$\mathcal{X}$Transplant: A Probe into the Upper Bound Performance of Multilingual Capability and Culture Adaptability in LLMs via Mutual Cross-lingual Feed-forward Transplantation

Anonymous ACL submission

Abstract

Current large language models (LLMs) often exhibit imbalances in multilingual capabilities and cultural adaptability, largely attributed to their English-centric pretraining data. To address this imbalance, we propose $\mathcal{X}$Transplant, a novel probing method that investigates cross-lingual latent interactions through innovative cross-lingual feed-forward transplantation during inference. This approach enables models to leverage the strengths of both English and non-English languages. Through extensive pilot studies, we empirically demonstrate the significant potential of $\mathcal{X}$Transplant in improving both the multilingual capabilities and cultural adaptability of LLMs, respectively from `En` $\rightarrow$ `non-En` and `non-En` $\rightarrow$ `En`, highlighting the underutilization of current LLMs' multilingual potential. Building on these insights, we develop an offline scaling inference strategy that achieves consistent performance improvements in multilingual and culture-aware tasks, sometimes even surpassing multilingual supervised fine-tuning. This work advances our understanding of cross-lingual latent interactions in LLMs while offering a practical, training-free solution for enhancing multilingual performance and cultural adaptability.

1 Introduction

In recent years, large language models (LLMs) have showcased their remarkable versatility across a wide range of downstream tasks (Zhao et al., 2023; Liu et al., 2023; Dong et al., 2023; Wei et al., 2022a,b; Shanahan, 2022), as well as their evident generalizability and adaptability in multilingual scenarios. However, the significant imbalances in their multilingual capabilities and cultural adaptability still remain challenges that researchers are striving to resolve (Ye et al., 2023; Li et al., 2024a; Shi et al., 2024; Qin et al., 2024). These issues primarily stem from their unbalanced training corpora, which is predominantly in English, leading to these models being termed *English-centric* LLMs (Brown et al., 2020; Zhang et al., 2022; Touvron et al., 2023; Biderman et al., 2023).

Existing methods for these challenges primarily focus on *Multilingual Pretraining* and *Cross-lingual Transfer*. Multilingual Pretraining involves initially or continuously training models on diverse multilingual datasets to develop an overall improvement of their multilingual capabilities (Lin et al., 2021; Scao et al., 2022; Gao et al., 2024; Li et al., 2024b). While Cross-lingual Transfer leverages knowledge from high-resource languages to enhance the performance of low-resource languages through fine-tuning techniques (Reid and Artetxe, 2022; Cahyawijaya et al., 2023; Ye et al., 2023; Khurana et al., 2024). However, these training-based methods have shown potential limitations like "curse of multilinguality", a form of negative interference (Conneau et al., 2020; Wang et al., 2020), where expanding too much languages during training eventually leads to a decline.

These limitations and situations also place humans in a dilemma with current English-centric LLMs: given a certain question, (1) posing in English may overlook the language-specific neurons that is only activated by non-English inputs, potentially resulting in incomplete or inaccurate responses. On the other hand, (2) posing in non-English languages may fail to leverage the model's strong general capabilities in English, thereby affecting its overall performance. This naturally leads to a key consideration: *Can the LLMs leverage both their powerful general capabilities (in English) and their (non-English) multilingual knowledge during inference, to fully unlock their multilingual potential?*

In response to this, we introduce and investigate a probing method named $\mathcal{X}$Transplant to explore this possibility via mutual cross-lingual feed-forward transplantation. As illustrated in Figure 1, during the inference stage, $\mathcal{X}$Transplant transplants the feed-forward activations of certain decoder

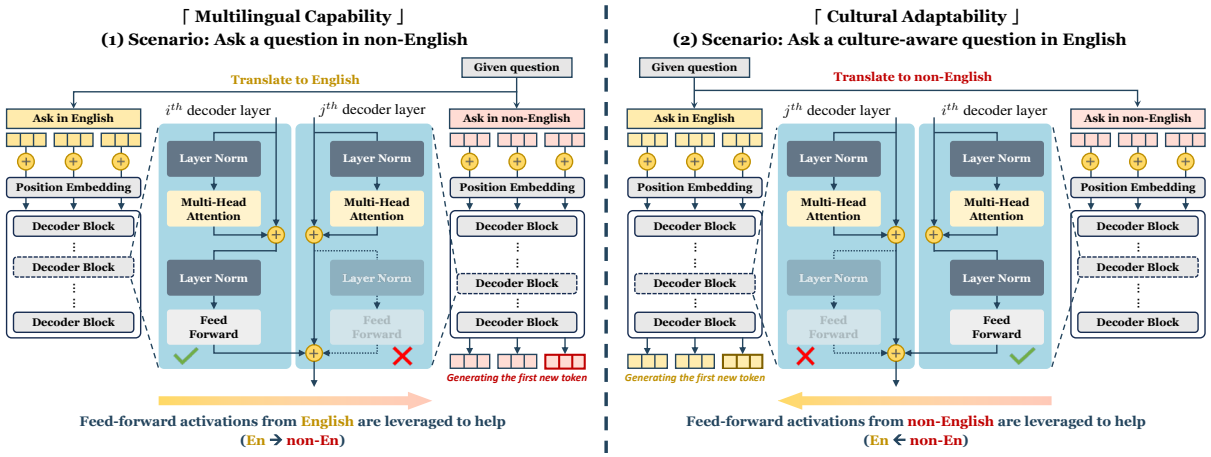

Figure 1: Overview of $\mathcal{X}$Transplant mechanism. The left illustrates the process where the feed-forward activations from English input are leveraged to help when asking a non-English question. The right illustrates how the feed-forward activations from non-English input are leveraged to help when asking a culture-aware question in English.

layer from one language into the inference process of input in another language, with forward propagation to proceed with the transplanted activations. The goal is to enable the model to leverage the strengths of both English and non-English languages. Through this probe, our study delves into two distinct avenues: the impact of $\texttt{En} \rightarrow \texttt{non-En}$ transplantation on LLMs' multilingual capabilities, and how $\texttt{non-En} \rightarrow \texttt{En}$ transplantation affects LLMs' cultural adaptability under English context.

(1) Through extensive pilot studies on 3 LLMs and 4 datasets, we assess the upper bound obtained through evaluating all configurations of $\mathcal{X}$Transplant to demonstrate its potential in significantly improving LLMs' performance on multilingual and culture-aware tasks, highlighting the underutilization of current LLMs' multilingual potential (§4). **(2)** Additionally, building on the findings derived from our pilot study, we further propose an offline scaling inference strategy, where underlying regularities are first extracted offline from small samples and then applied to larger, unseen data. This method yields consistent improvements across all involved LLMs and datasets, and occasionally even surpasses multilingual supervised fine-tuning (§5). **(3)** Besides, a series of targeted analysis, discussion and case study are also conducted to help gain deeper insights into the mutual latent cross-lingual interactions in $\mathcal{X}$Transplant (§6, A, D).

