Pruning Multilingual Large Language Models for Multilingual Inference

Hwichan Kim[1] **Jun Suzuki**[2] **Tosho Hirasawa**[1] **Mamoru Komachi**[3]
[1]Tokyo Metropolitan University, [2]Tohoku University, [3]Hitotsubashi University
{kim-hwichan@ed.tmu, jun.suzuki@tohoku, toshosan@tmu, mamoru.komachi@hit-u}.ac.jp

Abstract

Multilingual large language models (MLLMs), trained on multilingual balanced data, demonstrate better zero-shot learning performance in non-English languages compared to large language models trained on English-dominant data. However, the disparity in performance between English and non-English languages remains a challenge yet to be fully addressed. A distinctive characteristic of MLLMs is their high-quality translation capabilities, indicating an acquired proficiency in aligning between languages. This study explores how to enhance the zero-shot performance of MLLMs in non-English languages by leveraging their alignment capability between English and non-English languages. To achieve this, we first analyze the behavior of MLLMs when performing translation and reveal that there are large magnitude features that play a critical role in the translation process. Inspired by these findings, we retain the weights associated with operations involving the large magnitude features and prune other weights to force MLLMs to rely on these features for tasks beyond translation. We empirically demonstrate that this pruning strategy can enhance the MLLMs' performance in non-English language.[1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable language reasoning capabilities, particularly in English contexts. However, these LLMs limited their proficiency in non-English language (Ahuja et al., 2023, 2024). Recognizing this limitation, recent research endeavors have given rise to Multilingual Large Language Models (MLLMs), such as XGLM (Lin et al., 2022), mGPT (Shliazhko et al., 2024), and BLOOM (Scao et al., 2022), designed to address the challenges

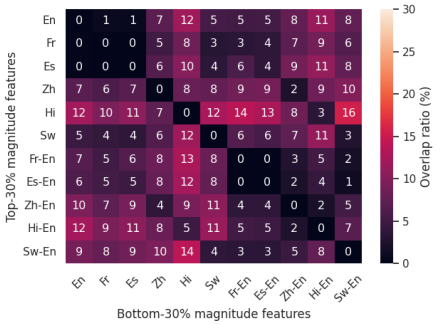


Figure 1: The overlap ratios among the top- and bottom-30% features in the 27-th layer of XGLM, ranked by their magnitude. The row and column labels correspond to languages and language pairs used in few-shot monolingual (En, Fr, etc.) and translation (Fr-En, Es-En, etc.) demonstrations, respectively. Each element represents the ratios of overlapping features between the top- and bottom-30% in magnitude within each demonstration. This figure shows that specific features are active only when inputting translation demonstrations.

posed by linguistic diversity. In the context of zero-shot learning, MLLMs have demonstrated superior proficiency across multiple languages compared to English LLMs (Etxaniz et al., 2024). Nevertheless, a discernible disparity persists in accuracy levels when comparing results between English and non-English languages. For example, in the Cross-lingual NLI (XNLI) task (Conneau et al., 2018), XGLM-2.9B achieves accuracies of 51.1 and 39.2 in English and Russian, respectively (see Tab. 2).

Achieving performance parity in non-English languages with English, which represents the upper bound of MLLMs, necessitates the alignment of English and non-English texts. Ahuja et al. (2023) has demonstrated the effectiveness of the translate-test, a methodology translating non-English texts into English using an external machine translation (MT) system and running inference over the translated texts, which serves as an approach to superficially align English and non-English. However, the translate-test approach increases inference

[1]The code used in our experiments is available at https://github.com/hwichan0720/pruning_for_multilinguality.

Findings of the Association for Computational Linguistics: EACL 2024, pages 9921–9942

costs due to its reliance on the MT system. Conversely, recent research indicated that MLLMs can perform translation by incorporating few-shot translation demonstrations into the contextual information (Lin et al., 2022; Etxaniz et al., 2024; Vilar et al., 2023). These findings suggest that MLLMs already have alignment capability between English and non-English, and it is manifested through the incorporation of few-shot translation demonstrations. We believe that if the alignment capability can be brought out in tasks other than translation, non-English performance elevates to the level equivalent to English.

Previous researches (Dettmers et al., 2022; Sun et al., 2024) have identified two distinctive characteristics within hidden state features of LLMs: first, the presence of large magnitude features, a small set of hidden state features that emerge with significantly larger magnitudes than the remaining ones; and second, the essential nature of these features for the predictive capabilities of LLMs. Referring to the researches, we analyzed magnitudes of features of MLLMs when inputting few-shot monolingual and translation demonstrations. Fig. 1 shows the overlap ratios within features between the monolingual and translation demonstrations, and indicates that specific features are predominantly active only when inputting translation demonstrations. In addition, we will see later that the large magnitude features are relevant for the translation performance.

Motivated by the results, we hypothesized that MLLMs carry out zero-shot inferences while accentuating their alignment capability by forcing them to use large magnitude features that are active when inputting few-shot translation demonstrations. To achieve this, we retained and pruned weights of MLLMs following Sun et al. (2024) that involve operations for the large magnitude features and others, respectively. We observed that the pruned MLLMs improved zero-shot performance in non-English languages compared to pre-pruned MLLMs. Our contributions in this study are threefold:

1. We demonstrated that specific features exhibit large magnitudes and are predominantly active only when inputting few-shot translation demonstrations. In addition, we showed that the large magnitude features are relevant for performing the translation task.

2. We conducted multilingual zero-shot learning in the Cross-lingual Natural Language Inference (XNLI) and Multilingual Amazon Review Corpus (MARC) tasks while forcing MLLMs to rely on the large magnitude features through pruning. The results indicated that the pruning enhances performance in XGLM and mGPT, but it did not improve the performance of BLOOM.

3. Since BLOOM was trained on programming language texts as well as multilingual natural language texts, it has the capability to generate programming language, unlike XGLM and mGPT. We hypothesized that the capability to generate programming language introduces noise. Based on the hypothesis, we attempted to prune weights associated operations for the large magnitude features activated when generating programming language texts. We observed that it enhances multilingual zero-shot learning performance in BLOOM.

2 Task Setting

In this study, we conduct zero-shot learning with MLLMs under Lin et al. (2022)'s scenario. We consider a language $l \in \mathcal{L}$ and its test example as x_l. To perform in-context learning, we convert x_l to a cloze-style format that contains a `[Mask]` symbol using a template $\mathcal{T}$ and map each candidate label $y \in \mathcal{Y}$ into a string using a verbalizer $v : \mathcal{Y} \rightarrow \mathcal{V}^*$. An input prompt $\mathcal{P}(x_l, y)$ is obtained by substituting the `[Mask]` symbol in $\mathcal{T}(x_l)$ with $v(y)$. In zero-shot in-context learning, a prediction $\hat{y}$ is the label with the maximum likelihood:

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(\mathcal{P}(x_l, y)|\boldsymbol{\theta}) \quad (1)$$

where $\boldsymbol{\theta}$ is parameters of the MLLM. The objective of our study is to enhance the non-English language performance of MLLMs in a zero-shot in-context learning setting.

3 Related Works

English LLM and their characteristics Large language models (LLMs) have shown strong performance in a wide range of downstream tasks. While fine-tuning has been a popular approach to adapt models to new tasks, it is often impractical to fine-tune very large models. Brown et al. (2020) proposed zero- and few-shot in-context learning as an alternative, which do not require any gradient updates. One of the problems with using LLMs is

that it requires a lot of computational resources for inference.

One of the problems with using LLMs is that it requires a lot of computational resources for inference. To overcome this problem, several studies have attempted quantization of the LLMs. Dettmers et al. (2022) proposed a novel quantization method called `LLM.int8()` that performs matrix multiplications for large magnitude features and others in 16-bit and 8-bit, respectively. They empirically demonstrated that `LLM.int8()` reduces performance degradation compared to performing all operations in 8-bit. This result suggests that the large magnitude features of LLMs are crucial for their prediction.

Motivated from the success of LLM.int8(), Sun et al. (2024) proposed a weights pruning approach named pruning by weights and activations (Wanda). Wanda drops weights that do not involve operation for the large magnitude features. Consider a linear layer's weight of k-th layer of a model $\boldsymbol{\theta}^k \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and hidden states output from previous layer $\boldsymbol{X}^{k-1} \in \mathbb{R}^{T \times d_{in}}$, where T denotes the number of tokens included in calibration data X. Wanda calculates importance scores $\boldsymbol{S}^k \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ for each element of the weight based on $\boldsymbol{\theta}^k$ and $\boldsymbol{X}^{k-1}$. Specifically, a score of i, j-th element $S_{i,j}$ is calculated as:

$$S_{i,j} = \left|\boldsymbol{\theta}^k_{i,j}\right| \cdot \|\boldsymbol{X}^{k-1}_j\|_2 \tag{2}$$

where $|\cdot|$ represents the absolute value operator, $\|\boldsymbol{X}^{k-1}_j\|_2$ evaluates the L2-norm of the j-th feature vector aggregated across T different tokens. Wanda prunes weights of the bottom $\alpha\%$ scores. Their experiments demonstrated that Wanda can prune weights of LLMs even mitigating degradation of their performance compared to other pruning methods.

Enhancing multilingual performance of multilingual pre-trained models As outlined in §1, several studies have attempted to train Multilingual Large Language Models (MLLMs) using datasets that exhibit a more balanced linguistic distribution compared to English-dominant data used in LLMs. Although MLLMs have impressive multilingual capability, their performances in non-English languages do not achieve those in the English level, which is an upper bound level of MLLMs.

This phenomenon has been observed in multilingual pre-trained masked language models, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). Several studies have enhanced the alignment of hidden states between target languages and English using bilingual resources during either the pre-training or fine-tuning phases, demonstrating improved performance in the target languages (Lample and Conneau, 2019; Cao et al., 2020; Yang et al., 2021; Chi et al., 2021; Dou and Neubig, 2021). Few-shot cross-lingual transfer, which involves fine-tuning the multilingual pre-trained models with a small amount of supervised data in the target language for a downstream task, is a promising approach, and several studies have validated its effectiveness (Lauscher et al., 2020; Kim and Komachi, 2023). However, the efficacy of these methodologies when applied to MLLMs remains unverified, and such training approaches require substantial computational resources for adaptation to MLLMs.

Xu et al. (2023) undertook research closely aligned with our objective, specifically aiming to enhance non-English performance of MLLMs in a zero-shot learning scenario without the necessity for fine-tuning. They introduced a novel methodology termed Language Representation Projection (LRP2). LRP2 adjusts the hidden states at the a-th layer by subtracting and adding vectors that correspond to the target language and English, respectively. Subsequently, the inverse operations are applied at the b-th layer. These vectors are derived from the mean-pooled vectors across all tokens in each respective language's dataset at the corresponding layers.

In this research, we provide a promising direction for enhancing non-English performance through comprehensive experiments on various MLLMs. Specifically, we identify large-magnitude features that are relevant for bringing out the inherent alignment capabilities of MLLMs (Lin et al., 2022; Etxaniz et al., 2024; Vilar et al., 2023; Chitale et al., 2024). Motivated by the results, we encourage MLLMs to leverage these prominent features by implementing Wanda (Sun et al., 2024). Subsequent sections will show that our approach improves the performance of non-English languages across multiple MLLMs, surpassing that of LRP2.

4 Detecting Translation Features

Our challenge is accentuating the alignment capability of MLLMs to enhance their non-English performance even when tasks other than transla-

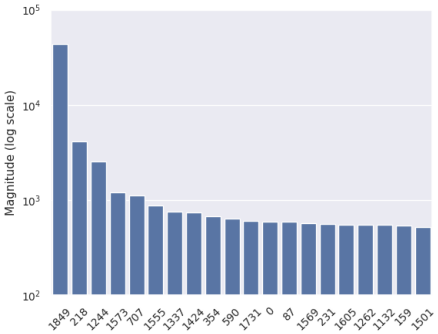


Figure 2: The top 20 dimensions with the largest magnitudes of 27-th layer's features of XGLM activated when inputting $X_{\text{Zh-En}}$.

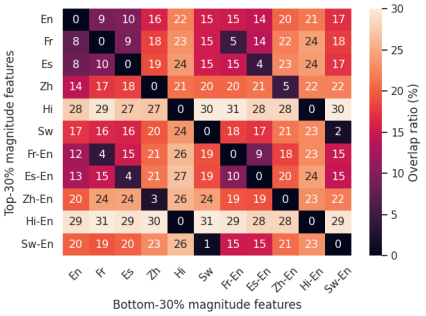


Figure 3: The overlap ratios among the top- and bottom-30% features in the 47th layer of XGLM, ranked by magnitude.

tion. To explore how to emphasize the alignment capability, we first analyze the behavior of MLLMs during translation. While previous studies have suggested that large magnitude features are important for inference, they did not reveal whether large magnitude features are the same or different across tasks. If the large magnitude features that activate exclusively during translation are instrumental for MLLMs' translation performance, the features may be key to bringing out the alignment capability. Therefore, in this section, we conduct analyses based on the following research questions to show step-by-step that there are large magnitude features that affect for translation performance.

RQ1: Do few-shot translation demonstrations activate specific features? Previous studies have suggested that MLLMs can achieve translation performance equivalent to supervised machine translation models by incorporating few-shot translation demonstrations into contextual information. We consider that salient features when inputting few-shot translation demonstrations play an important role in translation.

