Revealing the Parallel Multilingual Learning within Large Language Models

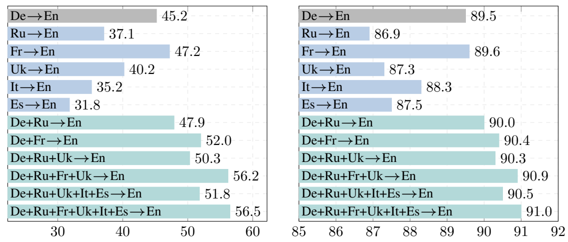
Anonymous ACL submission

Abstract

Large language models (LLMs) can handle multilingual and cross-lingual text within a single input; however, previous works leveraging multilingualism in LLMs primarily focus on using English as the pivot language to enhance language understanding and reasoning. Given that multiple languages are a compensation for the losses caused by a single language's limitations, it's a natural next step to enrich the model's learning context through the integration of the original input with its multiple translations. In this paper, we start by revealing that LLMs learn from **P**arallel **M**ultilingual **I**nput (**PMI**). Our comprehensive evaluation shows that PMI enhances the model's comprehension of the input, achieving superior performance than conventional in-context learning (ICL). Furthermore, to explore how multilingual processing affects prediction, we examine the activated neurons in LLMs. Surprisingly, involving more languages in the input activates fewer neurons, leading to more focused and effective neural activation patterns. Also, this neural reaction coincidently mirrors the neuroscience insight about synaptic pruning, highlighting a similarity between artificial and biological 'brains'.

1 Introduction

Many of the recent large language models (LLMs) are multilingual. Unlike language-specific NLP systems, such as machine translation systems specialized to a given language pair, these models are generally trained on large-scale multilingual datasets, using a unified vocabulary. Because of this training approach, it is possible to learn a universal representation of texts across different languages. Therefore, the resulting models can be directly applied to a variety of multilingual and cross-lingual tasks. For example, most commercialized LLMs can respond to user queries in different languages, without needing to specify what languages are used. More recently, the multilingual capabilities of these models have been shown to help cross-lingual in-context learning (ICL). By providing simple prompts involving cross-lingual thinking and reasoning, LLMs can understand and generate text in languages that were less represented in the training data (Qin et al., 2023; Huang et al., 2023; Zhang et al., 2023; Nguyen et al., 2023).



(a) BLEU (b) COMET

Figure 1: Comparing the effectiveness of our PMI versus direct and pivot translation on the Qwen-14B model and the FLORES-200 dataset. We also provide the results of ChatGPT in Table 1.

Despite the apparent usefulness of multilingualism in LLMs, previous work has primarily focused on using English as the pivot language in language understanding and reasoning. It is a natural next step to incorporate more languages and investigate how these languages are simultaneously processed in LLMs. In this paper we explore methods that make use of parallel multilingual input (PMI) in ICL and explain how neurons are activated in this processing. There are two major findings.

- LLMs can benefit from receiving parallel input in multiple languages. By transforming single-language input into multi-language input, we build a multi-source LLM that uses contexts from all these languages to make predictions. On the FLORES-200 machine translation benchmark, it achieves improvements of 11.3 BLEU points and 1.52 COMET points over the baseline.

- Somewhat surprisingly, as more languages are involved in the input, fewer neurons are activated in the LLMs, facilitating more targeted and effective neuron activation patterns. This result links multilingual representation learning to *synaptic pruning* in neuroscience (Huttenlocher et al., 1979; Huttenlocher, 1990): as a brain develops, some neural connections are strengthened, while others are deemed redundant and eliminated, making the transmission of neural signals more efficient.

More specifically, we find that in addition to the performance improvements from incorporating more languages, LLMs can gain advantages from extensive languages even involving ones that do not surpass baseline performances. With the help of high-quality machine translation, we efficiently acquire abundant parallel input, enabling us to apply this method to various tasks. Experimental results across 7 datasets, 7 languages, and 9 LLMs further demonstrate the effectiveness and applicability of PMI.

Since previous neuron activation statistics are primarily designed for the vanilla transformer model (Zhang et al., 2022; Li et al., 2023), we have extended these methods to analyze more advanced LLM architectures. When LLMs receive PMI, we observe simultaneous performance improvements and neuron inhibition. In addition, PMI selectively activates only a small portion of the most commonly used neurons while inhibiting the rest. Further analysis reveals that few-shot learning produces a similar effect on neuron activation, and integrating it with PMI enhances this neural reaction. These findings are consistently sustained across different models and tasks.

We introduce our PMI and evaluate it with human translation in Section 2.1. Subsequently, we comprehensively analyze the performance gains brought by PMI in Section 2.2 and explain its effectiveness from a view of neuron activation in Section 3. Moreover, we apply PMI to various tasks under real scenario setups in Section 4.

2 Parallel Multilingual Input

2.1 LLMs benefit from PMI

Given an input $\mathbf{X}$ of a task and a template $f(\cdot)$ to transform the input to an instruction, the conventional ICL can be expressed as follows:

$$\mathbf{Y} \;=\; \operatorname{argmax} P(y_t|f(\mathbf{X})) \qquad (1)$$

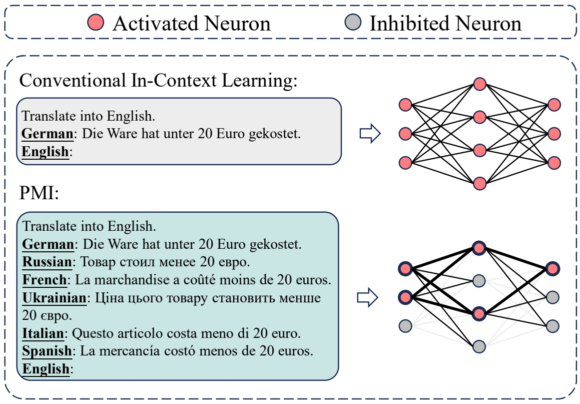


Figure 2: Compared to conventional ICL, PMI inhibits neurons and promotes more precise activation (i.e., the thickened line). Other prompts are shown in Table 20.

where $\mathbf{Y}$ denotes the target output of the task and y_t denotes the token generated at moment t. PMI extends beyond the conventional ICL approach of feeding LLMs solely with inputs in one language. Instead, it encompasses providing input in multiple languages, translated by professional human translators or sophisticated machine translation (MT) systems. The PMI can be shown as:

$$\mathbf{Y} \;=\; \operatorname{argmax} P(y_t|f(\mathbf{M}, \mathbf{X})) \qquad (2)$$

where $\mathbf{M} = \{m_1, m_2, ..., m_k\}$ is a parallel language set containing k translations of the input. The template $f(\cdot)$ we used is neutral for both the input $\mathbf{X}$ and its translations $\mathbf{M}$, making LLMs cannot distinguish them. Figure 2 shows the difference between the conventional ICL and our PMI when translating De $\rightarrow$ En.

Three aspects should be considered when constructing a PMI prompt, including the choice of languages, the choice of translators, and the display order of languages. As shown in Appendix D.1, our preliminary experiments suggest that: (1) choosing the language that LLMs understand better is crucial; (2) higher translation quality can lead to larger improvements; (3) it is preferable to place languages better understood at head and tail of the input sequence.

Experimental Settings. We conducted translation experiments on the FLORES-200 which allowed us to probe the upper bound of the performance by constructing PMI using human-translated parallel sentences. Direct and pivot translation were our baselines. We utilized two powerful multilingual LLMs, including ChatGPT (`gpt-3.5-turbo-0613`) and Qwen-14B

Method	Input	ChatGPT		Qwen-14B	
		BLEU	COMET	BLEU	COMET
	German → English				
Direct	De	44.3	89.8	45.2	89.5
Pivot	Fr	45.6	89.6	47.2	89.6
	Ru	35.2	87.0	37.1	86.9
PMI-1	De + Ru	46.2	90.0	47.9	90.0
PMI-3	De + Ru + Fr + Uk	49.2	90.4	56.2	90.9
PMI-5	De + Ru + Fr + Uk + It + Es	**50.2**	**90.6**	**56.5**	**91.0**
	English → German				
Direct	En	**40.5**	88.8	**35.0**	87.2
Pivot	Fr	30.4	86.5	25.9	84.7
	Ru	25.8	85.2	22.6	83.4
PMI-1	En + Ru	40.1	88.8	34.4	87.2
PMI-3	En + Ru + Fr + Uk	40.3	88.8	34.8	87.4
PMI-5	En + Ru + Fr + Uk + It + Es	**40.5**	**88.9**	34.6	**87.5**
	German → French				
Direct	De	37.2	86.2	35.2	85.3
Pivot	Ro	39.6	87.4	37.2	86.2
	Ru	29.5	84.0	30.7	83.6
PMI-1	De + Ru	39.3	86.7	36.6	85.7
PMI-3	De + Ru + Ro + Uk	41.4	87.1	40.7	86.5
PMI-5	De + Ru + Ro + Uk + It + Es	**42.4**	**87.3**	**42.3**	**86.9**

Table 1: Experiments of PMI, direct and pivot translation on the FLORES-200. We provide k parallel languages denoted as PMI-k. Pivot row reports the best performance among all pivot translations in the first line and the performance of Russian in the second line.

(`Qwen-14B-Chat`) (Bai et al., 2023) [1]. ChatGPT was prompted with one-shot for baseline and PMI prompts. While Qwen-14B exhibited confusion when processing PMI prompts, so we made some instruction training data of PMI and baseline prompts, and employed the LoRA technique (Hu et al., 2022) to fine-tune Qwen-14B. More details can be found in Appendix E. The translation performance was evaluated in terms of SacreBLEU (Post, 2018) and COMET-22 (`wmt22-comet-da`) (Rei et al., 2022).

Results and Analyses. Table 1 delineates the performance of direct translation (Direct), pivot translation (Pivot), and PMI in three translation directions. We see, first of all, PMI achieves the best result among all the baselines especially when more parallel languages are used. Despite that the COMET score of some baselines reaches as high as 90, PMI still beats both direct and pivot translation with significant improvements. Furthermore, we find that PMI even benefits from parallel languages which perform worse than direct translation. For example, integrating Russian into PMI achieves better performance than the baseline. Besides, when English becomes the original input, PMI leads to a small performance increase. We attribute this to the fact that LLMs have shown great success in understanding English input, leaving little room for improvement.

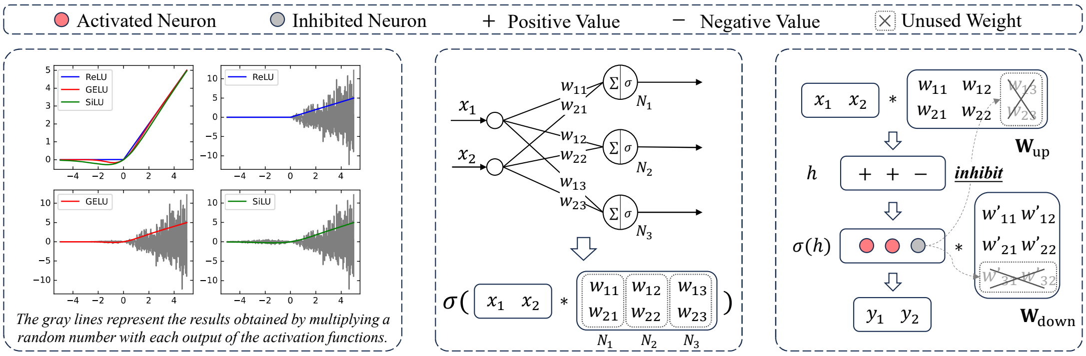
[1]We also tried Bloomz (Muennighoff et al., 2023), however, compared to the performance on WMT, it showed deviant high performance on FLORES-200 indicating a data leakage, which is also reported by Zhu et al. (2023).

