Yang (2023) employed the Kullback-Leibler (KL) divergence loss, minimizing the distributional differences between the original and unlearned models on the retained data. Our approach builds on these methods but differs in its focus. While the previous works targeted monolingual models like DistilBERT (Sanh et al., 2019) and T5 (Raffel et al., 2020), we extend the unlearning process to a multilingual context.

### 2.3 Cross-Lingual Transfer

Multilingual pretrained language models (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2021; Lin et al., 2022; Le Scao et al., 2023) have shown to exhibit remarkable cross-lingual transfer across various tasks by leveraging shared semantic spaces and joint training techniques to bridge language gaps. However, Xu et al. (2023) demonstrated that editing knowledge in one language does not propagate to others and thus introduced cross-lingual model editing, a technique using random sampling of languages to improve model adaptability and robustness in a multilingual context. Additionally, Qi et al. (2023) investigated the cross-lingual consistency of factual knowledge in multilingual models, finding that factual knowledge does not remain consistent across languages, but only when languages share a larger portion of vocabulary. Building on these advancements, our work employs multilingual language models to investigate the cross-lingual transfer of machine unlearning and proposes an effective method to unlearn specific information across languages, addressing the need for precise and reliable information removal in a multilingual context.

## 3 Methodology

### 3.1 Problem Definition

Given a token sequence $\mathbf{x} = \{x\}_{i=1}^{T}$ in the training dataset $\mathcal{D} = \{\mathbf{x}\}_{i=1}^{N}$, the task of knowledge unlearning is to safely remove the influence of a subset of data $\mathcal{D}_f$ from a trained machine learning model such that the model behaves as if the removed data had never been part of the training process, thus maintaining the model performance for the rest of the dataset. Conventionally, the data to be forgotten $\mathcal{D}_f$ is expressed as the *forget set*, while the data to be retained $\mathcal{D}_r$ is named as the *retain set*. For simplicity, we consider the standard case where $\mathcal{D}_f$ and $\mathcal{D}_r$ represent the whole training dataset and are mutually exclusive; that is, $\mathcal{D}_f \cup \mathcal{D}_r = \mathcal{D}$ and $\mathcal{D}_f \cap \mathcal{D}_r = \varnothing$. The objective is to adjust the model parameters $\theta$ such that the updated parameters $\theta' = S(\theta; \mathcal{D}_f)$ reflect the removal of $\mathcal{D}_f$. This unlearning (scrubbing) function $S$ ensures the model behaves as if trained solely on $\mathcal{D}_r$, effectively forgetting $\mathcal{D}_f$ while maintaining performance on $\mathcal{D}_r$. Extending to a multilingual context, the definition must hold for all datasets across languages $\mathcal{Z} = \{z\}_{i=1}^{Z}$.

## 3.2 Knowledge Unlearning

The primary objective in language modeling is to minimize the log-likelihood of token sequences, training the model to predict the next word in a sequence accurately. Knowledge unlearning (Jang et al., 2023) involves *negating* this objective to remove specific learned information from the model. Instead of reinforcing certain sequences, unlearning aims to decrease their probability by maximizing their log-likelihood:

$$\mathcal{L}_f = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(x_t|x_{<t}), \tag{1}$$

where $x$ comes from a sequence of tokens $\mathbf{x}_f \in \mathcal{D}_f$ and $p_\theta(x_t|x_{<t})$ denotes the conditional probability of predicting the next token given the model parameters $\theta$. This effectively reverses the learned patterns, reducing the probability of generating the targeted sequences and allowing the model to "forget" specific knowledge.

## 3.3 Language-Adaptive Unlearning

After forgetting a subset of data, many previous works highlight the critical need to retain the rest of the knowledge explicitly (Wang et al., 2023; Chen and Yang, 2023). This involves adjusting the model so that its performance on retained data aligns closely with the original model as if the forgotten samples never existed. Formally, this can be expressed as minimizing the KL divergence between the original model and the unlearned model on the retained data:

$$\mathcal{L}_{LT} = \frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}(p_{\theta_0}(\cdot|x_{<t}) \parallel p_\theta(\cdot|x_{<t})), \tag{2}$$

where $x$ represents a token from the sequence $\mathbf{x}_r \in \mathcal{D}_r$ and $\theta_0$ denotes the original (teacher) model with frozen weights, ensuring that the student model weights remain aligned with the teacher model on retained examples. This approach works optimally when the teacher model performs well on $\mathcal{D}_r$ at initialization. However, for a multilingual language model, performance may be suboptimal for languages that were insufficiently represented during pretraining. In such cases, it is more beneficial for the student model to learn independently when the teacher model's language capability is poor. The student model can do this by training on hard labels using a standard language modeling objective:

| Property | FLORES-200 | BMLAMA-53 |
|---|---|---|
| Train-Forget | 32 | 32 |
| Train-Retain | 32-128 | 32-128 |
| Validation | 357 | 1023 |
| Test | 1012 | 1024 |
| Languages | 10 / 206 | 9 / 53 |
| Data Type | Token Sequence | Factual Knowledge |

Table 1: Dataset statistics. Due to the unavailability of training data, we created our own training splits for our experiments. The number of retaining samples varies depending on the model (see Appendix A.1). We selected 10 languages for FLORES and 9 for BMLAMA, ensuring compatibility with the multilingual models used.

$$\mathcal{L}_{LM} = \frac{1}{T}\sum_{t=1}^{T} \log p_\theta(x_t|x_{<t}). \tag{3}$$

This represents the positive counterpart of Equation 1. Employing an adaptive weighting scheme, we can combine the language teaching loss and the language modeling loss:

$$\mathcal{L}_r = \kappa \cdot \mathcal{L}_{LT} + (1-\kappa)\cdot \mathcal{L}_{LM}, \tag{4}$$

where $\kappa = \frac{1}{T}\sum_{t=1}^{T} p_{\theta_0}(\cdot|x_{<t})$ represents the confidence of the teacher in token sequence $\mathbf{x}$ for the given language. This implies that the student learns from the teacher when the teacher's confidence is high; otherwise, the student learns independently to retain examples. Finally, combining all losses, we obtain

$$\mathcal{L} = \mathcal{L}_f + \lambda \cdot \mathcal{L}_r, \tag{5}$$

where $\lambda$ is a scaling hyperparameter. In practice, we follow Kurmanji et al. (2023), alternating the updates for the forget set and the retain set to optimize *min-max* terms in $\mathcal{L}$ more stably. Furthermore, this objective supports *token-level* unlearning, indicating that cross-lingual transfer will only occur if the languages share the same vocabulary, as noted by Qi et al. (2023). To facilitate fast and effective cross-lingual unlearning, we randomly sample languages for token sequences in both the forget and retain sets, following Xu et al. (2023). We demonstrate in Section 5 that this approach achieves comparable performance to unlearning one language at a time, with significantly improved efficiency.

# 4 Experimental Setup

## 4.1 Datasets

We evaluate our framework using two multilingual datasets FLORES-200 (NLLB Team, 2022)