Therefore, in this study, we analyze magnitudes of the hidden state features $\boldsymbol{X}^k_{src\text{-}tgt} \in \mathbb{R}^{T_{src\text{-}tgt} \times d_{\text{in}}}$, those of few-shot translation demonstrations $X_{src\text{-}tgt} = \{x^1_{src\text{-}tgt}, ..., x^N_{src\text{-}tgt}\}$ output from k-th layer of an MLLM. Here, $T_{src\text{-}tgt}$ is a total amount of tokens included in $X_{src\text{-}tgt}$. To construct a translation demonstration $x^1_{src\text{-}tgt}$, we use n-shot bilingual sentence pairs between src and tgt language randomly sampled from bilingual data. In addition, we find the hidden state features $\boldsymbol{X}^k_{src} \in \mathbb{R}^{T_{src} \times d_{\text{in}}}$ and $\boldsymbol{X}^k_{tgt} \in \mathbb{R}^{T_{tgt} \times d_{\text{in}}}$, those of few-shot monolingual demonstrations $X_{src} = \{x^1_{src}, ..., x^N_{src}\}$ and $X_{tgt} = \{x^1_{tgt}, ..., x^N_{tgt}\}$, respectively. Please refer to Appendix A for detailed descriptions on how to construct each demonstration ($x_{src\text{-}tgt}$, x_{src}, and x_{tgt}).

To investigate the magnitude of each feature of $\boldsymbol{X}^k_{src\text{-}tgt}$, $\boldsymbol{X}^k_{src}$, and $\boldsymbol{X}^k_{tgt}$, we find $\|\boldsymbol{X}^k_{src\text{-}tgt}\|_2, \|\boldsymbol{X}^k_{src}\|_2, \|\boldsymbol{X}^k_{tgt}\|_2 \in \mathbb{R}^{d_{\text{in}}}$ following Sun et al. (2024), which are vectors that each element is L2-norm of the feature aggregated across the tokens of the corresponding hidden states. Specifically, j-th feature of $\|\boldsymbol{X}^k_{src\text{-}tgt}\|_2$ is $\|(\boldsymbol{X}^k_{src\text{-}tgt})_j\|_2$. We examine whether there are features that have extremely large magnitude compared to others.

Furthermore, we measure overlap ratio between the top and bottom β% of features, ranked according to their magnitudes. Specifically, the overlap ratio between the features in the top β% of $\|\boldsymbol{X}^k_{src\text{-}tgt}\|_2$ and those in the bottom β% of $\|\boldsymbol{X}^k_{src}\|_2$ is calculated as follows:

$$\frac{|\{d^1_{src\text{-}tgt}, \ldots, d^{d_\beta}_{src\text{-}tgt}\} \cap \{d^{d_{\text{in}}-d_\beta}_{src}, \ldots, d^{d_{\text{in}}}_{src}\}|}{d_\beta}$$

where d_β ($= d_{\text{in}} \cdot \beta/100$) is the number of dimensions accounting for β% of the total dimensions. The sets $\{d^1_{src\text{-}tgt}, \ldots, d^{d_\beta}_{src\text{-}tgt}\}$ and $\{d^{d_{\text{in}}-d_\beta}_{src}, \ldots, d^{d_{\text{in}}}_{src}\}$ denote dimensions corresponding to the top- and bottom-d_β features. If the ratio between the top of $\|\boldsymbol{X}^k_{src\text{-}tgt}\|_2$ and the bottom of $\|\boldsymbol{X}^k_{src}\|_2$ (or $\|\boldsymbol{X}^k_{tgt}\|_2$) is high value, it suggests that the features prominently active during translation demonstrations diminish in importance during monolingual demonstrations. This implies the existence of large magnitude features that are active only when processing translation demonstrations.

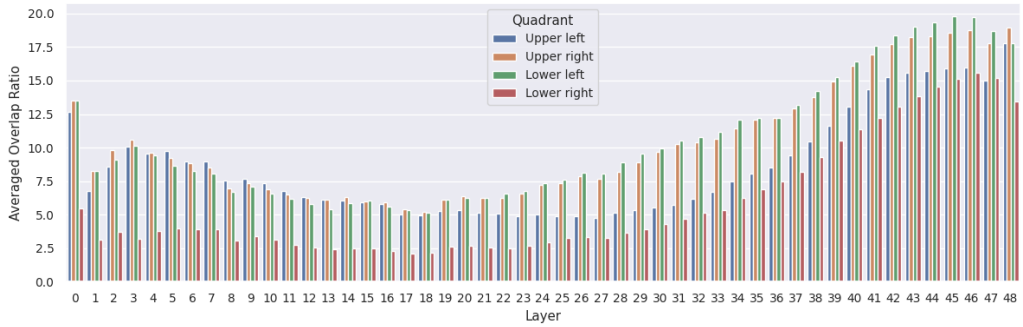


Figure 4: Averaged overlap ratios for each quadrant. This plot quantifies the overlap between monolingual and translation demonstrations in the upper-left, upper-right, lower-left, and lower-right quadrants across different layers.

RQ2: Are the large magnitude features relevant for translation performance? We investigate whether the large magnitude features when inputting few-shot translation demonstrations are relevant for translation performance. To reveal this, we retain weights of the MLLM that involve operation for the large magnitude features and prune other weights. It is accomplished through pruning based on Wanda using $X_{src\text{-}tgt}$ as calibration data. We denote the pruned weights as $\boldsymbol{\theta}_{src\text{-}tgt}$. We also perform pruning using X_{src} and X_{tgt} as calibration data, and denote each pruned weights as $\boldsymbol{\theta}_{src}$ and $\boldsymbol{\theta}_{tgt}$, respectively. We conduct translation with original weights before pruning $\boldsymbol{\theta}$ and each pruned weight. If $\boldsymbol{\theta}_{src\text{-}tgt}$ maintains the performance relative to $\boldsymbol{\theta}$, and also $\boldsymbol{\theta}_{src}$ and $\boldsymbol{\theta}_{tgt}$ decrease the performance relative to $\boldsymbol{\theta}$, it suggests that it is relevant for translation to use the large magnitude features activated by inputting translation demonstrations.

4.1 Experimental Settings

We experimented with XGLM-2.9B, mGPT-1.7B, and BLOOM-3B and employed Etxaniz et al. (2024)'s setting and implementation for the construction of demonstrations and translation. To construct $X_{src\text{-}tgt}$, X_{src}, X_{tgt}, we set N and n, the numbers of demonstrations and shots of each demonstration, as 100 and 4, respectively. As a result, we randomly sampled 400 (= $N \times n$) bilingual sentence pairs from development data of FLORES-200 (NLLB Team et al., 2022).

To evaluate translation performance for RQ2, we conducted 4-shot in-context learning. We constructed a 4-shot translation demonstration $x_{src\text{-}tgt}$ with the procedure described in RQ1. We incorporated a test example from the source language denoted as s_{test} at the end of $x_{src\text{-}tgt}$. Consequently, a translation was produced by inputting a concatenated string $x_{src\text{-}tgt} \oplus src : s_{\text{test}} \oplus tgt :$ into each model. We used the XNLI test set (Conneau et al., 2018) as evaluation data. The test sets of each language are bilingual to each other.

For pruning, we used the implementation of Sun et al. (2024) and set a pruning ratio α to 30%. [2]

4.2 Experimental Results

Answer to RQ1: Few-shot translation demonstrations activate specific features up to middle layer.[3] Fig. 2 shows that the top-20 dimensions with the largest magnitudes of $\|\boldsymbol{X}^{27}_{\text{Zh}-\text{En}}\|_2$. The figure shows that there are features that have extremely large magnitudes compared to others. We found that there are common features that have large magnitude independent of the demonstrations, such as 1849, 218, and 1244. On the other hand, we observed that there are features included in the top-20 magnitudes of $\|\boldsymbol{X}^{27}_{\text{Zh}-\text{En}}\|_2$ but not in those of $\|\boldsymbol{X}^{27}_{\text{Zh}}\|_2$ and $\|\boldsymbol{X}^{27}_{\text{En}}\|_2$, such as 354, 231, and 1262. We quantified the ratio of unique dimensions within the top-20, top-50, and top-100 magnitudes and showed the results in Appendix C.

To further investigate the commonalities and differences among features with large magnitudes, we quantified the ratio of overlaps among the top and bottom 30% of features, ranked by their magnitude, across each demonstration, as depicted in Figures 1 and 3. Up to the middle (27-th) layer, as shown in Fig. 1, the ratios among the monolingual demonstrations (the upper left quadrant) reveal minimal overlap within linguistically similar languages (En, Fr, and Es), yet exhibit many overlaps within linguistically distant languages, suggesting that the

[2] See the Appendix B for implementation details, including sources of the code and hyperparameters.

[3] We observed the similar trends as described in this section across other models (i.e. BLOOM and mGPT), layers, and languages. Please refer Appendix C and supplemental materials for the remaining results.

Weight	Source Language														
	Ar	Bg	De	El	Hi	Ru	Sw	Th	Tr	Ur	Vi	Fr	Es	Zh	Avg.
$\boldsymbol{\theta}$	19.4	31.3	34.3	35.4	16.7	22.5	21.4	17.3	14.8	13.2	25.0	35.2	37.3	16.1	24.3
$\boldsymbol{\theta}_{\mathrm{En}}$	17.9	30.7	33.5	34.8	16.2	21.9	20.6	17.1	14.1	13.0	24.6	34.3	36.7	15.5	23.5
$\boldsymbol{\theta}_{\mathrm{Fr}}$	18.0	30.2	33.3	34.4	16.1	21.8	20.4	17.1	14.2	13.1	24.7	34.1	36.2	15.0	23.5
$\boldsymbol{\theta}_{\mathrm{Es}}$	18.0	30.3	33.4	33.5	16.0	21.9	20.5	16.8	14.1	12.6	24.4	34.2	36.4	15.1	23.4
$\boldsymbol{\theta}_{\mathrm{Zh}}$	18.0	30.2	33.5	34.3	16.0	21.9	20.2	16.9	13.9	12.8	24.4	33.8	36.2	15.2	23.4
$\boldsymbol{\theta}_{\mathrm{Hi}}$	18.2	30.1	33.0	33.8	15.9	22.1	20.1	17.0	14.0	12.5	23.6	33.1	35.7	15.3	23.2
$\boldsymbol{\theta}_{\mathrm{Sw}}$	18.1	30.5	33.3	34.4	15.8	21.7	20.2	17.1	14.4	12.8	24.7	34.0	36.0	15.5	23.5
$\boldsymbol{\theta}_{\mathrm{Fr-En}}$	$18.3^{\dagger\ddagger}$	$31.1^{\dagger\ddagger}$	$33.9^{\dagger\ddagger}$	$35.2^{\dagger\ddagger}$	$16.5^{\dagger\ddagger}$	$22.4^{\dagger\ddagger}$	$20.9^{\dagger\ddagger}$	$17.7^{\dagger\ddagger}$	$14.5^{\dagger\ddagger}$	$13.3^{\dagger}$	$25.0^{\dagger}$	$34.6^{\dagger\ddagger}$	$37.0^{\dagger\ddagger}$	$15.7^{\dagger\ddagger}$	24.1
$\boldsymbol{\theta}_{\mathrm{Es-En}}$	$18.6^{\dagger\ddagger}$	$30.9^{\dagger\ddagger}$	$33.9^{\dagger\ddagger}$	$35.0^{\dagger\ddagger}$	$16.5^{\dagger\ddagger}$	$22.3^{\dagger}$	$20.9^{\dagger\ddagger}$	$17.3^{\dagger\ddagger}$	$14.5^{\dagger}$	$13.1^{\ddagger}$	$24.9^{\dagger\ddagger}$	$34.6^{\dagger\ddagger}$	$37.0^{\dagger\ddagger}$	$15.6^{\dagger\ddagger}$	23.9
$\boldsymbol{\theta}_{\mathrm{Zh-En}}$	$18.9^{\dagger\ddagger}$	$30.9^{\dagger\ddagger}$	$33.9^{\dagger}$	$34.9^{\ddagger}$	$16.5^{\dagger}$	$22.2^{\dagger}$	$20.8^{\ddagger}$	$17.6^{\dagger\ddagger}$	14.3	$13.4^{\dagger\ddagger}$	$25.2^{\dagger\ddagger}$	$34.8^{\dagger\ddagger}$	$36.7^{\ddagger}$	$15.8^{\ddagger}$	24.0
$\boldsymbol{\theta}_{\mathrm{Hi-En}}$	$18.7^{\dagger\ddagger}$	$31.1^{\dagger\ddagger}$	$33.8^{\dagger\ddagger}$	34.9	$16.3^{\dagger\ddagger}$	$22.2^{\dagger\ddagger}$	$20.7^{\dagger\ddagger}$	$17.5^{\dagger\ddagger}$	$14.4^{\dagger\ddagger}$	$12.9^{\dagger\ddagger}$	$25.1^{\dagger\ddagger}$	$34.4^{\dagger\ddagger}$	$36.9^{\dagger\ddagger}$	$15.7^{\dagger\ddagger}$	23.9
$\boldsymbol{\theta}_{\mathrm{Sw-En}}$	$18.6^{\dagger\ddagger}$	30.8	$33.8^{\dagger\ddagger}$	34.9	$16.2^{\dagger\ddagger}$	$22.2^{\dagger\ddagger}$	$20.9^{\dagger\ddagger}$	$17.5^{\dagger\ddagger}$	$14.4^{\dagger\ddagger}$	$13.2^{\dagger\ddagger}$	$24.9^{\dagger\ddagger}$	34.4	$36.9^{\dagger\ddagger}$	$15.6^{\dagger\ddagger}$	23.9

Table 1: BLEU scores of XGLM on original weights $\boldsymbol{\theta}$ and each pruned weights $\boldsymbol{\theta}_{src\text{-}tgt}$, $\boldsymbol{\theta}_{src}$, and $\boldsymbol{\theta}_{\mathrm{tgt}}$. † and ‡ denote statistical significance against $\boldsymbol{\theta}_{src}$ and $\boldsymbol{\theta}_{tgt}$, respectively. The details for the statistical significance test were shown in Appendix B.

activated features are influenced by linguistic characteristics. Additionally, two notable observations can be discerned from the figures. Firstly, the ratios between the monolingual and translation demonstrations (the lower left and upper right quadrants) are relatively high compared to the others, indicating that the features prominently activated by the translation or monolingual demonstrations become less significant in the alternate demonstrations. Secondly, the ratios within the translation demonstrations (the lower left quadrant) display little overlap, suggesting that similar features are activated by each translation demonstration. These findings suggest that specific features are uniquely activated during the input of translation demonstrations.

