Bridging the Language Gaps in Large Language Models with Inference-Time Cross-Lingual Intervention

Weixuan Wang[1†] **Minghao Wu**[2†] **Barry Haddow**[1] **Alexandra Birch**[1]

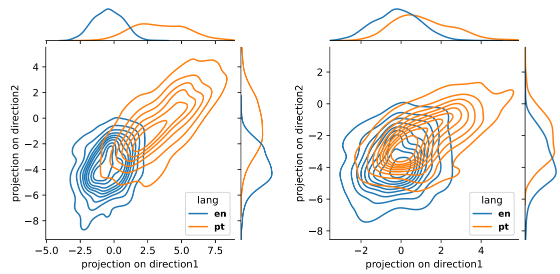
[1]School of Informatics, University of Edinburgh
[2]Monash University
{weixuan.wang, bhaddow, a.birch}@ed.ac.uk
minghao.wu@monash.edu

Abstract

Large Language Models (LLMs) have shown remarkable capabilities in natural language processing but exhibit significant performance gaps among different languages. Most existing approaches to address these disparities rely on pretraining or fine-tuning, which are resource-intensive. To overcome these limitations without incurring significant costs, we propose **Inference-Time Cross-Lingual Intervention (INCLINE)**, a novel framework that enhances LLM performance on low-performing (source) languages by aligning their internal representations with those of high-performing (target) languages during inference. INCLINE initially learns alignment matrices using parallel sentences from source and target languages through a Least-Squares optimization, and then applies these matrices during inference to transform the low-performing language representations toward the high-performing language space. Extensive experiments on nine benchmarks with five LLMs demonstrate that INCLINE significantly improves performance across diverse tasks and languages, compared to recent strong baselines. Our analysis demonstrates that INCLINE is highly cost-effective and applicable to a wide range of applications. In addition, we release the code to foster research along this line.[1]

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across a variety of natural language processing tasks, demonstrating strong capabilities in language understanding and generation (OpenAI, 2023; Dubey et al., 2024; Mesnard et al., 2024; Anthropic, 2024; OpenAI, 2024a,b). However, despite these advancements, most state-of-the-art LLMs remain predominantly English-centric, exhibiting significant performance gaps among different languages (Petrov et al., 2023; Kumar et al., 2024), which can adversely affect user experience and potentially exclude large portions of the global population from accessing advanced AI services (Lai et al., 2023a; Wang et al., 2024a).



(a) Before intervention (b) After intervention

Figure 1: Bivariate kernel density estimation plots displaying the representations (hidden states of the last token) from 100 random examples in English (blue) and their Portuguese translations (orange) from XCOPA (Ponti et al., 2020). After intervention using INCLINE, the Portuguese representations are aligned closer to the English representations.

Addressing the performance gaps across languages is highly challenging. Recent approaches are mostly data-driven, such as multilingual supervised fine-tuning or continued pre-training (Üstün et al., 2024; Cui et al., 2023; Kuulmets et al., 2024). However, collecting and annotating large-scale datasets for numerous underrepresented languages is both time-consuming and resource-intensive (Xue et al., 2021; Lai et al., 2023b). Furthermore, training LLMs on multilingual data requires substantial computational resources, limiting their practicality for widespread applications, especially in resource-constrained settings (Muennighoff et al., 2023; Li et al., 2023a). Given these limitations, a natural question arises: *How can we bridge the performance gaps between high-performing and low-performing languages without incurring prohibitive costs?*

Inspired by Lample et al. (2018) showing that

[†] Equal contribution.
[1]https://github.com/weixuan-wang123/INCLINE

word embeddings in different languages can be aligned to a shared representation space through learned rotations for word translation, we propose **Inference-Time Cross-Lingual Intervention (INCLINE)**. This novel framework utilizes a group of learned alignment matrices that transform the representations (e.g., hidden states) of a low-performing (source) language into those of a high-performing (target) language during inference. Our framework comprises two main steps. First, we train the alignment matrices for each layer of LLM using parallel sentences from the source and target languages. The learning process is formulated as a Least-Squares optimization problem, where these alignment matrices are learned by minimizing the distance between the projected source language representations and their corresponding target language representations, without the need for extensive retraining or fine-tuning the LLM. Second, we apply the learned alignment matrices to transform the source language input representations into the target language representation space at each layer during inference. By integrating these steps, INCLINE leverages the rich representations learned from high-performing languages to enhance performance on downstream tasks involving low-performing languages. As shown in Figure 1, INCLINE effectively aligns the input representations in Portuguese to their parallel representations in English.

In this study, we conduct extensive experiments to validate the effectiveness of INCLINE on nine widely used benchmarks using five LLMs. Our results demonstrate that aligning internal representations using INCLINE significantly improves performance on diverse tasks among languages.

Our contributions are summarized as follows:

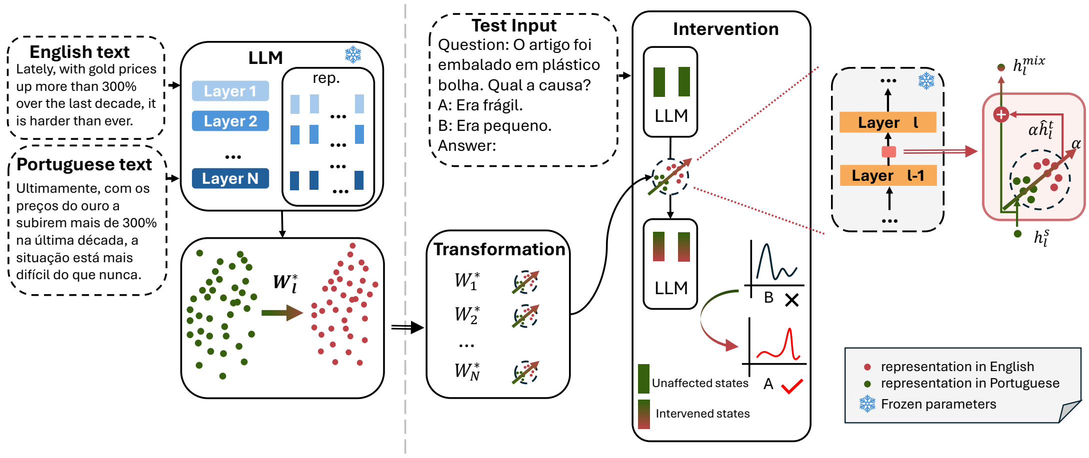
- We propose INCLINE, a cross-lingual intervention approach that enhances LLMs by transforming source language representations into a target language representation space during inference without requiring additional training of LLMs (see Section 3).
- We conduct extensive evaluations across five discriminative tasks and four generative tasks, covering 21 languages. Our experimental results show that INCLINE significantly improves model performance, boosting average accuracy by up to +4.96 compared to strong baselines (see Section 4).
- Our detailed analysis indicates that INCLINE is highly cost-effective, as it requires minimal computational resources while delivering substantial performance improvements (see Section 5). Moreover, we demonstrate that INCLINE is effective with regard to LLM backbones, model sizes, and in-context learning, underscoring its general applicability and potential for broader use in enhancing LLMs for underrepresented languages (see Section 6).