| Model | Method | Forget Set | | | | | | Test Set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | EN | | HIGH-SRC | | LOW-SRC | | EN | | HIGH-SRC | | LOW-SRC | |
| | | MA(↓) | PPL(↑) | MA(↓) | PPL(↑) | MA(↓) | PPL(↑) | MA | PPL | MA | PPL | MA | PPL |
| XGLM-564M | Original | 34.6 | 117.4 | 34.8 | 136.8 | 30.8 | 150.3 | 35.4 | 107.1 | 35.3 | 120.2 | 30.7 | 153.9 |
| | GradAscent+ | 22.7 | 4242.9 | 23.9 | **6274.9** | 19.8 | **16858.8** | 32.2 | 301.0 | 32.0 | 440.1 | 26.1 | 1169.8 |
| | NegTaskVector+ | 22.7 | 663.7 | 24.8 | 521.9 | 22.0 | 548.6 | 30.8 | 191.3 | 32.3 | 168.5 | 28.0 | 203.5 |
| | LINGTEA (ours) | **19.3** | **4261.8** | **22.2** | 816.5 | **19.5** | 1031.2 | 30.8 | 114.8 | 31.7 | 85.2 | 26.4 | 127.8 |
| | Oracle | *17.2* | *6295.8* | *18.7* | *4579.1* | *16.6* | *4780.3* | *32.4* | *114.4* | *33.3* | *86.9* | *29.0* | *113.2* |
| XGLM-2.9B | Original | 36.8 | 66.6 | 37.8 | 104.8 | 36.5 | 90.6 | 38.5 | 68.6 | 39.2 | 90.1 | 35.6 | 99.1 |
| | GradAscent+ | 26.3 | 16553.7 | 26.4 | **3504002.3** | **22.3** | **1613956.2** | 36.2 | 922.8 | 36.2 | 790.4 | 30.8 | 6934.2 |
| | NegTaskVector+ | 25.4 | 206.8 | 28.3 | 184.8 | 25.5 | 246.9 | 33.8 | 78.4 | 35.9 | 59.8 | 32.6 | 91.1 |
| | LINGTEA (ours) | **19.9** | **10216406.9** | **23.5** | 428687.2 | 23.1 | 56735.2 | 35.2 | 102.4 | 35.7 | 116.4 | 30.6 | 172.6 |
| | Oracle | *19.7* | *11605.0* | *22.4* | *38993.3* | *18.9* | *177055.2* | *38.2* | *70.5* | *38.7* | *51.3* | *34.6* | *71.2* |
| BLOOM-560M | Original | 28.4 | 81.0 | 27.9 | 86.4 | 19.9 | 603.4 | 29.5 | 73.2 | 28.8 | 78.4 | 19.4 | 565.7 |
| | GradAscent+ | 25.1 | 127.0 | 23.7 | 142.4 | 16.5 | 1993.1 | 29.7 | 72.4 | 28.6 | 80.9 | 19.1 | 686.0 |
| | NegTaskVector+ | 22.7 | 277.1 | 21.1 | 290.3 | 14.2 | 2682.3 | 28.6 | 83.0 | 27.9 | 89.6 | 18.8 | 723.4 |
| | LINGTEA (ours) | **18.2** | **2787.0** | **20.2** | **1793.0** | **13.8** | **6550.6** | 28.5 | 86.7 | 28.6 | 96.5 | 19.0 | 580.8 |
| | Oracle | *13.9* | *12702.6* | *13.3* | *93205.8* | *9.9* | *103180.6* | *31.0* | *71.8* | *30.2* | *86.4* | *20.6* | *435.2* |
| BLOOM-3B | Original | 35.8 | 42.4 | 35.1 | 51.5 | 27.2 | 149.2 | 36.6 | 42.7 | 35.7 | 45.4 | 27.0 | 154.9 |
| | GradAscent+ | 25.5 | 291.2 | 24.1 | 913.0 | **15.0** | 7348.4 | 35.6 | 54.5 | 34.5 | 65.9 | 25.2 | 311.8 |
| | NegTaskVector+ | 28.7 | 119.7 | 27.9 | 135.8 | 20.0 | 622.2 | 36.6 | 42.8 | 35.6 | 44.6 | 26.5 | 168.7 |
| | LINGTEA (ours) | **17.8** | **21063692.6** | **21.0** | **711058.2** | 17.0 | **63395.3** | 35.5 | 51.0 | 34.9 | 60.0 | 24.9 | 233.0 |
| | Oracle | *13.8* | *134342.4* | *13.4* | *321033.9* | *9.2* | *467830.4* | *35.7* | *49.5* | *35.5* | *51.8* | *26.9* | *162.0* |

Table 2: Performance of unlearning multilingual token sequences on FLORES-200. Oracle, serving as a reference, unlearns one language at a time. All other methods dynamically sample languages at runtime for multilingual unlearning, prioritizing the retention of PPL on the retain set. High-resource languages include FR, ES, ZH, AR, and VI, while low-resource languages include EU, UR, TE, and SW, with performance metrics averaged across these languages. Detailed results for each language are available in Appendix B.1.

and BMLAMA-53 (Qi et al., 2023). Detailed data statistics are presented in Table 1. FLORES-200 is a high-quality machine translation benchmark containing parallel sentences in 206 languages, including many extremely low-resource languages. BMLAMA-53 is a balanced version of the multilingual factual knowledge probing dataset mLAMA (Kassner et al., 2021), keeping only the parallel facts across languages. It is important to note that these datasets do not contain sensitive data, such as private or copyrighted information. High-quality multilingual parallel datasets are rare, especially in specific domains. Despite being general domain datasets, we consider them effective alternatives to sensitive data, as unlearning token sequences would function similarly.

### 4.2 Baselines

We compare our framework with several strong unlearning approaches and various baselines: **Original**: The "original" model without any unlearning applied. **GradAscent+**: This method begins with the original model and finetunes it on both the retain and forget sets, using gradient ascent on the latter. Previous work (Jang et al., 2023) examined a weaker baseline that only trains on the forget set with gradient ascent. We enhance GradAscent+ to achieve a better balance between retention and forgetting. **NegTaskVector+**: This approach also starts from the original model but finetunes two separate models, one on the forget set and another on the retain set. During inference, the weights of the forget-set-tuned model are negated, while the retained weights are added. Prior research (Ilharco et al., 2023) explored a weaker baseline training only on the forget set. Our refined version includes explicit retention tuning. **Oracle**: Serves as a reference point where our proposed method is applied one language at a time. This represents the "pseudo" upper bound performance of our approach, achieved inefficiently as the number of languages increases, i.e., $O(Z)$. We do not directly compare with other teacher-student frameworks for unlearning (Chundawat et al., 2023; Kurmanji et al., 2023), as their training objectives involve a classification loss to forget a class label. Instead, we evaluate our adaptive unlearning scheme against the general knowledge distillation framework to demonstrate its effectiveness, as detailed in Section 5.3.

### 4.3 Evaluation Metrics

Following Jang et al. (2023), we evaluate unlearning for token sequences using **Memorization Accuracy (MA)** as defined by Tirumala et al. (2022):

$$\mathrm{MA}(\mathbf{x}) = \frac{\sum_{t=1}^{T-1} \mathbb{1}\{\operatorname{argmax}(p_\theta(\cdot|x_{<t})) = x_t\}}{T-1}. \tag{6}$$

This metric quantifies the extent to which the model has memorized the given token sequence. For as-