In the proximity of final (47-th) layer, as depicted in Fig. 3, the ratios were higher across the board. Interestingly, the ratios between the translation and the source languages' monolingual demonstrations were minimal in these layers. This result indicates that the features that are active when inputting translation demonstrations are more dependent on non-English languages than on English.

To provide an overarching perspective of the varying patterns across layers, we computed the average overlap ratios for each quadrant of the heatmap, as illustrated in Fig. 4. The figure shows that a decline in each overlap ratio from the initial to the middle layers, followed by an increase from the middle to the last layers. Moreover, the figure illustrates that the average overlap ratios in the lower quadrant, those among the translation demonstrations, are lower compared to the others across each layer.

Answer to RQ2: The large magnitude features are relevant for maintaining translation performance. We conducted to-English translations with the original and pruned models. Tab. 1 shows the BLEU scores of each model. The BLEU scores for models pruned by monolingual ($\boldsymbol{\theta}_{\mathrm{En}}$, $\boldsymbol{\theta}_{\mathrm{Fr}}$, $\boldsymbol{\theta}_{\mathrm{Es}}$, etc.) and translation ($\boldsymbol{\theta}_{\mathrm{Fr-En}}$, $\boldsymbol{\theta}_{\mathrm{Es-En}}$, etc.) demonstrations were degraded approximately one and 0.3 points compared to the scores of the unpruned original model ($\boldsymbol{\theta}$), respectively. This result suggests that the features retained in the model pruned based on translation demonstrations, but omitted in the model pruned based on monolingual demonstrations, are important for maintaining the translation performance of the original model.

5 Multilinguality of Pruned MLLMs

Our main goal is to enhance the zero-shot performance of MLLMs in non-English contexts by leveraging the alignment capability between English and non-English languages. In the previous section, we demonstrated that the large magnitude features that are active only when processing few-shot translation demonstrations play an important role to maintain the translation performance of MLLMs. This implies that these prominent features are essential for bringing out the alignment capability. We hypothesize that zero-shot in-context learning is performed while accentuating the alignment capability by prioritizing the use of the large magnitude features from few-shot translation demonstrations. We can accomplish this by employing the pruned weights, denoted as $\boldsymbol{\theta}_{src\text{-}tgt}$, based on few-shot translation demonstrations. The pruned model focuses on the large magnitude features relevant for

Model	Weight	Ar	Bg	De	El	Hi	Ru	Sw	Th	Tr	Ur	Vi	Fr	Es	Zh	Avg.
XGLM	$\boldsymbol{\theta}$	44.0	41.8	41.6	44.0	44.7	39.5	42.5	44.2	41.0	42.7	45.2	45.0	35.2	43.9	42.5
	LRP2	-	-	-	-	44.6	-	42.4	-	-	-	-	46.4	36.0	45.1	-
	$\boldsymbol{\theta}_{\text{Rand}}$	32.4	33.5	34.6	33.9	33.2	33.5	34.3	34	33.2	34.8	33.5	36.2	34.4	33.9	33.9
	$\boldsymbol{\theta}_{\text{Fr-En}}$	**44.9**†	45.8†	42.9†	**46.2**†	44.9	43.0†	**43.5**†	45.4†	42.5†	42.8	**47.7**†	47.2†	39.7†	**47.2**†	**44.5**
	$\boldsymbol{\theta}_{\text{Es-En}}$	44.3	**45.9**†	42.5†	45.6†	44.7	43.2†	43.3†	**45.9**†	**42.8**†	42.5	47.5†	47.0†	36.3†	46.8†	44.1
	$\boldsymbol{\theta}_{\text{Zh-En}}$	44.1	**45.9**†	**43.0**†	45.9†	**45.0**	**43.6**†	42.5	45.5†	42.2†	**43.0**	47.5†	**47.7**†	**39.8**†	46.7†	44.4
	$\boldsymbol{\theta}_{\text{Hi-En}}$	43.7	42.8†	40.3	45.2†	44.7	42.5†	42.7	45.5†	41.0	42.2	45.5	46.2	37.2†	46.0†	43.3
	$\boldsymbol{\theta}_{\text{Sw-En}}$	43.8	44.5†	40.3	46.1†	43.7	41.7†	42.0	45.6†	41.7	42.0	45.4	46.4†	38.9†	46.1†	43.4
mGPT	$\boldsymbol{\theta}$	39.2	39.7	35.0	41.0	38.9	39.2	34.0	41.6	39.9	39.9	42.2	42.3	39.4	**41.8**	39.6
	LRP2	-	-	-	-	35.2	-	34.4	-	-	-	-	34.2	33.1	34.1	-
	$\boldsymbol{\theta}_{\text{Rand}}$	33.3	33.2	32.9	33.2	33.3	33.3	33.4	33.0	33.0	33.2	33.5	33.6	34.0	33.1	33.2
	$\boldsymbol{\theta}_{\text{Fr-En}}$	39.3	39.5	36.3†	41.9†	**40.2**†	39.5†	34.7†	42.6†	**40.2**†	40.0	42.8†	42.1	39.7†	41.7	40.0
	$\boldsymbol{\theta}_{\text{Es-En}}$	**40.6**†	**40.3**†	**36.6**†	**42.6**†	39.5†	**39.8**†	**35.2**†	42.9†	40.1	39.9	**43.5**†	**42.5**†	**40.4**†	41.2	**40.3**
	$\boldsymbol{\theta}_{\text{Zh-En}}$	39.1	39.9†	36.1†	41.5†	39.8†	39.0	34.4†	**43.4**†	40.1	40.2†	43.2†	42.2	39.9†	41.4	40.0
	$\boldsymbol{\theta}_{\text{Hi-En}}$	39.1	39.5	34.9	41.1	40.3†	38.9	34.5†	42.1†	40.0	**40.5**†	42.6	42.1	38.9	41.6	39.7
	$\boldsymbol{\theta}_{\text{Sw-En}}$	38.7	39.4	34.9	40.1	40.1†	38.8	34.5†	42.3†	39.8	**40.7**†	43.6†	42.5†	39.1	41.9	39.7
BLOOM	$\boldsymbol{\theta}$	46.7	40.4	41.9	38.6	44.9	40.9	36.8	36.2	35.9	41.4	42.9	45.0	41.1	45.4	41.2
	LRP2	-	-	-	-	44.6	-	37.3	-	-	-	-	**46.0**	**44.3**	**46.8**	-
	$\boldsymbol{\theta}_{\text{Rand}}$	32.8	33.1	33.3	32.9	33.5	32.8	33.2	33.5	33.1	33.3	33.7	33.9	34.1	33.3	33.3
	$\boldsymbol{\theta}_{\text{Fr-En}}$	47.0	40.3	42.2	39.5†	45.9†	41.3†	36.9	36.5	35.6	41.1	41.9	44.9	41.2	44.6	41.3
	$\boldsymbol{\theta}_{\text{Es-En}}$	47.0	40.6	42.2	38.9	45.5†	41.1	37.1	36.6†	35.7	40.9	41.9	44.5	41.3	44.4	41.2
	$\boldsymbol{\theta}_{\text{Zh-En}}$	46.7	40.3	41.9	39.2†	45.3†	41.1	37.0	36.7†	35.7	40.4	41.6	44.2	40.5	45.2	41.1
	$\boldsymbol{\theta}_{\text{Hi-En}}$	47.2†	40.7	40.6	40.1†	45.2	41.6†	36.0	**37.0**†	**36.2**	41.4	42.1	45.4†	42.0†	43.7	41.4
	$\boldsymbol{\theta}_{\text{Sw-En}}$	46.9	40.3	40.6	40.3†	43.8	41.3†	35.7	36.2	36.1	41.5	43.0	44.4	40.8	44.1	41.0
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr-En}}$	46.9	40.6	42.0	40.0†	45.8†	**42.3**†	**37.4**†	36.6†	35.6	41.9†	41.3	45.1	41.6†	45.2	41.6
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Es-En}}$	47.1†	40.8†	42.2†	39.9†	**46.5**†	41.3†	37.2†	36.3	35.5	41.2	41.1	45.1	42.2	45.6	41.6
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh-En}}$	47.2†	40.5	42.2	40.0 †	45.8†	41.2	37.1	36.0	35.8	41.2	42.1	45.8†	42.4†	46.3†	**41.7**
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Hi-En}}$	**47.3**†	40.3	41.2	**40.3**†	45.4†	41.6†	36.2	36.8†	36.1	41.3	42.2	45.2	42.1†	45.3	41.5
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Sw-En}}$	46.8	**41.1**†	**42.5**†	39.5†	45.5†	41.2	36.9	36.6†	35.5	**42.5**†	**43.3**†	45.2	41.3	45.0	41.6

Table 2: Accuracy scores on the XNLI task. The highest scores in each model are indicated in bold. † denotes statistical significance against the original model $\boldsymbol{\theta}$.

performing translation, disregarding other features during the prediction process, thereby accentuating alignment capability regardless of the target task. We conduct zero-shot in-context learning based on the pruned weights $\boldsymbol{\theta}_{src\text{-}tgt}$ as follows:

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(\mathcal{P}(x_l, y)|\boldsymbol{\theta}_{src\text{-}tgt}) \quad (3)$$

5.1 Experimental Settings

We conducted zero-shot in-context learning with the framework as shown in §2 and followed the Lin et al. (2022)'s setting. We employed the XNLI (Conneau et al., 2018) and Multilingual Amazon Review Corpus (MARC) (Keung et al., 2020) tasks to evaluate non-English performance. For the MARC task, we used the two-label (positive and negative) classification setting. Please refer to Appendix B for the details of a template and verbalizer.

In this experiment, the performances of MLLMs pruned by demonstrations from En, Fr, Es, Zh, Hi, and Sw data were evaluated. For the construction of demonstrations, MLLMs, and pruning ratio, we used the same setting as described in §4.1.

For the comparative analysis, we evaluated the performance of randomly pruned models $\boldsymbol{\theta}_{\text{Rand}}$. In addition to them, we used LRP2 (Xu et al., 2023). We aligned the data used for applying LRP2 with those used for the pruning process. Given that LRP2 necessitates data from both source and target languages, we restricted the evaluation to Fr, Es, Zh, Hi, and Sw in the LRP2 experiment. The hyperparameter (a and b) search for LRP2 was conducted using the XNLI development sets corresponding to the respective languages.

5.2 Experimental Results

Tables 2 and 3 show the accuracy scores of each model. In XGLM and mGPT, the models pruned by the translation demonstrations with high-resource languages ($\boldsymbol{\theta}_{\text{Fr-En}}$, $\boldsymbol{\theta}_{\text{Es-En}}$, and $\boldsymbol{\theta}_{\text{Zh-En}}$) outperformed the original model $\boldsymbol{\theta}$, the randomly

Model	Weight	De	Ja	Fr	Es	Zh	AVg.
XGLM	$\boldsymbol{\theta}$	64.0	64.6	60.2	60.8	67.6	63.4
	LRP2	-	-	60.4	60.1	66.3	-
	$\boldsymbol{\theta}_{\text{Rand}}$	49.3	52.3	55.4	53.4	51.1	52.3
	$\boldsymbol{\theta}_{\text{Fr-En}}$	64.4†	**66.9**†	**60.8**†	60.8	69.2†	**64.4**
	$\boldsymbol{\theta}_{\text{Es-En}}$	64.1	66.3†	**60.8**†	**61.0**†	68.7†	64.2
	$\boldsymbol{\theta}_{\text{Zh-En}}$	**64.5**†	64.4	60.5†	60.0	68.0†	63.5
	$\boldsymbol{\theta}_{\text{Hi-En}}$	63.8	66.1†	60.0	60.4	**69.5**†	63.9
	$\boldsymbol{\theta}_{\text{Sw-En}}$	63.9	64.8	59.8	60.3	68.5†	63.5
mGPT	$\boldsymbol{\theta}$	65.9	56.3	64.1	65.6	53.9	61.2
	LRP2	-	-	51.7	52.4	50.7	-
	$\boldsymbol{\theta}_{\text{Rand}}$	52.8	54.4	50.3	50.8	51.9	52.0
	$\boldsymbol{\theta}_{\text{Fr-En}}$	**66.4**†	57.1†	63.6	65.7	55.2†	61.6
	$\boldsymbol{\theta}_{\text{Es-En}}$	**66.4**†	**57.2**†	64.6†	**65.9**†	55.3†	**61.9**
	$\boldsymbol{\theta}_{\text{Zh-En}}$	66.1	56.9†	**64.7**†	**65.9**†	**55.5**†	61.8
	$\boldsymbol{\theta}_{\text{Hi-En}}$	66.2†	57.7†	63.7	65.6	55.4†	61.7
	$\boldsymbol{\theta}_{\text{Sw-En}}$	66.3†	57.4	64.2	65.8	55.2†	61.7
BLOOM	$\boldsymbol{\theta}$	52.7	59.3	62.4	63.6	63.1	60.2
	LRP2	-	-	60.4	**63.8**	66.3	-
	$\boldsymbol{\theta}_{\text{Rand}}$	51.4	53.4	52.2	52.3	52.9	52.4
	$\boldsymbol{\theta}_{\text{Fr-En}}$	52.8	59.7†	61.9	62.5	64.2†	60.3
	$\boldsymbol{\theta}_{\text{Es-En}}$	**53.9**†	59.8†	61.3	62.0	63.6†	60.1
	$\boldsymbol{\theta}_{\text{Zh-En}}$	53.4†	59.4	62.1	62.7	64.6†	60.4
	$\boldsymbol{\theta}_{\text{Hi-En}}$	52.5	59.6†	60.9	61.6	64.6†	59.8
	$\boldsymbol{\theta}_{\text{Sw-En}}$	53.0†	59.6†	61.1	62.0	63.2	59.8
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr-En}}$	**53.9**†	59.9†	**62.7**†	63.1	63.9†	60.7
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Es-En}}$	53.6†	**60.3**†	62.5	63.2	64.9†	60.9
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh-En}}$	**53.9**†	59.9†	62.5	63.1	**65.6**†	**61.0**
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Hi-En}}$	53.2†	60.2†	61.9	62.8	64.5†	60.5
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Sw-En}}$	53.3†	59.6†	61.8	62.9	63.4†	60.2

Table 3: Accuracy scores on the MARC task.

pruned models $\boldsymbol{\theta}_{\text{Rand}}$, and LRP2.[4] Additionally, those pruned models showed statistical significance against the original model in more than half of the languages. While the models pruned by the translation demonstrations with low-resource languages ($\boldsymbol{\theta}_{\text{Hi-En}}$, $\boldsymbol{\theta}_{\text{Sw-En}}$) also enhanced the performance, the improvements were marginal. This result indicates that non-English performances of XGLM and mGPT are effectively enhanced by pruning using translation demonstrations with high-resource languages.