2 Related Work

Multilingual LLMs LLMs are pivotal in multilingual NLP tasks, typically leveraging external parallel datasets for training (Xue et al., 2021; Muennighoff et al., 2023; Chung et al., 2024; Li et al., 2025). For low-resource languages, data augmentation techniques generate parallel data by mining sentence pairs or translating monolingual text using machine translation tools (Edunov et al., 2018; Zhao et al., 2021; Ranaldi et al., 2023). However, these methods heavily rely on robust parallel corpora. To reduce data costs, studies have shifted toward Parameter-Efficient Fine-Tuning (PEFT) techniques (Pfeiffer et al., 2020; Parović et al., 2022; Agrawal et al., 2023; Wu et al., 2024; Wang et al., 2025) and cross-lingual embeddings mapping methods (Mikolov et al., 2013; Ormazabal et al., 2019; Wang et al., 2022), which still demand considerable computational resources.

Multilingual Prompting There is a growing interest in methods that do not require parameter adjustments. Prompting techniques have emerged, utilizing LLMs with multilingual prompts (Lin et al., 2021c, 2022; Shi et al., 2022b; Huang et al., 2023). However, these strategies face challenges like poor translation quality and prompt framing interference (Wang et al., 2024c). Additionally, their effectiveness varies by task, as recent research indicates that few-shot learning may not outperform zero-shot learning in translation tasks (Hendy et al., 2023; Wang et al., 2024d).

Intervention To address these challenges, we explore inference-time intervention techniques as cost-effective and efficient alternatives to traditional fine-tuning. Prior research in style transfer (Subramani et al., 2022; Turner et al., 2023), knowledge editing (Meng et al., 2022), and truthfulness shifting (Li et al., 2023b; Rimsky et al., 2024; Wang et al., 2024e) demonstrates the potential of linear probe-based interventions. However, these methods have been largely limited to mono-



(a) Learning the Cross-Lingual Alignment

(b) Inference-Time Transformation

Figure 2: Framework of INCLINE. INCLINE involves two steps: (a) Learning the Cross-Lingual Alignment: sentence representations from a parallel dataset are used to train alignment matrices that map source (Portuguese) representations to the target (English) representations. (b) Inference-Time Transformation: this step adapts the source representations from downstream tasks into the target representation space using the alignment matrices.

lingual contexts. Our goal is to design a novel cross-lingual inference-time intervention that effectively aligns representations across languages, aiming to improve performance across multiple languages.

3 Methodology

In Figure 2, we illustrate the framework of INCLINE, which enhances LLMs through inference-time cross-lingual intervention. Our approach comprises two main steps:

- **Learning the Cross-Lingual Alignment**: Using parallel corpora, we train alignment matrices for each layer to map source language representations to target language representations (see Section 3.1).
- **Inference-Time Transformation**: During inference, we utilize the learned alignment matrices to transform input representations from the source language into the target language representation space, thereby improving the LLM's performance on tasks in the source language (see Section 3.2).

By minimizing the distance between the source language representations and their corresponding target language representations, we effectively reduce cross-lingual representation gaps and align representation spaces across languages.

3.1 Learning the Cross-Lingual Alignment

Inspired by Schuster et al. (2019) that align embeddings across languages with learned linear transformations, we aim to learn a cross-lingual alignment matrix $\boldsymbol{W}_l$ that aligns sentence representations from the source language to the target language at the l-th layer of LLM. Given a parallel dataset $D = \{(\boldsymbol{x}_i^{\mathrm{s}}, \boldsymbol{x}_i^{\mathrm{t}})\}_{i=1}^N$, where each $\boldsymbol{x}_i^{\mathrm{s}}$ is the i-th source sentence and $\boldsymbol{x}_i^{\mathrm{t}}$ is its corresponding translation in the target language. Both $\boldsymbol{x}_i^{\mathrm{s}}$ and $\boldsymbol{x}_i^{\mathrm{t}}$ are sequences of tokens. From these sequences, we extract sentence representations by taking the hidden state of the last token in each sequence, denoted as $\boldsymbol{h}_{i,l}^{\mathrm{s}} \in \mathbb{R}^d$ and $\boldsymbol{h}_{i,l}^{\mathrm{t}} \in \mathbb{R}^d$ for the source and target sentence, respectively, where d is the dimensionality of the hidden states.

To minimize the difference between the projected source sentence representations and the target sentence representations, our objective can be defined as a Least-Squares optimization problem:

$$\boldsymbol{W}_l^* = \underset{\boldsymbol{W}_l}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \boldsymbol{W}_l \boldsymbol{h}_{i,l}^{\mathrm{s}} - \boldsymbol{h}_{i,l}^{\mathrm{t}} \right\|^2 \tag{1}$$

This problem seeks the optimal $\boldsymbol{W}_l^*$ that aligns the source representations with the target representations by minimizing the distance between them. Hence, the closed-form solution to this optimization problem is:

$$\boldsymbol{W}_l^* = \left(\sum_{i=1}^{N} (\boldsymbol{h}_{i,l}^{\mathrm{s}})^\top \boldsymbol{h}_{i,l}^{\mathrm{s}} \right)^{-1} \left(\sum_{i=1}^{N} (\boldsymbol{h}_{i,l}^{\mathrm{s}})^\top \boldsymbol{h}_{i,l}^{\mathrm{t}} \right) \tag{2}$$

By applying the learned alignment matrix $\boldsymbol{W}_l^*$ to the source sentence representations, we effectively map them into the target language's representation space. This alignment reduces cross-lingual representation discrepancies, allowing the model to leverage knowledge from the target language to improve performance on tasks in the source language.

3.2 Inference-Time Transformation

With the learned alignment matrix $\boldsymbol{W}_l^*$, we can enhance the LLM's processing of source language inputs by transforming their representations to the target representation space during inference.