2.2 Multiple Languages or Information Sources?

Due to the parallel languages being translated by numerous human experts in the above experiments, one may argue that the improvement of PMI results from multiple information sources rather than languages. Specifically, multiple information sources can bring different perspectives of the original input, and translating inputs derived from human experts is like doing ensemble learning based on various strong translation systems. To separately quantify the effects of multiple languages and information sources, we decompose the PMI based on the human translations (PMI$_{GT}$) into three prompting strategies:

- **Mono-source and monolingual:** The original input is paraphrased into different versions without changing the semantics. We denote this prompt as PMI$_{PA}$.
- **Multi-source but monolingual:** The human translation texts used in PMI are translated into the language of the original input by one translator. This prompt integrates different information sources but expresses in one language, e.g., we provide "De + De (Ru) + De (Fr) + De (Uk) + De (It) + De (Es)" to LLMs where the language in parentheses represents the human translation text. We call it PMI$_{MS}$.
- **Multilingual but mono-source:** The original input is translated into different parallel languages by one translator. The source of this prompt is only the original input whereas the expression holds a multilingual form, like "De + Ru (De) + Fr (De) + Uk (De) + It (De) + Es (De)", which is represented by PMI$_{ML}$. We also illustrate these prompts in Figure 6.

Experimental Settings. With access to Qwen-14B, ChatGPT and GPT-4 (`gpt-4-0613`), we conducted experiments on two translation directions of FLORES-200. The translation system used by both PMI$_{MS}$ and PMI$_{ML}$ prompt was the NLLB-54B model (Costa-jussà et al., 2022). We derived the paraphrased sentences by requesting ChatGPT. Notably, Qwen-14B used in this experiment is different from the one in the previous experiment, as we have to fine-tune Qwen-14B with extra training data based on the PMI$_{MS}$ prompt for fairness.



(a) (b) (c)

Figure 3: The impact of ReLU-like activation functions on neurons during the forward process of transformer models. Figure (a) shows that activation function $\sigma(\cdot)$ like ReLU and some of its variants, when encountering negative inputs, saturate to zero and weaken the values multiplied by their outputs. Figure (b) details the equivalence between artificial neurons and the linear-transform matrix of MLPs. Figure (c) illustrates that ReLU-like activation functions inhibit neurons in $\mathbf{W}_{\text{up}}$ and some weights of $\mathbf{W}_{\text{down}}$ when the input is negative.

System		BLEU	COMET	BLEU	COMET
Direction		*De → En*		*De → Fr*	
ChatGPT	Direct	**44.3**	**89.8**	37.2	86.2
	PMI_{PA}	36.4$^{\downarrow 7.9}$	88.6$^{\downarrow 1.1}$	34.8$^{\downarrow 2.4}$	85.5$^{\downarrow 0.7}$
	PMI_{MS}	42.6$^{\downarrow 1.7}$	89.4$^{\downarrow 0.3}$	37.1$^{\downarrow 0.1}$	86.0$^{\downarrow 0.2}$
	PMI_{ML}	44.1$^{\downarrow 0.2}$	89.7$^{\downarrow 0.1}$	**39.7**$^{\uparrow 2.5}$	**86.6**$^{\uparrow 0.4}$
	PMI_{GT}	*50.2*	*90.6*	*42.4*	*87.3*
Qwen-14b	Direct	45.5	89.6	35.4	85.4
	PMI_{PA}	40.4$^{\downarrow 5.1}$	89.0$^{\downarrow 0.6}$	31.8$^{\downarrow 3.6}$	84.6$^{\downarrow 0.8}$
	PMI_{MS}	**46.6**$^{\uparrow 1.1}$	**90.0**$^{\uparrow 0.4}$	36.5$^{\uparrow 1.1}$	**86.1**$^{\uparrow 0.7}$
	PMI_{ML}	44.9$^{\downarrow 0.6}$	89.6$^{\uparrow 0.0}$	**37.6**$^{\uparrow 2.2}$	86.0$^{\uparrow 0.6}$
	PMI_{GT}	*56.3*	*91.1*	*42.8*	*87.0*
GPT-4	Direct	44.9	**89.9**	39.0	86.5
	PMI_{MS}	43.6$^{\downarrow 1.3}$	89.8$^{\downarrow 0.1}$	39.6$^{\uparrow 0.6}$	**87.0**$^{\uparrow 0.5}$
	PMI_{ML}	**45.4**$^{\uparrow 0.5}$	89.7$^{\downarrow 0.1}$	**40.1**$^{\uparrow 1.1}$	86.8$^{\uparrow 0.2}$
	PMI_{GT}	*52.9*	*90.9*	*45.9*	*88.1*

Table 2: The ablation study of the mono-source and monolingual (PMI_{PA}), multi-source but monolingual (PMI_{MS}), multilingual but mono-source (PMI_{ML}), multi-source and multilingual (PMI_{GT}) prompts on the FLORES-200. The best results are in bold among all the prompts except for PMI_{GT}.

Results and Analyses. From Table 2, we can see that both PMI_{MS} and PMI_{ML} prompt achieve improvement most of the time, while none of them can reach the same performance as the PMI_{GT} prompt. In addition, the PMI_{ML} prompt far outperforms the PMI_{PA} prompt, which demonstrates that multilingual input helps LLMs again. Also, we see that despite the similar baseline performance, GPT-4 always outperforms ChatGPT significantly when being armed with PMI, suggesting that stronger LLMs benefit more from the PMI.

3 PMI Can Help: From a View of Neuron Activation

Although LLMs benefit from PMI, there is still no idea about how this mechanism works. Considering that knowledge is memorized in different neurons in transformer models (Dai et al., 2022), hence a straightforward hypothesis is that giving the input in multiple languages may increase the number of activated neurons in the inference process. To quantify how many neurons in transformer models are activated during inference, some works propose to make statistics of the nonzero values in the intermediate output of multi-layer perceptrons (MLPs) after a ReLU activation function (Zhang et al., 2022; Li et al., 2023). This is based on the idea that, in matrix multiplication, zero can be omitted; therefore, neurons that output zero are considered inhibited while others are activated. Next, we will explain this statistical method.

3.1 Method of Counting Activated Neurons

ReLU controls the life and death of neurons. In transformer models, the activation function $\sigma(\cdot)$ lays in the middle of the two-layer MLPs, like this:

$$\mathbf{Y} = \sigma\left(\mathbf{X}\mathbf{W}_{\text{up}}\right)\mathbf{W}_{\text{down}} \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ stand for input and output, respectively. $\mathbf{W}_{\text{up}}$ and $\mathbf{W}_{\text{down}}$ represent different MLP layers containing artificial neurons. The vanilla transformer uses ReLU as the activation function (Vaswani et al., 2017), i.e., $\max(x, 0)$. In Figure 3 (b) and (c), ReLU outputs zero value means two aspects: the neuron in $\mathbf{W}_{\text{up}}$ is inhibited and stripped from the whole neural network; the weight in $\mathbf{W}_{\text{down}}$ that accepts the zero value is inhibited.

Counting activated neurons in MLPs with ReLU variants. Despite the success of ReLU, recent works find that making a ReLU-like non-linearity

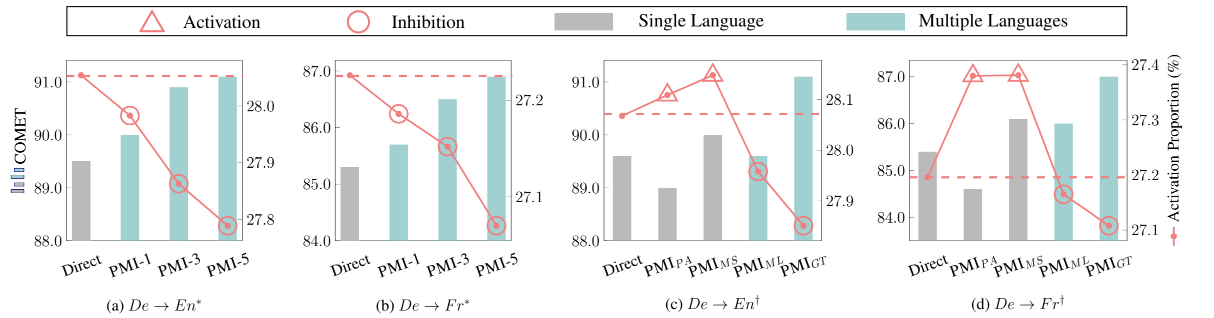


Figure 4: The COMET score and the activation proportion of Qwen-14B armed with different prompts on FLORES-200. Notably, whether a method inhibits or activates neurons depends on its activation proportion being below or above the baseline level. Thus, a point on the curves suggests inhibition ◯ if it falls below the first point, and activation △ if it exceeds the first point. ∗ and † indicates the model used in Section 2.1 and 2.2, respectively.

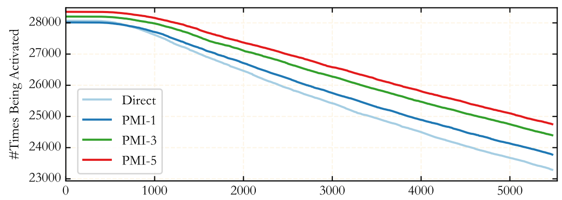


Figure 5: The distribution of the top 1% of activated neurons in Qwen-14B on FLORES-200 De → En. The horizontal axis represents different neurons arranged in descending order of the number of times being activated.

to output negative values can increase training speed (Clevert et al., 2016; Hendrycks and Gimpel, 2016). Hence, as shown in Table 8, these variants of ReLU become popular among LLMs. We draw ReLU, GELU and SiLU in Figure 3 (a). We see despite both GELU and SiLU performing as smooth ReLU, they remain the basic character, i.e., saturating to zero at negative input values and protecting positive input values. In other words, these ReLU variants significantly reduce the absolute value of any negative input to a level that is close to or equal to zero. As a result, some neurons and weights are inhibited as before. This motivates us to make statistics of activated neurons in MLPs with ReLU variants by *counting the output values of the activation function that are greater than zero*.