On the contrary, no discernible enhancement was observed through the application of pruning in the context of BLOOM regardless of the languages.

5.3 Analysis

Eliminating programming language generation ability. While XGLM and mGPT were trained by multilingual natural language texts, BLOOM was trained by both multilingual natural language and programming language texts. Scao et al. (2022) showed that BLOOM has the ability to generate programming language text comparable to models such as PolyCoder (Xu et al., 2022), which is trained using only programming language data. If the ability for programming language generation persists within the pruned model, it may introduce undesired noise into the model's predictions, such as inference on natural language understanding tasks.

Weight	XGLM	mGPT	BLOOM
$\boldsymbol{\theta}$	0.741	0.537	0.561
$\boldsymbol{\theta}_{\text{Fr-En}}$	0.744	0.593	0.569 (0.567)
$\boldsymbol{\theta}_{\text{Es-En}}$	0.749	0.593	**0.576** (0.569)
$\boldsymbol{\theta}_{\text{Zh-En}}$	**0.751**	**0.603**	0.569 (0.568)

Table 4: Averaged RankC scores with English across each language. The BLOOM scores reported both inside and outside of parentheses reflect the results obtained with and without the application of pruning using texts from programming language, respectively.

In this study, we refine the scoring metric (Eq. 2) of Wanda to reduce the weights that involve operation for such noisy features. Specifically, we reformulate Eq. 2 as follows:

$$S_{i,j} = \left|\boldsymbol{\theta}^{k}_{i,j}\right| \cdot \|\boldsymbol{X}^{k-1}_{j}\|_2 \cdot \frac{\|\boldsymbol{X}^{k-1}_{j}\|_2}{\|\boldsymbol{Z}^{k-1}_{j}\|_2} \quad (4)$$

where $\boldsymbol{Z}^{k-1}_{j} \in \mathbb{R}^{T_Z \times d_{\text{in}}}$ represents $(k-1)$-th layer's hidden state features of another calibration data Z. This reformulated equation assigns a small score for i,j-th elements if the j-th features of $\boldsymbol{Z}^{k-1}_{j}$ have large magnitudes. By selecting calibration data Z that accentuates a model's capacity to perform a specific task, we can eliminate weights associated with operations on features that are activated when executing the task. In our scenario, we use programming language texts as Z to eliminate BLOOM's programming language generation capability.

We employed $X_{src\text{-}tgt}$, X_{src}, and X_{tgt} as X and Python codes from huggingface as Z, and denoted each pruned model as $\boldsymbol{\theta}^{\text{Prog}}_{src\text{-}tgt}$, $\boldsymbol{\theta}^{\text{Prog}}_{src}$, and $\boldsymbol{\theta}^{\text{Prog}}_{tgt}$, respectively. The models pruned by our reformulated metric demonstrated superior performance compared to those pruned using the original metric proposed by Wanda as shown in Tables 2 and 3. Furthermore, the table reveals that the pruned

[4] We observed that the models pruned by the translation demonstrations outperformed those pruned by the monolingual demonstrations. Additionally, we confirmed that this pruning enhances the performance of a larger-scale model. Please refer Appendix D for the detail.

models, $\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr-En}}$, $\boldsymbol{\theta}^{\text{Prog}}_{\text{Es-En}}$, and $\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh-En}}$, surpassed the performance of $\boldsymbol{\theta}$. These results indicate that to effectively enhance performance in non-English languages, it is important to selectively retain and prune weights for translation and programming language generation, respectively. Although the pruned models did not exceed the performance of LRP2, it is noteworthy that the pruning strategy consistently improved non-English performance, unlike LRP2, which significantly degraded the performance in mGPT. Therefore, our experimental results suggest that pruning is a promising strategy to enhance the non-English performance.

Evaluation of cross-lingual consistency. The purpose of pruning MLLMs through translation demonstrations is to elicit alignment ability between English and non-English languages to utilize English inference capability for non-English inference. Therefore, finally, we measured Ranking based Consistency (RankC) (Qi et al., 2023), which is a metric to measure the consistency of a model's predictions across each language.[5] If the RankC scores between English and non-English languages are high, predictions are consistent across the languages, i.e., the model utilizes inference capability in English for non-English inference.

Tab. 4 shows the averaged RankC scores with English across each language and employed the XNLI dataset for this experiment. It illustrates that models pruned using translation demonstrations ($\boldsymbol{\theta}_{\text{Fr}-\text{En}}$, $\boldsymbol{\theta}_{\text{Es}-\text{En}}$, and $\boldsymbol{\theta}_{\text{Zh}-\text{En}}$) achieve superior scores relative to the original model $\boldsymbol{\theta}$. This result indicates that the improvement in zero-shot performance for non-English languages stems from more effective utilization of the English inference capabilities than the original model.

6 Conclusion

In this study, we showed that there are large magnitude features activated when inputting few-shot translation demonstrations and the pruned MLLMs (i.e. XGLM and mGPT) based on the features enhances zero-shot performance on non-English languages by utilizing the English inference capabilities. Additionally, we reformulated the scoring metric to eliminate weights associated with operations for large magnitude features in programming language generation and demonstrated that the pruned BLOOM based on the reformulated metric enhances the non-English performance.

The observation from the result of pruning based on the reformulated metric paves the way for further inquiry into the selective pruning of model weights to optimize performance across diverse linguistic tasks. For future work, we would like to delve deeper into this aspect to identify weights that should be retained or pruned to enhance the performance of MLLMs.

Limitations

Lack of experiments on various hyperparameters and demonstrations. In the experiments, the pruning ratio α was fixed to 0.3, and no experiments were conducted with varying ratios. By exploring how the performance evolves with different pruning ratios, it's possible to identify the optimal pruning ratio that improves multilingual capabilities. Furthermore, if increasing the pruning ratio enhances performance, it allow for further model size reduction, enabling inference operations with lower memory requirements. Therefore, this experiment is crucial for investigating the potential for enhancing the performance of MLLMs and reducing computational costs. Additionally, we set the parameters N and n to 100 and 4, respectively. As the consequence, we used a total of 400 bilingual sentence pairs to construct few-shot translation demonstrations. The experiments were conducted under the assumption that the bilingual data were well-prepared, although, in reality, languages with well-prepared bilingual data are scarce. If capable of enhancing performance with an even more limited number of bilingual sentence pairs are proved, we can apply the pruning strategy across several language pairs. Moreover, since this study did not explor the effect of pruning by varying the values of N and n while keeping the size of the bilingual sentences, the optimal value of each hyperparameter remains unknown.

Previous studies (Vilar et al., 2023; Chitale et al., 2024) have demonstrated that the quality of few-shot demonstrations significantly impacts the translation performance of multilingual large language models (MLLMs). However, our research did not analyze how changes in the translation demonstrations for the pruning affect the performance. By conducting this analysis, we will provide deeper insights into the sensitivity of MLLMs to variations in translation demonstrations during the pruning

[5] See Appendix E for the detailed descriptions of RankC.

process.

Lack of analysis concerning the architectural differences between each model. In §5.3, our focus was on the differences in pre-training data. However, an analysis based on architectural differences was not conducted. Notably, BLOOM's architecture is distinctive, especially due to its use of ALiBi, which operates directly on attention scores influenced by token positions. This method of integrating positional information sets BLOOM apart from other models such as XGLM and mGPT. By analyzing these architectural differences, we may reveal the reasons for the distinct trends observed in each model. A detailed analysis focusing on these architectural variances will be undertaken.

Lack of experiments using larger-scale or additionally fine-tuned MLLMs. In this study, we focused on evaluating MLLMs with sizes up to 3B parameters. While we observed the effectiveness of the translation demonstration based pruning on XGLM-7.5B, we have not yet evaluated its effectiveness on other larger-scale MLLMs. As a result, it remains uncertain whether the performance improvements observed through the pruning with few-shot translation demonstrations extend to larger models.

In addition, the MLLMs we focused on our study were pre-trained solely on causal language modeling task. Recent developments have shown that models fine-tuned through instruction tuning (Wei et al., 2022) and Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) can produce outputs more aligned with human preferences, and it is such models that end users are likely to use. Demonstrating the utility of pruning on these types of models is considered important.

In future studies, we aim to investigate the effectiveness of pruning on larger models and those that have been fine-tuned, to assess its utility further.

Ethical Considerations

Potential risks for bias. In recent years, several studies have issued warnings about the potential risks associated with pre-trained language models, notably their propensity for generating biased statements. Previous researches (Zhao et al., 2020; Reusens et al., 2023; Goldfarb-Tarrant et al., 2023) have shown that biases in multilingual pre-trained models can be transferred across languages. Intuitively, enhancing the models' alignment capability between languages, i.e., strengthening the connections between languages, make the transfer of bias more straightforward. In this research, we improved the alignment capability between English and non-English languages in MLLMs through pruning, aiming to boost their zero-shot performance in non-English languages. Consequently, there is a potential risk that the pruned models might produce statements that include English biases, even when generating content in non-English languages. This issue was not considered in our study. Therefore, when using the pruned models, sufficient attention should be paid to the problem of bias.

Acknowledgments

This work was partly supported by JST, PRESTO Grant Number JPMJPR2366, Japan.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in LLMs for machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7384–7406, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Massimo Poesio, and Dirk Hovy. 2022. Hard and soft evaluation of NLP models with BOOtSTrap SAmpling - BooStSa. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–134, Dublin, Ireland. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Björn Ross, and Adam Lopez. 2023. Cross-lingual transfer can worsen bias in sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5691–5704, Singapore. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Hwichan Kim and Mamoru Komachi. 2023. Enhancing few-shot cross-lingual transfer with target language peculiar examples. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 747–767, Toronto, Canada. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

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.

NLLB Team, Marta R. Costa-jussà, 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, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896, Singapore. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat,

Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Frank F. Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, MAPS 2022, page 1–10, New York, NY, USA. Association for Computing Machinery.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Ziqing Yang, Wentao Ma, Yiming Cui, Jiani Ye, Wanxiang Che, and Shijin Wang. 2021. Bilingual alignment pre-training for zero-shot cross-lingual transfer. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 100–105, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Hyperparameters	Value
Max new tokens	64
Beam size	1
Temperature	0.8
Top k	100
Top p	0.75

Table 5: Hyperparameters used in translation experiments.

A How to Construct Few-shot Demonstrations

In this study, we investigated the hidden states from few-shot translation demonstrations $X_{src\text{-}tgt} = \{x^1_{src\text{-}tgt}, ..., x^N_{src\text{-}tgt}\}$. Each few-shot demonstration demonstration were constructed using Etxaniz et al. (2024)'s template function $f_{\text{mt}}(\cdot)$. Therefore, a demonstration $x^1_{src\text{-}tgt}$ is as follow:

$$\begin{aligned} x^1_{src\text{-}tgt} &= f_{\text{mt}}(s^1_1, t^1_1) \oplus ... \oplus f_{\text{mt}}(s^1_n, t^1_n) \\ f_{\text{mt}}(s, t) &= src : s \oplus tgt : t \oplus \text{[EOS]} \end{aligned}$$

where ":" and "[EOS]" denote the tokens corresponding to a colon and a end-of-sentence, respectively, and $\oplus$ denotes the concatenation operator. src and tgt represent source and target language names. For example, when English to Chinese translation, src and tgt are "English" and "Chinese".

In addition, we examined the the hidden states from few-shot monolingual demonstrations $X_{src} = \{x^1_{src}, ..., x^N_{src}\}$ and $X_{tgt} = \{x^1_{tgt}, ..., x^N_{tgt}\}$. Here, a demonstration x^1_{src} is as follow:

$$\begin{aligned} x^1_{src} &= f_{\text{lm}}(s^1_1) \oplus ... \oplus f_{\text{lm}}(s^1_n) \\ f_{\text{lm}}(s) &= src : s \oplus \text{[EOS]} \end{aligned}$$

where $f_{\text{lm}}(\cdot)$ is a template function for the monolingual demonstration.