We denote the hidden state of the last token of the test input $\boldsymbol{q}^s$ in the source language at the l-th layer of the LLM as $\boldsymbol{h}^{\mathrm{s}}_{q,l}$ and then project this source language representation into the target representation space using the alignment matrix $\boldsymbol{W}_l^*$:

$$\hat{\boldsymbol{h}}^{\mathrm{t}}_{q,l} = \boldsymbol{W}_l^* \boldsymbol{h}^{\mathrm{s}}_{q,l} \tag{3}$$

To perform the cross-lingual intervention at the l-th layer using the intervention vector $\hat{\boldsymbol{h}}^{\mathrm{t}}_{q,l}$, we adjust the original hidden state in source language $\boldsymbol{h}^{\mathrm{s}}_{q,l}$ by blending it with the projected hidden state in target language $\hat{\boldsymbol{h}}^{\mathrm{t}}_{q,l}$. This adjustment is controlled by a hyperparameter α, which balances the influence between the source and target hidden states:

$$\boldsymbol{h}^{\mathrm{mix}}_{q,l} = \boldsymbol{h}^{\mathrm{s}}_{q,l} + \alpha \hat{\boldsymbol{h}}^{\mathrm{t}}_{q,l} \tag{4}$$

Here, Equation 4 represents a shift of representation of source language towards target language representation by a magnitude of α times.

Decoding with Minimal Intervention In this work, we only conduct one single intervention on the last token of $\boldsymbol{q}^s$ by replacing $\boldsymbol{h}^{\mathrm{s}}_{q,l}$ with $\boldsymbol{h}^{\mathrm{mix}}_{q,l}$ for the test input $\boldsymbol{q}^s$ at the l-th layer of LLM. In such a way, we can effectively intervene the model output while preserve the features in the source language.

Comparison with ITI and CAA Recently, ITI (Li et al., 2023b) and CAA (Rimsky et al., 2024) have been proposed as interventions in the model behaviors by manipulating the selected attention heads and hidden states, respectively. INCLINE is distinct from ITI and CAA due to three primary differences. Firstly, ITI and CAA utilize a learned *static* intervention vector to alter model behaviors, whereas INCLINE leverages a set of alignment matrices to *dynamically* align input representations from the source language to the target language. Secondly, ITI and CAA apply the intervention vector across all token positions following the instruction, potentially causing excessive perturbation during inference. In contrast, INCLINE performs a single intervention solely on the last token of the input. Additionally, unlike ITI and CAA, which target on only a limited number of layers, INCLINE modifies the representations across all layers. These modifications enable the LLMs to comprehensively leverage their target language capabilities for multilingual prediction.

4 Experiments

In this section, we introduce our experimental setup (Section 4.1) and present our results in Section 4.2.

4.1 Experimental Setup

We present our evaluation tasks, model backbones, implementation details of INCLINE, and baselines in this section.

Evaluation Tasks We conduct extensive evaluations across nine diverse downstream tasks, categorized into two groups:

- **Discriminative Tasks**: `XCOPA` (Ponti et al., 2020), `XStoryCloze` (Lin et al., 2021b), `XWinograd` (Lin et al., 2021b), `XCSQA` (Lin et al., 2021a), `XNLI` (Conneau et al., 2018);
- **Generative Tasks**: `MZsRE` (Wang et al., 2024b), `Flores` (Goyal et al., 2021), `WMT23` (Kocmi et al., 2023), `MGSM` (Shi et al., 2022a).

These tasks covers 21 languages including English (en), Arabic (ar), German (de), Greek (el), Spanish (es), Estonian (et), French (fr), Hindi (hi), Indonesian (id), Italian (it), Japanese (ja), Dutch (nl), Portuguese (pt), Russian (ru), Swahili (sw), Tamil (ta), Thai (th), Turkish (tr), Ukrainian (uk), Vietnamese (vi), and Chinese (zh). We include more details of these tasks in Appendix A.

Model Backbones In this work, we mainly use BLOOMZ-7B1-MT as our model backbone for all the baseline approaches, unless otherwise specified. To demonstrate the effectiveness of INCLINE across various model backbones, we include four additional LLMs: LLAMA3.1-8B-INSTRUCT (Dubey et al., 2024), LLAMA2-7B-CHAT (Touvron et al., 2023), MISTRAL-7B-INSTRUCT (Jiang et al., 2023), FALCON-7B-INSTRUCT (Almazrouei et al., 2023). We present these results in Section 6. For the `MGSM` task, we employ the MATHOCTOPUS (Chen et al., 2023),[2]

[2]https://huggingface.co/Mathoctopus/Parallel_7B

	XCOPA			XStoryCloze			XWinograd			XCSQA			XNLI		
	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}
BASELINE	61.62	69.00	52.40	74.96	77.83	57.78	57.05	59.71	53.06	47.35	55.31	34.62	46.48	50.04	41.48
MT-GOOGLE	73.31†	73.52	73.05†	76.63	76.05	80.08†	57.63	57.12	57.90†	58.52†	54.84	64.40†	50.72	49.80	52.00
MT-LLM	59.84	67.16	50.70	79.41	82.23	62.48	43.02	41.67	45.04	30.73	35.38	23.30	43.64	47.83	37.77
Intervention Methods															
ITI	60.91	67.56	52.60	76.38	79.33	58.70	48.24	58.37	33.06	46.32	55.33	31.92	46.32	49.51	41.84
CAA	63.96	71.80	54.15	78.16	80.92	61.61	58.42	60.70	55.01	47.97	56.01	35.10	46.17	50.92	39.52
INCLINE	**65.22 (+3.60)**	**72.56 (+3.56)**	**56.05 (+3.65)**	**79.92 (+4.96)**	**82.03 (+4.20)**	**67.24 (+9.46)**	**59.35† (+2.30)**	**62.04† (+2.33)**	**55.32 (+2.26)**	**48.45 (+1.10)**	**56.45† (+1.14)**	**35.64 (+1.02)**	**48.12 (+1.64)**	**51.44 (+1.40)**	**43.47 (+1.99)**
SFT	66.89	76.84	54.45	87.36	89.50	74.52	43.78	48.63	36.50	42.18	47.95	32.96	69.68	76.76	59.76
SFT +INCLINE	**69.24 (+2.35)**	**79.28† (+2.44)**	**61.22 (+6.77)**	**88.11† (+0.75)**	**90.00† (+0.50)**	**76.77 (+2.25)**	**49.84 (+6.06)**	**57.58 (+8.95)**	**38.24 (+1.74)**	**42.55 (+0.37)**	**48.38 (+0.43)**	**33.22 (+0.26)**	**71.17† (+1.49)**	**77.83† (+1.07)**	**61.84† (+2.08)**

Table 1: Main results of discriminative tasks. All the tasks are evaluated using accuracy. † denotes the best results. μ_{ALL}, μ_{SEEN}, and μ_{UNSEEN} indicate the macro-average of results across all the languages, the seen languages, and the unseen languages, respectively.

a specialized model fine-tuned from LLAMA2-7B for mathematical reasoning tasks, as the backbone.

INCLINE (Ours) In this work, we mainly focus on aligning the low-performing language (source) representations closer to the English (target) representations, as LLMs are predominantly English-centric. For training the alignment matrices between languages, we randomly sample 500 parallel sentence pairs for each language pair involving English and other languages. These pairs are sourced from the News Commentary v16 dataset (Barrault et al., 2019), and for languages not covered by this dataset, we use the CCAligned dataset (El-Kishky et al., 2020). Following Rimsky et al. (2024), the value of the α controlling the intervention strength is in the range from -1 to 1 and determined by the validation results for each language across tasks. We use one A100 GPU (40G) for all experiments.