Other works combine GELU and SiLU with the gated linear units (Shazeer, 2020) like this:

$$\mathbf{Y} = \left(\sigma\left(\mathbf{X}\mathbf{W}_{\text{up}}\right) \odot \left(\mathbf{X}\mathbf{V}_{\text{up}}\right)\right) \mathbf{W}_{\text{down}} \tag{4}$$

where $\odot$ is the element-wise product and a new matrix V_{up} is introduced to perform the gate. If we transform the formula into this:

$$\mathbf{Y} = \sigma\left(\mathbf{X}\mathbf{W}_{\text{up}}\right) \left(\mathbf{X}\mathbf{V}_{\text{up}} \odot {\mathbf{W}_{\text{down}}}^{\top}\right)^{\top} \tag{5}$$

then we can consider $\mathbf{X}\mathbf{V}_{\text{up}} \odot {\mathbf{W}_{\text{down}}}^{\top}$ as a whole, and both inhibiting neurons and weights happen as before. Thus, our statistical method of activated neurons remains unchanged.

3.2 Experiments and Results

Figure 4 shows performances and the proportion of activated neurons[2] on Qwen-14B models. From the results, we get the following observations:

Activated neurons are far fewer than inhibited ones. Despite performing dense computations, only a small number of neurons around 27% are activated in Qwen-14B during the inference stage, which is similar to the sparse activation phenomenon observed by Li et al. (2023). Besides, the differences in the proportion of activated neurons are small in numerical terms, we attribute this to the finding that few parameters are in charge of linguistic competence in LLMs (Zhao et al., 2023).

More languages, more inhibited neurons, more performance gain. As shown in Figure 4 (a) and (b), if we add more parallel languages in PMI, then the proportion of activated neurons becomes small meanwhile LLM yields better translations, indicating a consistent correlation between inhibiting neurons and performance improvements.

Multilingual input inhibits neurons whereas monolingual input activates neurons. Figure 4 (c) and (d) show the proportion of activated neurons caused by monolingual and multilingual input. We see that, compared to direct translation, though monolingual and multilingual input can achieve better performance, their influence on neurons is the opposite, i.e., monolingual input activates neurons whereas multilingual input inhibits neurons. More-

[2]Note that the proportion mentioned is derived by averaging the percentages of activated neurons for each token generated by an LLM across the dataset. We discuss this implementation in detail in Appendix B.

System		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Direction		*De → En*		*Zh → En*		*De → Fr*		*En → De*		*En → Zh*		*Is → En*	
Parallel Languages		*Es Ru Fr Zh Ja Cs*		*Es Ru Fr Ja Cs De*		*En Ru Es Zh It Cs*		*Es Ru Fr Zh Ja Cs*		*Es Ru Fr Ja Cs De*		*Es Ru Fr It Cs De*	
ChatGPT*	Direct	29.8	82.7	**24.7**	81.9	38.6	84.1	**34.5**	87.2	**43.8**	**87.2**	35.6	84.5
	Pivot	28.5	84.0	21.6	81.9	40.4	84.0	30.0	86.4	40.3	86.0	35.0	85.6
	PMI-1	**32.4**	85.3	24.6	**82.8**	40.9	**84.5**	34.0	87.3	41.8	86.5	38.0	86.4
	PMI-3	32.1	85.4	23.4	82.6	41.1	**84.5**	**34.5**	87.5	41.7	86.9	38.2	86.6
	PMI-6	31.6	**85.5**	18.6	82.4	**41.3**	**84.5**	**34.5**	**87.6**	41.7	86.9	**38.5**	**86.7**
LLaMA3-8B*	Direct	**30.4**	84.0	21.4	80.2	29.2	79.8	27.3	83.2	**35.8**	83.7	22.1	76.7
	Pivot	27.4	83.4	21.3	81.4	31.7	80.8	22.8	81.8	29.3	81.7	31.0	84.6
	PMI-1	30.3	85.0	23.2	82.1	33.4	81.5	26.1	83.4	32.5	82.8	34.7	85.2
	PMI-3	30.1	**85.1**	23.4	82.4	33.9	82.3	**27.4**	84.6	35.1	83.5	**36.6**	**86.0**
	PMI-6	29.9	**85.1**	**24.1**	**82.7**	**34.5**	**82.5**	27.3	**84.9**	34.1	**84.1**	36.0	85.8
Qwen-14B†	Direct	30.4	84.4	23.7	80.8	34.2	81.9	29.6	85.3	**45.2**	**87.6**	18.4	69.7
	Pivot	28.2	84.0	22.4	81.8	37.4	82.7	26.9	84.7	41.2	86.3	34.1	85.4
	PMI-1	31.3	84.8	**24.3**	**82.0**	38.0	83.1	29.7	85.4	45.1	**87.6**	35.6	85.1
	PMI-3	**31.6**	**84.9**	23.5	**82.0**	37.7	**83.4**	**30.0**	**85.8**	44.9	**87.6**	37.2	85.6
	PMI-6	31.0	**84.9**	22.0	81.3	**38.4**	**83.4**	29.9	85.5	**45.2**	**87.6**	**37.9**	**85.7**
ALMA-13B†	Direct	28.1	83.8	21.6	79.6	27.1	79.2	29.6	85.5	36.9	85.8	34.0	85.8
	Pivot	26.0	83.3	21.7	81.2	29.9	80.3	26.4	84.8	32.3	84.6	32.7	85.2
	PMI-1	29.9	84.6	**23.8**	**81.8**	31.1	80.8	29.7	85.3	36.9	85.9	37.0	86.3
	PMI-3	**30.8**	**85.0**	22.9	**81.8**	**33.3**	**81.5**	**29.9**	**86.0**	36.9	**86.0**	**38.3**	**86.5**
	PMI-6	30.0	84.9	18.1	79.5	**33.3**	**81.5**	**29.9**	85.9	**37.2**	**86.0**	38.2	86.3
mT0-13B*	Direct	25.1	82.2	13.7	76.2	27.9	78.5	**17.6**	77.3	26.0	83.1	29.9	83.9
	Pivot	24.5	82.5	19.3	**80.7**	30.5	80.0	17.4	**78.5**	23.8	82.1	30.8	84.6
	PMI-1	27.0	83.4	18.3	79.9	29.9	79.4	17.4	76.5	25.5	82.4	33.0	84.9
	PMI-3	**27.6**	**83.5**	**19.6**	**80.7**	**32.4**	**80.4**	16.0	74.4	27.5	82.9	33.8	**85.4**
	PMI-6	26.8	83.3	19.5	80.5	32.2	**80.4**	15.5	74.5	**28.5**	**83.3**	**33.9**	85.3
Bloomz-176B*	Direct	24.0	78.4	16.0	76.4	27.3	77.1	13.0	70.7	29.5	83.9	5.6	53.8
	Pivot	25.0	82.8	20.8	81.3	34.6	82.1	9.5	66.2	27.6	82.6	31.5	**84.6**
	PMI-1	25.4	80.7	17.3	77.6	33.1	80.4	11.9	70.0	28.0	82.4	23.5	75.8
	PMI-3	28.2	**83.9**	21.1	81.2	**35.7**	**82.2**	**16.0**	**73.9**	31.7	83.8	31.8	83.7
	PMI-6	**28.3**	83.8	**21.7**	**81.4**	36.6	82.9	15.0	73.5	**32.4**	**84.7**	**34.0**	84.2

Table 3: Experiments on the WMT dataset. Note that the pivot row displays the maximum scores among all pivot prompts, and the order of the parallel languages indicates the priority when being integrated into PMI-k prompts. † and $*$ represent the model is fine-tuned or not respectively.

over, PMI$_{GT}$ inhibits more neurons than PMI$_{ML}$ and PMI$_{MS}$ activates more neurons than PMI$_{PA}$.

PMI simulates one-off synaptic pruning. During the maturation of biological brains, synaptic pruning is a necessary process that removes less commonly used neural connections, thus making frequently-used neural pathways more powerful and efficient (Huttenlocher et al., 1979; Huttenlocher, 1990). In other words, the brain benefits from little and precise neuron activation. We show that PMI simulates one-off synaptic pruning during the inference from two aspects: (1) as demonstrated above, PMI *inhibits neurons;* (2) PMI *promotes more precise neuron activation.* Figure 5 records the activation state of the most commonly used neurons. It shows that compared to the baseline prompt, PMI promotes the activation of the top 1% of neurons commonly used. Meanwhile, other neurons rarely used are activated fewer times to achieve an overall effect of inhibition, as shown in Figure 7. This indicates that more targeted and effective neuron activation patterns—where some important neurons are activated more while others less often—could be facilitated by PMI. Notably, as more languages are used, both neuron inhibition and precise activation are enhanced, potentially leading to more remarkable synaptic pruning.

4 Wide Evaluation of PMI Without Human Translations

Next, we focus on evaluating PMI method without human translation across sentence and paragraph levels, natural language understanding (NLU), and generation (NLG) tasks.

4.1 Tasks and Evaluation

We totally evaluated PMI on five tasks. **(1) Machine Translation:** We conducted experiments on five high-resource directions of WMT22 and one low-resource direction of WMT21. **(2) Nature Language Inference:** We chose RTE (Wang et al., 2019) and three languages in XNLI (Conneau et al., 2018). The metric was accuracy. **(3) Reading Comprehension:** We did evaluation on this long sequence task using BoolQ[3] (Clark et al.,

[3]This dataset is also leaked to Bloomz-176B.

System		Accuracy				
		RTE	XNLI			BoolQ
Source Language		*En*	*Fr*	*De*	*Zh*	*En*
Parallel Languages		*Es Fr De*	*Es Ru De*	*Es Ru Fr*	*Es Fr De*	*Es*
Qwen-7B†	Direct	91.3	79.9	76.7	78.2	86.0
	Pivot	86.6	78.9	80.2	74.2	83.3
	PMI	**91.7**	**80.7**	**80.6**	**80.7**	**86.7**
Qwen-14B†	Direct	91.3	81.5	78.2	80.6	88.5
	Pivot	90.6	80.5	79.8	74.2	86.0
	PMI	**92.4**	**81.6**	**80.7**	**80.7**	**89.0**
Qwen-72B†	Direct	91.7	**86.4**	84.4	**84.6**	91.2
	Pivot	92.4	85.8	85.5	80.6	89.1
	PMI	**92.4**	**86.4**	**85.6**	**84.6**	**91.9**
ALMA-13B†	Direct	89.5	82.1	79.3	77.5	86.5
	Pivot	84.5	82.0	80.8	75.9	81.1
	PMI	**90.3**	**83.8**	**81.9**	**78.8**	**87.4**
Yi-34B†	Direct	92.1	70.0	66.8	72.0	89.6
	Pivot	85.9	71.5	72.6	68.1	86.8
	PMI	**93.1**	**73.1**	**73.7**	**72.6**	**90.2**
Bloomz-176B*	Direct	76.5	53.9	50.5	53.9	-
	Pivot	77.6	53.1	**53.3**	53.7	-
	PMI	**82.0**	**57.3**	52.5	**54.9**	-

Table 4: Experiments on NLU tasks. We apply PMI-3 across all tasks, with the exception of the reading comprehension task, for which we apply PMI-1.