B Implementation Details

We conducted the experiments with XGLM-2.9B[6], mGPT-1.7B[7], and BLOOM-3B[8] from huggingface and used a single Quadro RTX 8000 in the all experiments.

[6] https://huggingface.co/facebook/xglm-2.9B
[7] https://huggingface.co/ai-forever/mGPT
[8] https://huggingface.co/bigscience/bloom-3b

		a	b
XGLM	Fr	1	2
	Es	8	48
	Zh	1	7
	Hi	1	4
	Sw	7	29
mGPT	Fr	0	2
	Es	21	28
	Zh	5	16
	Hi	3	9
	Sw	7	13
BLOOM	Fr	1	22
	Es	21	28
	Zh	22	29
	Hi	8	12
	Sw	19	26

Table 6: Optimal layer configurations in LRP2 evaluation. The configurations were searched using the development sets of the corresponding languages.

	$\|\|\boldsymbol{X}^{27}_{\text{En}}\|\|_2$	$\|\|\boldsymbol{X}^{27}_{\text{Zh}}\|\|_2$
Top-20	30	25
Top-50	34	28
Top-100	36	25

Table 7: The ratio (%) of unique dimensions within the top-20, top-50, and top-100 magnitudes of $\|\|\boldsymbol{X}^{27}_{\text{Zh}-\text{En}}\|\|_2$ that are absent in the corresponding dimensions of $\|\|\boldsymbol{X}^{27}_{\text{Zh}}\|\|_2$ and $\|\|\boldsymbol{X}^{27}_{\text{En}}\|\|_2$.

To construct each few-shot translation demonstrations and conduct translation experiments, we employed the Etxaniz et al. (2024)'s implementation[9] and the license was not stated. Tab. 5 shows the hyperparameters used in our translation experiments and they are the default ones of the implementation.

For pruning using Wanda, we used the official implementation of Sun et al. (2024)[10] published by MIT license. For pruning using refined equation (Eq. 4), we used our original implementation and attached the code in supplemental materials.

In our experiments, we used datasets of FLORES-200[11], XNLI[12], MARC[13], and Python

[9] https://github.com/juletx/self-translate
[10] https://github.com/locuslab/wanda
[11] https://huggingface.co/datasets/facebook/flores
[12] https://huggingface.co/datasets/xnli
[13] https://huggingface.co/datasets/SetFit/

codes[14] from huggingface for constructing of demonstrations, evaluating MLLMs' performance, and measuring the refined importance score, respectively.

To perform the XNLI task in cloze-style format, we used a template that converts a preliminary x^{pre} and a hypothesis x^{hyp} to "x^{pre}, right?, [Mask], x^{hyp}," and a verbalizer that maps each candidate label (Entailment, Contradiction, and Neutral) to 'Yes', 'No', and 'Also', respectively. For the MARC task, we used a template that converts a review x^{rev} to "x^{rev}It is [Mask]" and a verbalizer that maps each candidate label (negative and positive) to 'negative', and 'positive', respectively. When selecting a language of template and verbalizer, the language of the test example is expected to be the most intuitive and effective, but previous studies (Lin et al., 2022; Ahuja et al., 2023; Enomoto et al., 2024) demonstrated that English template and verbalizer achieves the best performance for most test languages. Therefore, we adopted English template and verbalizer regardless of test languages.

For the evaluation in the translation task, we employed BLEU (Papineni et al., 2002) score and used the huggingface's implementation[15]. For the evaluation in the XNLI and MARC tasks, We employed accuracy score and used the scikit-learn implementation[16]. We also performed statistical significance tests through bootstrapping sampling. We used the Koehn (2004)[17] and Fornaciari et al. (2022)[18]'s implementations for the statistical significance tests for the translation and other tasks, respectively. In this study, we let that there are statistical significance between the performances when the p-value is less than 0.1.

C Detecting Translation Features

Figures 5, 7, and 10 show the top-20 dimensions with the largest magnitudes of k-th layer's features of $\|\boldsymbol{X}^k_{\text{Zh}}\|_2$, $\|\boldsymbol{X}^k_{\text{En}}\|_2$, and $\|\boldsymbol{X}^k_{\text{Zh-En}}\|_2$. Each figure corresponds to XGLM, mGPT, and BLOOM, respectively. The figures demonstrate that there are features that have extremely large magnitudes compared to others regardless of each MLLM.

Figures 6, 8 and 11 are heatmaps illustrated the overlap ratios within top- and bottom 30% of features $\|\boldsymbol{X}^k_{src}\|_2$, $\|\boldsymbol{X}^k_{tgt}\|_2$, and $\|\boldsymbol{X}^k_{src\text{-}tgt}\|_2$ ranked by magnitude. Each figure corresponds to XGLM, mGPT and BLOOM, respectively. In XGLM and BLOOM, the ratios indicate that the overlaps within the monolingual demonstrations (the upper left quadrant) and translation demonstrations (the lower right quadrant) are smaller compared to those observed between the monolingual and translation demonstrations (the lower left and upper right quadrants). In mGPT, while the ratios associated with English, French, Spanish, and Chinese languages exhibit similar trends to those in XGLM and BLOOM, the ratios linked with Hindi and Swahili demonstrate high values irrespective of the demonstration type. One potential explanation for this discrepancy may be the inadequate training of mGPT, as these languages were insufficiently represented in its training dataset. Furthermore, these figures demonstrates that the ratio between the translation demonstrations involving identical language pairs is minimal, suggesting that the features activated are consistent regardless of the direction of translation. Consequently, this evidence suggests that the activation of certain features exclusively in response to the translation demonstrations is a model-independent phenomenon.[19]

Tab. 8 presents the BLEU scores of mGPT and BLOOM. In mGPT, the BLEU scores were very low compared to XGLM and BLOOM in overall, and the pruned models outperformed the original model. In BLOOM, the performances of models pruned by few-shot translation demonstrations have been preserved relative to the performances of the original model before pruning. However, the statistical significance between the models pruned by translation demonstrations and those pruned by monolingual demonstrations was not consistently observed, as in the case of XGLM.

D Multilinguality of Pruned MLLMs

Table 9 presents the accuracy scores for each model. The table reveals that the models pruned through translation demonstrations surpass those pruned by monolingual demonstrations and the original mod-

amazon_reviews_multi_ja

[14] thomwolf/github-python

[15] https://huggingface.co/spaces/evaluate-metric/bleu

[16] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

[17] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/analysis/bootstrap-hypothesis-difference-significance.pl

[18] https://github.com/fornaciari/boostsa

[19] We observed consistent trends across other layers. To avoid redundancy of the main text, we have included the remaining results in the supplemental materials.

els in performance. While LRP2 also improved the accuracy scores in XGLM and BLOOM, it did not enhance performance in mGPT. Although the scores for the models pruned using translation demonstrations are lower than those for LRP2 in BLOOM, the table suggests that this pruning strategy robustly enhances performance in non-English languages across different models, unlike LRP2.

Table 10 presents the accuracy scores of XGLM-7.5B, the largest model in the XGLM series, on the MARC task, and indicates that the pruning based on translation demonstrations also enhances the performance of XGLM-7.5B on non-English languages. This finding suggests the potential effectiveness of the translation demonstration-based pruning in larger-scale models.

E Evaluation of Cross-lingual Consistency

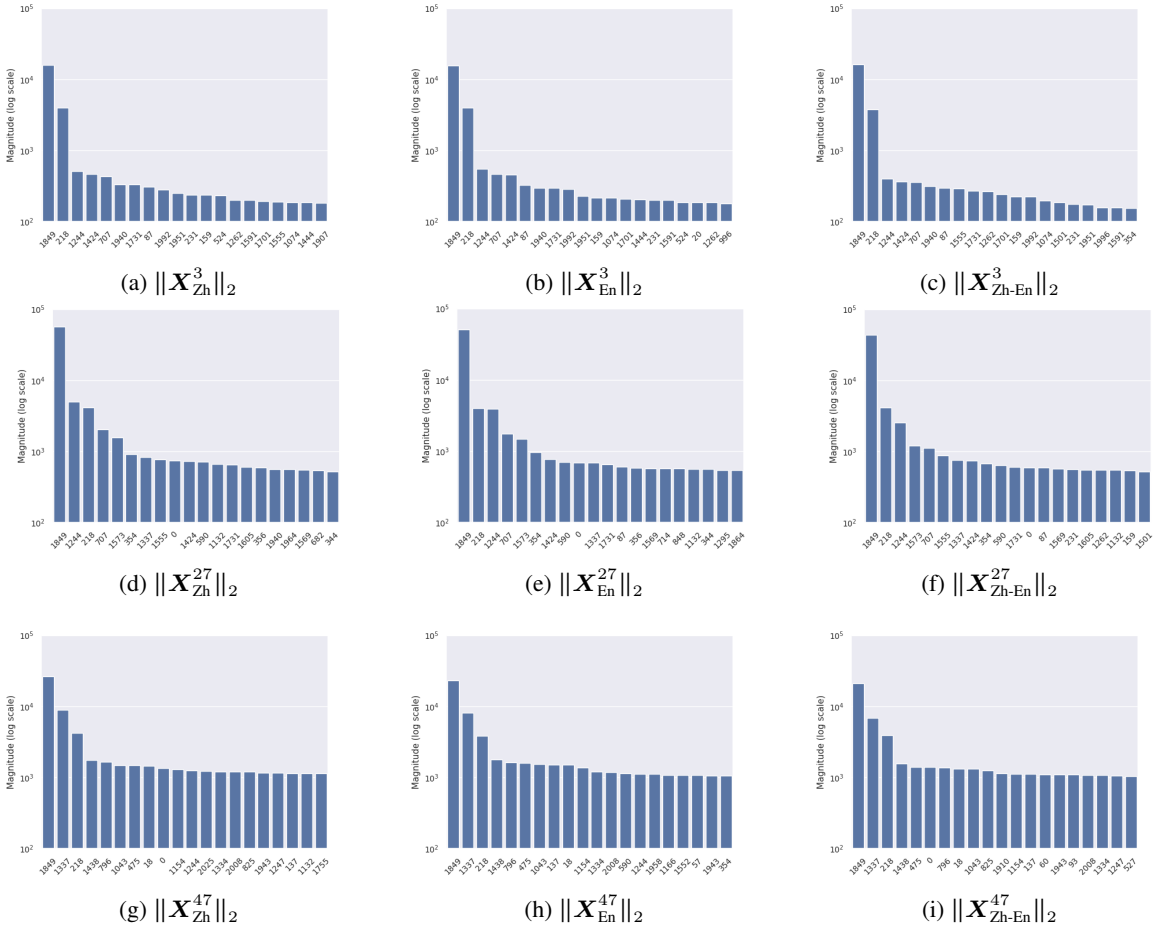
We measured consistency between the predictions across languages using Ranking based Consistency (RankC) (Qi et al., 2023). In this section, we explain calculation of RankC in detail.

Let us consider test examples denoted by $x_i \in X_l$ for language l and $x_i^{'} \in X_l^{'}$ for another language $l^{'}$, along with their corresponding sets of candidate labels $\{y_i^1, ..., y_i^{|\mathcal{Y}|}\}$ and $\{y_i^{'1}, ..., y_i^{'|\mathcal{Y}|}\}$, respectively. Here, y_i^1 has the highest prediction probability and $y_i^{|\mathcal{Y}|}$ has the lowest. RankC is defined as:

$$
\begin{aligned}
\mathrm{RankC}(l, l^{'}) &= \frac{\sum_{i=1}^{|X_l|} \mathrm{consist}(x_i, x_i^{'})}{|X_l|} \\
\mathrm{consist}(x_i, x_i^{'}) &= \sum_{j=1}^{|\mathcal{Y}|} w_j \cdot P@j \\
P@j &= \frac{1}{j} \left| \{y_i^1, ..., y_i^j\} \cap \{y_i^{'1}, ..., y_i^{'j}\} \right| \\
w_j &= \frac{e^{|\mathcal{Y}|-j}}{\sum_{k=1}^{|\mathcal{Y}|} e^{|\mathcal{Y}|-k}}
\end{aligned}
$$

The RankC metric assign high scores when the predicted labels consistent across across languages.

Tab. 11 presents the RankC scores of each model. This table demonstrates that the models pruned by few-shot translation demonstrations achieve the highest scores. Therefore, the pruning using few-shot translation demonstrations improve the cross-lingual consistency of their prediction compared to the original model.