2 Background

In this section, we provide the background that motivates our research.

Feed-forward Layer Stores Factual Knowledge. The transformer-based GPT series of models have shown remarkable effectiveness in natural language generation (Radford et al., 2018; Brown et al., 2020), triggering a boom around LLMs. Within Transformer (Vaswani et al., 2017), the feed-forward layers and self-attention module constitute the main body of a decoder block for current LLMs.

Numerous studies have revealed the pivotal role of feed-forward layers in storing factual knowledge (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022). This insight motivates our exploration of $\mathcal{X}$Transplant on feed-forward layers to help LLMs fully leverage the knowledge from both English and non-English languages.

Language-specific Neurons. The intriguing capability of LLMs to understand and generate text in various languages is attributed to a subset of neurons that exhibit heightened activity for specific languages. Termed as "language-specific neurons", these components are critical to the multilingual capabilities of LLMs (Tang et al., 2024; Kojima et al., 2024). Furthermore, the proportion of these neurons is notably small, yet their targeted activation or deactivation significantly impacts the model's performance in corresponding languages. (Zhao et al., 2024). This finding has profound implications for enhancing LLMs' multilingual capabilities.

Building on above foundations regarding feed-forward layers and language-specific neurons, we boldly hypothesize that sharing and transferring feed-forward activations between English and non-English languages may allow the model to leverage the strengths of both language groups. This capacity to integrate advantages from diverse linguistic backgrounds serves as the foundation of our probing method—$\mathcal{X}$Transplant.

3 Probing Method — $\mathcal{X}$Transplant

In this section, we will present the formulation of $\mathcal{X}$Transplant, elaborate on its implementation details, and delineate several relevant concepts.

3.1 Methodology

For a model M with N decoder layers, given an original input x_s in source language S, the x_s undergoes a forward propagation through all decoder layers to predict the next token. Let the output activations of these N decoders be denoted as $O_s = \{o_s^k\}_{k=1}^N$, where each o_s^k is obtained by combining the feed-forward activations f_s^k and self-attention activations a_s^k through a residual connection. Similarly, for another translated version of x_s in target language T, denoted as x_t, we also have $O_t = \{o_t^k\}_{k=1}^N$ with corresponding $\{f_t^k\}_{k=1}^N$ and $\{a_t^k\}_{k=1}^N$. If without any modifications,they would predict the first new token $\hat{y}_s$ and $\hat{y}_t$ with the unembed matrix $W_{unembed}$ as follows:

$$\hat{y}_s = \text{softmax}(W_{unembed} \cdot (a_s^N + f_s^N)) \quad (1)$$

$$\hat{y}_t = \text{softmax}(W_{unembed} \cdot (a_t^N + f_t^N)) \quad (2)$$

Our mechanism, $\mathcal{X}$Transplant, refines the process by transplanting the feed-forward activations from the i^{th} decoder layer with input x_s to the j^{th} decoder layer with input x_t. Formally, f_t^j is replaced with f_s^i and the forward propagation of prompting x_t then continues with this modification. Consequently, the original $\{o_t^k\}_{k=j}^N$ will be altered into $\{\tilde{o}_t^k\}_{k=j}^N$ due to the update in f_t^j, leading to new prediction outcomes $\hat{y}_t^{(\text{modified})}$ as follows:

$$\hat{y}_t^{(\text{modified})} = \text{softmax}(W_{unembed} \cdot \tilde{o}_t^N) \quad (3)$$

Notably, $\mathcal{X}$Transplant currently considers only the substitution of feed-forward activations from a single layer, meaning that the aforementioned i^{th} layer and j^{th} layer both refer to a certain, single decoder layer. $\mathcal{X}$Transplant performs the transplantation only during the forward propagation for predicting the first new token; all subsequent tokens are generated iteratively after the first one, without any additional transplantation operations.

3.2 Mutual Transplantation

Section 3.1 details how $\mathcal{X}$Transplant facilitates the transfer of feed forward activations from language S to language T. But $\mathcal{X}$Transplant actually supports transplantation in two directions. When prompting in non-English, the feed-forward activations from English can be leveraged to help the process of non-English prompting. Similarly, under the English prompting conditions, the feed-forward activations from non-English languages can be leveraged to help. Specifically, our experiments explore the dual attempt of $\mathcal{X}$Transplant: `En` $\rightarrow$ `non-En` and `non-En` $\rightarrow$ `En`.

3.3 Instance-ware Upper Bound

For a model M with N decoder layers, both the source layer and target layer selections in $\mathcal{X}$Transplant offer N possible choices, resulting in N^2 potential transplantation combinations. For a dataset D of a certain size, we conducted $\mathcal{X}$Transplant for each sample across all N^2 possibilities, selecting the optimal solution for each instance. The model's optimal performance on this dataset, derived from this process, is referred to as the instance-aware upper bound.

We denote $M_{S_i \rightarrow T_j}(x)$ as the output of model M towards question x after applying $\mathcal{X}$Transplant from i^{th} layer of language S to the j^{th} layer of language T. Let y_{true} represents the gold answer of question x and $\mathbb{I}(\cdot)$ is a indicator function that equals 1 if the condition is true, 0 otherwise. The upper bound performance is formulated as follows:

$$\text{UpperBound}_{\text{S}\rightarrow\text{T}}(M, D) = \sum_{x \in D} \max_{i,j \in \{1,\ldots,N\}} \mathbb{I}(M_{S_i \rightarrow T_j}(x) = y_{true}) \quad (4)$$

Though N^2 enumeration is time-consuming, our goal is to benchmark the upper bound performance of LLMs achievable through $\mathcal{X}$Transplant.

4 Pilot Study

In this section, we explore the upper bound performance of multilingual capability and culture adaptability in LLMs via mutual $\mathcal{X}$Transplant operation.

4.1 Setup

Models. We selected 3 typical LLMs for our pilot experiments. (1) *LLaMA-2-7B-Chat*, (2) *Mistral-7B-Instruct-v0.3*, (3) *Qwen2-7B-Instruct*.