System		SARI	R2 / RL	
		Wiki-Auto	XLSum	
Source Language		*En*	*Es*	*Ru*
Parallel Languages		*Es Fr De*	*Fr*	*Es*
Qwen-7B†	Direct	45.6	10.7 / 23.5	**45.4 / 41.6**
	Pivot	43.2	9.4 / 22.7	41.1 / 38.6
	PMI	**47.6**	**11.0 / 23.6**	45.3 / 41.1
Qwen-14B†	Direct	46.2	12.2 / 24.7	46.6 / 42.7
	Pivot	43.8	9.0 / 21.4	40.2 / 38.3
	PMI	**48.9**	**12.7 / 25.4**	**47.9 / 43.1**
ALMA-13B†	Direct	45.7	**12.1 / 24.8**	47.7 / 43.5
	Pivot	43.2	10.4 / 22.9	44.3 / 41.2
	PMI	**47.5**	11.5 / 24.5	**47.7 / 43.9**
Yi-34B†	Direct	45.4	11.8 / 24.6	45.4 / 41.5
	Pivot	43.5	10.6 / 23.3	41.7 / 38.8
	PMI	**47.2**	**12.0 / 24.6**	**45.5 / 41.8**

Table 5: Experiments on other NLG tasks. We employ PMI-3 and PMI-1 for the text simplification and abstractive summarization task respectively.

2019). Our metric was accuracy. **(4) Text Simplification:** We used Wiki-auto (Jiang et al., 2020), and SARI (Alva-Manchego et al., 2020) was chosen as the metric. **(5) Abstractive Summarization:** For this paragraph-level task, we mainly reported the performance on two languages in XL-Sum (Hasan et al., 2021). The metric was F1-Rouge (Lin, 2004). To streamline computation, we reconstructed our test set by randomly selecting 1000 samples from BoolQ, Wiki-auto, and XLSum, along with 3000 samples from XNLI, leaving other tasks unchanged.

4.2 Models

The experiment was conducted on 9 instruction-tuned commonly used multilingual LLMs whose parameters range from 7B to 176B, including ChatGPT, LLaMA3-8B (AI@Meta, 2024), Bloomz-176B (Muennighoff et al., 2023), Qwen-7B, -14B, -72B (Bai et al., 2023), ALMA-13B (Xu et al., 2023), Yi-34B (01-ai, 2023) and mT0-13B (Scao et al., 2022). All of them are pre-trained with multilingual corpus except for ALMA-13B which is specially fine-tuned for the MT task based on LLaMA2-13B (Touvron et al., 2023). Other details about models, training, and decoding setups can be found in Appendix E.

4.3 Baselines

Direct Prompt means that given the original input, LLMs accomplish the task directly. Here, we report the results of one-shot on ChatGPT while zero-shot on others for the best performance.

Pivot Prompt indicates that the original input is translated into a parallel language, and LLMs are fed with the translation to accomplish the task. To ensure the high-quality of translations and the reproducibility of our study, we utilized publicly and easily accessible GPT-4 to translate the source sentence of WMT and ChatGPT to translate other datasets. We display the maximum scores of pivot prompts, see Appendix F for full results.

4.4 Results and Analyses

PMI effectively pushes the boundaries across various tasks and languages. Table 3 suggests that PMI achieves superior results across 6 translation directions including high-resource and low-resource source languages. Additionally, Tables 4 and 5 show PMI's competitive edge against baselines in various tasks, irrespective of text length. Furthermore, in Table 11, we see PMI outperforms few-shot learning, especially in terms of the COMET score.

Weak model augments strong model. Table 6 shows that when we utilize parallel multilingual translations from GPT-4 to augment a stronger LLM like GPT-4o, the performance of GPT-4o+PMI surpasses two exceptional baselines, including GPT-4 and GPT-4o. It underscores the necessity of using PMI instead of relying solely on a remarkable MT system. Also, this demonstrates that PMI still yields better performance when the parallel translations come from a weak model, further validating its effectiveness and practicality.

Automatic translation triggers learning from PMI. Since the lack of high-quality human translation, all the translations used in experiments come

System		BLEU	COMET	BLEU	COMET
Direction		*De → Fr*		*Zh → En*	
GPT-4		39.0	84.3	23.2	81.6
GPT-4o	Direct	39.2	83.1	23.1	**82.4**
	PMI	**42.5**	**84.8**	**23.6**	**82.4**
Direction		*En → De*		*En → Zh*	
GPT-4		35.5	87.2	42.5	86.4
GPT-4o	Direct	**36.8**	87.5	**44.5**	87.6
	PMI	36.3	**88.0**	**45.5**	**87.7**

Table 6: Experiments of GPT-4o on WMT. We report the best performance among PMI-1, PMI-3, and PMI-6 in the PMI lines.

Qwen-14B				Bloomz-176B			
XNLI (De)		Wiki-Auto		RTE			
Direct	PMI-3	Direct	PMI-3	Direct	PMI-3	5-shot	5-shot + PMI-3
Accuracy		*SARI*		*Accuracy*			
78.2	80.7	46.2	49.0	76.5	82.0	80.1	81.2
Activation Proportion (%)				*Activation Proportion (%)*			
29.5	29.3	28.7	28.4	4.4	4.3	4.1	3.9

Table 7: The performance and activation proportion of conventional ICL and PMI on NLU and NLG tasks.

from GPT-4 or ChatGPT. We see, on the one hand, PMI powered by MT outperforms pivot prompts. Even though some pivot prompts have inferior performance than the direct prompt, integrating these languages into PMI still boosts the comprehension of LLMs. On the other hand, Figure 11 shows that PMI armed with MT achieves improvements by inhibiting neurons and promoting more precise activation.

Few-shot learning performs similarly as PMI. Table 7 and Figure 7 suggest that few-shot learning also inhibits neurons and facilitates more precise activation, and combining few-shot learning and PMI further enhances this neuron reaction.

Superiority of PMI remains when English is the original or parallel language. Despite the subtle improvements on FLORES-200 En → De in Section 2.1, results of RTE, BoolQ, and WMT De → Fr show that PMI not only achieves prime performance on English-source inputs but also outperforms all pivot prompts when we choose English as one of the parallel languages.

We discuss the fine-tuning demands of PMI in Appendix D.3, self-augmentation in Appendix D.4, and the trade-off between the inference speed and improvements in Appendix D.5.

5 Related Work

Multi-way Neural Machine Translation. Multi-way input is a successful method to enhance multilingual neural machine translation (MNMT) systems by providing the source language and its translations in different languages (Och and Ney, 2001). In the inference stage, most works rely on high-quality translations from human experts (Zoph and Knight, 2016; Firat et al., 2016; Nishimura et al., 2018; Choi et al., 2018). However, this ground truth multilingual data is scarce in reality, limiting the application of multi-way input. Different from multi-way MNMT, we find that LLMs benefit from PMI even when parallel multilingual input is derived from automatic MT systems, enabling us to apply PMI on a wide range of tasks.

Statistics of Activated Neurons in Transformer Models. Recently, statistics of activated neurons in transformer models by counting nonzero values in the output of ReLU is introduced by Zhang et al. (2022). Moreover, Li et al. (2023) show that the sparse activation of neurons is a ubiquitous phenomenon. In this work, we extend the statistical method to advanced transformer architectures. We hope this effort can help deepen our insights into the learning mechanism behind LLMs.

Cross-lingual In-context Learning. Several works have investigated cross-lingual prompts (Wang et al., 2023; Shi et al., 2023). One line of research requests LLMs to address the input problem in multiple languages orderly, then emphasizes self-consistency by aligning results of these languages to improve performance on reasoning tasks (Qin et al., 2023). To augment LLMs' performance with multilingual input, other works encourage LLMs to rephrase the input in English and then perform step-by-step analysis, indeed turning English into a pivot language (Huang et al., 2023; Zhang et al., 2023; Nguyen et al., 2023). Our work, in contrast, explores the behavior of LLMs that learns from parallel input in multiple languages simultaneously, revealing a new ICL capability.

6 Conclusions

We reveal that LLMs can learn from parallel multilingual input. Firstly, comprehensive experiments across 7 typical datasets, 9 commonly used multilingual LLMs, and 7 languages demonstrate the effectiveness and applicability of PMI. Secondly, statistics of activated neurons indicate that PMI optimizes performance by inhibiting neurons and promoting more precise neuron activation, which performs like one-off synaptic pruning.

7 Limitations

In fact, during the inference, LLMs will inevitably refer to the semantics of the translation in PMI to understand the input comprehensively. As a result, though our extensive experiments have demonstrated that LLMs can benefit from PMI, the quality of translation will influence the final performance. On the other hand, we do not discuss the effect of cross-language such as code-switch multilingual prompts because it deviates from the intention of PMI, i.e., providing parallel input. However, it is still a potential direction and we leave it for future work.

References

01-ai. 2023. A series of large language models trained from scratch by developers at 01-ai. https://github.com/01-ai/Yi.

AI@Meta. 2024. Llama 3 model card. *Github*.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4668–4679. Association for Computational Linguistics.

Duarte Alves, Nuno Miguel Guerreiro, João Alves, José Pombal, Ricardo Rei, José Guilherme Camargo de Souza, Pierre Colombo, and André Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11127–11148. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Gyu-Hyeon Choi, Jong-Hun Shin, and Young Kil Kim. 2018. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2924–2936. Association for Computational Linguistics.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics.

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, Loïc 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. *CoRR*, abs/2207.04672.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 268–277. The Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Samin Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4693–4703. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

hiyouga. 2023. Llama factory. https://github.com/hiyouga/LLaMA-Factory.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12365–12394. Association for Computational Linguistics.

Peter R Huttenlocher. 1990. Morphometric study of human cerebral cortex development. *Neuropsychologia*, 28(6):517–527.

Peter R Huttenlocher et al. 1979. Synaptic density in human frontal cortex-developmental changes and effects of aging. *Brain Res*, 163(2):195–205.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7943–7960. Association for Computational Linguistics.

Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J. Reddi, Ke Ye, Felix Chern, Felix X. Yu, Ruiqi Guo, and Sanjiv Kumar. 2023. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *CoRR*, abs/2306.11372.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with data augmentation. In *Proceedings of the 15th International Conference on Spoken Language Translation, IWSLT 2018, Bruges, Belgium, October 29-30, 2018*, pages 48–53. International Conference on Spoken Language Translation.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Proceedings of Machine Translation Summit VIII, MTSummit 2001, Santiago de Compostela, Spain, September 18-22, 2001*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2695–2709. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022. COMET-22: unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 578–585. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, 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, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Noam Shazeer. 2020. GLU variants improve transformer. *CoRR*, abs/2002.05202.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Cross-lingual summarization via chatgpt. *CoRR*, abs/2302.14229.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. Moefication: Transformer feed-forward layers are mixtures of experts. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 877–890. Association for Computational Linguistics.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2023. PLUG: leveraging pivot language in cross-lingual instruction tuning. *CoRR*, abs/2311.08711.

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023. Unveiling A core linguistic region in large language models. *CoRR*, abs/2310.14928.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 30–34. The Association for Computational Linguistics.

A Design of Prompts

To prohibit LLMs from skewing towards any particular languages in the input, we don't point out the original input of tasks in our prompts. All of the prompts are listed in Table 20. In this table, the content that is italicized and highlighted in gray indicates variable elements, which should be replaced according to the specific task requirements.