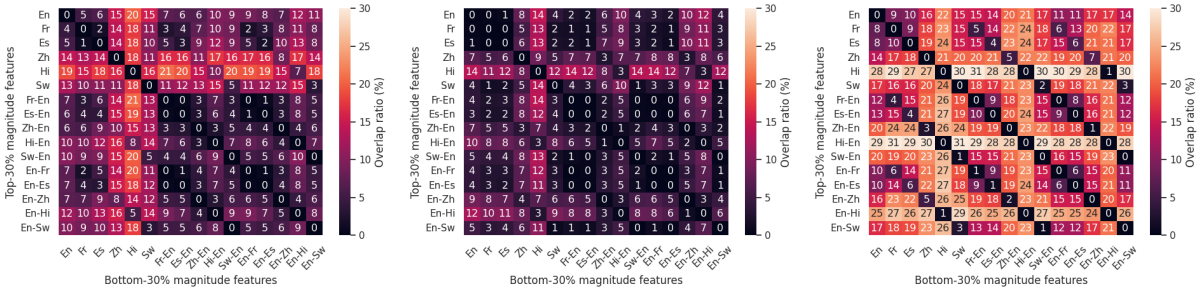


(a) $||\boldsymbol{X}^{3}_{\text{Zh}}||_2$ (b) $||\boldsymbol{X}^{3}_{\text{En}}||_2$ (c) $||\boldsymbol{X}^{3}_{\text{Zh-En}}||_2$

(d) $||\boldsymbol{X}^{27}_{\text{Zh}}||_2$ (e) $||\boldsymbol{X}^{27}_{\text{En}}||_2$ (f) $||\boldsymbol{X}^{27}_{\text{Zh-En}}||_2$

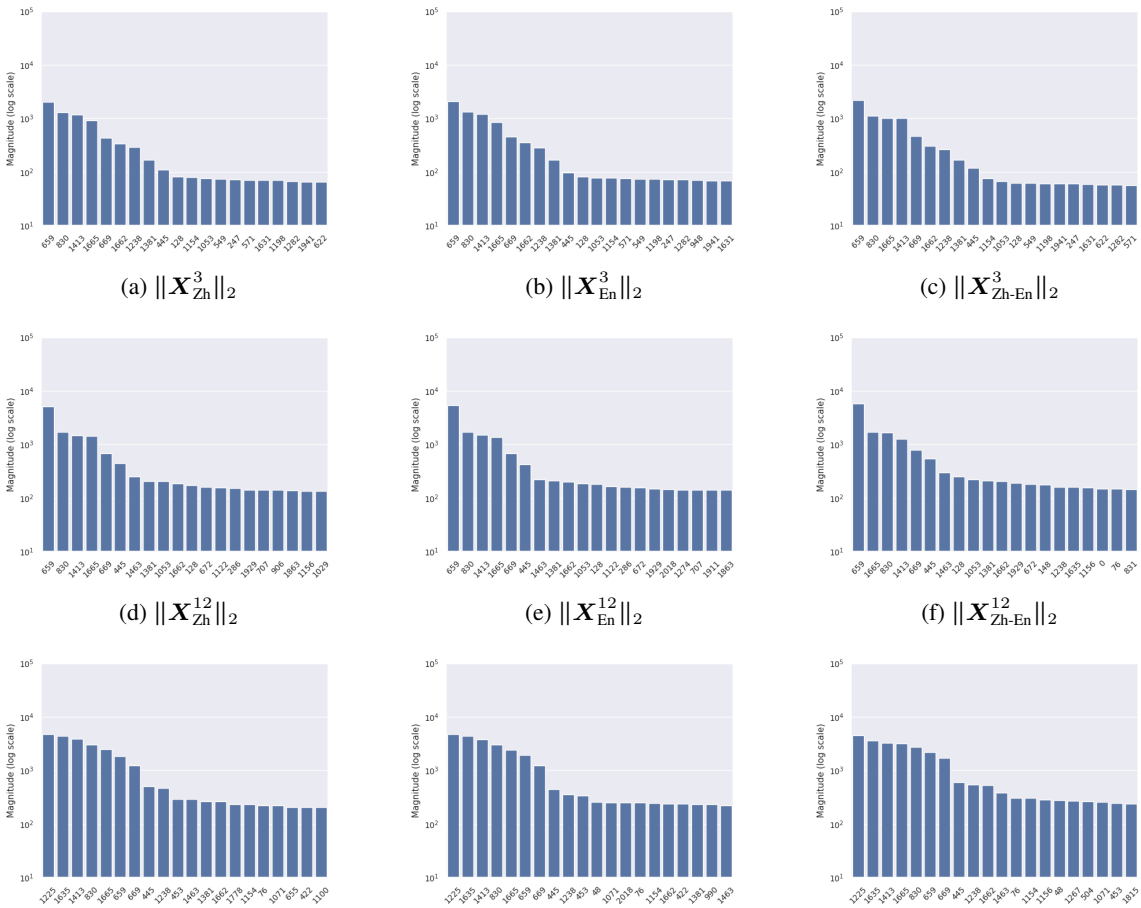
(g) $||\boldsymbol{X}^{47}_{\text{Zh}}||_2$ (h) $||\boldsymbol{X}^{47}_{\text{En}}||_2$ (i) $||\boldsymbol{X}^{47}_{\text{Zh-En}}||_2$

Figure 5: The top 20 dimensions of k-th layer's features of $||\boldsymbol{X}^{k}_{\text{Zh}}||_2$, $||\boldsymbol{X}^{k}_{\text{En}}||_2$, and $||\boldsymbol{X}^{k}_{\text{Zh-En}}||_2$ of XGLM.



(a) Third layer (b) 20-th layer (c) 47-th layer

Figure 6: The overlap ratios among the top- and bottom-30% features of XGLM ranked by magnitude.

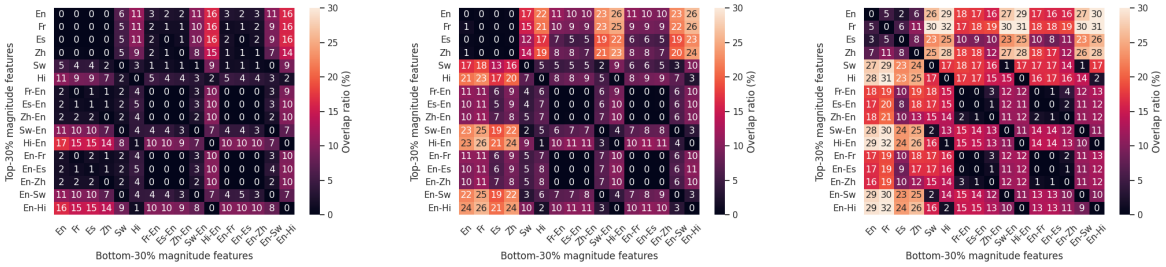


(a) $\|\boldsymbol{X}^3_{\text{Zh}}\|_2$ (b) $\|\boldsymbol{X}^3_{\text{En}}\|_2$ (c) $\|\boldsymbol{X}^3_{\text{Zh-En}}\|_2$

(d) $\|\boldsymbol{X}^{12}_{\text{Zh}}\|_2$ (e) $\|\boldsymbol{X}^{12}_{\text{En}}\|_2$ (f) $\|\boldsymbol{X}^{12}_{\text{Zh-En}}\|_2$

(g) $\|\boldsymbol{X}^{23}_{\text{Zh}}\|_2$ (h) $\|\boldsymbol{X}^{23}_{\text{En}}\|_2$ (i) $\|\boldsymbol{X}^{23}_{\text{Zh-En}}\|_2$

Figure 7: The top 20 dimensions of k-th layer's features of $\|\boldsymbol{X}^k_{\text{Zh}}\|_2$, $\|\boldsymbol{X}^k_{\text{En}}\|_2$, and $\|\boldsymbol{X}^k_{\text{Zh-En}}\|_2$ of mGPT.



(a) Third layer (b) 12-th layer (c) 23-th layer

Figure 8: The overlap ratios within the top- and bottom-30% features of mGPT ranked by magnitude.

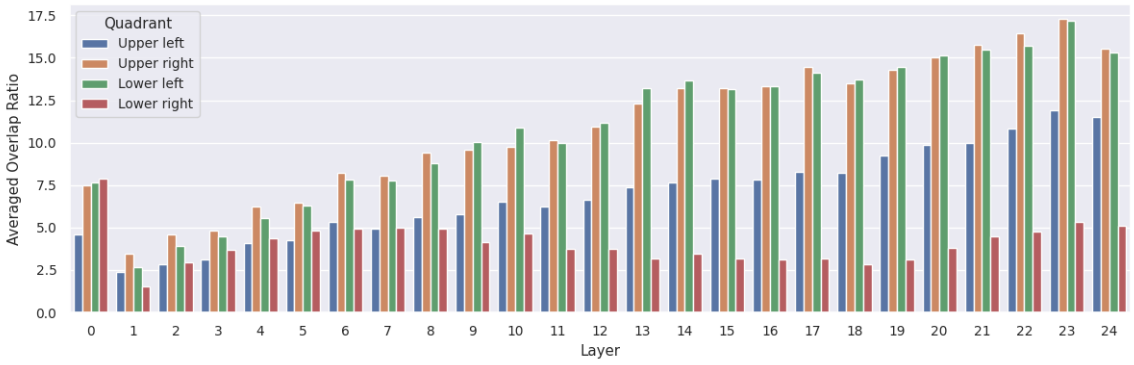
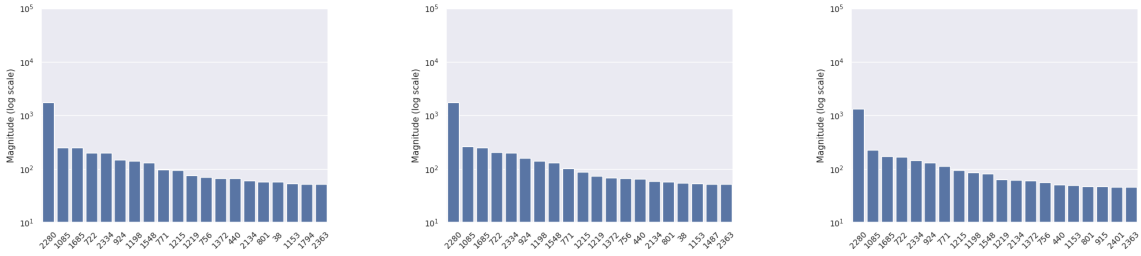
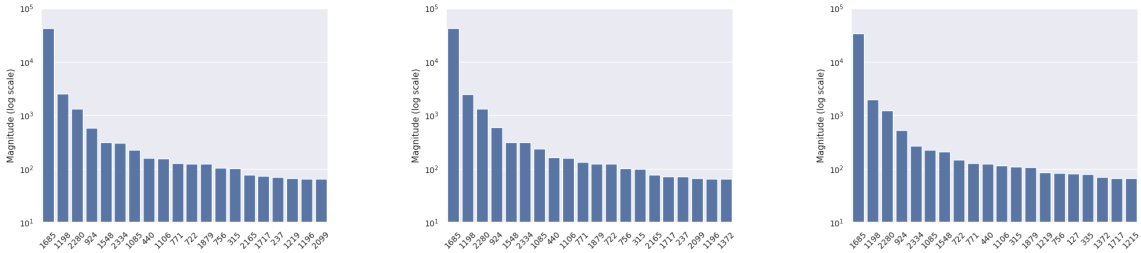


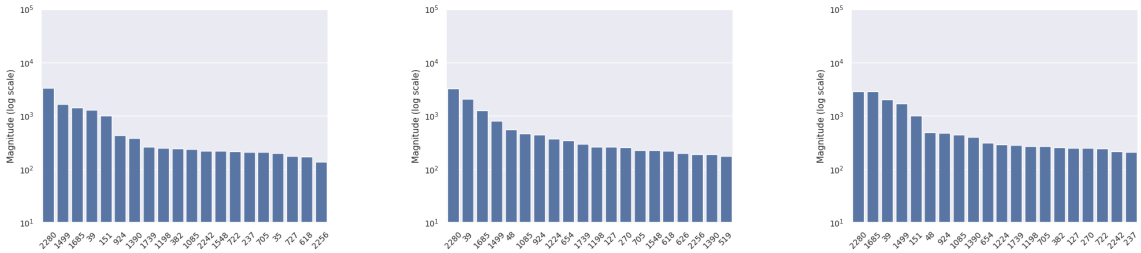
Figure 9: Averaged overlap ratios for each quadrant in mGPT.



(a) $||\boldsymbol{X}^{3}_{\text{Zh}}||_2$ (b) $||\boldsymbol{X}^{3}_{\text{En}}||_2$ (c) $||\boldsymbol{X}^{3}_{\text{Zh-En}}||_2$

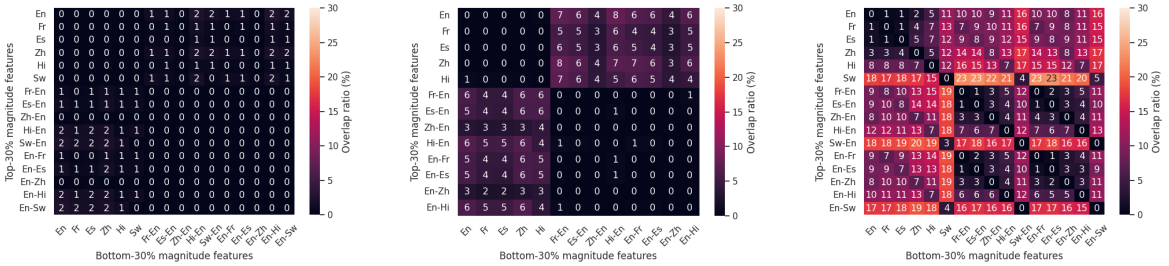


(d) $||\boldsymbol{X}^{15}_{\text{Zh}}||_2$ (e) $||\boldsymbol{X}^{15}_{\text{En}}||_2$ (f) $||\boldsymbol{X}^{15}_{\text{Zh-En}}||_2$



(g) $||\boldsymbol{X}^{28}_{\text{Zh}}||_2$ (h) $||\boldsymbol{X}^{28}_{\text{En}}||_2$ (i) $||\boldsymbol{X}^{28}_{\text{Zh-En}}||_2$

Figure 10: The top 20 dimensions of k-th layer's features of $||\boldsymbol{X}^{k}_{\text{Zh}}||_2$, $||\boldsymbol{X}^{k}_{\text{En}}||_2$, and $||\boldsymbol{X}^{k}_{\text{Zh-En}}||_2$ of BLOOM.