Baselines We compare INCLINE against several established techniques on the zero-shot setting: (1) **BASELINE** indicates the predictions given by the original BLOOMZ-7B1-MT; (2) **MT-GOOGLE** utilizes GOOGLE TRANSLATE to translate non-English questions into English; (3) **MT-LLM** leverages BLOOMZ-7B1-MT to translate questions in non-English languages into English, employing the structured prompt template "`{Source Language}: {Inputs} English:`"; (4) **SFT** represents the task-specific supervised fine-tuning (SFT) involving updating all parameters of the LLM on the English training set for each downstream task individually with the hyperparameters described in Appendix B and evaluating the resulting model on the multilingual test sets; (5) **ITI** (Li et al., 2023b) is an intervention method that identifies attention heads with high linear probing accuracy for truthfulness and adjusts activations along these truth-correlated directions during inference. Originally used to shift models from generating false statements to truthful ones, we adapt it to encourage the generation of English text over non-English text. (6) **CAA** (Rimsky et al., 2024) employs the mean difference in hidden states between positive and negative examples from additional training data as an intervention vector to adjust the model's behavior towards the desired direction. Initially designed for monolingual alignment-relevant tasks, we utilize it to shift the model's output from non-English to English.

4.2 Results

In this section, we present our results on the discriminative tasks (Table 1) and generative tasks (Table 2). Furthermore, we also categorize the languages involved in the downstream tasks into two groups based on the training data of BLOOMZ-7B1-MT: *seen languages* (ar, es, fr, hi, id, pt, sw, ta, vi, and zh) and *unseen languages* (de, el, et, it, ja, nl, ru, th, tr, and uk). The breakdown results are provided in Table 7 (see Appendix C).

INCLINE significantly improves discriminative task performance. The experimental results in Table 1 clearly demonstrate the effectiveness of INCLINE. Although methods like SFT, MT-GOOGLE, and MT-LLM achieve high performance, they come with substantial costs, including the need for extensive fine-tuning of LLMs and reliance on third-party tools. Activation intervention methods, such as ITI and CAA, offer a more cost-effective solution but yield only minimal improvements, indicating a potential inadequacy in capturing the complexities of multilingual tasks. In contrast, INCLINE provides significant perfor-

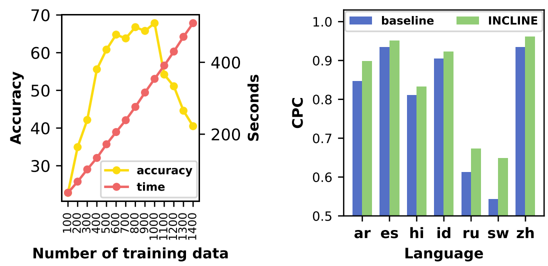
	MZsRE			Flores			WMT23			MGSM		
	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}	μ_{ALL}	μ_{SEEN}	μ_{UNSEEN}
BASELINE	39.96	45.79	32.67	46.09	58.57	21.12	13.78	14.39	13.63	39.35	39.80	38.90
MT-GOOGLE	73.56†	72.76†	74.56†	-	-	-	-	-	-	46.70†	47.70†	45.70†
MT-LLM	33.18	39.25	25.61	-	-	-	-	-	-	21.40	30.00	12.80
Intervention Methods												
ITI	36.31	41.72	29.54	2.85	2.97	1.95	2.34	3.16	2.13	40.50	41.90	39.10
CAA	42.88	50.17	33.78	47.87	60.63	16.75	13.74	14.86	13.46	39.43	40.85	38.00
INCLINE	**43.22 (+3.26)**	**50.21 (+4.42)**	**34.49 (+1.82)**	**48.19† (+2.10)**	**61.28† (+2.71)**	**22.00† (+0.88)**	**14.23† (+0.45)**	**15.05† (+0.66)**	**14.02† (+0.39)**	**42.85 (+3.50)**	**43.30 (+3.50)**	**42.40 (+3.50)**

Table 2: Main results of generative tasks. † denotes the best results. μ_{ALL}, μ_{SEEN}, and μ_{UNSEEN} indicate the macro-average of results across all the languages, the seen languages, and the unseen languages, respectively. We use Exact Match (EM) to evaluate `MZsRE`, use BLEU to evaluate `Flores` and `WMT23`, and use accuracy to evaluate `MGSM`.

mance gains by enhancing multilingual representation alignment at inference time without requiring extensive resources or dependencies. This results in a more efficient improvement in multilingual performance. For example, INCLINE increases the average accuracy by +4.96 on `XStoryCloze`. Additionally, it delivers improvements of +4.20 and +9.46 for seen and unseen languages, respectively. Moreover, INCLINE can further improve the performance of the task-specific SFT.

INCLINE significantly enhances generative task performance. The experimental results presented in Table 2 suggest the effectiveness of INCLINE in enhancing performance across generative tasks. Unlike ITI and CAA, which show only marginal improvements similar to those observed in discriminative tasks, INCLINE appears to achieve substantial advancements. Notably, ITI seems to struggle significantly in machine translation tasks, such as `Flores` and `WMT23`, highlighting its limitations. Furthermore, INCLINE reportedly boosts accuracy in the `MGSM` task by up to +3.50 across various languages. This finding suggests that, although the mathematical capabilities are independent from the languages, understanding the questions written in different languages still requires language-specific knowledge. INCLINE successfully transfers the LLMs' natural language understanding capabilities from English to other languages. It is important to note that SFT is not evaluated on generative tasks because there are no training sets associated with these tasks.

In summary, these results demonstrate that INCLINE offers a significant improvement in both discriminative and generative tasks by effectively aligning multilingual representations.



(a) Training cost (b) Prediction consistency

Figure 3: (a) Training costs of INCLINE with regard to the number of parallel sentences and time used for training alignment matrices. INCLINE is evaluated on `XStoryCloze` in Swahili. (b) Correct Prediction Consistency (CPC) between non-English and English on `XStoryCloze` for the model using INCLINE.

5 Analysis

In this section, we conduct an in-depth analysis of INCLINE, focusing on four key aspects: computational costs, enhanced consistency after intervention, the impacts of the intervened components of LLMs, and the choice of intervention strength α. This analysis provides a comprehensive understanding of how INCLINE operates and its implications for model performance and efficiency.