| | | Forget Set | | | | | | Test Set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | EN | | HIGH-SRC | | MID-SRC | | EN | | HIGH-SRC | | MID-SRC | |
| **Model** | **Method** | **PA(↓)** | **PPL(↑)** | **PA(↓)** | **PPL(↑)** | **PA(↓)** | **PPL(↑)** | **PA** | **PPL** | **PA** | **PPL** | **PA** | **PPL** |
| XGLM-564M | Original | 28.1 | 122.0 | 13.8 | 116.9 | 18.8 | 78.7 | 29.9 | 152.8 | 17.0 | 135.0 | 17.5 | 95.9 |
| | GradAscent+ | 27.1 | **187.2** | 7.3 | 173.9 | 12.5 | 99.8 | 30.3 | 187.8 | 16.7 | 162.8 | 16.6 | 100.8 |
| | NegTaskVector+ | 28.1 | 150.7 | 7.3 | 142.1 | 11.8 | 90.7 | 30.1 | 145.9 | 18.0 | 130.0 | 17.5 | 90.3 |
| | LINGTEA (ours) | **25.0** | 185.3 | **5.4** | **179.6** | **10.8** | **102.1** | 29.4 | 165.5 | 17.3 | 138.8 | 16.9 | 88.7 |
| | Oracle | *5.2* | *71681.6* | *2.5* | *3185.4* | *2.4* | *738.5* | *28.6* | *1249.1* | *16.3* | *367.5* | *15.0* | *198.8* |
| XGLM-2.9B | Original | 34.4 | 90.9 | 15.6 | 82.7 | 25.0 | 48.6 | 34.7 | 112.7 | 21.9 | 95.3 | 19.5 | 59.1 |
| | GradAscent+ | 29.2 | 133.5 | 11.0 | 205.9 | **11.8** | 188.8 | 35.4 | 127.5 | 21.2 | 174.5 | 17.7 | 156.3 |
| | NegTaskVector+ | 29.2 | 124.6 | 9.4 | 127.2 | 12.5 | 72.2 | 33.4 | 120.2 | 20.4 | 109.7 | 18.8 | 64.8 |
| | LINGTEA (ours) | **14.6** | **908.6** | **6.9** | **678.3** | 12.8 | **480.0** | 37.1 | 156.5 | 24.4 | 137.5 | 21.6 | 131.4 |
| | Oracle | *13.5* | *1274.7* | *5.4* | *552.8* | *4.5* | *2982.7* | *43.3* | *176.1* | *27.0* | *107.9* | *24.1* | *133.6* |
| BLOOM-560M | Original | 31.3 | 145.8 | 18.8 | 145.0 | 10.4 | 267.5 | 28.5 | 202.6 | 17.3 | 159.7 | 12.4 | 257.0 |
| | GradAscent+ | 15.6 | 238.8 | 11.3 | 220.5 | 6.9 | 364.6 | 28.5 | 237.6 | 16.7 | 184.6 | 11.7 | 280.8 |
| | NegTaskVector+ | 22.9 | 184.9 | 12.9 | 168.1 | 7.3 | 331.4 | 29.0 | 204.7 | 17.3 | 148.7 | 12.1 | 253.5 |
| | LINGTEA (ours) | **9.4** | **267.5** | **6.9** | **267.7** | **5.6** | **492.0** | 27.4 | 206.4 | 17.0 | 162.5 | 12.2 | 308.3 |
| | Oracle | *7.3* | *629.3* | *2.7* | *6814.3* | *1.0* | *822.7* | *29.6* | *204.4* | *18.1* | *199.6* | *11.7* | *265.1* |
| BLOOM-3B | Original | 50.0 | 68.9 | 24.4 | 74.8 | 14.6 | 95.0 | 46.6 | 89.5 | 26.8 | 79.2 | 16.1 | 99.9 |
| | GradAscent+ | **16.7** | 645.1 | 7.7 | 617.7 | **5.9** | 402.4 | 40.8 | 258.6 | 23.6 | 168.2 | 14.9 | 173.8 |
| | NegTaskVector+ | 35.4 | 110.8 | 16.0 | 128.4 | 7.3 | 183.4 | 47.2 | 104.4 | 24.7 | 93.4 | 14.6 | 119.9 |
| | LINGTEA (ours) | 19.8 | **1077.3** | **6.3** | **781.8** | 6.9 | **725.9** | 47.1 | 137.0 | 32.0 | 90.5 | 18.3 | 176.9 |
| | Oracle | *17.7* | *2708.9* | *7.3* | *385.1* | *2.4* | *1778.2* | *46.1* | *136.5* | *34.9* | *63.6* | *20.2* | *139.6* |

Table 3: Performance of unlearning multilingual factual knowledge on BMLAMA-53. High-resource languages consist of FR, ES, PT, AR, and VI, while mid-resource languages consist of CA, HI, and BN. The performance metrics presented are averaged across these languages, with detailed results for each language provided in Appendix B.2.

sessing the unlearning of factual knowledge, we adopt the approach of Petroni et al. (2019) and report **Probing Accuracy (PA)**, which is a rank-based metric that calculates the mean precision at $k$ ($P@k$) across all relations, with $k$ set to 1. This means that for a given fact, the value is 1 if the object is ranked among the top $k$ results, and 0 otherwise. Additionally, we measure the **Perplexity (PPL)** of token sequences to determine how surprised the model is by the data.

### 4.4 Implementation Details

All experiments were conducted using PyTorch and Huggingface's Transformers library (Wolf et al., 2020). We employed two multilingual language models: XGLM (564M, 2.9B) (Lin et al., 2022) and BLOOM (560M, 3B) (Le Scao et al., 2023). Model weights were optimized using AdamW (Loshchilov and Hutter, 2019), and hyperparameters were tuned to minimize MA/PA on the forget set while maintaining the original PPL on the validation set. Note that this differs from Jang et al. (2023), focusing only on minimizing MA due to the lack of a retaining procedure, whereas our priority is retaining model utility after unlearning. To match the number of samples to forget, we set the batch size to 32 to facilitate simultaneous forgetting. Detailed hyperparameter settings are provided in Appendix A.1. Each experiment was repeated with three different random seeds, and the results were averaged for reporting.

## 5 Results and Analyses

### 5.1 Token Sequence Unlearning

We compare the token sequence unlearning results across various methods and report them in Table 2. For each method, we aimed to identify the configuration where PPL remains close to the validation PPL of the original model. Otherwise, while achieving a 0% MA on the forget set is possible, it would significantly degrade the model performance on other tasks. In that sense, the effectiveness of an approach in retaining the remaining information determines the extent of unlearning that can be applied safely to remove specific information. At the point where GradAscent+ and NegTaskVector+ retain the performance of the test set, the models cannot be unlearned further to preserve the model utility, limiting their capacity for more robust unlearning. In contrast, our method, LINGTEA, achieves better unlearning performance due to maintaining adaptive proximity to the teacher model. Additionally, LINGTEA demonstrates comparable performance to Oracle for XGLM models; however, single-language unlearning shows significantly lower values for the BLOOM models, indicating room for improvement. We leave the exploration of varying behaviors across multilingual LMs to future work.

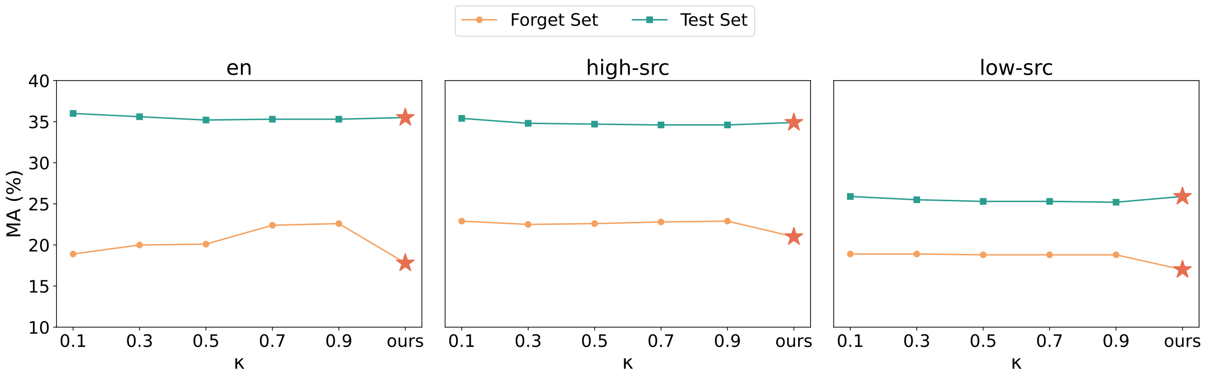


Figure 3: Comparison of the forget set and test set performance of BLOOM-3B after unlearning on FLORES-200 for EN, HIGH-SRC, and LOW-SRC across different $\kappa$ values. Our adaptive unlearning scheme yields the lowest MA on the forget set and maintains a competitive MA on the test set, highlighting the superiority of the approach.

## 5.2 Factual Knowledge Unlearning

We present the results of factual knowledge unlearning across various methods in Table 3. Factual knowledge is probed using fill-in-the-blank cloze statements like "Paris is the capital of [MASK]", where the language model predicts the masked token. Although this is also a token sequence, the unlearning process differs as we focus on removing information about the answer token(s) in the context, preventing the model from generating the correct answer, "France". This approach may lead to hallucinations when dealing with actual factual knowledge, where editing might be more suitable. However, we argue that it relates to unlearning specific *parts* of information, such as the names of copyrighted characters in multiple languages. We measure the PPL of the entire answer sentence, as measuring PPL only on the answer token(s) can result in disproportionately high values. Our method, similar to unlearning token sequences, generally outperforms other methods across various metrics, showcasing its effectiveness. It is worth noting that English factual knowledge is hardly removed from XGLM-564M. We believe that techniques like weighted random sampling of languages, which we did not explore in this study, may help reduce memorization.

## 5.3 Effect of Adaptive Unlearning

To evaluate the effectiveness of our adaptive unlearning scheme, we fix various $\kappa$ values and compare them against our proposed method. As illustrated in Figure 3, the adaptive unlearning approach implemented in LINGTEA consistently achieves the lowest MA on the forget set across all categories, including English, high-resource, and low-resource languages. Moreover, LINGTEA exhibits competitive performance on the test set, indicating its ability to retain knowledge effectively. These findings demonstrate that selectively adapting to the teacher's strengths in specific languages enhances the overall multilingual unlearning process.