(a) Third layer (b) 20-th layer (c) 28-th layer

Figure 11: The overlap ratios within the top- and bottom-30% features of BLOOM ranked by magnitude.

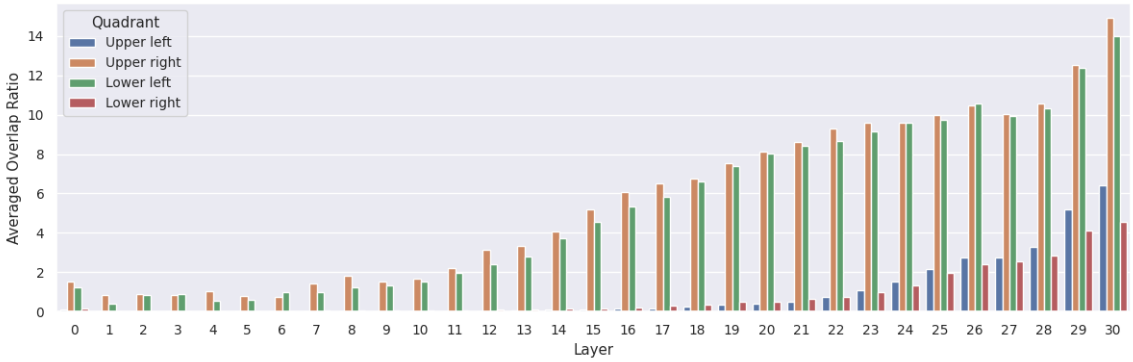


Figure 12: Averaged overlap ratios for each quadrant in BLOOM.

Model	Weight	Source Language														
		Ar	Bg	De	El	Hi	Ru	Sw	Th	Tr	Ur	Vi	Fr	Es	Zh	Avg.
mGPT	$\boldsymbol{\theta}$	3.53	6.14	8.75	8.76	1.37	5.16	3.13	1.55	2.17	2.62	5.14	9.16	9.46	2.02	4.92
	$\boldsymbol{\theta}_{\text{En}}$	4.42	6.54	6.10	9.86	3.00	4.75	5.28	4.50	3.63	2.97	6.17	6.93	8.36	5.71	5.58
	$\boldsymbol{\theta}_{\text{Fr}}$	4.35	7.02	6.98	10.35	3.24	4.61	5.35	4.56	3.19	2.94	6.15	6.42	8.39	5.37	5.63
	$\boldsymbol{\theta}_{\text{Es}}$	3.71	6.84	6.40	9.74	3.19	4.55	5.42	4.46	3.73	3.02	5.75	6.19	9.09	5.37	5.53
	$\boldsymbol{\theta}_{\text{Zh}}$	4.07	7.11	6.77	9.92	3.21	4.06	5.16	4.23	3.68	2.75	5.65	6.61	8.50	5.74	5.53
	$\boldsymbol{\theta}_{\text{Hi}}$	4.07	7.66	6.98	9.60	3.53	4.26	5.07	4.38	3.69	2.94	5.02	6.18	8.28	5.53	5.51
	$\boldsymbol{\theta}_{\text{Sw}}$	4.00	6.94	6.69	9.75	3.22	4.49	4.91	4.48	3.40	3.01	6.37	6.55	8.49	5.51	5.55
	$\boldsymbol{\theta}_{\text{Fr}-\text{En}}$	4.31	$7.31^{\dagger\ddagger}$	$6.67^{\dagger}$	$10.11^{\dagger}$	3.34	4.72	5.19	4.52	$3.66^{\ddagger}$	3.19	5.99	6.30	8.33	$5.77^{\ddagger}$	5.67
	$\boldsymbol{\theta}_{\text{Es}-\text{En}}$	$4.23^{\ddagger}$	$7.14^{\dagger\ddagger}$	$6.45^{\ddagger}$	9.72	3.21	$4.94^{\dagger\ddagger}$	5.05	4.36	3.55	2.96	$6.17^{\ddagger}$	$6.45^{\ddagger}$	8.19	$5.86^{\ddagger}$	5.59
	$\boldsymbol{\theta}_{\text{Zh}-\text{En}}$	$4.33^{\ddagger}$	$7.22^{\dagger\ddagger}$	$6.38^{\dagger}$	9.79	$3.32^{\dagger}$	$4.86^{\ddagger}$	$5.34^{\ddagger}$	4.39	3.30	2.94	$6.04^{\ddagger}$	6.27	$8.62^{\dagger}$	5.60	5.60
	$\boldsymbol{\theta}_{\text{Hi}-\text{En}}$	4.13	$7.17^{\dagger}$	$6.34^{\dagger}$	9.93	$3.55^{\dagger}$	4.15	5.02	4.25	3.58	2.89	$5.55^{\ddagger}$	$6.39^{\ddagger}$	$8.50^{\ddagger}$	5.55	5.50
	$\boldsymbol{\theta}_{\text{Sw}-\text{En}}$	3.96	$6.99^{\dagger}$	$6.57^{\dagger}$	9.46	$3.45^{\dagger\ddagger}$	4.49	5.02	4.68	3.56	3.01	5.63	6.33	$8.79^{\dagger\ddagger}$	5.57	5.54
BLOOM	$\boldsymbol{\theta}$	20.73	4.80	18.67	4.36	13.93	9.21	13.9	1.65	0.95	11.06	24.54	32.81	34.16	17.54	14.87
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{En}}$	16.67	2.74	15.35	2.86	10.97	6.66	10.21	1.33	1.00	8.06	20.65	29.39	29.9	14.36	12.15
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr}}$	20.93	4.14	18.56	4.32	13.11	8.9	14.24	1.48	1.06	10.76	24.63	33.01	33.74	17.04	14.70
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Es}}$	20.45	4.42	18.25	4.18	13.35	9.35	14.15	1.32	1.09	10.64	24.67	33.06	33.82	17.1	14.70
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh}}$	20.89	4.53	18.41	4.23	13.52	9.07	14.03	1.48	0.96	10.65	24.64	32.99	33.82	13.52	14.48
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Hi}}$	20.64	4.29	18.52	4.67	14.55	9.72	14.00	1.43	1.25	10.96	24.82	33.14	34.4	17.16	14.96
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Sw}}$	15.54	2.52	13.77	2.77	10.65	6.10	8.37	0.99	0.73	7.01	18.56	28.34	29.22	12.86	11.24
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr}-\text{En}}$	$21.01^{\dagger}$	$4.45^{\dagger\ddagger}$	$18.53^{\dagger}$	4.00	$13.22^{\dagger\ddagger}$	$9.12^{\dagger\ddagger}$	$13.99^{\dagger}$	1.59	1.11	$10.69^{\dagger}$	$24.66^{\dagger}$	$33.98^{\dagger\ddagger}$	$33.76^{\dagger}$	17.10	14.80
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Es}-\text{En}}$	$20.78^{\dagger\ddagger}$	$4.43^{\dagger}$	$18.23^{\dagger}$	$4.12^{\dagger}$	$13.33^{\dagger}$	$9.54^{\dagger\ddagger}$	$14.27^{\dagger\ddagger}$	1.54	1.09	$10.64^{\dagger}$	$24.70^{\dagger\ddagger}$	$33.24^{\dagger\ddagger}$	$33.85^{\dagger}$	$16.88^{\dagger}$	14.76
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh}-\text{En}}$	$20.77^{\dagger}$	$4.64^{\dagger}$	$18.74^{\dagger\ddagger}$	$4.02^{\dagger}$	$13.30^{\dagger}$	$9.09^{\dagger}$	$14.31^{\dagger\ddagger}$	1.56	0.97	$10.61^{\dagger}$	$24.85^{\dagger\ddagger}$	$32.67^{\dagger}$	$33.85^{\dagger}$	$17.09^{\dagger\ddagger}$	14.74
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Hi}-\text{En}}$	$20.97^{\dagger\ddagger}$	$4.33^{\dagger}$	$18.57^{\dagger}$	$4.57^{\dagger}$	$14.40^{\dagger}$	$9.65^{\dagger}$	$14.95^{\dagger\ddagger}$	1.38	1.16	$10.95^{\dagger}$	$24.76^{\dagger}$	$32.87^{\dagger\ddagger}$	$34.59^{\dagger\ddagger}$	$17.83^{\dagger\ddagger}$	15.07
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Sw}-\text{En}}$	$15.92^{\ddagger}$	2.69	$14.50^{\ddagger}$	3.00	$11.06^{\ddagger}$	$6.49^{\ddagger}$	$8.63^{\ddagger}$	1.24	0.91	7.21	$19.58^{\ddagger}$	$29.20^{\ddagger}$	$30.24^{\dagger\ddagger}$	$13.50^{\ddagger}$	11.72

Table 8: BLEU scores on original weights and each pruned weights.