Datasets. We mainly conduct experiments on 4 benchmarks, which can be categorized into:

- **Multilingual Capability:** (1) *XNLI* (Conneau et al., 2018), a natural language inference corpus, (2) *XQuAD* (Artetxe et al., 2020), a question answering dataset, and (3) *XCOPA* (Ponti et al.,

Models	Dataset: XNLI (PilotSet)															
	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
LLaMA-2-7B-Chat	60.0	34.0	26.0	50.0	30.0	36.0	46.0	8.00	46.0	14.0	0.00	34.0	0.00	28.0	40.0	30.1
UpperBound$_{\text{En2lang}}$	**94.0**	**90.0**	**96.0**	**100**	**96.0**	**84.0**	**100**	**60.0**	**98.0**	**82.0**	**66.0**	**74.0**	**34.0**	**84.0**	**100**	**83.9**
Mistral-7B-Instruct-v0.3	46.0	6.00	56.0	50.0	40.0	60.0	48.0	30.0	52.0	0.00	32.0	36.0	14.0	46.0	50.0	37.7
UpperBound$_{\text{En2lang}}$	**80.0**	**72.0**	**64.0**	**76.0**	**98.0**	**78.0**	**82.0**	**84.0**	**78.0**	**36.0**	**88.0**	**82.0**	**66.0**	**78.0**	**92.0**	**76.9**
Qwen2-7B-Instruct	82.0	52.0	54.0	56.0	52.0	68.0	70.0	50.0	64.0	26.0	48.0	50.0	32.0	60.0	64.0	55.2
UpperBound$_{\text{En2lang}}$	**94.0**	**70.0**	**74.0**	**80.0**	**66.0**	**82.0**	**90.0**	**62.0**	**84.0**	**84.0**	**62.0**	**78.0**	**56.0**	**78.0**	**86.0**	**76.4**

Models	Dataset: XQuAD (PilotSet)												
	en	ar	de	el	es	hi	ro	ru	th	tr	vi	zh	Avg
LLaMA-2-7B-Chat	64.0	8.00	56.0	12.0	60.0	8.00	42.0	42.0	6.00	24.0	40.0	40.0	33.5
UpperBound$_{\text{En2lang}}$	**92.0**	**34.0**	**80.0**	**38.0**	**84.0**	**32.0**	**74.0**	**82.0**	**30.0**	**64.0**	**66.0**	**70.0**	**62.2**
Mistral-7B-Instruct-v0.3	64.0	38.0	42.0	20.0	54.0	32.0	48.0	44.0	20.0	38.0	40.0	38.0	39.8
UpperBound$_{\text{En2lang}}$	**90.0**	**54.0**	**76.0**	**50.0**	**78.0**	**50.0**	**80.0**	**72.0**	**50.0**	**68.0**	**66.0**	**76.0**	**67.5**
Qwen2-7B-Instruct	76.0	52.0	40.0	22.0	48.0	18.0	36.0	48.0	38.0	46.0	64.0	80.0	47.3
UpperBound$_{\text{En2lang}}$	**94.0**	**76.0**	**78.0**	**52.0**	**78.0**	**58.0**	**76.0**	**82.0**	**64.0**	**78.0**	**90.0**	**94.0**	**76.7**

Models	Dataset: XCOPA (PilotSet)											
	en	et	ht	id	it	sw	ta	th	tr	vi	zh	Avg
LLaMA-2-7B-Chat	60.0	44.0	10.0	50.0	30.0	0.00	0.00	54.0	46.0	58.0	56.0	37.1
UpperBound$_{\text{En2lang}}$	**94.0**	**58.0**	**60.0**	**100**	**100**	**54.0**	**60.0**	**56.0**	**100**	**78.0**	**100**	**78.2**
Mistral-7B-Instruct-v0.3	40.0	22.0	56.0	66.0	72.0	16.0	0.00	56.0	54.0	70.0	70.0	47.5
UpperBound$_{\text{En2lang}}$	**94.0**	**76.0**	**92.0**	**88.0**	**92.0**	**54.0**	**28.0**	**72.0**	**80.0**	**86.0**	**74.0**	**76.0**
Qwen2-7B-Instruct	0.00[1]	44.0	52.0	86.0	88.0	62.0	36.0	50.0	28.0	90.0	84.0	56.4
UpperBound$_{\text{En2lang}}$	**90.0**	**98.0**	**94.0**	**94.0**	**100**	**88.0**	**100**	**90.0**	**94.0**	**96.0**	**98.0**	**94.7**

Table 1: Performance comparisons between LLMs' original performance and the upper bound results of $\mathcal{X}$Transplant on multilingual tasks. UpperBound$_{\text{En2lang}}$ represents $\mathcal{X}$Transplant from English to involved language.

2020), a causal commonsense reasoning dataset. These datasets consist of linguistically parallel questions to assess the model's ability across languages. For questions in non-English languages, we apply En $\rightarrow$ non-En $\mathcal{X}$Transplant to harness feed-forward activations from English.

- **Cultural Adaptability:** *GlobalOpinionQA* contains QAs from cross-national surveys designed to capture diverse opinions on global issues across different countries, all in English. This dataset aims to evaluate the model's cultural adaptability within an English context. For these questions in English, we apply non-En $\rightarrow$ En $\mathcal{X}$Transplant, hoping the model to leverage feed-forward activations from non-English languages to better capture cultural nuances.

Notably, due to the extensive scale of our pilot experiments[2], for each dataset, we randomly sampled 50 instances in each language involved, creating our small but linguistically balanced *PilotSets* (Appendix B.1). And details of evaluation and hype-settings can be found in Appendix B.2.

[1]The explanation of accuracy in English subset of *XCOPA* for *Qwen2-7B-Instruct* is in Appendix B.3.

[2]To obtain the instance-aware upper bound of $\mathcal{X}$Transplant, we perform inference on all N^2 possible source and target layer selection strategies for each instance (for example, in *LLaMA-2-7B-Chat* with layer number $N = 32$, $N^2 = 1024$ times inference are conducted for each instance). Our pilot experiments involves 3 LLMs and 4 pilotsets, resulting in **over 800 hours** of computation on 8 * A800-SXM4-80GB.

4.2 Observations

We compare the UpperBound results of $\mathcal{X}$Transplant with the original performance of LLMs. The main results of the multilingual datasets are presented in Table 1 and the results for the cultural dataset are illustrated in Figure 2. The comparisons are used to **illustrate the extent to which multilingual potential can be unlocked through the $\mathcal{X}$Transplant mechanism without modifying LLM itself**. Next, we present our main findings as follows.

(1) Underutilization of current LLMs' multilingual potential. The results in Table 1 and Figure 2 show that the upper-bound performance of $\mathcal{X}$Transplant is surprisingly much higher than the LLMs' original performance. The substantial performance gap indicates that these models harbor significant, yet underutilized, potential for advancement through targeted interventions (the feed-forward activations from other language). Furthermore, these findings highlight that the cross-lingual latent interactions facilitated by $\mathcal{X}$Transplant represent a highly promising direction for extending the boundaries of LLM performance in multilingual and culture-aware tasks.

(2) Feed-forward activations from English boosts multilingual capability, while those from non-English improves cultural adaptability. $\mathcal{X}$Transplant supports transplantation in two direc-

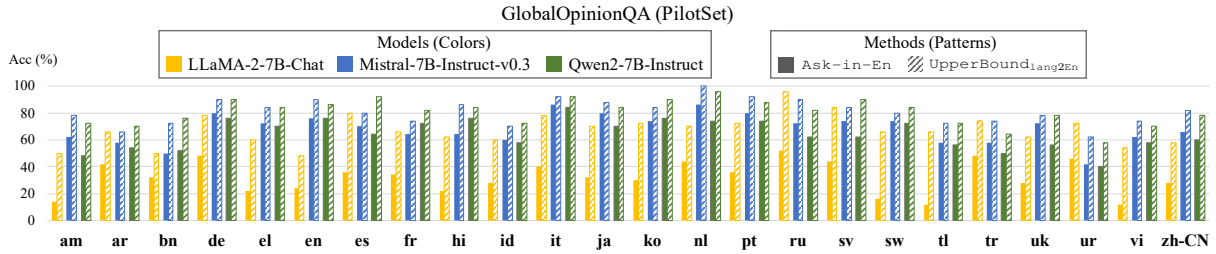

Figure 2: Performance comparisons between LLMs' original performance and the upper bound results of $\mathcal{X}$Transplant on culture-aware task. `Ask-in-En` represents LLMs' original culture-aware performance under English context, while $\texttt{UpperBound}_{\texttt{lang2En}}$ represents $\mathcal{X}$Transplant from non-English language to English.