INCLINE is highly efficient for training and introduces only marginal overhead for inference. To analyze the relationship between computational costs and accuracy, we measure both the training and inference costs of our method, INCLINE, using the `XStoryCloze` task in Swahili. As shown in Figure 3(a), increasing the amount of training data does not necessarily lead to improved accuracy, even though the training time is directly proportional to the number of samples. In our study, we empirically determine that using 500 samples for training the alignment matrices provides the best

	XCOPA	XCSQA	Flores	MGSM
BASELINE	61.60	47.35	46.09	39.35
INCLINE				
├ INCLINE-HIDDEN	**65.22**	**48.45**	**48.19**	**42.85**
├ INCLINE-ATTN	63.87	48.18	47.54	41.55
├ INCLINE-FFN	64.20	47.96	46.10	41.80
└ INCLINE-EMB	63.16	47.59	39.23	40.90

Table 3: The averaged results of `XStoryCloze`, `XCSQA`, `Flores`, `MGSM` tasks with four configurations for INCLINE given by BLOOMZ-7B1-MT.

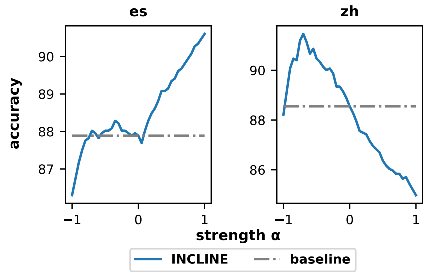


Figure 4: The accuracy changed with hyperparameter α on the `XStoryCloze` task with BLOOMZ-7B1-MT.

	ar	es	hi	id	ru	sw	zh	AVG
BLOOMZ-7B1-MT								
BASELINE	79.22	87.89	76.37	84.45	57.78	50.50	88.55	74.96
INCLINE	**83.12**	**90.60**	**81.47**	**86.10**	**67.24**	**59.70**	**91.20**	**79.92**
LLAMA3.1-8B-INSTRUCT								
BASELINE	86.50	91.73	84.84	37.46	66.98	54.00	92.39	73.41
INCLINE	**87.36**	**92.39**	**85.31**	**64.53**	**73.73**	**55.66**	**92.72**	**78.81**
LLAMA2-7B-CHAT								
BASELINE	49.37	47.25	39.25	48.18	34.94	0.93	55.53	39.35
INCLINE	**51.42**	**56.65**	**47.25**	**49.97**	**41.03**	**17.67**	**60.69**	**46.38**
MISTRAL-7B-INSTRUCT								
BASELINE	18.33	81.34	24.95	76.64	83.65	2.58	90.07	53.94
INCLINE	**36.71**	**84.23**	**35.77**	**80.18**	**85.13**	**25.71**	**90.34**	**62.58**
FALCON-7B-INSTRUCT								
BASELINE	53.61	58.31	53.21	55.59	54.60	51.16	54.00	54.35
INCLINE	**54.33**	**61.81**	**54.33**	**58.04**	**57.91**	**53.47**	**59.70**	**57.09**

Table 4: The results of `XStoryCloze` dataset with five LLM backbones.

balance between performance gains and computational costs. Consequently, the training process takes only 172 seconds. During inference, our approach involves a single intervention at the last token, resulting in a time complexity of $O(1)$. This method incurs only a 12% increase in inference time, taking 0.80 seconds per item compared to 0.71 seconds without it, thereby maintaining a low inference cost. More results are provided in Appendix G.

INCLINE effectively enhances the consistency of correct predictions between non-English languages (source) and English (target). Recent non-English test sets are commonly translated from their English versions, either by humans or machines, creating parallel datasets. To quantify the alignment between non-English languages (source) and English (target), we propose using the Correct Prediction Consistency (CPC) rate. This metric measures the proportion of questions correctly answered in both languages, with a higher CPC rate indicating better alignment. The results in Figure 3(b) demonstrate that CPC significantly improves after intervention by INCLINE, suggesting that INCLINE effectively aligns non-English representations with English ones for more accurate predictions. Notably, CPC for Swahili (sw) increases from 0.54 to 0.65 with INCLINE, showing its effectiveness for low-resource languages.

Intervening on hidden states yields the greatest performance improvements. We apply INCLINE to various components of LLMs, including the hidden states (INCLINE-HIDDEN), the outputs of attention heads (INCLINE-ATTN), the outputs of FFN blocks (INCLINE-FFN), and the embeddings (INCLINE-EMB). The results presented in Table 3 indicate that intervening on the hidden states (INCLINE-HIDDEN) leads to the most significant improvements across multilingual tasks. This finding suggests that hidden states can capture comprehensive semantic information that is crucial for cross-lingual alignment. While INCLINE-ATTN, INCLINE-FFN, and INCLINE-EMB also enhance performance, their performance gains vary across different tasks. These findings justify our design choice of using hidden states in INCLINE.

The value of α varies across languages and depends on language relatedness. In this study, we introduce α to control the strength of intervention in Equation 4. To investigate the impact of α, we conduct a grid search to find the optimal α values across the languages in `XStoryCloze`. We present the results for Spanish and Chinese in Figure 4. We observe that the optimal α values for these two languages are opposite: positive for Spanish and negative for Chinese. These findings suggest that the value of α is likely to depend on language relatedness, as both Spanish and English belong to the Indo-European language family, while Chinese belongs to the Sino-Tibetan language family. Results for more languages are provided in Appendix D.

6 Discussions

In this section, we conduct a series of experiments to investigate how variations in LLMs, model sizes,

	ar	el	es	fr	hi	ru	tr	vi	zh	AVG
BASELINE	66.59	15.30	48.52	67.86	71.97	35.66	12.38	40.40	56.11	46.09
INCLINE	68.68	15.63	50.79	69.93	76.92	37.95	12.42	43.11	58.27	48.19
INCLINE-FDEV	**73.95**	**15.76**	**56.11**	**75.84**	**77.85**	**39.33**	**12.92**	**46.49**	**60.19**	**50.94**

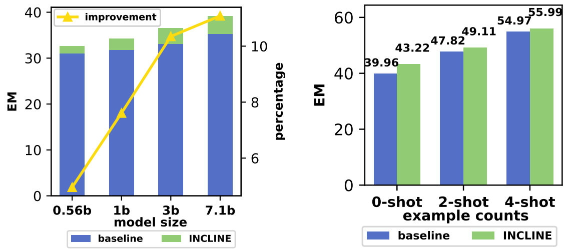
Table 5: The BLEU results of Flores dataset with INCLINE and INCLINE-FDEV.

in-context learning, and the data used for training alignment matrices affect our results. Additionally, we also explore using French as the target language (Appendix E).

INCLINE consistently enhances performance across multiple LLMs. To demonstrate the versatility of INCLINE across different LLMs, we apply it to another four high-performing models on the XStoryCloze task. As shown in Table 4, INCLINE consistently enhances performance compared to the BASELINE. Specifically, we observe increases of +4.96 for BLOOMZ-7B1-MT, +5.40 for LLAMA3.1-8B-INSTRUCT, +7.03 for LLAMA2-7B-CHAT, +8.64 for MISTRAL-7B-INSTRUCT, and +2.74 for FALCON-7B-INSTRUCT.