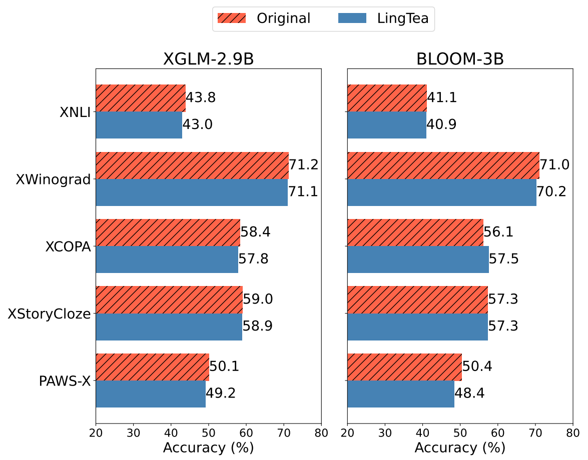


Figure 4: Zero-shot performance comparison between the original model and our LINGTEA framework across five multilingual language understanding tasks. The results demonstrate that LINGTEA retains world knowledge on par with the original model, ensuring the safety and efficacy of our unlearning approach.

## 5.4 Retaining World Knowledge

While our unlearning approach may succeed in retaining the test set, it is equally important to assess whether it has preserved the original multilingual language model capabilities. To verify the retention of world knowledge, we compare our framework with the original model across five multilingual language understanding tasks: natural language inference (XNLI) (Conneau et al., 2018), coreference

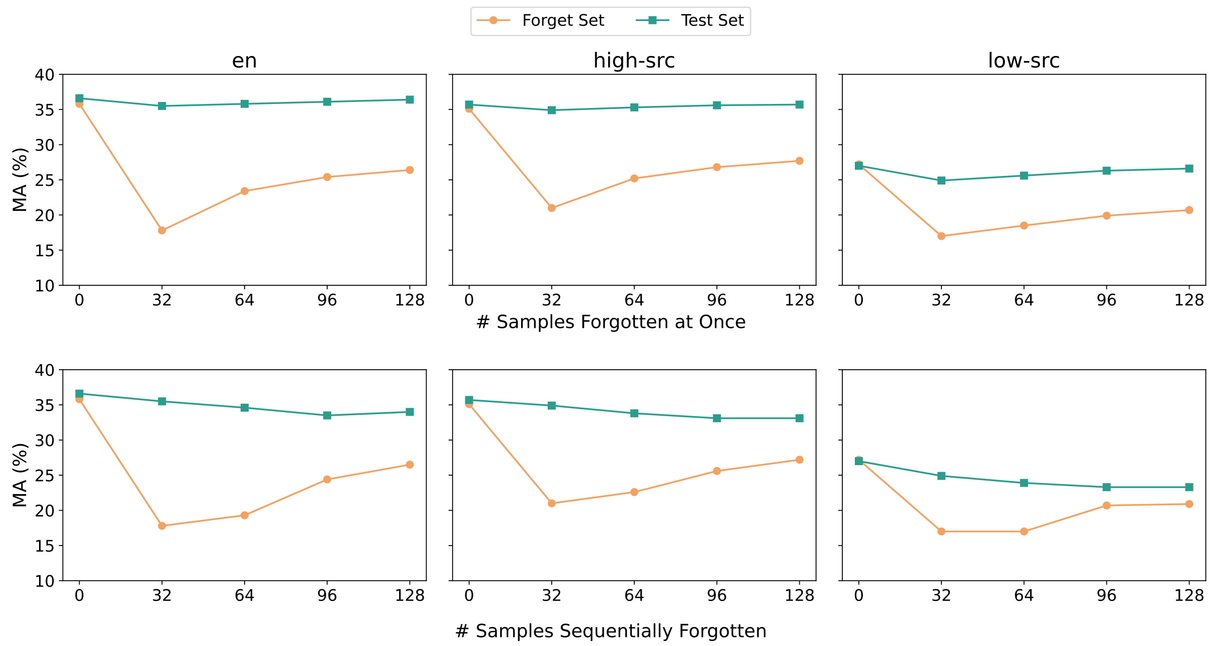


Figure 5: Performance of BLOOM-3B after unlearning token sequences in FLORES-200, shown by scaling the number of samples to be forgotten. The first row illustrates results for unlearning samples at once (**Batch Unlearning**), while the second row depicts results for unlearning samples sequentially (**Sequential Unlearning**).

resolution (XWinograd) (Tikhonov and Ryabinin, 2021), causal reasoning (XCOPA) (Ponti et al., 2020), sentence completion (XStoryCloze) (Lin et al., 2022), and paraphrase identification (PAWS-X) (Yang et al., 2019). We evaluate 3B models to ensure fair zero-shot performance, presenting the results in Figure 4. Our observations indicate that our method, LINGTEA, performs on par with the original model, thereby demonstrating the reliability of our approach. Although NLP benchmark results may not capture all aspects of world knowledge, they at least indicate the retention of information in domains outside our unlearning data.

### 5.5 Scaling the Number of Samples to Forget

To examine the scalability of our unlearning approach, we illustrate the impact of increasing the number of samples to forget by up to four-fold in Figure 5. Consistent with previous findings on unlearning monolingual models (Jang et al., 2023), forgetting larger quantities of samples simultaneously proves to be more challenging, leading to no further reduction in MA. We also investigate whether sequential unlearning could mitigate this issue; however, unlike with monolingual models, we observe no significant improvement. On a positive note, the retention performance remains stable even as the number of samples to forget increases, highlighting the reliability of multilingual unlearning. We hypothesize that forgetting numerous samples in a multilingual context is inherently more complex, as the total number of samples to forget effectively multiplies by the number of languages. For instance, in the FLORES study, the increase isn't merely four-fold but rather forty-fold due to the involvement of ten languages. Exploring the scalability of multilingual unlearning presents a non-trivial challenge, and we leave this as a direction for future research.

## 6 Conclusion

In response to rising privacy concerns and regulatory demands, our study pioneers a method for machine unlearning in multilingual language models. We introduce an adaptive unlearning scheme using a multilingual teacher model to address performance disparities across languages, ensuring the effective removal of sensitive information while maintaining overall model performance. Our empirical results, validated on multilingual parallel datasets, demonstrate significant improvements over existing unlearning methods. This approach not only mitigates vulnerabilities to low-resource language attacks but also offers a practical, efficient alternative to retraining models from scratch, aligning with modern privacy regulations and advancing the field of NLP.

## Limitations

Despite the robust findings presented in our paper, certain limitations warrant discussion. The datasets used to explore multilingual unlearning in this study, namely FLORES and BMLAMA, are in the general domain. This is due to the scarcity of multilingual parallel datasets, especially within specific domains such as privacy data. This challenge mirrors those seen in computer vision, where datasets like CIFAR and MNIST, although unrelated to privacy, are used due to the difficulty in obtaining privacy-specific data. Future research should focus on inventing and benchmarking real or synthetic privacy data in multilingual settings to address these gaps. Additionally, our research was constrained by GPU resources, preventing us from testing models with 7B parameters or more. Investigating whether our conclusions hold for larger-scale models is a promising avenue for future work.

## References


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. IEEE.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052, Singapore. Association for Computational Linguistics.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. 2023. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7210–7217.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020a. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020b. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 383–398. Springer.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. 2019. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European*

*Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2023. Towards unbounded machine unlearning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N Ravi. 2022. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10422–10431.

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

Stuart L Pardau. 2018. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Jeffrey Rosen. 2011. The right to be forgotten. *Stan. L. Rev. Online*, 64:88.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Alexey Tikhonov and Max Ryabinin. 2021. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3534–3546, Online. Association for Computational Linguistics.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023. KGA: A general machine unlearning framework based on knowledge gap alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13264–13276, Toronto, Canada. Association for Computational Linguistics.

WGA. 2023. Summary of the 2023 wga mba.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5554–5569, Toronto, Canada. Association for Computational Linguistics.

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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023. Low-resource languages jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2023. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *arXiv preprint arXiv:2307.03941*.