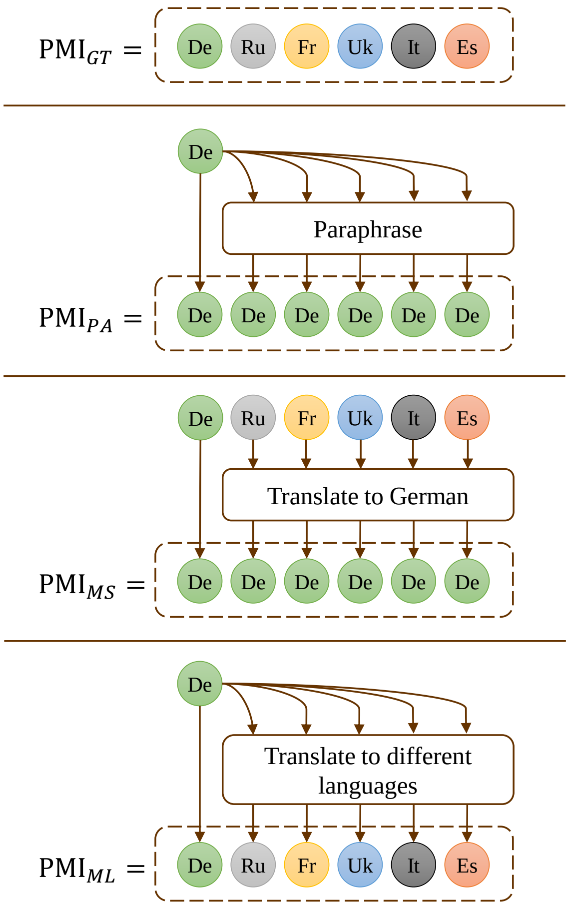


Figure 6: An illustration of different strategies for constructing parallel inputs in Section 2.2. Taking De $\rightarrow$ En translation as an example, PMI_{GT} consists of multilingual human translations from several experts; PMI_{PA} is made up of monolingual sentences paraphrased from the original German input; PMI_{MS} is composed of German translations where their source language texts are from different experts; and PMI_{ML} includes multilingual translations of the original German input derived from a single translator.

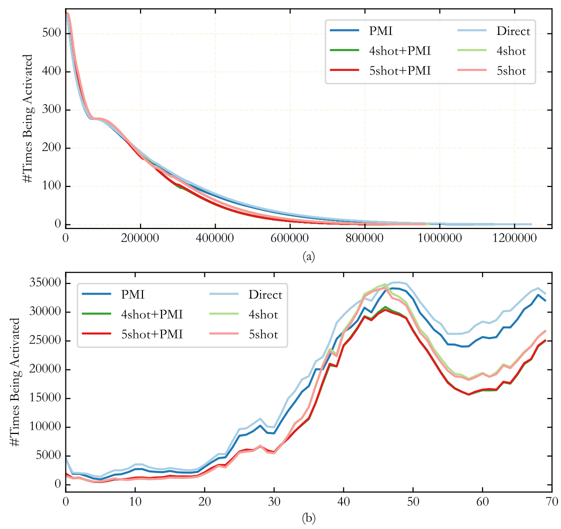


Figure 7: Distribution of all activated neurons in Bloomz-176B on RTE. The horizontal axis of the figure (a) represents different neurons arranged in descending order of the number of times being activated, and the horizontal axis of the figure (b) stands for the number of transformer layers.

B More Details About Statistical Method of Activated Neurons

Implementation of Counting Activated Neurons. During the inference stage, each time LLMs calculate the representation of a token including input and output, the intermediate result of MLPs stands for an activation state of neurons. It is essential to note that *we only make statistics of activated neurons based on the intermediate result corresponding to the output tokens.* This implementation is based on two concerns: (1) only the activation state of neurons corresponding to the output tokens directly contributes to the model-generated results. (2) since different prompting strategies differ in the length of input significantly, if the statistics are made based on both input and output tokens, then the results will be disturbed by the factor of length but not the actual impact of prompts, resulting in misdirected conclusions.

Activation Functions Used in LLMs. Table 8 records some popular LLMs and the activation functions they used.

C Supplementary Results About Neuron Activation

In Figure 7 (a), we can see that: (1) in the interval from 0 to 200000, the curves of PMI, few-shot learning and their combination are above that of

Activation Function	Formula	Model
ReLU	$\max(x, 0)$	Vanilla Transformer
GELU	$0.5x\left(1 + \text{erf}\left(x/\sqrt{2}\right)\right)$	Bloom, Falcon
SiLU	$x/\left(1 + e^{-x}\right)$	\
GEGLU	$\text{GELU}\left(XW_{up}\right) \odot \left(XV_{up}\right)$	mT0
SwiGLU	$\text{SiLU}\left(XW_{up}\right) \odot \left(XV_{up}\right)$	LLaMA, Qwen, ALMA, Yi

Table 8: The activation functions of some commonly used multilingual LLMs. In GELU, the $\text{erf}(\cdot)$ stands for the Gauss Error Function. Note that our extended statistical method can be applied to all LLMs shown in this table.

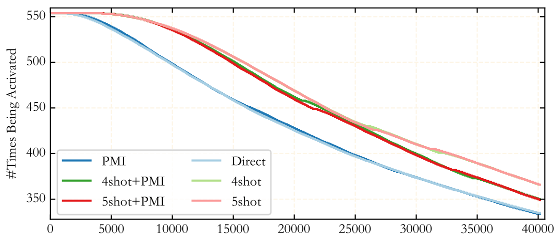


Figure 8: The distribution of the top 1% of activated neurons in Bloomz-176B on RTE.

Method		COMET	AP	COMET	AP
Direction		*De → En*		*De → Fr*	
w/o FT	0-shot	89.0	28.7	84.8	27.7
	5-shot	89.3	28.5	85.0	27.6
w/ FT	0-shot	89.5	28.1	85.3	27.2
	5-shot	89.3	27.8	84.9	27.1

Table 9: The translation performance and activation proportion (AP) of zero-shot and few-shot on Qwen-14B w/ or w/o fine-tuning (FT).

baseline (i.e., Direct), indicating that they activate top 200,000 commonly used neurons; (2) beyond the 200,000 mark, these curves are below the curve of baseline, demonstrating that these prompts perform inhibiting other less used neurons. Furthermore, in Figure 7 (b), we can see that the inhibited neurons concentrate in the back two-thirds of model layers. Figures 10 and 8 report the distribution of the top 1% of activated neurons in Bloomz-176B where PMI shows a clear impact of activation on most commonly used neurons.

To visualize the activation happening in each neuron, in Figure 9, we draw heat maps of Qwen-14B and Bloomz-176B when using the PMI-5 to translate De → En in the FLORES-200 and WMT dataset, respectively. It suggests that the neurons of Qwen-14B are more active while those of Bloomz-176B seem lazy and are activated fewer times. Furthermore, in each model, there are significant differences in the number of times being activated among different layers.

In Figure 11, we also make statistics of activated neurons in Bloomz-176B and Qwen-14B during the inference on the WMT dataset.

Table 9 shows the results of few-shot learning, which suggests that it also inhibits neurons and more neurons are inhibited after the LLM is fine-tuned.

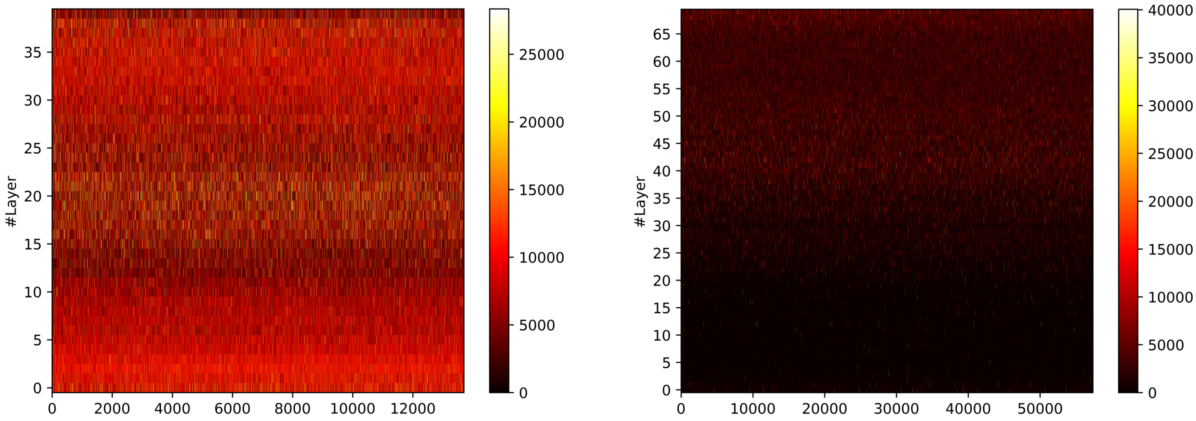
D More Analyses

D.1 Preliminary Experiments of Constructing PMI

Choose the parallel language that LLMs can understand. We test the impact of selecting parallel languages on the PMI-1 translating De → En of the FLORES-200, where Zh, Fr, Uk, and It are selected as the parallel languages. By comparing the results of translating them to English, we examine the model's understanding of these languages. In Figure 12, experimental results show that PMI-1 achieves better performance when the score of pivot translation is high and returns worse results when the score of pivot translation is low. This suggests that choosing parallel languages that the model comprehends better can bring more benefits for PMI.

Provide the highest quality translations as far as you can. Here, we utilize some translation systems with different performances to construct the parallel input of PMI in various qualities, including NLLB-1.3B, NLLB-54B, Qwen-14B, ChatGPT, and GPT-4. Experiments are conducted on both Qwen-14B and ChatGPT. In Figure 13, translation systems are arranged in the ascending order of their translation performance according to the curve, and the results show that higher quality of translations can result in larger improvements.

Place better understood language at the head and tail of the input sequence. We test the performance of PMI prompts with identical parallel



(a) Qwen-14B

(b) Bloomz-176B

Figure 9: The heat maps of activated neurons in MLPs of Qwen-14B and Bloomz-176B when using the PMI-5 to translate De → En in the FLORES-200 and WMT dataset, respectively. The horizontal axis represents the dimension of the middle outputs in MLPs (i.e., each neuron). The vertical axis represents the number of layers in the model. And each element in the map stands for the number of times of was activated during the inference stage.

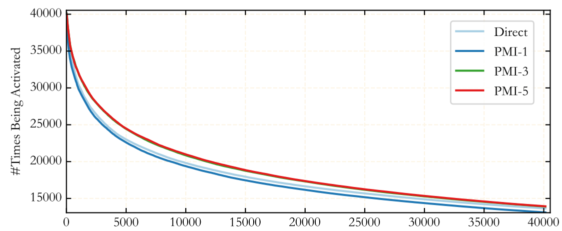


Figure 10: The distribution of the top 1% of activated neurons in Bloomz-176B on WMT22 De → En. The horizontal axis represents different neurons arranged in descending order of the number of times being activated.