Larger LLMs benefit more from INCLINE. Building on the work of Wang et al. (2024b), who demonstrates a scaling relationship between the size of backbone models and their performance, we evaluate the impact of different model sizes within the BLOOMZ series on the MZsRE dataset. Our findings, illustrated in Figure 5(a), show that the relative performance gain of INCLINE over the baseline increases with the size of the backbone model. Specifically, the Exact Match (EM) scores (in the stacked columns) and the improvement percentages (in the line chart) indicate that larger models in the BLOOMZ series exhibit more significant enhancements when INCLINE is applied. This observation demonstrates that larger LLMs can benefit more from INCLINE.

INCLINE can further enhance model performance when combined with in-context learning. In-context learning (ICL) has been shown to improve the performance of LLMs on the MZsRE task (Wang et al., 2024b). Building upon this finding, we evaluate the effectiveness of combining INCLINE with ICL. As illustrated in Figure 5(b), INCLINE demonstrates enhanced performance, achieving an additional increase of +1.02 in average Exact Match (EM) score with four in-context examples compared to the baseline using ICL alone. While this improvement is smaller than the +3.26



(a) Various model sizes (b) In-context learning

Figure 5: (a) Exact Match (left y-axis) and relative improvements over the baseline (right y-axis) on MZsRE with respect to various model sizes of BLOOMZ. (b) Exact Match score for MZsRE dataset with INCLINE based on the zero-shot setting and few-shot settings given by BLOOMZ-7B1-MT.

increase observed in the zero-shot setting, it suggests that the benefits of INCLINE and ICL are complementary, with both methods capturing features from different perspectives. This highlights the versatility of INCLINE in various applications.

High-quality parallel sentences improve alignment in INCLINE. We explore how the quality of parallel sentences affects the performance of INCLINE. By default, the alignment matrices of INCLINE are trained using 500 random samples from the News Commentary dataset. To assess the impact of sentence quality, we also train the alignment matrices using 500 high-quality parallel sentences from the development set of Flores, which are carefully translated by professional human translators. We refer to this variant as INCLINE-FDEV. In Table 5, INCLINE-FDEV significantly outperforms both the standard INCLINE and BASELINE, highlighting the importance of high-quality parallel sentences.

7 Conclusion

In this paper, we introduce **Inference-Time Cross-Lingual Intervention (INCLINE)**, an innovative framework that bridges the performance gaps between high-performing and low-performing languages in LLMs. By training alignment matrices to transform source low-performing language representations into the target high-performing language representation space, INCLINE enhances performance on underrepresented languages without requiring additional training or fine-tuning of LLMs. Extensive experiments across nine benchmarks and five LLMs demonstrate that, INCLINE delivers significant improvements by up to +4.96 in terms of accuracy compared to strong baselines, while it only requires minimal computational costs.

8 Limitations

While INCLINE demonstrates significant enhancement for the multilingual tasks with cross-lingual intervention, the alignment matrices are trained for specific pairs of source and target languages. Future work will focus on developing multilingual alignment matrices that can accommodate multiple languages simultaneously, reducing the need for language pair-specific training and enhancing scalability. Implementing INCLINE requires access to the internal layers and representations of LLMs. For proprietary or closed-source models, or models accessible only through APIs without exposure of internal representations (e.g., GPT-4o), applying this method may not be feasible. How to perform cross-lingual alignment as a plug-and-play tool for all LLMs, including those with restricted access, requires further investigation.

Acknowledgement

This project has received funding from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant numbers 10039436].

The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

References

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. Qameleon: Multilingual qa with only 5 examples. *Transactions of the Association for Computational Linguistics*, 11:1754–1771.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *CoRR*, abs/2311.16867.

Anthropic. 2024. Claude 3.5 sonnet.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *CoRR*, abs/2310.20246.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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*. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *CoRR*, abs/2304.08177.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco GuzmÃ¡n, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne 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. *arXiv preprint arXiv:2305.07004*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popovic, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 1–42. Association for Computational Linguistics.

Somnath Kumar, Vaibhav Balloli, Mercy Ranjit, Kabir Ahuja, Tanuja Ganu, Sunayana Sitaram, Kalika Bali, and Akshay Nambi. 2024. Bridging the gap: Dynamic learning strategies for improving multilingual performance in llms. *CoRR*, abs/2405.18359.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3309–3325. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023a. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13171–13189. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 318–327. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*

2023, New Orleans, LA, USA, December 10 - 16, 2023.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, 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 T. Diab, Veselin Stoyanov, and Xian Li. 2021b. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021c. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024a. Hello gpt-4o.

OpenAI. 2024b. Learning to reason with llms.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. *arXiv preprint arXiv:1906.05407*.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. Bad-x: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Leonardo Ranaldi, Giulia Pucci, and André Freitas. 2023. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. *CoRR*, abs/2308.14186.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 15504–15522. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613. Association for Computational Linguistics.

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. 2022a. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022b. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 566–581. Association for Computational Linguistics.

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.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15894–15939. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 370–390. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024b. Retrieval-augmented multilingual knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. 2024c. Assessing factual reliability of large language model knowledge. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 805–819. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024d. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Demystifying multilingual chain-of-thought in process reward modeling. *arXiv preprint arXiv:2502.12663*.

Weixuan Wang, Jingyuan Yang, and Wei Peng. 2024e. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. *arXiv preprint arXiv:2203.09435*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

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.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, *SEM 2021, Online, August 5-6, 2021*, pages 229–240. Association for Computational Linguistics.

A Details of Datasets

The tasks and the corresponding output format, prompt template, evaluation metrics, the number of languages are shown in Table 6.

B Hyperparameters for SFT

We fine-tune all parameters of LLMs using the AdamW optimizer with a learning rate of 2×10^{-6} and a batch size of 4. This process is conducted over three epochs on four NVIDIA A100 GPUs (80GB). During training, we use a linear learning rate schedule with a warm-up phase that constitutes 10% of the total training steps.

C Detailed Results of Intervention

The detailed results of BASELINE, MT-GOOGLE, MT-LLM, SFT, ITI, CAA, INCLINE and SFT +INCLINE for each languages across discriminative and generative tasks are shown in Table 7.

D The value of α across languages

We explore the optimal value of α for each language in `XStoryCloze` using grid search, as shown in Figure 6.

E Projection to Non-English

We have demonstrated the effectiveness of INCLINE in aligning representations from non-English to English. To further prove the generalizability of INCLINE with another high-performing language, we conduct an ablation study aligning representations of various languages with French. As shown in Table 8, INCLINE enhances translation performance to non-English languages, with an average BLEU score increase of +5.35. This further demonstrates that INCLINE can effectively align representations across different languages.