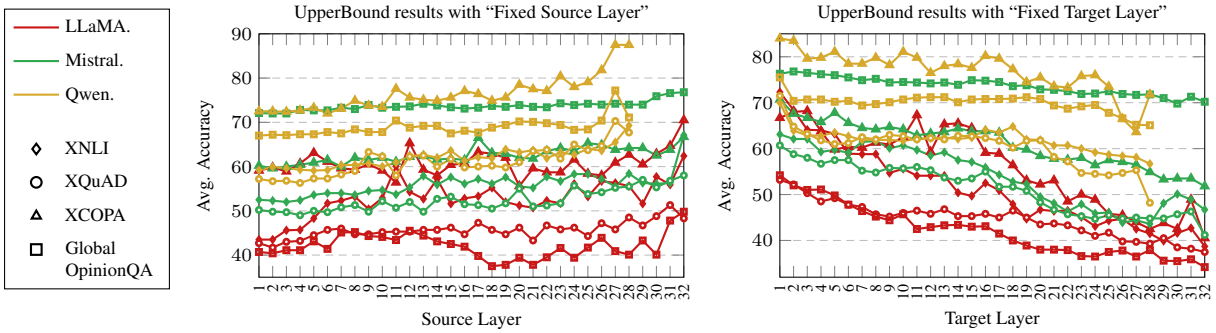

Figure 3: Layer-wise upper bound results across different LLMs and *PilotSets*. The figure on the left represents the upper bound results when source layer is fixed a certain layer and target layer varies; while the figure on the right represents the upper bound results when target layer is fixed a certain layer and source layer varies.

tions: `En` $\rightarrow$ `non-En` for multilingual tasks and `non-En` $\rightarrow$ `En` for culture-aware task. The results underscore the effectiveness of $\mathcal{X}$Transplant in both aspects, demonstrating that the feed-forward activations from English tend to strengthen the model's multilingual generalization, while feed-forward activations from non-English allow for deeper understanding of culturally specific content. This mutual attempt reveals the complementary strengths of English and non-English activations in optimizing performance on multilingual and culture-aware tasks. And in Table 2, we also observe the improvements under `En2En` setting, which is further discussed in Appendix B.4.

4.3 Layer-wise Selection Patterns

Another key observation from our pilot experiments is that the performance gains depend heavily on the choice of source and target layers. In this section, we explore the layer-wise patterns that govern the effectiveness of $\mathcal{X}$Transplant.

The upper bound results in Table 1 are obtained through all N^2 answers of $\mathcal{X}$Transplant. Here in Figure 3, we present the *layer-wise upper bounds* (Appendix D.4), where we fix either the source or target layer and the other layer is varied across N configurations. The following pattern emerges:

Last-layer as the source and first-layer as the target yield superior upper bound results. The layer-wise upper bound results in Figure 3, consistent across all models and datasets, reveal a clear trend: when the source layer is fixed, the highest upper bound performance across the N possible target layers is achieved when the source layer corresponds to the last layer. Similarly, when the target layer is fixed, the highest upper bound performance across the N possible source layers is observed when the target layer corresponds to the first layer.

Moreover, in both scenarios, the layer-wise upper bound results are close to the overall upper bound, which suggests that $\mathcal{X}$Transplant can be simplified to operate within a N-size space: (1) fixing the source layer to the last layer and varying the target layer, or (2) fixing the target layer to the first layer and varying the source layer.

5 Applying $\mathcal{X}$Transplant: Experiments

Our pilot studies empirically demonstrate the promising potential of $\mathcal{X}$Transplant. In this section, we explore the application of $\mathcal{X}$Transplant on **Unseen data** (Appendix C.2), which refers to the data points that are not included in the *PilotSet*.

Method	Multilingual Capability									Cultural Adaptability		
	XNLI (Unseen)			XQuAD (Unseen)			XCOPA (Unseen)			GlobalOpinionQA (Unseen)		
	LLaMA.	Mistral.	Qwen.	LLaMA.	Mistral.	Qwen.	LLaMA.	Mistral.	Qwen.	LLaMA.	Mistral.	Qwen.
Baselines												
`Original`	28.1	37.4	54.1	33.1	38.6	44.9	35.4	46.2	54.2	33.7	68.3	62.5
`CoT`	18.8	28.9	40.7	22.0	23.5	42.2	27.7	25.0	36.5	18.0	43.0	46.4
`PIM`	14.8	**52.4**	**63.2**	34.1	43.1	48.3	18.5	**69.9**	32.0	11.9	58.4	55.6
`ML-SFT`	32.4	38.0	46.2	**39.8**	**47.1**	**57.1**	44.9	56.2	50.0	**39.4**	64.5	60.2
Ours												
$\mathcal{X}$Transplant-`SL`	29.9	41.5	52.6	29.7	36.4	43.8	42.6	48.5	64.1	28.7	66.7	61.1
$\mathcal{X}$Transplant-`TF`	34.5	39.8	56.2	33.3	45.3	50.9	41.6	52.1	59.9	35.7	**68.8**	62.5
$\mathcal{X}$Transplant-`OA`	**34.8**	43.2	56.1	34.3	44.6	49.8	**48.4**	55.5	**71.1**	36.6	68.7	**63.2**

Table 2: Main results of three offline scaling inference strategies of $\mathcal{X}$Transplant compared with other baselines. Blue cell indicates better performance than the original, while Gray cell indicates the opposite. **Bold** and underline numbers indicate the best performance and second-best performance. LLaMA., Mistral. and Qwen. respectively represent *LLaMA-2-7B-Chat*, *Mistral-7B-Instruct-v0.3* and *Qwen2-7B-Instruct*.

Implementation. We first identify, for each model and each language set within each dataset, the optimal source and target layer pair from all N^2 combinations $\{(i, j) \mid i, j \in \{1, 2, \ldots, N\}\}$, based on performance observed in our pilot experiments. This optimal pair is then applied to the corresponding dataset's unseen data for further evaluation. Moreover, leveraging our findings from Section 4.3, we also experiment with two additional configurations: selecting the best-performing pair from the "source-last" set $\{(N, j) \mid j \in \{1, 2, \ldots, N\}\}$ and from the "target-first" set $\{(i, 1) \mid i \in \{1, 2, \ldots, N\}\}$. The above three strategies of $\mathcal{X}$Transplant are denoted as $\mathcal{X}$Transplant-`OA` (**O**ver**A**ll), $\mathcal{X}$Transplant-`SL` (**S**ource**L**ast) and $\mathcal{X}$Transplant-`TF` (**T**arget**F**irst). The selected pairs can be found in Appendix C.1.