# A Additional Details for LingTea

## A.1 Hyperparameters

We have performed a grid search to find the best hyperparameter configuration and report the tuning range used for our experiments in Table 4. For all experiments, we have incorporated `bfloat16` mixed precision training, a linear warmup scheduler followed by decay to 0, and early stopping with a max tolerance of 5.

## A.2 Amount of Data Trained Per Language

The categories of high-resource, mid-resource, and low-resource languages are determined by the amount of data used to pretrain the corresponding multilingual language model. Specifically, we follow tables in Lin et al. (2022) and Le Scao et al. (2023) and report the statistics for the languages used in our study in Table 5.

| Model | Hyperparameter | Range | Best |
|---|---|---|---|
| XGLM-564M | learning rate | { 5e-4, 3e-4, 1e-4, 5e-5, 3e-5 } | 5e-4 |
| | warm-up ratio | { 0.0, 0.1 } | 0.1 |
| | retaining samples | { 32, 64, 96, 128 } | 96 |
| | $\lambda$ | { 0.1, 0.5, 1.0 } | 1.0 |
| XGLM-2.9B | learning rate | { 5e-4, 3e-4, 1e-4, 5e-5, 3e-5 } | 1e-4 |
| | warm-up ratio | { 0.0, 0.1 } | 0.0 |
| | retaining samples | { 32, 64, 96, 128 } | 96 |
| | $\lambda$ | { 0.1, 1.0, 10 } | 1.0 |
| BLOOM-560M | learning rate | { 5e-4, 3e-4, 1e-4, 5e-5, 3e-5 } | 3e-5 |
| | warm-up ratio | { 0.0, 0.1 } | 0.0 |
| | retaining samples | { 32, 64, 96, 128 } | 96 |
| | $\lambda$ | { 0.1, 1.0, 10 } | 1.0 |
| BLOOM-3B | learning rate | { 5e-4, 3e-4, 1e-4, 5e-5, 3e-5 } | 3e-5 |
| | warm-up ratio | { 0.0, 0.1 } | 0.0 |
| | retaining samples | { 32, 64, 96, 128 } | 128 |
| | $\lambda$ | { 0.1, 1.0, 10 } | 1.0 |

Table 4: Hyperparameter tuning range and best values used in the experiments.

| Language | XGLM | BLOOM |
|---|---|---|
| English (en) | 3,324.45 | 484.95 |
| HIGH-SRC | | |
| French (fr) | 303.76 | 208.24 |
| Spanish (es) | 363.83 | 175.10 |
| Chinese (zh) | 485.32 | 261.02 |
| Portuguese (pt) | 147.12 | 79.28 |
| Arabic (ar) | 64.34 | 74.85 |
| Vietnamese (vi) | 50.45 | 43.71 |
| MID-SRC | | |
| Catalan (ca) | 26.90 | 17.79 |
| Hindi (hi) | 26.63 | 24.62 |
| Bengali (bn) | 11.19 | 18.61 |
| LOW-SRC | | |
| Basque (eu) | 0.35 | 2.36 |
| Urdu (ur) | 7.77 | 2.78 |
| Telugu (te) | 5.28 | 2.99 |
| Swahili (sw) | 3.19 | 0.24 |

Table 5: Amount of pretraining data in gigabytes (GB) used to train each multilingual model.

# B Per-Language Performance

## B.1 Token Sequence Unlearning Results for Each Language

We report the per-language performance of unlearning token sequences in FLORES-200 across compared models in Tables 6, 7, 8, and 9.

## B.2 Factual Knowledge Unlearning Results for Each Language

We report the per-language performance of unlearning factual knowledge in BMLAMA-53 across compared models in Tables 10, 11, 12, and 13.

| Model | Method | EN | High-Resource FR | ES | ZH | AR | VI | Low-Resource EU | UR | TE | SW |
|---|---|---|---|---|---|---|---|---|---|---|---|
| XGLM-564M | Original | 34.6 | 39.2 | 35.8 | 32.3 | 30.6 | 36.2 | 31.8 | 30.8 | 30.6 | 30.0 |
| | GradAscent+ | 22.7 | 29.3 | 25.7 | **19.4** | 21.1 | 24.3 | 20.9 | 19.6 | 20.7 | **18.2** |
| | NegTaskVector+ | 22.7 | 29.3 | 25.5 | 21.9 | 22.7 | 24.5 | 21.6 | 22.3 | 22.1 | 22.1 |
| | LINGTEA (ours) | **19.3** | **25.6** | **21.7** | 22.3 | **18.8** | **22.6** | **19.7** | **19.5** | **19.9** | 18.8 |
| | Oracle | 17.2 | 23.7 | 19.9 | 16.7 | 14.3 | 19.2 | 14.8 | 17.4 | 17.3 | 17.1 |
| XGLM-2.9B | Original | 36.8 | 43.0 | 36.7 | 35.5 | 34.1 | 40.0 | 38.6 | 34.8 | 37.6 | 34.9 |
| | GradAscent+ | 26.3 | 30.2 | 29.0 | **22.7** | 22.0 | 28.2 | 23.3 | **21.4** | **21.4** | 23.0 |
| | NegTaskVector+ | 25.4 | 33.2 | 28.5 | 25.8 | 25.5 | 28.6 | 25.8 | 25.2 | 26.5 | 24.7 |
| | LINGTEA (ours) | **19.9** | **27.6** | **23.6** | 22.8 | **20.6** | **23.1** | **22.6** | 23.6 | 24.7 | **21.4** |
| | Oracle | 19.7 | 25.8 | 22.2 | 22.1 | 18.2 | 23.4 | 19.6 | 19.3 | 17.7 | 19.0 |
| BLOOM-560M | Original | 28.4 | 32.3 | 29.0 | 21.4 | 24.3 | 32.2 | 19.9 | 23.3 | 20.7 | 15.9 |
| | GradAscent+ | 25.1 | 28.4 | 25.3 | 16.9 | 20.3 | 27.5 | 17.8 | 19.9 | 17.0 | 11.4 |
| | NegTaskVector+ | 22.7 | 25.6 | 22.5 | **15.1** | 17.9 | 24.4 | 14.8 | 17.7 | **14.1** | 10.2 |
| | LINGTEA (ours) | **18.2** | **24.5** | **20.8** | 16.2 | **16.9** | **22.6** | **13.1** | **16.8** | 15.5 | **9.8** |
| | Oracle | 13.9 | 17.3 | 14.9 | 8.8 | 8.6 | 17.0 | 7.8 | 11.3 | 10.7 | 9.7 |
| BLOOM-3B | Original | 35.8 | 39.6 | 37.3 | 30.3 | 28.2 | 40.1 | 30.3 | 28.8 | 28.1 | 21.5 |
| | GradAscent+ | 25.5 | 29.3 | 26.4 | 18.7 | 18.6 | 27.3 | 16.2 | **18.4** | **15.2** | **10.2** |
| | NegTaskVector+ | 28.7 | 32.8 | 29.6 | 24.4 | 20.8 | 31.7 | 22.2 | 21.9 | 20.8 | 15.3 |
| | LINGTEA (ours) | **17.8** | **23.2** | **22.4** | **18.3** | **17.6** | **23.5** | **15.5** | 19.2 | 20.3 | 13.0 |
| | Oracle | 13.8 | 15.2 | 14.6 | 9.4 | 9.7 | 17.8 | 9.2 | 11.3 | 9.3 | 6.9 |

Table 6: Memorization Accuracy (%) of Forget Set in FLORES-200.