F Details of Visualizing

Following Li et al. (2023b), we use Linear Regression to examine multilingual input representations. For each English and corresponding Portuguese sample from the News Commentary dataset (a total of 500 items), we extract the hidden states at the last token to create a probing dataset for each layer. We randomly divide this dataset into training and validation sets in a 4:1 ratio and fit a binary linear classifier to the training set. Similar to principal component analysis (PCA), we train a second linear probe on the same dataset, constrained to be orthogonal to the first probe. This orthogonality ensures that the two probes capture distinct aspects of the data. Finally, we project the hidden states of each sample in the `MZsRE` test set onto the directions defined by the probes from the last layer, allowing us to visualize and analyze the multilingual representations effectively.

G Supplementary Results of Training Data Volume

We find that increasing the amount of training data for learning alignment matrices can unexpectedly degrade performance in Figure 3(a). To investigate this observation, we examine the learned alignment matrices from the last layer, using varying numbers of parallel sentences. Interestingly, as shown in Table 9, the absolute values of the learned alignment weights consistently increase as the number of parallel sentences increases. It is well known that large weights provide the capacity for the network to fit the training data closely. As a result, the learned alignment matrices tend to fit more closely to the distribution of the parallel sentences. Consequently, the distribution shift between the parallel sentences and downstream tasks is enlarged as the number of parallel sentences grows.

Data	200	500	700	1000	1200	1400
max	0.57	0.84	1.01	1.31	1.43	1.57
min	-0.54	-0.73	-0.88	-1.29	-1.49	-1.71
avg	2.9e-7	1.0e-6	1.4e-6	1.9e-6	2.9e-6	3.6e-6
std	0.005	0.009	0.011	0.014	0.016	0.019

Table 9: The learned alignment matrices from the last layer with using varying numbers of parallel sentences.

Dataset	Output	prompt	Metric	|L|
XCOPA	2-way class	Here is a premise: "{premise}". A: "{choice1}" B: "{choice2}" What is the {question}? "A" or "B"?	acc.	10
XStoryCloze	2-way class	{input} What is a possible continuation for the story given the following options? A: {quiz1} B: {quiz2}’	acc.	8
XWinograd	2-way class	{input} Replace the _ in the above sentence with the correct option: - {option1} - {option2}	acc.	6
XNLI	3-way class	Take the following as truth: {premise} Then the following statement: "{hypothesis}" is "true", "false", or "inconclusive"?	acc.	13
XCSQA	multi-choice	Question: {question} {choice} Answer:	acc.	14
MZsRE	answer	{context} Quesion: {question} Answer:	EM	10
Flores	answer	Translate the following sentence from {language} to English: {input}	BLEU	10
WMT23	answer	Translate the following sentence from {language} to English: {input}	BLEU	5
MGSM	answer	Write a response that appropriately completes the request in {language}. Please answer in {language}. ### Instruction: {query}### Response:	EM	9

Table 6: The nine datasets used to evaluate multilingual intervention. |L| indicates the number of languages. EM is the Exact Match score and acc. represents the accuracy.

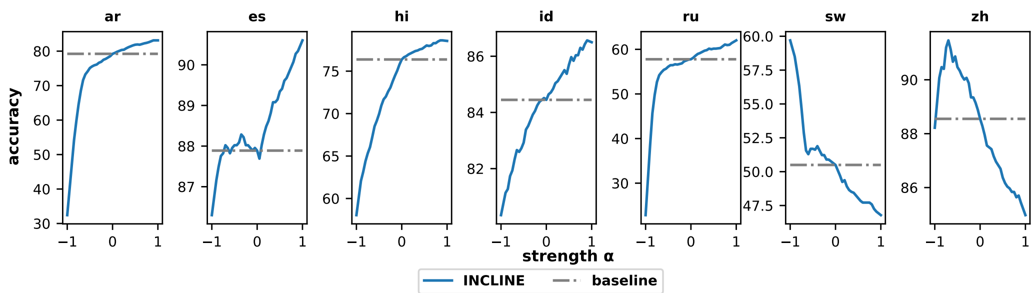


Figure 6: The accuracy changed with hyperparameter α on the XStoryCloze task.