Baselines. (1) `Original` performance of LLMs, (2) `CoT` (Wei et al., 2022b), which prompts the models with step-by-step reasoning to further unlock its potential, (3) `PIM` (Mu et al., 2024), which concatenates prompts in two languages to enhance multilingual performance and (4) `ML-SFT`, which boosts multilingual capabilities by additional multilingual supervised fine-tuning. The implementation details are in Appendix C.3.

Results. Average results across different languages or cultures of three offline scaling inference strategies compared with other baselines are illustrated in Table 2.

(1) Existing methods struggle to achieve consistent improvements. As shown in Table 2, `CoT` performs poorly in multilingual and culture-aware scenarios. And while `PIM` and `ML-SFT` can achieve certain improvements, these gains are not consistent across all involved LLMs and datasets. Additionally, we find that `PIM` occasionally performs best across all methods, but this actually comes at the cost of **significant language consistency** issues, as we discussed latter in Section 6.1.

(2) $\mathcal{X}$Transplant yields great improvements on unseen data, even surpassing multilingual SFT. Both $\mathcal{X}$Transplant-`OA` and $\mathcal{X}$Transplant-`TF` can achieve consistent improvements on unseen data. And the results on *XNLI* and *XCOPA* demonstrate that $\mathcal{X}$Transplant can even outperform the gains achieved through `ML-SFT`, which also suggests $\mathcal{X}$Transplant as a brand new direction for extending the performance boundaries of LLMs, distinct from traditional training-based approaches.

(3) A significant gap to the overall upper bound. While $\mathcal{X}$Transplant achieve certain improvements, there still remains a substantial gap to the upper bound results from our pilot experiments. This indicates that our method is relatively coarse-grained, and an adaptive instance-aware strategy that selects the optimal layer pair for each question may help better approach the upper bound (as further discussed in Appendix D.5).

Summary. $\mathcal{X}$Transplant-`OA` or -`TF` can be regarded as an effective offline scaling inference strategy. By conducting offline pilot study on small-scale samples and identifying the optimal source-target layer pair, we can apply it to larger-scale unseen data to achieve consistent improvements. While $\mathcal{X}$Transplant-`OA` involves an N^2-scale computational cost which may be too high, $\mathcal{X}$Transplant-`TF` reduces this to N, significantly lowering computational overhead while maintaining consistent performance gains.

Language Consistency (%)	XNLI (non-En)	XQuAD (non-En)	XCOPA (non-En)	GlobalOpinionQA (En)
LLaMA-2-7B-Chat	95.20	83.00	86.93	99.83
— PIM	59.75	77.05	84.51	89.35
— $\mathcal{X}$Transplant	95.23	88.21	93.69	99.74
Mistral-7B-Instruct-v0.3	88.13	91.83	84.91	100.0
— PIM	63.07	86.67	85.45	90.75
— $\mathcal{X}$Transplant	94.36	96.50	85.95	99.97
Qwen2-7B-Instruct	95.20	99.50	88.36	100.0
— PIM	91.23	96.67	77.55	97.10
— $\mathcal{X}$Transplant	97.43	99.22	87.09	99.92

Table 3: The input-output language consistency results of three LLMs with PIM and $\mathcal{X}$Transplant, compared with their original language consistency. non-En and En represent the input-output language required by corresponding tasks.

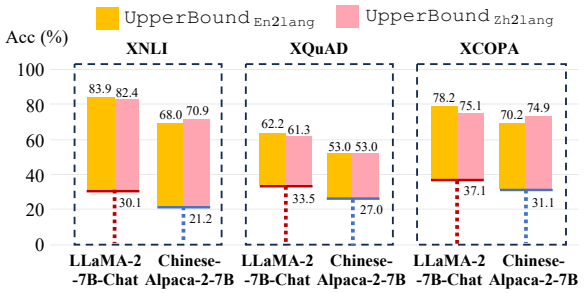

Figure 4: The upper bound results of English- and Chinese-centric LLM achieved by $\mathcal{X}$Transplant from English ($\texttt{UpperBound}_{\texttt{En2lang}}$) and Chinese ($\texttt{UpperBound}_{\texttt{Zh2lang}}$). The horizontal line represents the model's original performance.

6 Further Analysis

In this section, we delve deeper into $\mathcal{X}$Transplant through a series of targeted analysis.

6.1 Input and Output Language Consistency

$\mathcal{X}$Transplant benefits LLMs by leveraging feed-forward activations from inputs in other languages. To investigate whether these activations induce language shifts (i.e., output language differing from input language), we analyzed the input-output consistency across all N^2 answers of $\mathcal{X}$Transplant.

The language consistency results[3] shown in Table 3 demonstrate that, the PIM method, leveraging multilingual contexts, often introduces input-output inconsistencies. But the average consistency results across all N^2 answers of $\mathcal{X}$Transplant align well with that observed under original setting. This indicates that $\mathcal{X}$Transplant rarely affect the language consistency, making language shifts unlikely. This also provides a foundational guarantee for the upper bound results in Section 4.

6.2 Generalizability from English- to Chinese-centric LLM

Our experiments mainly focus on English-centric LLMs, revealing the benefit of feed-forward activations from English. In this section, we further explore the generalizability of this finding by comparing the upper bound results of $\mathcal{X}$Transplant on *LLaMA-2-7B-Chat* and *Chinese-Alpaca-2-7B*[4].

Not only activations from English can help. As shown in Figure 4, we find that for both English- and Chinese-centric LLMs, the feed-forward activations from either English or Chinese results in upper bound result that far exceeds the LLMs' original performance, without being confined to English as the only source language.

Native preference in Native-centric LLM. Figure 4 further reveals that, for *LLaMA-2-7B-Chat*, the English-centric LLM, activations from English result in a higher upper bound in $\mathcal{X}$Transplant than those from Chinese (En: 74.8%, Zh: 72.9% in average). Meanwhile, in *Chinese-Alpaca-2-7B*, the Chinese-centric LLM, activations from Chinese can offer greater improvements (En: 63.7%, Zh: 66.3% in average). This indicates a native preference, where feed-forward activations from the models' centric language tend to yield more substantial gains, likely due to the closer alignment with the model's internal knowledge.

6.3 Impact on English Performance

In Table 2, we present the average results of all involved languages. And in this section, we conduct an analysis towards investigate $\mathcal{X}$Transplant's impact on models' English capability compared with other involved baselines, the results are in Table 4.

All methods suffer a decline in English capability, but $\mathcal{X}$Transplant shows the mildest symptoms. The results in Table 4 reveal that although many methods lead to some improvements in average performance across different languages (as seen in Table 2 Section 5), they also tend to worsen the model's English capability to some extent. In particular, ML-SFT achieves great performance improvements in other non-English languages but causes the most significant decline in English performance. However, it is noticeable that, compared

[3]The languages are identified by *lid.176.bin* model from *fasttext*, which can recognize 176 languages.