Method	Input	COMET
Direct	De	89.5
	Es	87.4
	Ru	86.9
	Zh	86.9
	German → English	
PMI-3	**De + Zh + Ru + Es**	**90.5**
	De + Zh + Es + Ru	90.4
	De + Ru + Es + Zh	90.3
	Chinese → English	
PMI-3	**Zh + Ru + De + Es**	**90.3**
	Zh + Ru + Es + De	90.2
	Zh + Es + De + Ru	90.0

Table 10: Examining the factor of language order for PMI. The experiment is conducted on FLORES-200 and Qwen-14B.

texts but in different language order, and conduct experiments on De → En and Zh → En of the FLORES-200 using Qwen-14B. Results in Table 10 show that an LLM yields superior results when German is placed at the beginning and Spanish is placed at the end. Considering German and Spanish achieve higher score than other languages, we can infer that it is better to place the language better understood by the model at both ends of the input sequence.

D.2 Comparing the Performance Between Few-shot Learning and PMI

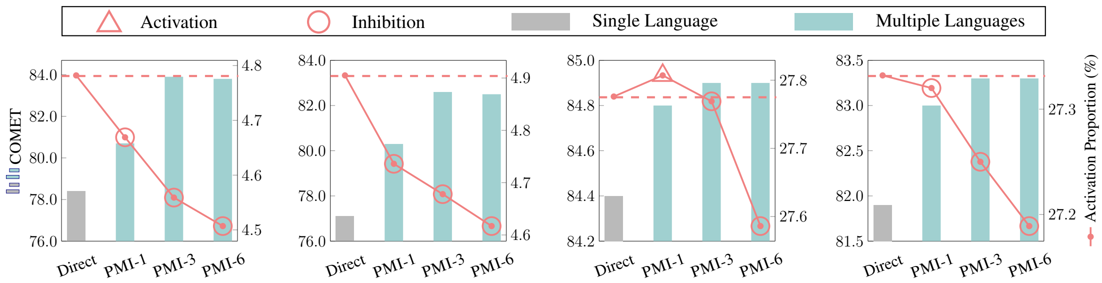
To further evaluate the effectiveness of our PMI, here we compare the results of PMI with those of few-shot learning. Notably, since our fine-tuning data is constructed by zero-shot instructions, which hurts the performance of few-shot learning evaluated on these fine-tuned models (Alves et al., 2023), hence we conduct experiments of few-shot learning on original models, i.e., the officially released weights without being fine-tuned. As shown in Table 11, PMI also outperforms few-shot learning.

D.3 Effectiveness of PMI on more modern LLMs

As LLMs develop further, we anticipate that more and more LLMs will benefit from PMI in the future. Here, we make experiments on Qwen1.5-14B, a successor of Qwen-14B. The latter is fine-tuned with PMI prompts in our paper, while the former is the original official version. From Table 13, we can see that Qwen1.5-14B responds to PMI prompts without prior fine-tuning and exhibits performance enhancements due to PMI.

D.4 Self-augmentation

In Table 14, we report the experimental results of prompting Qwen-14B with PMI while the parallel sentence pairs are translated by Qwen-14B itself. Although the improvements resulting from PMI are

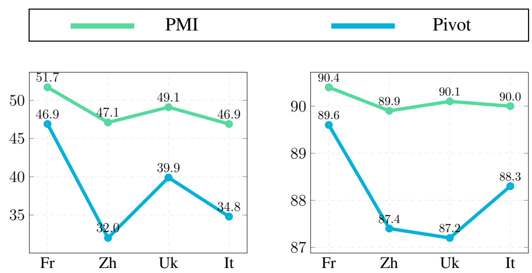


(a) $De \rightarrow En^{*}$ (b) $De \rightarrow Fr^{*}$ (b) $De \rightarrow En^{\dagger}$ (b) $De \rightarrow Fr^{\dagger}$

Figure 11: The translation performance and the activation proportion of different prompts on WMT dataset. $*$ and † stand for Bloomz-176B and Qwen-14B, respectively.

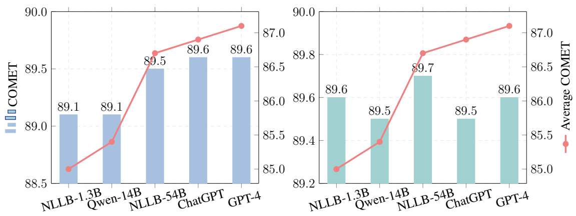
System		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Direction		*De → En*		*Zh → En*		*De → Fr*		*En → De*		*En → Zh*		*Is → En*	
Parallel Languages		*Es Ru Fr Zh Ja Cs*		*Es Ru Fr Ja Cs De*		*En Ru Es Zh It Cs*		*Es Ru Fr Zh Ja Cs*		*Es Ru Fr Ja Cs De*		*Es Ru Fr it Cs De*	
ChatGPT	Direct (1-shot) *	29.8	82.7	24.7	81.9	38.6	84.1	34.5	87.2	43.8	87.2	35.6	84.5
	Direct (5-shot) *	**32.9**	85.6	**25.4**	82.6	40.5	84.5	34.7	87.4	44.4	**87.4**	37.9	85.9
	PMI (5-shot) *	32.8	**85.7**	24.9	**82.9**	**41.5**	**84.7**	**34.8**	**87.6**	**45.1**	87.3	**39.3**	**86.7**
Qwen-14B	Direct (0-shot) †	30.4	84.4	23.7	80.8	34.2	81.9	29.6	85.3	45.2	**87.6**	18.4	69.7
	Direct (5-shot) *	31.5	84.7	24.0	80.8	33.0	81.8	29.3	84.9	**45.4**	87.3	19.6	71.9
	PMI (0-shot) †	**31.6**	**84.9**	**24.3**	**82.0**	**38.4**	**83.4**	**30.0**	**85.8**	45.1	**87.6**	**37.9**	**85.7**
ALMA-13B	Direct (0-shot) †	28.1	83.8	21.6	79.6	27.1	79.2	29.6	85.5	36.9	85.8	34.0	85.8
	Paper Reported *	30.7	84.4	**24.7**	79.9	-	-	31.4	85.5	**39.1**	85.8	36.5	86.3
	PMI (0-shot) †	**30.8**	**85.0**	23.8	**81.8**	**33.3**	**81.5**	**29.9**	**86.0**	36.9	**86.0**	**38.3**	**86.5**
Bloomz-176B	Direct (0-shot) *	24.0	78.4	16.0	76.4	27.3	77.1	13.0	70.7	29.5	83.9	5.6	53.8
	Direct (5-shot) *	23.1	79.7	14.5	77.3	25.9	77.2	**16.1**	**74.1**	**33.5**	**85.2**	5.1	56.1
	PMI (0-shot) *	**28.2**	**83.9**	**21.7**	**81.4**	**36.6**	**82.9**	16.0	73.9	32.4	84.7	**34.0**	**84.2**

Table 11: Comparing the performance of few-shot and PMI. In fairness, the results of few-shot come from models without fine-tuning, i.e., the official release. † and $*$ represent whether the prompt is fed to a model that has been fine-tuned or not, respectively.



(a) BLEU (b) COMET

Figure 12: Examining the factor of selecting parallel languages for PMI. The experiment is conducted on FLORES-200 De → En in PMI-1.



(a) Qwen-14B (b) ChatGPT

Figure 13: Examining the factor of translation quality for PMI. This experiment is conducted on FLORES-200 De → En in PMI-3. Each point on the red line represents the average COMET score of translating German to the three parallel languages by a translation system, reflecting the different translation qualities of parallel languages.

Method	Time Cost	Increase Rate (%)	BLEU	Increase Rate (%)
Direct	189.4s	-	45.2	-
PMI-1	249.7s	31.8	47.9	5.9
PMI-3	397.9s	110.1	56.2	24.3
PMI-5	507.3s	167.8	56.5	25.0

Table 12: The inference speed and performance gain of PMI with different amount of parallel languages.

System	BLEU	COMET	BLEU	COMET
Direction	*De → En*		*Zh → En*	
Direct	24.8	83.0	12.1	76.8
Pivot	23.4	83.4	**17.2**	80.7
PMI	**25.2**	**84.4**	17.0	**81.1**
Direction	*En → De*		*En → Zh*	
Direct	22.9	81.5	36.1	85.9
Pivot	21.0	82.1	35.7	85.2
PMI	**23.2**	**83.4**	**39.8**	**86.5**

Table 13: Experiments of Qwen1.5-14B on the WMT dataset.

System	BLEU	COMET	BLEU	COMET
Direction	*Zh → En*		*De → Fr*	
Direct	**23.7**	80.8	34.2	81.9
Pivot	15.9	78.7	36.2	81.3
PMI	22.1	**80.9**	37.6	**82.7**
Direction	*En → De*		*En → Zh*	
Direct	**29.6**	85.3	45.2	87.6
Pivot	25.8	83.5	39.7	86.2
PMI	**29.6**	**85.5**	**45.4**	**87.7**

Table 14: Augmenting Qwen-14B by the translations from Qwen-14B itself on the WMT dataset.

not as large as those reported in Table 3, PMI still outperforms baselines, especially at the COMET score. This further demonstrates the applicability of PMI. We attribute the diminished performance gains to the lower quality of translations produced by Qwen-14B compared to those from GPT-4.

D.5 Inference Speed

Since the inference speed of LLMs inevitably slows down as the input sequence lengthens, we also focus on the trade-off between performance and inference speed when increasing the number of parallel languages in the PMI. Here, we conduct experiments on the FLORES-200 De → En and Qwen-14B model. Table 12 indicates that for every additional parallel language integrated into the PMI input, there is an approximate 30% increase of time cost, along with a 5% improvement of performance. Notably, when the number of parallel languages reaches three, the improvement can reach up to 24.34%. Despite the increased inference cost, it is reasonable and acceptable considering the substantial performance gain.

System		BLEU	COMET	BLEU	COMET
Direction		*Fr → De*		*Fr → Es*	
ChatGPT	Direct	**30.4**	**86.5**	25.3	**86.3**
	PMI_{PA}	26.0$^{\downarrow 4.4}$	85.7$^{\downarrow 0.8}$	24.7$^{\downarrow 0.6}$	86.0$^{\downarrow 0.3}$
	PMI_{MS}	30.0$^{\downarrow 0.4}$	85.6$^{\downarrow 0.9}$	**26.1**$^{\uparrow 0.8}$	86.2$^{\downarrow 0.1}$
	PMI_{ML}	**30.4**$^{\uparrow 0.0}$	86.3$^{\downarrow 0.2}$	25.5$^{\uparrow 0.2}$	**86.3**$^{\uparrow 0.0}$
	PMI_{GT}	*32.4*	*86.9*	*27.0*	*86.8*
Qwen-14b	Direct	25.9	84.8	24.0	85.6
	PMI_{PA}	**28.1**$^{\uparrow 2.2}$	**86.0**$^{\uparrow 1.2}$	23.5$^{\downarrow 0.5}$	85.5$^{\downarrow 0.1}$
	PMI_{MS}	27.6$^{\uparrow 1.7}$	85.5$^{\uparrow 0.7}$	**25.4**$^{\uparrow 1.4}$	**86.0**$^{\uparrow 0.4}$
	PMI_{ML}	26.8$^{\uparrow 0.9}$	85.0$^{\uparrow 0.2}$	24.1$^{\uparrow 0.1}$	85.8$^{\uparrow 0.2}$
	PMI_{GT}	*29.6*	*86.0*	*27.3*	*86.4*
GPT-4	Direct	30.4	86.5	25.6	86.4
	PMI_{MS}	**32.1**$^{\uparrow 1.7}$	**87.1**$^{\uparrow 0.5}$	**26.3**$^{\uparrow 0.7}$	**87.0**$^{\uparrow 0.6}$
	PMI_{ML}	**32.1**$^{\uparrow 1.7}$	86.7$^{\uparrow 0.2}$	25.9$^{\uparrow 0.3}$	86.5$^{\uparrow 0.1}$
	PMI_{GT}	*35.8*	*87.7*	*28.4*	*87.3*

Table 15: Supplement results of the ablation study.