Model	Weight	Ar	Bg	De	El	Hi	Ru	Sw	Th	Tr	Ur	Vi	Fr	Es	Zh	Avg.
XGLM	$\boldsymbol{\theta}$	44.0	41.8	41.6	44.0	44.7	39.5	42.5	44.2	41.0	42.7	45.2	45.0	35.2	43.9	42.5
	LRP2	-	-	-	-	-	-	-	-	-	-	-	46.4	36.0	45.1	-
	$\boldsymbol{\theta}_{\text{En}}$	44.6	43.3	42.8	44.3	45.0	41.4	42.2	44.1	42.5	42.7	45.9	47.1	39.4	46.5	43.7
	$\boldsymbol{\theta}_{\text{Fr}}$	44.5	43.5	41.9	45.4	44.6	42.2	42.6	44.5	41.7	**43.0**	46.9	47.3	38.0	46.6	43.7
	$\boldsymbol{\theta}_{\text{Es}}$	44.1	44.0	42.4	44.3	44.3	42.5	42.2	44.6	41.5	42.7	45.4	46.8	35.4	46.1	43.3
	$\boldsymbol{\theta}_{\text{Zh}}$	44.7	44.4	41.5	**46.2**	**45.0**	42.1	42.9	44.9	42.9	42.2	46.5	46.8	38.5	46.6	43.9
	$\boldsymbol{\theta}_{\text{Hi}}$	43.6	41.1	39.9	35.9	43.7	42.1	41.9	45.3	40.8	42.3	45.8	44.4	38.9	44.8	42.1
	$\boldsymbol{\theta}_{\text{Sw}}$	43.6	41.8	39.7	44.6	44.8	42.0	41.4	45.0	41.1	42.1	45.1	45.9	37.4	45.7	42.8
	$\boldsymbol{\theta}_{\text{Fr-En}}$	**44.9**	45.8	42.9	**46.2**	44.9	43.0	**43.5**	45.4	42.5	42.8	**47.7**	47.2	39.7	**47.2**	44.5
	$\boldsymbol{\theta}_{\text{Es-En}}$	44.3	45.9	42.5	45.6	44.7	43.2	43.3	**45.9**	42.8	42.5	47.5	47.0	36.3	46.8	44.1
	$\boldsymbol{\theta}_{\text{Zh-En}}$	44.1	45.9	43.0	45.9	**45.0**	**43.6**	42.5	45.5	42.2	**43.0**	47.5	47.7	39.8	46.7	44.4
	$\boldsymbol{\theta}_{\text{Hi-En}}$	43.7	42.8	40.3	45.2	44.7	42.5	42.7	45.5	41.0	42.2	45.5	46.2	37.2	46.0	43.3
	$\boldsymbol{\theta}_{\text{Sw-En}}$	43.8	44.5	40.3	46.1	43.7	41.7	42.0	45.6	41.7	42.0	45.4	46.4	38.9	46.1	43.4
	$\boldsymbol{\theta}_{\text{En-Fr}}$	**44.9**	**46.8**	**43.2**	45.9	44.6	42.7	43.4	45.7	42.5	42.9	47.3	47.6	**40.0**	46.8	**44.6**
	$\boldsymbol{\theta}_{\text{En-Es}}$	44.8	45.7	43.1	46.0	44.6	43.0	43.4	45.6	**43.1**	42.6	47.6	47.6	36.9	47.1	44.4
	$\boldsymbol{\theta}_{\text{En-Zh}}$	44.2	45.5	42.5	46.1	44.7	43.4	42.4	45.2	41.8	42.9	46.5	**47.8**	38.8	45.7	44.1
mGPT	$\boldsymbol{\theta}$	39.2	39.7	36.0	41.0	38.9	39.2	34.0	43.0	39.9	39.9	42.2	42.3	39.4	**41.8**	39.7
	LRP2	-	-	-	-	-	-	-	-	-	-	-	34.2	33.1	34.1	-
	$\boldsymbol{\theta}_{\text{En}}$	38.5	38.0	35.7	40.4	39.6	37.7	**35.8**	41.4	39.9	39.3	40.3	40.7	37.7	39.4	38.8
	$\boldsymbol{\theta}_{\text{Fr}}$	38.4	39.3	36.0	42.0	39.5	38.7	34.3	42.5	39.8	**40.2**	42.5	42.0	39.0	40.3	39.6
	$\boldsymbol{\theta}_{\text{Es}}$	38.5	39.0	35.8	41.1	39.3	39.3	35.4	42.4	39.8	39.7	42.6	42.1	39.5	40.6	39.6
	$\boldsymbol{\theta}_{\text{Zh}}$	38.7	38.7	36.4	40.6	39.8	37.9	34.2	42.5	39.5	40.0	41.5	41.8	38.9	39.9	39.3
	$\boldsymbol{\theta}_{\text{Hi}}$	39.0	39.0	34.7	40.8	40.2	38.3	34.3	41.6	38.7	40.6	42.8	42.1	38.8	40.3	39.3
	$\boldsymbol{\theta}_{\text{Sw}}$	38.6	38.1	34.4	40.4	39.5	38.2	34.0	42.1	39.4	40.2	43.2	41.7	39.2	40.7	39.2
	$\boldsymbol{\theta}_{\text{Fr-En}}$	39.3	39.5	36.3	41.9	40.2	39.5	34.7	42.6	**40.2**	40.0	42.8	42.1	39.7	41.7	40.0
	$\boldsymbol{\theta}_{\text{Es-En}}$	**40.6**	**40.3**	**36.6**	**42.6**	39.5	**39.8**	35.2	42.9	40.1	39.9	**43.5**	**42.5**	**40.4**	41.2	**40.3**
	$\boldsymbol{\theta}_{\text{Zh-En}}$	39.1	39.9	36.1	41.5	39.8	39.0	34.4	**43.4**	40.1	40.2	43.2	42.2	39.9	41.4	40.0
	$\boldsymbol{\theta}_{\text{Hi-En}}$	39.1	39.5	34.9	41.1	**40.3**	38.9	34.5	42.1	40.0	40.5	42.6	42.1	38.9	41.6	39.7
	$\boldsymbol{\theta}_{\text{Sw-En}}$	38.7	39.4	34.9	40.1	40.1	38.8	34.5	42.3	39.8	**40.7**	43.6	42.5	39.1	**41.9**	39.7
	$\boldsymbol{\theta}_{\text{En-Fr}}$	39.6	39.3	36.1	42.0	39.1	39.4	34.5	42.8	39.5	39.6	42.6	42.4	40.2	40.2	39.8
	$\boldsymbol{\theta}_{\text{En-Es}}$	39.5	39.7	36.4	42.0	39.2	39.1	35.4	42.4	**40.4**	40.2	42.6	42.2	39.7	40.6	40.0
	$\boldsymbol{\theta}_{\text{En-Zh}}$	39.4	39.4	36.2	42.1	39.7	39.4	34.5	42.6	39.3	40.2	42.3	42.4	38.8	40.5	39.8
BLOOM	$\boldsymbol{\theta}$	46.7	40.4	41.9	38.6	44.9	40.9	36.8	36.2	35.9	41.4	42.9	45.0	41.1	45.4	41.2
	LRP2	-	-	-	-	-	-	-	-	-	-	-	**46.0**	**44.3**	**46.8**	-
	$\boldsymbol{\theta}_{\text{En}}$	46.4	40.3	42.5	39.1	45.2	41.2	36.8	35.9	35.6	41.4	40.9	44.5	41.2	43.8	41.0
	$\boldsymbol{\theta}_{\text{Fr}}$	46.4	40.5	42.2	39.0	44.9	40.9	37.1	35.5	35.1	41.2	40.9	44.5	41.5	44.1	40.9
	$\boldsymbol{\theta}_{\text{Es}}$	46.5	40.7	42.0	39.5	45.3	41.1	36.9	36	35.3	40.8	40.8	44.3	41.2	43.6	41.0
	$\boldsymbol{\theta}_{\text{Zh}}$	46.5	40.5	42.3	39.2	45.2	40.8	36.9	35.8	35.6	41.1	40.8	44.4	40.8	44.7	41.0
	$\boldsymbol{\theta}_{\text{Hi}}$	46.3	40.3	40.2	40.4	44.5	41.5	35.5	35.2	35.5	41.3	41.4	43.6	40.6	43.1	40.6
	$\boldsymbol{\theta}_{\text{Sw}}$	46.8	40.2	40.6	40.5	43.7	41.8	35.5	35.4	**36.1**	41.0	42.3	44.0	40.8	44.0	40.9
	$\boldsymbol{\theta}_{\text{Fr-En}}$	47.0	40.3	42.2	39.5	45.9	41.3	36.9	36.5	35.6	41.1	41.9	44.9	41.2	44.6	41.3
	$\boldsymbol{\theta}_{\text{Es-En}}$	47.0	40.6	42.2	38.9	45.5	41.1	37.1	36.6	35.7	40.9	41.9	44.5	41.3	44.4	41.2
	$\boldsymbol{\theta}_{\text{Zh-En}}$	46.7	40.3	41.9	39.2	45.3	41.1	37.0	36.7	35.7	40.4	41.6	44.2	40.5	45.2	41.1
	$\boldsymbol{\theta}_{\text{Hi-En}}$	47.2	40.7	40.6	**41.1**	45.2	41.6	36.0	37.0	36.2	41.4	42.1	45.4	42.0	43.7	41.4
	$\boldsymbol{\theta}_{\text{Sw-En}}$	46.9	40.3	40.6	40.3	43.8	41.3	35.7	36.2	**36.1**	41.5	43.0	44.4	40.8	44.1	41.0
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{En}}$	45.9	40.6	41.8	40.1	45.1	41.1	37.0	36.2	35.9	41.3	41.6	45.0	41.1	44.7	41.2
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr}}$	46.9	40.1	41.7	39.5	45.8	40.8	36.8	35.8	35.2	40.9	40.8	44.7	41.2	45.2	41.1
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Es}}$	47.0	40.1	41.5	39.7	46.4	41.0	36.5	35.8	35.6	41.1	41.0	44.4	42.1	45.1	41.2
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh}}$	46.8	40.4	41.5	39.5	45.2	40.9	36.6	35.5	35.3	41.2	42.3	45.3	41.8	45.9	41.3
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Hi}}$	47.0	40.0	40.5	40.5	44.3	41.0	35.5	35.6	35.5	35.8	40.9	42.1	41.9	43.1	40.2
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Sw}}$	46.5	39.5	38.6	40.1	45.1	41.2	35.7	35.6	35.9	**42.5**	42.0	42.9	39.4	41.6	40.4
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr-En}}$	46.9	40.6	42.0	40.0	45.8	**42.3**	**37.4**	36.6	35.6	41.9	41.3	45.1	41.6	45.2	41.6
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Es-En}}$	47.1	**40.8**	42.2	39.9	**46.5**	41.3	37.2	36.3	35.5	41.2	41.1	45.1	42.2	45.6	41.6
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh-En}}$	**47.2**	40.5	42.2	40.0	45.8	41.2	37.1	36.0	35.8	41.2	42.1	45.8	42.4	46.3	**41.7**
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Hi-En}}$	47.3	40.3	41.2	40.3	45.4	41.6	36.2	36.8	**36.1**	41.3	42.2	45.2	42.1	45.3	41.5
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Sw-En}}$	46.8	41.1	42.5	39.5	45.5	41.2	36.9	**36.6**	35.5	**42.5**	**43.3**	45.2	41.3	45.0	41.6
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{En-Es}}$	46.9	40.3	42.3	39.8	46.3	41.0	37.1	36.2	35.4	41.3	41.2	44.9	41.9	45.2	41.4
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{En-Zh}}$	46.2	**40.8**	**42.6**	39.9	45.9	41.3	36.9	36.1	35.8	41.0	42.3	45.7	42.0	45.5	41.6

Table 9: Accuracy scores on the XNLI task. The highest scores in each model are indicated in bold.

Model	Weight	De	Ja	Fr	Es	Zh	AVg.
XGLM-7.5B	$\boldsymbol{\theta}$	76.3	73.0	73.6	74.7	74.0	74.3
	$\boldsymbol{\theta}_{\text{Fr-En}}$	77.7†	**74.2**†	74.8†	75.7†	76.6†	**75.8**
	$\boldsymbol{\theta}_{\text{Es-En}}$	**77.8**†	73.7†	75.1†	75.6†	76.1†	75.6
	$\boldsymbol{\theta}_{\text{Zh-En}}$	77.0†	73.9†	75.2†	75.8†	76.4†	75.6
	$\boldsymbol{\theta}_{\text{Hi-En}}$	76.2	73.1	**75.5**†	74.3	**76.9**†	75.2
	$\boldsymbol{\theta}_{\text{Sw-En}}$	76.3	73.4†	74.9†	**75.9**†	75.6†	75.2

Table 10: Accuracy scores on the MARC task.

Model	Weight	Ar	Bg	De	El	Hi	Ru	Sw	Th	Tr	Ur	Vi	Fr	Es	Zh	Avg.
XGLM	$\boldsymbol{\theta}$	0.790	0.786	0.441	0.775	0.520	0.811	0.824	0.784	0.770	0.751	0.791	0.820	0.792	0.723	0.741
	$\boldsymbol{\theta}_{\text{Fr}}$	0.799	0.800	0.449	0.779	0.580	0.81	0.816	0.79	0.784	0.744	0.788	0.806	0.794	0.688	0.744
	$\boldsymbol{\theta}_{\text{Es}}$	0.804	0.795	0.464	0.79	0.626	0.811	0.809	0.793	0.776	0.764	0.79	0.811	0.769	0.683	0.748
	$\boldsymbol{\theta}_{\text{Zh}}$	0.801	0.785	0.418	0.785	0.601	0.814	0.826	0.802	0.786	0.770	0.793	0.819	0.796	0.694	0.749
	$\boldsymbol{\theta}_{\text{Fr-En}}$	0.795	0.799	0.492	0.777	0.583	0.814	0.814	0.788	0.761	0.740	0.783	0.801	0.795	0.682	0.744
	$\boldsymbol{\theta}_{\text{Es-En}}$	0.801	0.804	0.494	0.789	0.604	0.818	0.816	0.795	0.773	0.757	0.790	0.805	0.774	0.674	0.749
	$\boldsymbol{\theta}_{\text{Zh-En}}$	0.797	0.803	0.481	0.785	0.585	0.823	0.825	0.802	0.781	0.755	0.795	0.804	0.792	0.681	0.751
mGPT	$\boldsymbol{\theta}$	0.535	0.556	0.603	0.554	0.465	0.520	0.525	0.512	0.569	0.462	0.546	0.569	0.586	0.520	0.537
	$\boldsymbol{\theta}_{\text{Fr}}$	0.571	0.601	0.477	0.605	0.496	0.614	0.591	0.564	0.609	0.550	0.662	0.678	0.673	0.591	0.591
	$\boldsymbol{\theta}_{\text{Es}}$	0.589	0.600	0.485	0.602	0.493	0.605	0.573	0.565	0.599	0.557	0.653	0.659	0.658	0.581	0.587
	$\boldsymbol{\theta}_{\text{Zh}}$	0.595	0.621	0.461	0.626	0.481	0.627	0.581	0.561	0.626	0.556	0.670	0.688	0.698	0.590	0.598
	$\boldsymbol{\theta}_{\text{Fr-En}}$	0.573	0.593	0.464	0.610	0.502	0.624	0.579	0.581	0.606	0.549	0.662	0.673	0.689	0.603	0.593
	$\boldsymbol{\theta}_{\text{Es-En}}$	0.589	0.601	0.480	0.608	0.492	0.621	0.583	0.567	0.620	0.550	0.666	0.676	0.682	0.580	0.593
	$\boldsymbol{\theta}_{\text{Zh-En}}$	0.600	0.625	0.468	0.620	0.488	0.628	0.585	0.581	0.627	0.566	0.671	0.689	0.700	0.604	0.603
BLOOM	$\boldsymbol{\theta}$	0.618	0.514	0.546	0.511	0.576	0.520	0.548	0.487	0.531	0.548	0.597	0.636	0.631	0.595	0.561
	$\boldsymbol{\theta}_{\text{Fr}}$	0.613	0.525	0.560	0.527	0.572	0.534	0.546	0.494	0.533	0.544	0.595	0.644	0.630	0.596	0.565
	$\boldsymbol{\theta}_{\text{Es}}$	0.612	0.526	0.560	0.533	0.571	0.532	0.544	0.494	0.534	0.539	0.594	0.643	0.635	0.601	0.566
	$\boldsymbol{\theta}_{\text{Zh}}$	0.616	0.522	0.559	0.527	0.576	0.530	0.548	0.494	0.532	0.541	0.596	0.645	0.632	0.597	0.565
	$\boldsymbol{\theta}_{\text{Fr-En}}$	0.615	0.535	0.564	0.535	0.573	0.535	0.540	0.505	0.530	0.546	0.598	0.640	0.634	0.590	0.567
	$\boldsymbol{\theta}_{\text{Es-En}}$	0.617	0.534	0.567	0.535	0.575	0.538	0.548	0.500	0.533	0.542	0.599	0.642	0.638	0.596	0.569
	$\boldsymbol{\theta}_{\text{Zh-En}}$	0.617	0.533	0.564	0.535	0.574	0.536	0.544	0.500	0.534	0.543	0.598	0.640	0.634	0.597	0.568
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr}}$	0.614	0.539	0.562	0.532	0.570	0.534	0.547	0.507	0.539	0.547	0.607	0.645	0.639	0.592	0.569
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Es}}$	0.618	0.537	0.564	0.543	0.568	0.534	0.555	0.513	0.549	0.546	0.616	0.650	0.651	0.594	0.574
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh}}$	0.617	0.536	0.558	0.535	0.569	0.538	0.540	0.504	0.535	0.550	0.605	0.640	0.634	0.592	0.568
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Fr-En}}$	0.624	0.527	0.560	0.530	0.570	0.530	0.554	0.501	0.540	0.540	0.607	0.642	0.642	0.604	0.569
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Es-En}}$	0.626	0.548	0.564	0.544	0.568	0.534	0.557	0.516	0.542	0.546	0.610	0.651	0.655	0.605	0.576
	$\boldsymbol{\theta}^{\text{Prog}}_{\text{Zh-En}}$	0.621	0.533	0.563	0.535	0.568	0.538	0.545	0.503	0.534	0.548	0.604	0.639	0.638	0.598	0.569

Table 11: RankC scores between English and each language.