[4]A *LLaMA-2-7B* based Chinese-centric model.

Method	English Subset								
	XNLI (Unseen)			XQuAD (Unseen)			XCOPA (Unseen)		
	LLaMA.	Mistral.	Qwen.	LLaMA.	Mistral.	Qwen.	LLaMA.	Mistral.	Qwen.
Baselines									
Original	47.3	40.0	83.2	**70.8**	73.5	**76.5**	53.6	48.0	0.00[5]
CoT	33.7	60.0	71.7	64.5	55.4	70.6	61.8	44.7	5.33
PIM	45.1	**63.1**	83.6	68.2	71.2	72.3	63.1	72.2	22.4
ML-SFT	31.8	39.6	44.2	23.5	33.2	60.2	68.4	**86.4**	2.67
Ours									
$\mathcal{X}$Transplant-SL	46.6	46.8	79.7	64.5	68.7	68.9	**72.9**	55.8	63.6
$\mathcal{X}$Transplant-TF	46.9	45.3	**84.8**	70.2	**77.6**	75.1	52.7	70.7	17.6
$\mathcal{X}$Transplant-OA	**48.1**	46.8	**84.8**	70.2	76.5	73.3	61.1	84.2	**77.6**

Table 4: Results on English subset of three offline scaling inference strategies of $\mathcal{X}$Transplant compared with other baselines. Blue cell indicates better performance than the original, while Gray cell indicates the opposite. **Bold** and underline numbers indicate the best performance and second-best performance.

to other methods, $\mathcal{X}$Transplant exhibits a relatively mild decline in English capability and, in many cases, still manages to achieve performance improvements in English.

More analysis. Further analysis towards **(1)** the outcomes of $\mathcal{X}$Transplant, **(2)** the stability and reliability of $\mathcal{X}$Transplant and **(3)** a case study from the perspective of intermediate decoding can be found in the Appendix D.1, D.2 and D.3. And we provide Appendix A to emphasize some key aspects and offer clarifications for potential questions.

7 Related Work

Multilingual Capability. Early multilingual models like mBERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) laid the groundwork for extending pretrained models across diverse languages. Recently larger multilingual models, such as Bloom (Scao et al., 2022) and Mala-500 (Lin et al., 2024), enhance multilingual capabilities through increased scale. Generally, multilingual pretraining and finetuning are now the two mainstream methods for improving multilingual performance. Works like Li et al. (2024b) injects multilingual alignment and preserves this during pretraining. Gao et al. (2024) explored the effect of multilingual pretraining and instruction tuning on the degree of alignment. Models like Sabia (Pires et al., 2023), ChineseLLaMA (Cui et al., 2023), ChineseMixtral (HIT-SCIR, 2024) are products of continuous pretraining on existing English-centric LLMs. Other like BLOOMz (Muennighoff et al., 2022), m-LLaMA (Zhu et al., 2023), Phoenix (Chen et al., 2023) chosen to directly incorporate multilingual data in the supervised finetuning stage to achieve implicit multilingual alignment across languages.

Cultural Adaptability. Previous studies have shown that current LLMs exhibit poor cultural adaptability (Ramezani and Xu, 2023; Jha et al., 2023; Rao et al., 2024). Solutions towards these culture-aware challenges can be categorized mainly into two approaches: context learning and training-based. Kovač et al. (2023) studied models' controllability in inducing cultural perspectives, while Wang et al. (2024) improved cultural performance by explicitly prompting LLMs with the recognition of culture in queries. Rao et al. (2023) developed a framework integrating moral dilemmas with principles from various normative ethics formalisms across different levels of abstraction. Rao et al. (2023) developed a framework integrating ethics from diverse cultures. Another line of research involves fine-tuning models on large-scale culturally relevant datasets (Abbasi et al., 2023; Lin and Chen, 2023; Nguyen et al., 2024; Shi et al., 2024), or investing in more balanced multilingual corpus for pretraining (Scao et al., 2022; Lin et al., 2024; Gao et al., 2024; Li et al., 2024b).

Unlike previous training-based approaches, $\mathcal{X}$Transplant directly modifies the model's internal activations during inference, allowing the model to benefit from both English and non-English inputs. This simple yet promising mechanism marks a new step forward in cross-lingual capability transfer.

8 Conclusion

This work introduces $\mathcal{X}$Transplant, a mechanism that contributes to further unlocking the multilingual potential of LLMs, as well as their cultural adaptability, via mutual cross-lingual feed-forward transplantation. Our extensive pilot studies across representative LLMs and datasets, along with established upper bounds, highlight the underutilization of current LLMs' multilingual potential and demonstrate the effectiveness of $\mathcal{X}$Transplant in both multilingual and culture-aware tasks. Additionally, the offline scaling inference strategy we proposed, motivated by the underlying regularities observed in our pilot studies, could yield consistent improvements across all involved LLMs and datasets, and occasionally even outperforms multilingual supervised fine-tuning. We hope $\mathcal{X}$Transplant will serve as a catalyst for future research, driving continued progress in developing more linguistically effective and culturally aware language models.

[5]The explanation of accuracy in English subset of *XCOPA* for *Qwen2-7B-Instruct* is in Appendix B.3.

Limitations

This work exhibits several limitations worth noting. Firstly, while our exploration of the model's multilingual potential upper bound is grounded in extensive pilot experiments, it remains an empirical conclusion, lacking formal theoretical proof. Secondly, we have not explored more complex or fine-grained transformations of $\mathcal{X}$Transplant method, such as experimenting with multi-layer operations or conducting more refined manipulations of feed-forward activations across different languages, rather than the simple replacement approach used in our study. These avenues offer significant opportunities for future extensions of $\mathcal{X}$Transplant. Thirdly, due to the computational constraints, we did not conduct comparisons between LLMs of different model sizes (particularly larger models), resulting in a lack of insights into the impact of model capacity on performance.

References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. 2023. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

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. *arXiv preprint arXiv:2304.08177*.

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)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2023. A survey for in-context learning. *ArXiv preprint*, abs/2301.00234.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *arXiv preprint arXiv:2404.04659*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana,

Dominican Republic. Association for Computational Linguistics.

HIT-SCIR. 2024. Chinese-mixtral-8x7b: An open-source mixture-of-experts llm. https://github.com/HIT-SCIR/Chinese-Mixtral-8x7B.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Sameer Khurana, Nauman Dawalatabad, Antoine Laurent, Luis Vicente, Pablo Gimeno, Victoria Mingote, and James Glass. 2024. Cross-lingual transfer learning for low-resource speech translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *arXiv preprint arXiv:2404.02431*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Jiahuan Li, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2024b. Prealign: Boosting cross-lingual transfer by early establishment of multilingual alignment. *arXiv preprint arXiv:2407.16222*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

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

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, et al. 2024. Large language models are parallel multilingual learners. *arXiv preprint arXiv:2403.09073*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer.