E Details of Experiment Setups

E.1 Downstream tasks

We introduce the details of the downstream tasks we used here:

Machine Translation In this task, a source language text is input into the model, which then translates it into a target language.

Nature Language Inference This task involves inputting a pair of sentences into the model, which then determines and outputs their relational status, such as contradiction, entailment, or neutrality.

Reading Comprehension This task give a passage and a question to the model, and then the model answers the question with a 'Yes' or 'No' based on its comprehension.

Text Simplification This task is to input a complex sentence into the model, and then the model generates a simplified version of the sentence without losing important information or altering its original intent.

Abstractive Summarization In this task, a long text is input into the model, which then produces a summary in one or two sentences that captures the essence and most critical information of the text.

E.2 Multilingual LLMs

Here, we introduce the multilingual LLMs used in our main experiment.

ChatGPT: the most capable GPT-3.5 model which performs impressively on rich-resource languages. We use the `gpt-3.5-turbo-0613` API.

Model	Task	Training Super Parameters			Training Data	
		Batch Size	Epoch	Learning Rate	Ratio	Size
Qwen-7B	Machine Translation	16	1	2e-5	1:9	4985
	Nature Language Inference	16	2	5e-5	1:7	2000
	Reading Comprehension	16	8	8e-5	1:5	2000
	Text Simplification	16	7	7e-5	1:5	2000
	Abstractive Summarization	16	4	1e-5	1:9	1200
Qwen-14B	Machine Translation	16	1	2e-5	1:9	4985
	Nature Language Inference	16	1	5e-5	1:7	2000
	Reading Comprehension	16	9	8e-5	1:7	2000
	Text Simplification	16	7	7e-5	1:5	2000
	Abstractive Summarization	16	4	7e-5	1:7	1200
ALMA-13B	Machine Translation	16	1	5e-5	1:9	4985
	Nature Language Inference	16	6	5e-5	1:7	2000
	Reading Comprehension	16	6	8e-5	1:7	2000
	Text Simplification	16	8	7e-5	1:9	2000
	Abstractive Summarization	16	3	2e-4	1:9	1200
Yi-34B	Nature Language Inference	16	3	1e-5	1:7	2000
	Reading Comprehension	16	7	8e-5	1:9	2000
	Text Simplification	16	7	5e-5	1:9	2000
	Abstractive Summarization	16	5	7e-5	1:9	1200
Qwen-72B	Nature Language Inference	16	8	1e-5	1:7	2000
	Reading Comprehension	16	5	6e-5	1:7	2000

Table 16: Our training setups. Each model is trained to ensure optimal performance for both the baseline and PMI.

Model	WikiAuto		XLSum			
	En		*Es*		*Ru*	
	Pivot	SARI	Pivot	R2/RL	Pivot	R2/RL
Qwen-7B	**Fr**	**43.2**	**Fr**	**9.4/22.7**	**Es**	**41.1/38.5**
	De	43.1	-	-	-	-
	Es	43.0	-	-	-	-
Qwen-14B	Fr	43.6	**Fr**	**9.0/21.4**	**Es**	**40.2/38.3**
	De	43.1	-	-	-	-
	Es	**43.8**	-	-	-	-
ALMA-13B	Fr	43.1	**Fr**	**10.4/23.0**	**Es**	**44.3/41.2**
	De	**43.2**	-	-	-	-
	Es	**43.2**	-	-	-	-
Yi-34B	**Fr**	**43.5**	**Fr**	**10.6/23.3**	**Es**	**41.7/38.8**
	De	43.3	-	-	-	-
	Es	42.4	-	-	-	-

Table 17: Full experimental results of pivot prompts on WikiAuto and XLSum dataset. The best results of each group are in **bold**.

LLaMA3: a latest open-source multilingual LLM which is pre-trained with 15 trillion tokens and demonstrated superior performance across multiple benchmarks (AI@Meta, 2024).

Bloomz: a fine-tuned version of Bloom (Scao et al., 2022), and we conduct experiments on the largest `bloomz` containing 176B parameters.

Qwen: open-source models which are trained up to 3 trillion tokens of multilingual data with competitive performance on various tasks (Bai et al., 2023). We do evaluations on three models, including Qwen-7B (`Qwen-7B-Chat`), Qwen-14B (`Qwen-14B-Chat`) and Qwen-72B (`Qwen-72B-Chat`).

ALMA: a multilingual LLaMA-2 (Touvron et al., 2023) produced by continually pre-training and specially instruction-tuning on the MT task (Xu et al., 2023). We conduct experiments on `ALMA-13B`.

Yi: an open-source model which is mainly trained on English and Chinese corpus achieving competitive performance on multilingual tasks (01-ai, 2023). We assess the effectiveness of PMI on Yi-34B (`Yi-34B-Chat`).

mT0: an instruction-tuned version of mT5 (Xue et al., 2021), we choose the mT0-13B (`mt0-xxl`) as it supports 46 languages.

E.3 Training Setups

Limited by parameters and training data, it might be a challenge for every LLM to understand PMI prompts inherently. To address this, we conducted training data and fine-tuned the models which seemed confused when facing the PMI prompt. Specifically, we leveraged LLaMA-Factory[4] (hiyouga, 2023) and the LoRA technology to train models, where we set the LoRA-rank to 8, LoRA-alpha to 32 and dropout to 0.1. Since the different amount of trainable parameters in the LoRA module, we applied different training strategies to ensure that every model can adequately understand prompts of various tasks. These settings are detailed in Table 16.

E.4 Details of the Fine-tuning Datasets

We constructed our fine-tuning dataset based on the training or development datasets of these tasks for both conventional and PMI prompts in zero-shot style. There are two factors, including the ratio of baseline to PMI data and the size of the training dataset, which are detailed in Table 16.

E.5 Decoding Setups

We kept consistent super parameters during the inference stage of every LLM, i.e., we set the decoding batch size to 4 and the temperature to 0.01 in order to ensure the reproducibility of the results.

F Full Experimental Results of Pivot Prompts

We have reported the results of pivot prompts with the highest score in the table of the main experiment. In Tables 17, 18 and 19, we list all the results of the pivot prompts.

[4]https://github.com/hiyouga/LLaMA-Factory

Model	Pivot	BLEU	COMET	Pivot	BLEU	COMET	Pivot	BLEU	COMET	Pivot	BLEU	COMET	Pivot	BLEU	COMET	Pivot	BLEU	COMET
Direction	*De → En*			*Zh → En*			*De → Fr*			*En → De*			*En → Zh*			*Is → En*		
ChatGPT	**Es**	**28.5**	**84.0**	**Es**	**21.6**	**81.9**	**En**	**40.4**	**84.0**	Es	30.0	85.6	**Es**	**40.3**	**86.0**	Es	34.6	85.4
	Ru	25.2	83.6	Ru	18.4	80.7	Ru	33.1	82.6	Ru	27.4	86.2	Ru	35.9	85.6	Ru	30.5	84.6
	Fr	27.3	82.6	Fr	16.3	76.9	Es	37.0	83.3	**Fr**	**30.0**	**86.4**	Fr	36.9	85.1	Fr	31.2	84.1
	Zh	19.5	82.4	Ja	18.5	80.1	Zh	25.0	80.9	Zh	21.7	85.0	Ja	33.4	85.0	It	33.0	85.0
	Ja	19.5	81.7	Cs	18.6	80.2	It	37.3	83.3	Ja	20.4	84.8	Cs	37.2	85.4	Cs	27.7	81.9
	Cs	25.6	81.8	De	20.1	81.0	Cs	34.8	82.5	Cs	29.0	86.1	De	37.9	85.9	**De**	**35.0**	**85.6**
LLaMA3-8B	Es	26.4	83.3	**Es**	**21.3**	**81.4**	**En**	**31.7**	**80.8**	Es	22.8	81.8	**Es**	**30.2**	79.9	**Es**	**32.5**	**84.9**
	Ru	23.3	82.7	Ru	17.8	79.9	Ru	24.3	79.6	Ru	19.6	82.1	Ru	26.4	81.0	Ru	27.6	83.5
	Fr	**27.4**	**83.4**	Fr	20	80.9	Es	30.7	80.5	**Fr**	**24**	**83.3**	Fr	28.8	81.0	Fr	32.2	85.0
	Zh	18.1	81.2	Ja	17.1	79.2	Zh	18.1	77.3	Zh	14.2	80.7	Ja	25.2	80.4	It	31	84.6
	Ja	16.6	80.2	Cs	18.2	79.7	It	31.5	80.7	Ja	13.5	80.5	Cs	28.2	81.1	Cs	27.9	83.4
	Cs	25.5	82.4	De	19.8	80.7	Cs	27.5	78.8	Cs	21.7	82.5	De	29.3	**81.7**	De	32.4	84.8
Qwen-14B	Es	28.1	83.8	**Es**	**22.4**	**81.8**	**En**	**37.4**	**82.7**	Es	26.5	83.7	**Es**	**41.2**	**86.3**	Es	33.7	85.2
	Ru	25.0	82.9	Ru	19.8	80.6	Ru	29.8	81.2	Ru	23.5	84.1	Ru	38.7	86.3	Ru	30.3	84.1
	Fr	**28.2**	**84.0**	Fr	21.5	81.5	Es	34.5	82.1	**Fr**	**26.9**	**84.7**	Fr	40.4	86.6	**Fr**	**34.1**	**85.4**
	Zh	20.5	82.1	Ja	19.1	79.8	Zh	24.7	79.9	Zh	20.5	83.2	Ja	35.6	85.5	It	33.0	85.0
	Ja	19.2	81.3	Cs	19.6	80.2	It	34.3	82.1	Ja	17.5	82.5	Cs	38.5	85.5	Cs	29.9	84.1
	Cs	25.1	82.6	De	20.7	81.2	Cs	30.5	80.3	Cs	24.3	83.8	De	39.1	86.3	De	33.8	85.2
ALMA-13B	Es	25.5	83.0	**Es**	**21.7**	**81.2**	**En**	**29.9**	**80.3**	Es	26.2	83.7	Es	32.3	83.9	**Es**	**32.7**	**85.2**
	Ru	22.8	82.5	Ru	18.9	80.1	Ru	24.8	78.8	Ru	24.6	84.8	Ru	31.4	84.5	Ru	28.1	84.1
	Fr	**26.0**	**83.3**	Fr	20.9	80.9	Es	29.4	79.9	**Fr**	**26.4**	**84.8**	Fr	32.3	84.5	Fr	31.7	85.0
	Zh	18.1	81.0	Ja	16.7	78.4	Zh	18.0	76.6	Zh	18.8	82.9	Ja	28.0	82.5	It	31.3	84.7
	Ja	16.3	79.9	Cs	19.0	79.8	It	30.2	80.0	Ja	15.8	81.2	Cs	32.2	84.4	Cs	28.5	84.0
	Cs	24.0	82.6	De	20.2	80.9	Cs	25.7	78.2	Cs	25.4	84.6	**De**	**32.3**	**84.6**	De	31.8	85.1
mT0-13B	**Es**	**24.5**	**82.5**	**Es**	**19.3**	**80.7**	En	30.9	79.8	Es	17.2	77.1	Es	23.4	81.9	**Es**	**30.8**	**84.6**
	Ru	21.3	81.5	Ru	16.0	79.1	Ru	25.7	78.6	Ru	15.6	77.5	Ru	23.1	82.3	Ru	25.9	82.9
	Fr	24.5	82.4	Fr	18.5	80.2	**Es**	**30.5**	**80.1**	Fr	16.8	77.2	Fr	23.1	82.1	Fr	29.3	84.0
	Zh	16.6	79.8	Ja	12.9	76.8	Zh	18.8	76.3	Zh	12.2	75.8	Ja	22.3	81.9	It	29.6	84.1
	Ja	15.6	79.3	Cs	16.5	79.1	It	30.3	80.0	Ja	12.1	76.4	Cs	22.9	81.6	Cs	27.1	83.5
	Cs	22.7	81.5	De	17.4	79.7	Cs	26.6	78.2	**Cs**	**17.4**	**78.5**	**De**	**23.8**	**82.1**	De	29.8	84.0
Bloomz-176B	**Es**	**25.0**	**82.8**	**Es**	**20.8**	**80.9**	**En**	**34.6**	**82.1**	Es	6.1	63.6	Es	27.3	82.8	**Es**	**31.5**	**84.6**
	Ru	17.5	76.0	Ru	14.8	75.2	Ru	22.2	75.1	**Ru**	**9.5**	**66.2**	Ru	22.2	79.1	Ru	20.4	77.5
	Fr	24.9	82.6	Fr	19.7	80.2	Es	33.5	81.5	Fr	8.9	67.1	**Fr**	**27.6**	**82.6**	Fr	29.9	84.3
	Zh	17.1	79.2	Ja	13.2	74.5	Zh	21.0	78.0	Zh	7.3	66.3	Ja	17.2	78.9	It	28.9	82.4
	Ja	13.0	74.3	Cs	10.7	66.4	It	32.2	80.3	Ja	4.9	60.9	Cs	15.1	68.8	Cs	14.5	67.8
	Cs	13.6	64.7	De	17.3	77.7	Cs	15.1	64.0	Cs	2.5	51.9	De	25.5	79.6	De	26.8	81.5