| Model | Method | EN | High-Resource FR | ES | ZH | AR | VI | Low-Resource EU | UR | TE | SW |
|---|---|---|---|---|---|---|---|---|---|---|---|
| XGLM-564M | Original | 35.4 | 40.4 | 36.5 | 34.0 | 30.8 | 34.8 | 32.4 | 29.5 | 32.1 | 28.9 |
| | GradAscent+ | 32.2 | 37.5 | 34.0 | 28.4 | 28.6 | 31.2 | 28.2 | 24.8 | 27.3 | 24.0 |
| | NegTaskVector+ | 30.8 | 37.6 | 33.6 | 30.0 | 29.1 | 31.1 | 28.4 | 27.6 | 29.8 | 26.3 |
| | LINGTEA (ours) | 30.8 | 36.7 | 32.8 | 30.5 | 27.5 | 30.8 | 27.3 | 25.9 | 27.4 | 24.9 |
| | Oracle | 32.4 | 38.1 | 34.5 | 31.8 | 29.3 | 32.5 | 30.3 | 28.6 | 29.6 | 27.7 |
| XGLM-2.9B | Original | 38.5 | 43.7 | 39.6 | 37.5 | 36.2 | 39.2 | 38.1 | 33.9 | 37.1 | 33.1 |
| | GradAscent+ | 36.2 | 41.3 | 37.4 | 33.0 | 33.2 | 36.2 | 34.0 | 28.5 | 30.9 | 29.7 |
| | NegTaskVector+ | 33.8 | 40.8 | 36.6 | 33.4 | 34.0 | 34.8 | 34.3 | 31.6 | 34.4 | 30.1 |
| | LINGTEA (ours) | 35.2 | 40.7 | 37.1 | 34.1 | 31.5 | 35.0 | 32.3 | 29.8 | 31.8 | 28.6 |
| | Oracle | 38.2 | 43.2 | 39.1 | 37.5 | 35.0 | 38.9 | 36.5 | 33.7 | 35.3 | 32.9 |
| BLOOM-560M | Original | 29.5 | 33.6 | 30.9 | 21.6 | 26.9 | 31.0 | 20.5 | 22.6 | 20.2 | 14.4 |
| | GradAscent+ | 29.7 | 33.6 | 30.9 | 21.4 | 26.4 | 30.7 | 20.9 | 21.9 | 19.5 | 14.1 |
| | NegTaskVector+ | 28.6 | 33.0 | 30.1 | 20.3 | 26.4 | 29.7 | 19.9 | 22.3 | 19.2 | 13.9 |
| | LINGTEA (ours) | 28.5 | 32.9 | 30.7 | 22.0 | 27.0 | 30.3 | 20.0 | 22.4 | 20.0 | 13.8 |
| | Oracle | 31.0 | 34.8 | 32.0 | 24.4 | 27.2 | 32.3 | 21.2 | 23.4 | 21.5 | 16.0 |
| BLOOM-3B | Original | 36.6 | 40.7 | 37.4 | 29.4 | 32.0 | 38.8 | 30.0 | 28.5 | 26.7 | 22.8 |
| | GradAscent+ | 35.6 | 39.7 | 36.5 | 27.9 | 31.0 | 37.3 | 28.8 | 26.7 | 24.4 | 21.1 |
| | NegTaskVector+ | 36.6 | 40.5 | 37.5 | 29.4 | 32.1 | 38.7 | 29.4 | 28.7 | 25.9 | 22.1 |
| | LINGTEA (ours) | 35.5 | 39.5 | 36.6 | 29.4 | 31.6 | 37.2 | 26.9 | 27.6 | 25.7 | 19.7 |
| | Oracle | 35.7 | 40.2 | 37.2 | 29.6 | 32.4 | 37.9 | 28.6 | 28.9 | 27.6 | 22.7 |

Table 7: Memorization Accuracy (%) of Test Set in FLORES-200.

| Model | Method | EN | High-Resource | | | | | Low-Resource | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FR | ES | ZH | AR | VI | EU | UR | TE | SW |
| XGLM-564M | Original | 117.4 | 71.1 | 118.1 | 209.6 | 151.8 | 133.4 | 162.8 | 122.6 | 85.1 | 230.7 |
| | GradAscent+ | 4242.9 | **935.0** | **1242.0** | **22843.9** | **2428.6** | **3925.0** | **4142.2** | **21813.6** | **8914.5** | **32565.1** |
| | NegTaskVector+ | 663.7 | 253.8 | 484.0 | 853.6 | 473.8 | 544.5 | 654.5 | 412.6 | 254.2 | 872.9 |
| | LINGTEA (ours) | **4261.8** | 547.2 | 1188.4 | 882.2 | 720.2 | 744.8 | 1185.5 | 609.1 | 843.2 | 1487.1 |
| | Oracle | 6295.8 | 2336.1 | 2375.9 | 9107.0 | 5546.9 | 3529.5 | 5402.9 | 4823.6 | 3879.0 | 5015.6 |
| XGLM-2.9B | Original | 66.6 | 44.9 | 57.6 | 231.4 | 122.6 | 67.3 | 67.8 | 120.2 | 60.6 | 114.0 |
| | GradAscent+ | 16553.7 | 35430.7 | 6453.4 | **17446637.7** | 12230.1 | 19259.8 | 119347.7 | **2133682.2** | **4157203.2** | **45591.5** |
| | NegTaskVector+ | 206.8 | 91.3 | 157.5 | 221.4 | 298.0 | 156.0 | 266.7 | 213.5 | 215.9 | 291.5 |
| | LINGTEA (ours) | **10216406.9** | **89399.2** | **7333.6** | 1923080.6 | **55752.9** | **67869.4** | **200656.9** | 11703.3 | 5415.2 | 9165.5 |
| | Oracle | 11605.0 | 6285.3 | 4634.5 | 171286.7 | 9716.7 | 3043.3 | 8136.9 | 682569.5 | 11299.8 | 6214.6 |
| BLOOM-560M | Original | 81.0 | 48.5 | 59.9 | 151.6 | 115.7 | 56.3 | 300.0 | 183.0 | 647.0 | 1283.4 |
| | GradAscent+ | 127.0 | 76.4 | 104.6 | 239.6 | 189.3 | 102.4 | 530.7 | 389.4 | 2723.0 | 4329.3 |
| | NegTaskVector+ | 277.1 | 151.4 | 204.4 | 548.3 | 369.3 | 178.0 | 1049.8 | 561.3 | 3282.7 | 5835.6 |
| | LINGTEA (ours) | **2787.0** | **1719.4** | **2527.2** | **2712.2** | **1072.7** | **933.3** | **6191.8** | **1268.3** | **7125.7** | **11616.8** |
| | Oracle | 12702.6 | 6706.1 | 350794.6 | 70639.8 | 31163.5 | 6724.8 | 321866.9 | 38910.3 | 45578.8 | 6366.4 |
| BLOOM-3B | Original | 42.4 | 29.9 | 35.1 | 81.5 | 82.3 | 28.8 | 119.5 | 85.3 | 123.6 | 268.5 |
| | GradAscent+ | 291.2 | 348.5 | 246.9 | 3039.6 | 737.9 | 191.9 | 3872.6 | 654.7 | 5163.4 | 19702.8 |
| | NegTaskVector+ | 119.7 | 77.7 | 90.1 | 210.7 | 227.3 | 73.2 | 496.9 | 255.1 | 453.9 | 1283.0 |
| | LINGTEA (ours) | **21063692.6** | **276395.7** | **7120.2** | **2312623.1** | **859031.7** | **100120.3** | **203023.3** | **16307.8** | **5384.3** | **28865.8** |
| | Oracle | 134342.4 | 33805.2 | 168509.4 | 1163537.3 | 219896.4 | 19421.4 | 1275923.6 | 156499.3 | 141361.8 | 297537.1 |

Table 8: Perplexity of Forget Set in FLORES-200.