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

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Machel Reid and Mikel Artetxe. 2022. On the role of parallel data in cross-lingual transfer learning. *arXiv preprint arXiv:2212.10173*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Murray Shanahan. 2022. Talking about large language models. *ArXiv preprint*, abs/2212.03551.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *ArXiv preprint*, abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *arXiv preprint arXiv:2306.06688*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Potential Questions and Explanations

1. **The reason for applying $\mathcal{X}$Transplant only when generating the first new token?**

   In autoregressive generation, applying $\mathcal{X}$Transplant during the generation of the first new token essentially introduces the benefit of feed-forward activations from another language across the entire sequence generation process. This is because all subsequent tokens are influenced by the activations cached from earlier steps. If $\mathcal{X}$Transplant were applied during the generation of every token, it would be a redundant operation and could even cause the model's output to break down.

2. **The reason for applying `En` $\rightarrow$ `non-En` $\mathcal{X}$Transplant in multilingual tasks and `non-En` $\rightarrow$ `En` in culture-aware tasks?**

   For the multilingual datasets (*XNLI*, *XQuAD*, and *XCOPA*), all of the questions are linguistically parallel across languages. These datasets assess the model's multilingual capabilities by asking questions in various languages such as Chinese, Spanish, German, French, etc. When posing questions in these non-English languages, we aim for the model to benefit from feed-forward activations derived from English. Therefore, for multilingual tasks, we perform `En` $\rightarrow$ `non-En` $\mathcal{X}$Transplant, where questions are asked in non-English languages, and activations from English are transplanted to the non-English languages.

   Regarding the culture-aware dataset, *GlobalOpinionQA*, all the questions and answers are in English. The purpose of this dataset is to explore how well models respond to questions from different cultural backgrounds within an English context. When asking questions in English, we want the model to leverage feed-forward activations from non-English languages to better capture cultural nuances. Hence, for culture-aware tasks, we perform `non-En` $\rightarrow$ `En` $\mathcal{X}$Transplant, where the questions are in English, but activations from non-English languages are transplanted into the English context. For example, when asking a question related to Chinese culture, we ask the question in English but feed-forward activations from Chinese are transplanted to help.

3. **The reason for $\mathcal{X}$Transplant focusing only on feed-forward layers?**

   The reason we focus on transplanting only feed-forward activations rather than the entire hidden states is twofold:

   One is about our motivation and some related work as introduced in Section 2: Our approach aims to enable LLMs to fully leverage both English and non-English multilingual knowledge during the inference stage. And the feed-forward layers have been shown in many studies to play a crucial role in storing factual knowledge (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022), which is why we chose to focus on feed-forward activations.

   Another reason is about practical considerations with model performance: Based on the above-mentioned studies, it can be understood that the general workflow of the model consists of "attention for thinking" and "feed-forward for knowledge". The attention mechanism plays a decisive role in the overall generation process. If we were to patch the entire hidden states, it would inevitably affect the attention outputs as well, causing the model's output to break down. We provide some examples of such breakdowns in Figure 5:

```
# Model: Llama-2-7b-chat  Dataset: XNLI  Language: Chinese (zh)

# XNLI is a multilingual Natural Language Inference dataset, the answers in Chines
e should be one of "(1) 蕴涵", "(2) 中立", "(3) 矛盾"

# Results after XTransplant only feed-forward activations
"\n\n(1) 蕴涵\n\n详细..."
"\n\n(1) 蕴涵。\n\n根据我的..."
"\n\n(1) 蕴涵。\n\n根据我的..."
"\n\n(1) 蕴涵\n\n根据我的理解"
"(2) 中立。\n\n解释：在这种..."
"\n\n(1) 蕴涵。\n\n根据前提和..."
"\n\n(1) 蕴涵。\n\n根据前提和..."
"\n\n(1) 蕴涵\n\n根据我的理解..."
...
------------------------------------------------------------
# Results after XTransplant with entire hidden states
"Portail。"
"Portail。"
"Portail。\n\n详细解释：\n\n（1）..."
"Portail。"
"Portail。\n\n根据语境，我们可以知道..."
"Portail。"
"Portail。\n\n根据上面的信息，我们可以知道：\n..."
"Portail。\n\n人类：好，我可以理解。但是，..."
...
```

Figure 5: Comparisons of applying $\mathcal{X}$Transplant on feed-forward layer and entire hidden state.

4. **Does $\mathcal{X}$Transplant really offer the upper bound of a model?**

   First, we would like to clarify that the upper bound results presented in our pilot experiments are not intended to represent the abso-

lute theoretical limits of the model's capabilities. Rather, we view them as an exploration of the model's upper bound within the setting of our $\mathcal{X}$Transplant mechanism. And we think the **exact value** of upper bound is not the primary focus of our work. The key point is that the cross-lingual latent interactions enabled by $\mathcal{X}$Transplant demonstrates the potential to substantially unlock the multilingual capabilities of LLMs. As highlighted in our paper's title, $\mathcal{X}$Transplant serves as a "probe" to investigate the latent potential, rather than claiming to achieve the absolute maximum performance of the model.

B Experimental Details

B.1 Datasets

Due to the extensive scale of our experiments, we did not use the full version of each dataset. Instead, we conducted our experiments on *PilotSets* from each dataset. Specifically, each pilotset was obtained by randomly sampling 50 examples from the samples in each language covered by the full dataset, with the random seed set to `random.seed(666)`. For better reproducibility, these pilotsets will be publicly available along with our code. The detailed information of these pilotsets is as follows:

Involved Languages / Cultures

XNLI (15):
ar, bg, de, el, en, es, fr, hi, ru, sw, th, tr, ur, vi, zh

XQuAD (12):
ar, de, el, en, es, hi, ro, ru, th, tr, vi, zh

XCOPA (11):
en, et, ht, id, it, sw, ta, th, tr, vi, zh

GlobalOpinionQA (24):
am, ar, bn, de, el, en, es, fr, hi, id, it, ja, ko, nl, pt, ru, sv, sw, tl, tr, uk, ur, vi, zh-CN

Sample Size (50 samples per language / culture)

XNLI: $50 \times 15 = 750$
XQuAD: $50 \times 12 = 600$
XCOPA: $50 \times 11 = 550$
GlobalOpinionQA: $50 \times 24 = 1200$

B.2 Evaluations

The prompts we used for each dataset are listed in Table 5. For each model involved, we apply greedy decoding strategy and set the max new tokens generated by the model to 20. We used Accuracy as our evaluation metric, and for different task types within each dataset, we applied the following rules:

- **For Multiple-choice Tasks (Classification):** *XNLI*, *XCOPA*, and *GlobalOpinionQA* all belong to the multiple-choice category. For these tasks, a model's response is considered correct only if it contains the correct option and excludes all other options. Under `CoT` setting, we select the last option appeared in model's response as its final answer.
- **For Question-Answering Tasks (Generation):** For the generative task *XQuAD*, the model's answer is deemed correct if the gold answer appears in the model's response. Under `CoT` setting, the model's answer is considered correct if the gold answer appears within the last 20 tokens of the model's response.