Table 18: Full experimental results of pivot prompts on WMT dataset. The best results of each group are in **bold**.

Model	RTE		XNLI						BoolQ	
	En		*Fr*		*De*		*Zh*		*En*	
	Pivot	Accuracy	Pivot	Accuracy	Pivot	Accuracy	Pivot	Accuracy	Pivot	Accuracy
Qwen-7B	De	85.9	**De**	**78.9**	**Es**	**80.2**	**De**	**74.2**	**Es**	**81.6**
	Es	**86.6**	Es	77.9	Fr	79.2	Es	74.1	-	-
	Fr	85.6	Ru	77.2	Ru	77.2	Fr	72.3	-	-
Qwen-14B	De	89.2	De	80.1	Es	79.5	De	73.3	**Es**	**86.0**
	Es	**90.6**	**Es**	**80.5**	**Fr**	**79.8**	**Es**	**74.2**	-	-
	Fr	88.8	Ru	79.1	Ru	77.7	Fr	72.8	-	-
ALMA-13B	De	84.1	**De**	**82.0**	Es	79.6	**De**	**75.9**	**Es**	**77.7**
	Es	**84.5**	Es	81.7	**Fr**	**80.8**	Es	74.3	-	-
	Fr	80.1	Ru	79.4	Ru	79.8	Fr	74.6	-	-
Yi-34B	De	79.1	De	70.0	**Es**	**72.6**	De	64.7	**Es**	**84.2**
	Es	**85.9**	**Es**	**71.5**	Fr	71.9	**Es**	**68.1**	-	-
	Fr	84.8	Ru	66.6	Ru	64.8	Fr	66.6	-	-
Qwen-72B	De	91.3	**De**	**85.8**	**Es**	**85.5**	De	78.9	**Es**	**88.7**
	Es	**92.4**	Es	85.0	Fr	85.2	**Es**	**80.6**	-	-
	Fr	90.6	Ru	83.9	Ru	83.5	Fr	79.5	-	-
Bloomz-176B	De	74.4	De	50.0	Es	53.0	De	49.6	-	-
	Es	73.3	**Es**	**53.1**	Fr	50.5	**Es**	**53.7**	-	-
	Fr	**77.6**	Ru	50.8	**Ru**	**53.3**	Fr	52.0	-	-

Table 19: Full experimental results of pivot prompts on RTE, XNLI and BoolQ dataset. The best results of each group are in **bold**.

Dataset		Prompt
FLORES-200 **WMT**	Direct	Translate into *target-language* . *source-language* : *source-sentence* *target-language* :
	PMI	Translate into *target-language* . *source-language* : *source-sentence* *parallel-language(1)* : *parallel-sentence(1)* *parallel-language(2)* : *parallel-sentence(2)* ...... *parallel-language(n)* : *parallel-sentence(n)* *target-language* :
	PMI_{MS} PMI_{PA}	There are six sentences in *source-language* , I need you to fully understand all of them and then translate to one *target-language* sentence. *source-language* : 1. *paraphrase-sentence1* 2. *paraphrase-sentence2* 3. *paraphrase-sentence3* 4. *paraphrase-sentence4* 5. *paraphrase-sentence5* *target-language* :
Asset **WikiAuto**	Direct	You will be presented with a complex sentence. Your task is to simplify this sentence to make it easier to understand, while maintaining its core meaning and factual content. The goal is to generate a simplified version of the sentence without losing important information or altering its original intent. Please provide a single simplified sentence as your response, without any explanation. Here is the complex sentence: Complex Sentence: *sentence* Your simplified version:
	PMI	You will be presented with the same sentence in four different languages: *source-language* , *parallel-language1* , *parallel-language2* , and *parallel-language3* . These sentences convey the exact same meaning. Your task is to simplify the sentence into *source-language* to make it easier to understand, while maintaining its core meaning and factual content. It is important to note that since all sentences have the same meaning, you only need to provide one simplified *source-language* version. Please generate a single simplified *source-language* sentence as your response, without any explanation. Here are the sentences: *source-language* Sentence: *source-sentence* *parallel-language1* Sentence: *parallel-sentence1* *parallel-language2* Sentence: *parallel-sentence2* *parallel-language3* Sentence: *parallel-sentence3* Your simplified *source-language* version:

Continued on next page

<table>
<tr><th>Dataset</th><th></th><th>Prompt</th></tr>
<tr><td rowspan="2">RTE</td><td>Direct</td><td>You will be presented with a pair of sentences.Your task is to determine the relationship between these two sentences. There are two possible relationships: entailment, not_entailment. 'entailment' means the first sentence logically implies the second one. 'not_entailment' means the first sentence logically conflicts with the second one. Please provide a single prediction for the relationship based on these sentence pairs, without any explanation. Here is the sentence pair:
Premise: src-premise
Hypothesis: src-hypothesis
Your prediction:</td></tr>
<tr><td>PMI</td><td>You will be provided with a set of sentence pairs that are semantically identical but presented in four different languages: src-language, parallel-language1, parallel-language2, and parallel-language3. Each pair consists of a premise and a hypothesis. Despite the language differences, the meaning of these sentences is the same across all languages. Your task is to analyze these sentence pairs and determine the relationship between the premise and the hypothesis. There are two possible relationships: entailment and not_entailment. 'entailment' means the first sentence logically implies the second one. 'not_entailment' means the first sentence logically conflicts with the second one. Please provide a single prediction for the relationship based on these sentence pairs, without any explanation. Here are the sentence pairs:
src-language :
Premise: src-premise
Hypothesis: src-hypothesis
parallel-language1 :
Premise: para1-premise
Hypothesis: para1-hypothesis
parallel-lang2 :
Premise: para2-premise
Hypothesis: para2-hypothesis
parallel-lang3 :
Premise: para3-premise
Hypothesis: para3-hypothesis
Your prediction:</td></tr>
<tr><td rowspan="2">XLSum</td><td>Direct</td><td>You will be presented with a long text. Your task is to summarize this text in 1-2 sentences in source-language, capturing the most important and core content. The summary should distill the essence of the article concisely and accurately. Please provide a single summary for the text without any explanation. Here is the text:
source-text
Your summary:</td></tr>
<tr><td>PMI</td><td>You will be presented with two texts, each in a different language: source-language, parallel-language. These texts convey the same meaning in their respective languages. Your task is to summarize the core content of these texts in one summary (1-2 sentences) in source-language, capturing the most important and central idea. Please provide a single summary for the texts without any explanation. Here are the texts:
source-language Text: source-text
parallel-language Text: parallel-text
Your summary in source-language :</td></tr>
</table>

Continued on next page

Dataset		Prompt
BoolQ	Direct	You will be provided with a passage and a yes/no question based on that passage. Your task is to read the passage and then answer the question with a simple 'Yes' or 'No' based on the information in the passage. Please do not provide any explanations or reasoning for your answer. Passage: *source-passage* Question: *source-question* Please respond with 'Yes' or 'No' only. Your answer:
	PMI	You will be provided with two passages, each in a different language: *source-language* , *parallel-language* . These passages convey the same meaning. Your task is to understand the content of these passages and then answer a yes/no question based on them. It's important to note that you only need to make one prediction as the semantic content across all the passages is identical. Please do not provide any explanations or reasoning for your answer. *source-language* Passage: *source-sentence* *parallel-language* Passage: *parallel-sentence* Question: *source-question* Please respond with 'Yes' or 'No' only. Your answer:
XNLI	Direct	You will be presented with a pair of sentences. Your task is to determine the relationship between these two sentences. There are three possible relationships: entailment, contradiction, or neutral. Please provide a single prediction for the relationship based on these sentence pairs, without any explanation. Here is the sentence pair: Premise: *premise-sentence* Hypothesis: *hypothesis-sentence* Your prediction:
	PMI	You will be given a premise in multiple languages (*source-language* , *parallel-language1* , *parallel-language2* , *parallel-language3*) and a hypothesis in *source-language* . Your task is to determine the relationship between the multilingual premises and the *source-language* hypothesis. There are three possible relationships: entailment, contradiction, or neutral. Please provide a single prediction for the relationship, without any explanation. Here are the premises and the hypothesis: *source-sentence* Premise: *source-premise* *parallel-language1* Premise: *parallel-premise1* *parallel-language2* Premise: *parallel-premise2* *parallel-language3* Premise: *parallel-premise3* Hypothesis: *source-hypothesis* Your prediction:

Table 20: All the prompts used in experiments.