Discriminative tasks															
XCOPA	en	et	id	it	sw	ta	th	tr	vi	zh					AVG
BASELINE	76.40	50.80	69.60	58.60	55.20	71.60	50.60	49.60	71.20	77.40					61.62
MT-GOOGLE	-	75.40†	75.00	76.00†	76.20†	62.20	62.40†	78.40†	76.40	77.80					73.31†
MT-LLM	-	44.80	69.80	59.40	60.20	71.20	47.40	51.20	61.60	73.00					59.84
SFT	86.40	50.60	78.40	67.80	59.00	77.20	47.60	53.00	83.00	84.60					66.80
ITI	-	50.80	70.80	60.00	55.40	63.20	49.00	50.60	69.00	79.40					60.91
CAA	-	51.20	72.20	61.20	59.20	73.00	52.20	52.00	74.80	79.80					63.96
INCLINE	-	**55.40**	**73.40**	**62.80**	**59.80**	**73.40**	**52.60**	**53.40**	**76.20**	**80.00**					**65.22**
SFT +INCLINE	-	**53.20**	**81.20†**	**65.80**	**60.80**	**85.00†**	**54.40**	**53.40**	**84.40†**	**85.00†**					**69.24**
XStoryCloze	en	ar	es	hi	id	ru	sw	zh							AVG
BASELINE	91.46	79.22	87.89	76.37	84.45	57.78	50.50	88.55							74.96
MT-GOOGLE	-	79.48	81.34	50.69	80.81	80.08†	77.04	86.96							76.63
MT-LLM	-	81.80	86.83	82.59	83.59	62.48	73.66	84.91							79.41
SFT	94.11	90.47	92.85	88.22	91.59	74.52	81.14	92.72							87.36
ITI	-	78.23	90.54	80.28	85.70	58.70	52.55	88.68							76.38
CAA	-	86.04	90.47	79.15	88.22	61.61	52.61	89.01							78.16
INCLINE	-	**83.12**	**90.60**	**81.47**	**86.10**	**67.24**	**59.70**	**91.20**							**79.92**
SFT +INCLINE	-	**90.93†**	**92.98†**	**89.08†**	**91.99†**	**76.77**	**81.93†**	**93.05†**							**88.11†**
XWinograd	en	fr	ja	pt	ru	zh									AVG
BASELINE	73.76	59.04	51.51	57.80	54.60	62.30									57.05
MT-GOOGLE	-	61.45	58.39†	59.32†	57.41	50.60									57.63
MT-LLM	-	54.22	47.86	33.08	42.22	37.70									43.02
SFT	78.06	62.65	14.91	43.35	58.09	39.89									43.78
ITI	-	54.22	51.51	57.79	14.60	63.10									48.24
CAA	-	60.24	52.87	58.17	57.14	63.69									58.42
INCLINE	-	**63.86†**	**53.18**	**58.56**	**57.46**	**63.69†**									**59.35†**
SFT +INCLINE	-	**63.86†**	**16.48**	**46.39**	**60.00†**	**62.50**									**49.84**
XCSQA	en	ar	de	es	fr	hi	it	ja	nl	pt	ru	sw	vi	zh	AVG
BASELINE	76.50	52.40	33.90	64.30	63.30	48.50	41.30	36.00	28.70	61.30	33.20	40.50	55.20	57.00	47.35
MT-GOOGLE	-	61.60†	65.00†	68.00†	67.20†	32.10	68.70†	57.30†	66.50†	66.90†	64.50†	19.60	62.90†	60.40†	58.52†
MT-LLM	-	32.30	26.30	42.70	42.30	30.40	25.60	25.60	17.40	39.90	21.60	24.00	31.60	39.80	30.73
SFT	65.70	48.20	32.90	54.10	53.60	43.10	40.40	32.60	29.00	53.60	29.90	31.80	48.40	50.80	42.18
ITI	-	52.10	34.20	64.50	63.70	48.10	40.00	25.90	26.00	61.20	33.50	40.90	54.90	57.20	46.32
CAA	-	52.80	34.10	64.50	63.30	48.40	42.20	36.40	29.30	62.80	33.50	41.90	56.00	58.40	47.97
INCLINE	-	**53.20**	**34.90**	**65.00**	**63.80**	**48.80†**	**42.90**	**36.80**	**29.80**	**62.60**	**33.80**	**42.20†**	**57.30**	**58.70**	**48.45**
SFT +INCLINE	-	**48.50**	**33.30**	**54.40**	**53.70**	**43.90**	**40.60**	**33.00**	**29.30**	**53.70**	**29.90**	**32.50**	**49.10**	**51.20**	**42.55**
XNLI	en	ar	de	el	es	fr	hi	ru	sw	th	tr	vi	zh		AVG
BASELINE	54.81	53.63	43.33	41.04	51.36	50.54	50.16	47.80	45.01	40.32	34.93	49.68	49.92		46.48
MT-GOOGLE	-	51.46	53.13	52.71	51.84	50.82	41.58	51.68	50.54	50.50	52.00	51.94	50.42		50.72
MT-LLM	-	46.87	43.25	36.29	52.12	51.40	45.31	42.08	43.43	34.07	33.17	47.23	48.42		43.64
SFT	86.37	77.17	68.10	59.48	82.71	81.48	72.42	66.87	67.15	54.55	49.80	77.62	78.76		69.68
ITI	-	53.69	45.37	41.36	50.18	51.20	50.34	47.74	43.35	38.98	35.77	48.96	48.86		46.32
CAA	-	53.59	44.67	41.62	52.83	52.75	50.28	34.40	45.75	40.48	36.41	50.32	50.92		46.17
INCLINE	-	**53.89**	**47.74**	**41.96**	**54.33**	**53.11**	**50.50†**	**49.22**	**45.99**	**41.28**	**37.17**	**51.12**	**51.16†**		**48.12**
SFT +INCLINE	-	**78.44†**	**71.02†**	**61.22†**	**83.07†**	82.14†	**73.85†**	**69.68†**	**69.14†**	55.69†	**51.60†**	**78.64†**	**79.52†**		**71.17†**
Generative tasks															
MZsRE	en	de	es	fr	pt	ru	th	tr	vi	zh					AVG
BASELINE	96.23	55.05	48.86	49.53	45.49	30.55	6.33	38.76	51.68	33.38					39.96
MT-GOOGLE	-	78.73†	76.18†	75.50†	71.74†	63.66†	78.47†	77.39†	60.97†	79.41†					73.56†
MT-LLM	-	49.13	54.78	51.28	6.86	2.69	9.69	40.92	34.72	48.59					33.18
ITI	-	53.84	44.41	43.34	41.99	19.11	6.59	38.63	46.70	32.17					36.31
CAA	-	57.07	53.30	52.36	52.76	31.49	7.13	39.43	55.05	38.36					42.99
INCLINE	-	**57.20**	**53.30**	**51.82**	**52.09**	**31.49**	**7.40**	**41.86**	**55.32**	**38.49**					**43.22**
Flores	en	ar	el	es	fr	hi	ru	tr	vi	zh					AVG
BASELINE	-	66.59	15.30	48.52	67.86	71.97	35.66	12.38	40.40	56.11					46.09
ITI	-	2.39	2.34	3.71	4.40	3.31	2.44	3.03	3.64	0.37					2.85
CAA	-	67.88	15.92	54.85	68.16	72.98	38.99	12.09	43.01	56.93					47.87
INCLINE	-	**73.95†**	**15.79†**	**56.11†**	**75.84†**	**77.85†**	**39.33†**	**12.92†**	**48.62†**	**60.19†**					**51.18†**
WMT23	en	de	ja	ru	uk	zh									AVG
BASELINE	-	18.26	10.17	14.73	11.36	14.39									11.78
ITI	-	2.75	1.79	2.32	1.66	3.16									2.34
CAA	-	16.96	10.22	15.11	11.54	14.86									13.74
INCLINE	-	**18.85†**	**10.30†**	**15.24†**	**11.71†**	**15.05†**									**14.23†**
MGSM	en	de	es	fr	ja	ru	sw	th	zh						
BASELINE	51.20	46.40	42.40	42.40	35.20	38.40	34.80	35.60	39.60						39.35
MT-GOOGLE	-	46.00	50.40†	47.20†	44.40†	46.80†	45.60†	45.60†	47.60†						46.70†
MT-LLM	-	20.40	38.80	32.40	10.80	18.40	22.00	1.60	26.80						21.40
ITI	-	46.00	43.20	44.80	35.60	40.00	36.80	34.80	42.80						40.50
CAA	-	42.40	42.00	40.00	34.40	40.80	36.20	34.40	45.20						39.43
INCLINE	-	**48.40†**	**46.80**	**45.20**	**37.60**	**44.80**	**38.00**	**38.80**	**43.20**						**42.85**

Table 7: The overall results of nine NLP tasks with multilingual intervention. † denotes the best results.

	en	ar	el	es	hi	ru	tr	vi	zh	AVG
BASELINE	45.11	44.70	15.37	39.37	50.18	36.99	10.51	38.77	42.20	35.91
INCLINE	**52.36**	**52.33**	**15.62**	**51.37**	**55.40**	**39.69**	**10.94**	**46.48**	**47.14**	**41.26**

Table 8: INCLINE on the Many-to-French translation task.