| Model | Method | EN | High-Resource | | | | | Low-Resource | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FR | ES | ZH | AR | VI | EU | UR | TE | SW |
| XGLM-564M | Original | 107.1 | 56.5 | 93.6 | 199.1 | 124.9 | 126.6 | 163.7 | 135.6 | 86.5 | 229.9 |
| | GradAscent+ | 301.0 | 102.1 | 162.8 | 1153.5 | 409.6 | 372.2 | 563.6 | 1745.8 | 664.0 | 1705.7 |
| | NegTaskVector+ | 191.3 | 74.5 | 125.9 | 288.8 | 160.9 | 192.3 | 226.5 | 181.7 | 111.2 | 294.4 |
| | LINGTEA (ours) | 114.8 | 46.2 | 72.4 | 107.9 | 116.7 | 82.9 | 121.6 | 124.6 | 96.4 | 168.7 |
| | Oracle | 114.4 | 46.8 | 65.0 | 121.2 | 120.4 | 81.1 | 113.2 | 104.6 | 93.8 | 141.3 |
| XGLM-2.9B | Original | 68.6 | 35.7 | 48.7 | 192.3 | 92.6 | 81.5 | 71.6 | 131.5 | 64.3 | 128.8 |
| | GradAscent+ | 922.8 | 184.5 | 258.7 | 2573.7 | 403.9 | 531.5 | 873.0 | 15338.7 | 10443.3 | 1081.6 |
| | NegTaskVector+ | 78.4 | 30.7 | 52.6 | 79.3 | 76.2 | 60.3 | 80.9 | 89.3 | 81.5 | 112.6 |
| | LINGTEA (ours) | 102.4 | 40.3 | 59.0 | 171.9 | 221.3 | 89.6 | 111.5 | 254.0 | 155.5 | 169.5 |
| | Oracle | 70.5 | 32.7 | 47.4 | 62.5 | 73.9 | 40.1 | 70.2 | 65.5 | 60.9 | 88.2 |
| BLOOM-560M | Original | 73.2 | 39.8 | 48.3 | 154.6 | 97.1 | 52.2 | 311.4 | 178.8 | 558.7 | 1213.8 |
| | GradAscent+ | 72.4 | 39.5 | 48.3 | 158.3 | 103.0 | 55.3 | 319.6 | 212.1 | 826.5 | 1385.6 |
| | NegTaskVector+ | 83.0 | 43.9 | 54.3 | 170.3 | 114.8 | 64.6 | 350.9 | 217.7 | 806.3 | 1518.8 |
| | LINGTEA (ours) | 86.7 | 47.9 | 59.9 | 199.9 | 106.7 | 67.8 | 376.5 | 225.9 | 559.9 | 1160.7 |
| | Oracle | 71.8 | 39.1 | 62.5 | 164.7 | 108.1 | 57.6 | 390.4 | 217.2 | 468.8 | 664.5 |
| BLOOM-3B | Original | 42.7 | 24.6 | 30.1 | 83.2 | 60.2 | 28.8 | 114.8 | 89.9 | 119.5 | 295.4 |
| | GradAscent+ | 54.5 | 28.5 | 33.1 | 143.5 | 86.5 | 38.1 | 179.3 | 155.3 | 309.9 | 602.7 |
| | NegTaskVector+ | 42.8 | 24.7 | 30.5 | 78.4 | 60.6 | 28.7 | 122.6 | 93.2 | 128.6 | 330.7 |
| | LINGTEA (ours) | 51.0 | 27.5 | 33.6 | 114.8 | 83.9 | 40.3 | 153.9 | 127.1 | 184.7 | 466.5 |
| | Oracle | 49.5 | 25.7 | 32.2 | 99.7 | 68.0 | 33.6 | 129.5 | 108.2 | 141.8 | 268.3 |

Table 9: Perplexity of Test Set in FLORES-200.

| Model | Method | EN | High-Resource FR | ES | PT | AR | VI | Mid-Resource CA | HI | BN |
|---|---|---|---|---|---|---|---|---|---|---|
| XGLM-564M | Original | 28.1 | 12.5 | 12.5 | 12.5 | 12.5 | 18.8 | 15.6 | 21.9 | 18.8 |
| | GradAscent+ | 27.1 | 3.1 | **3.1** | **6.3** | 6.3 | 17.7 | 10.4 | 15.6 | **11.5** |
| | NegTaskVector+ | 28.1 | 3.1 | 6.3 | **6.3** | **3.1** | 17.7 | 10.4 | **11.5** | 13.5 |
| | LINGTEA (ours) | **25.0** | **0.0** | **3.1** | **6.3** | 4.2 | **13.5** | **4.2** | 15.6 | 12.5 |
| | Oracle | 5.2 | 0.0 | 3.1 | 6.3 | 0.0 | 3.1 | 1.0 | 6.3 | 0.0 |
| XGLM-2.9B | Original | 34.4 | 6.3 | 12.5 | 15.6 | 15.6 | 28.1 | 25.0 | 31.3 | 18.8 |
| | GradAscent+ | 29.2 | 7.3 | 8.3 | 10.4 | 7.3 | 21.9 | 12.5 | 15.6 | **7.3** |
| | NegTaskVector+ | 29.2 | 7.3 | 6.3 | 8.3 | 8.3 | 16.7 | **8.3** | **17.7** | 11.5 |
| | LINGTEA (ours) | **14.6** | **4.2** | **4.2** | **6.3** | **6.3** | **13.5** | 13.5 | 12.5 | 12.5 |
| | Oracle | 13.5 | 4.2 | 5.2 | 10.4 | 6.3 | 1.0 | 2.1 | 4.2 | 7.3 |
| BLOOM-560M | Original | 31.3 | 18.8 | 21.9 | 18.8 | 6.3 | 28.1 | 9.4 | 12.5 | 9.4 |
| | GradAscent+ | 15.6 | 12.5 | 11.5 | 7.3 | **5.2** | 19.8 | **4.2** | 7.3 | 9.4 |
| | NegTaskVector+ | 22.9 | 15.6 | 9.4 | 11.5 | **5.2** | 22.9 | 5.2 | 7.3 | 9.4 |
| | LINGTEA (ours) | **9.4** | **8.3** | **8.3** | **4.2** | **5.2** | **8.3** | 5.2 | **6.3** | **5.2** |
| | Oracle | 7.3 | 3.1 | 3.1 | 4.2 | 0.0 | 3.1 | 3.1 | 0.0 | 0.0 |
| BLOOM-3B | Original | 50.0 | 28.1 | 28.1 | 18.8 | 15.6 | 31.3 | 18.8 | 9.4 | 15.6 |
| | GradAscent+ | **16.7** | 10.4 | 8.3 | 6.3 | **3.1** | 10.4 | **3.1** | 6.3 | 8.3 |
| | NegTaskVector+ | 35.4 | 16.7 | 19.8 | 12.5 | 7.3 | 24.0 | 5.2 | **5.2** | 11.5 |
| | LINGTEA (ours) | 19.8 | **6.3** | **6.3** | **4.2** | 7.3 | **7.3** | 5.2 | 9.4 | **6.3** |
| | Oracle | 17.7 | 6.3 | 12.5 | 8.3 | 1.0 | 8.3 | 4.2 | 3.1 | 0.0 |

Table 10: Probing Accuracy (%) of Forget Set in BMLAMA-53.

| Model | Method | EN | High-Resource FR | ES | PT | AR | VI | Mid-Resource CA | HI | BN |
|---|---|---|---|---|---|---|---|---|---|---|
| XGLM-564M | Original | 29.9 | 15.9 | 17.1 | 16.8 | 16.1 | 18.9 | 21.4 | 15.9 | 15.3 |
| | GradAscent+ | 30.3 | 15.7 | 17.3 | 15.3 | 15.8 | 19.3 | 20.5 | 14.3 | 14.9 |
| | NegTaskVector+ | 30.1 | 16.9 | 18.0 | 18.0 | 16.5 | 20.6 | 21.0 | 15.3 | 16.2 |
| | LINGTEA (ours) | 29.4 | 17.5 | 18.1 | 16.9 | 15.9 | 18.2 | 20.2 | 13.8 | 16.6 |
| | Oracle | 28.6 | 14.0 | 16.1 | 14.9 | 13.9 | 22.6 | 19.7 | 12.0 | 13.4 |
| XGLM-2.9B | Original | 34.7 | 19.3 | 24.2 | 25.5 | 18.1 | 22.6 | 27.1 | 15.6 | 15.8 |
| | GradAscent+ | 35.4 | 21.1 | 23.0 | 23.4 | 16.4 | 22.4 | 26.3 | 14.0 | 12.8 |
| | NegTaskVector+ | 33.4 | 19.0 | 21.6 | 21.1 | 18.5 | 21.8 | 24.6 | 14.8 | 17.1 |
| | LINGTEA (ours) | 37.1 | 25.2 | 29.6 | 25.4 | 17.8 | 24.3 | 29.4 | 17.5 | 17.8 |
| | Oracle | 43.3 | 23.4 | 28.3 | 30.6 | 20.1 | 32.8 | 35.9 | 18.7 | 17.6 |
| BLOOM-560M | Original | 28.5 | 16.6 | 21.0 | 17.7 | 11.2 | 20.2 | 16.0 | 11.2 | 9.9 |
| | GradAscent+ | 28.5 | 15.3 | 18.9 | 17.3 | 11.2 | 20.9 | 15.3 | 10.1 | 9.6 |
| | NegTaskVector+ | 29.0 | 16.8 | 20.4 | 16.6 | 11.3 | 21.5 | 16.0 | 10.5 | 9.8 |
| | LINGTEA (ours) | 27.4 | 16.7 | 19.3 | 17.4 | 11.2 | 20.2 | 16.2 | 10.5 | 9.8 |
| | Oracle | 29.6 | 18.5 | 19.3 | 18.1 | 10.9 | 23.5 | 15.2 | 9.8 | 10.1 |
| BLOOM-3B | Original | 46.6 | 26.4 | 31.3 | 29.5 | 16.8 | 30.2 | 24.1 | 13.3 | 10.8 |
| | GradAscent+ | 40.8 | 21.9 | 25.2 | 25.5 | 15.2 | 30.1 | 21.4 | 12.1 | 11.1 |
| | NegTaskVector+ | 47.2 | 23.0 | 30.0 | 25.9 | 15.0 | 29.5 | 21.7 | 11.4 | 10.6 |
| | LINGTEA (ours) | 47.1 | 34.2 | 36.1 | 33.8 | 19.8 | 36.0 | 30.5 | 13.5 | 10.9 |
| | Oracle | 46.1 | 42.2 | 39.6 | 33.7 | 19.8 | 39.0 | 32.2 | 16.3 | 12.0 |

Table 11: Probing Accuracy (%) of Test Set in BMLAMA-53.

| Model | Method | | High-Resource | | | | | Mid-Resource | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | EN | FR | ES | PT | AR | VI | CA | HI | BN |
| XGLM-564M | Original | 122.0 | 108.7 | 151.4 | 100.6 | 110.4 | 113.1 | 114.7 | 73.5 | 48.0 |
| | GradAscent+ | **187.2** | 173.2 | 262.3 | 156.6 | 118.1 | 159.5 | 170.1 | 77.8 | 51.3 |
| | NegTaskVector+ | 150.7 | 136.3 | 194.0 | 116.4 | **131.9** | 132.0 | 122.2 | **90.6** | **59.2** |
| | LINGTEA (ours) | 185.3 | **177.6** | **282.1** | **168.0** | 103.2 | **167.1** | **176.4** | 72.7 | 57.1 |
| | Oracle | 71681.6 | 192.6 | 196.4 | 137.0 | 195.8 | 15205.2 | 1241.5 | 568.3 | 405.8 |
| XGLM-2.9B | Original | 90.9 | 80.7 | 109.4 | 84.7 | 45.3 | 93.6 | 77.6 | 36.4 | 31.7 |
| | GradAscent+ | 133.5 | 125.8 | 244.5 | 166.2 | **352.6** | 140.5 | 126.6 | 195.5 | 244.3 |
| | NegTaskVector+ | 124.6 | 121.4 | 179.3 | 140.8 | 60.9 | 133.9 | 118.8 | 54.0 | 43.9 |
| | LINGTEA (ours) | **908.6** | **744.5** | **673.2** | **766.7** | 163.9 | **1043.4** | **483.2** | **356.8** | **600.0** |
| | Oracle | 1274.7 | 100.9 | 250.8 | 290.4 | 85.9 | 2035.9 | 1228.4 | 6126.5 | 1593.2 |
| BLOOM-560M | Original | 145.8 | 112.4 | 170.3 | 227.0 | 85.7 | 129.4 | 239.6 | 183.3 | 379.6 |
| | GradAscent+ | 238.8 | **174.5** | 293.0 | 369.0 | 98.5 | 167.7 | 403.7 | 225.9 | 464.2 |
| | NegTaskVector+ | 184.9 | 128.0 | 211.3 | 272.2 | 91.0 | 137.8 | 289.7 | 236.3 | 468.2 |
| | LINGTEA (ours) | **267.5** | 174.2 | **354.9** | **486.0** | **113.4** | **209.9** | **471.3** | **313.8** | **691.1** |
| | Oracle | 629.3 | 159.6 | 316.1 | 536.6 | 168.9 | 32890.0 | 257.1 | 506.5 | 1704.6 |
| BLOOM-3B | Original | 68.9 | 71.1 | 80.6 | 115.9 | 45.8 | 60.8 | 99.2 | 77.5 | 108.3 |
| | GradAscent+ | 645.1 | 515.1 | 1045.2 | 1090.3 | 157.3 | 280.6 | **682.4** | 229.9 | 294.9 |
| | NegTaskVector+ | 110.8 | 118.3 | 150.4 | 221.1 | 64.5 | 87.9 | 179.1 | 168.3 | 202.8 |
| | LINGTEA (ours) | **1077.3** | **585.5** | **1243.3** | **1606.7** | **185.9** | **287.7** | 621.1 | **400.6** | **1156.0** |
| | Oracle | 2708.9 | 398.0 | 555.1 | 506.2 | 61.5 | 405.0 | 304.1 | 3234.7 | 1795.7 |

Table 12: Perplexity of Forget Set in BMLAMA-53.

| Model | Method | | High-Resource | | | | | Mid-Resource | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | EN | FR | ES | PT | AR | VI | CA | HI | BN |
| XGLM-564M | Original | 152.8 | 122.2 | 170.4 | 113.9 | 115.2 | 153.2 | 122.5 | 99.3 | 66.0 |
| | GradAscent+ | 187.8 | 154.3 | 223.5 | 140.6 | 110.8 | 184.8 | 149.0 | 91.4 | 62.0 |
| | NegTaskVector+ | 145.9 | 120.4 | 165.0 | 102.5 | 116.4 | 145.6 | 105.7 | 99.0 | 66.3 |
| | LINGTEA (ours) | 165.5 | 139.4 | 196.3 | 112.6 | 92.6 | 153.2 | 131.5 | 73.8 | 60.7 |
| | Oracle | 1249.1 | 141.7 | 185.1 | 135.9 | 148.6 | 1226.1 | 224.1 | 222.0 | 150.3 |
| XGLM-2.9B | Original | 112.7 | 95.1 | 121.5 | 94.1 | 51.6 | 114.2 | 85.7 | 52.9 | 38.8 |
| | GradAscent+ | 127.5 | 116.1 | 170.1 | 125.8 | 314.0 | 146.4 | 108.5 | 204.4 | 156.0 |
| | NegTaskVector+ | 120.2 | 110.1 | 142.7 | 109.7 | 57.3 | 128.8 | 95.6 | 58.9 | 39.9 |
| | LINGTEA (ours) | 156.5 | 104.5 | 129.0 | 128.3 | 73.7 | 251.9 | 94.8 | 135.0 | 164.4 |
| | Oracle | 176.1 | 83.3 | 92.2 | 67.4 | 79.5 | 216.8 | 80.9 | 235.6 | 84.3 |
| BLOOM-560M | Original | 202.6 | 134.3 | 210.6 | 220.8 | 93.3 | 139.1 | 267.7 | 169.5 | 333.8 |
| | GradAscent+ | 237.6 | 163.8 | 243.2 | 274.5 | 95.9 | 145.8 | 336.2 | 170.2 | 336.1 |
| | NegTaskVector+ | 204.7 | 130.4 | 188.1 | 211.1 | 86.4 | 127.5 | 265.0 | 170.1 | 325.3 |
| | LINGTEA (ours) | 206.4 | 134.1 | 195.1 | 241.0 | 91.2 | 151.1 | 317.3 | 189.1 | 418.4 |
| | Oracle | 204.4 | 101.3 | 139.5 | 163.4 | 84.2 | 509.5 | 237.1 | 164.2 | 393.9 |
| BLOOM-3B | Original | 89.5 | 86.4 | 92.6 | 105.9 | 47.1 | 63.9 | 98.5 | 83.7 | 117.5 |
| | GradAscent+ | 258.6 | 208.3 | 218.2 | 228.3 | 77.4 | 108.9 | 235.2 | 116.6 | 169.6 |
| | NegTaskVector+ | 104.4 | 106.4 | 109.3 | 128.8 | 51.1 | 71.2 | 120.6 | 101.8 | 137.2 |
| | LINGTEA (ours) | 137.0 | 84.6 | 118.4 | 107.3 | 56.0 | 86.3 | 114.3 | 119.2 | 297.3 |
| | Oracle | 136.5 | 65.7 | 65.1 | 81.9 | 37.6 | 67.8 | 74.8 | 150.5 | 193.5 |

Table 13: Perplexity of Test Set in BMLAMA-53.