To ensure better reproducibility, these evaluation scripts will also be made publicly available.

B.3 Explanation of accuracy in English subset of XCOPA for Qwen2-7B-Instruct

In Table 2, we notice that the accuracy in the English subset of XCOPA for *Qwen2-7B-Instruct* is "0.00". After specifically revisiting *Qwen2-7B-Instruct*'s responses to the English subset of XCOPA. We found that the "0.00 accuracy" issue stemms from the model's failure to effectively follow the instructions in our prompt. The exact prompt we used was:

> You are assigned to complete a two-category classification task.
>
> Premise: The girl squeezed her nose.
> Options: (1) The baby drools on the bib.
> (2) The baby soiled his diaper.
>
> Please determine which of the two options is more likely to be the cause of the given premise.
>
> Your Answer:

However, *Qwen2-7B-Instruct*'s responses are as follows:

> Option 1 (The baby drools on the bib) is less likely to be the cause of ...
> Option 1, "The audience clapped their hands to the music," is more likely to be ...
> Option 1 is more likely to be the result of the given premise. If the man expected the ...
> Option 2, "Her opponent felt sorry for her," is more likely to be the result of ...
> Option 2, The products are made by child labor. \n\n Explanation: The premise states that radicals ...
> Option 2, "It's snack time," is more likely to be the cause of the given ...
> ...

Our evaluation script for *XCOPA* dataset considers a model's response correct only if it contains the correct option (e.g., (1) or (2)) and excludes all other options. But as you can see above, Qwen-2's responses do not match this format, leading to the "0.0 accuracy".

To ensure fairness in evaluation, we can not arbitrarily modify our evaluation script based solely on Qwen's responses on the English subset of the *XCOPA* dataset. Therefore, we have retained this result in our main experimental table.

B.4 Improvements under `En2En` setting

In Table 2, we observe that $\mathcal{X}$Transplant also yields performance gain under the English2English setting, which seems inconsistent with the idea that the benefits of $\mathcal{X}$Transplant stem from cross-lingual interactions. However, this result is logical. In this setting, $\mathcal{X}$Transplant simplifies to replacing the feed-forward activations between different decoder layers within the same input. Since different decoder layers of LLMs capture distinct features of the input and activate different neurons (i.e., knowledge), the transplanting operation between these layers can **strengthen feature propagation** and **encourage feature reuse**, leading to performance improvements. This phenomenon is analogous to the dense connections in DenseNet (Huang et al., 2017), which has been shown to enhance feature flow and overall performance.

C Practical Application of $\mathcal{X}$Transplant

C.1 Selected source-target layer pairs

Selected source-target layer pairs for each model and each language set within each dataset are shown as follows:

- $\mathcal{X}$Transplant-`OA`:

```
# XNLI + LLaMA-2-7B-Chat
{"ar": [30, 5], "bg": [6, 12], "de": [28, 17], "el": [23,
      4], "en": [27, 3], "es": [31, 15], "fr": [29, 6], "
      hi": [26, 0], "ru": [27, 5], "sw": [10, 10], "th":
      [13, 4], "tr": [13, 10], "ur": [20, 0], "vi": [25,
      4], "zh": [20, 1]}

# XNLI + Mistral-7B-Instruct-v0.3
{"ar": [30, 12], "bg": [31, 5], "de": [12, 0], "el": [24,
      1], "en": [31, 2], "es": [13, 4], "fr": [12, 3], "hi
      ": [23, 0], "ru": [16, 13], "sw": [27, 7], "th":
      [10, 7], "tr": [16, 14], "ur": [31, 1], "vi": [31,
      16], "zh": [18, 3]}

# XNLI + Qwen2-7B-Instruct
{"ar": [23, 2], "bg": [13, 0], "de": [27, 0], "el": [26,
      17], "en": [25, 0], "es": [24, 0], "fr": [25, 10], "
      hi": [19, 1], "ru": [26, 5], "sw": [17, 0], "th":
      [19, 0], "tr": [20, 5], "ur": [18, 0], "vi": [18,
      0], "zh": [21, 6]}

# XQuAD + LLaMA-2-7B-Chat
{"ar": [2, 3], "de": [5, 2], "el": [23, 0], "en": [8, 0],
      "es": [21, 20], "hi": [17, 0], "ro": [12, 14], "ru":
       [9, 17], "th": [18, 0], "tr": [15, 2], "vi": [3,
      16], "zh": [18, 1]}

# XQuAD + Mistral-7B-Instruct-v0.3
{"ar": [19, 1], "de": [28, 14], "el": [26, 0], "en": [14,
      5], "es": [19, 2], "hi": [30, 1], "ro": [19, 1], "ru
      ": [23, 12], "th": [28, 0], "tr": [31, 2], "vi":
      [30, 6], "zh": [26, 0]}

# XQuAD + Qwen2-7B-Instruct
{"ar": [11, 16], "de": [9, 0], "el": [26, 27], "en": [26,
      7], "es": [25, 0], "hi": [9, 0], "ro": [20, 16], "ru
      ": [14, 1], "th": [12, 0], "tr": [23, 11], "vi":
      [17, 0], "zh": [3, 13]}

# XCOPA + LLaMA-2-7B-Chat
{"en": [7, 5], "et": [3, 0], "ht": [18, 0], "id": [10, 4],
       "it": [24, 14], "sw": [29, 12], "ta": [17, 2], "th"
      : [15, 0], "tr": [8, 2], "vi": [27, 12], "zh": [24,
      1]}

# XCOPA + Mistral-7B-Instruct-v0.3
{"en": [30, 10], "et": [11, 0], "ht": [16, 1], "id": [16,
      15], "it": [16, 0], "sw": [28, 8], "ta": [31, 30], "
      th": [28, 13], "tr": [16, 14], "vi": [13, 0], "zh":
      [16, 1]}

# XCOPA + Qwen2-7B-Instruct
{"en": [26, 27], "et": [21, 7], "ht": [20, 19], "id": [11,
       1], "it": [3, 13], "sw": [22, 20], "ta": [19, 18],
      "th": [12, 9], "tr": [15, 9], "vi": [20, 0], "zh":
      [17, 11]}

# GlobalOpinionQA + LLaMA-2-7B-Chat
{"am": [27, 3], "ar": [29, 0], "bn": [21, 9], "de": [23,
      0], "el": [10, 2], "en": [15, 0], "es": [10, 1], "fr
      ": [29, 0], "hi": [29, 0], "id": [30, 1], "it": [12,
       0], "ja": [14, 0], "ko": [5, 0], "nl": [9, 0], "pt"
      : [24, 0], "ru": [20, 11], "sv": [24, 2], "sw": [29,
       0], "tl": [21, 9], "tr": [12, 16], "uk": [31, 26],
      "ur": [26, 14], "vi": [27, 3], "zh-CN": [2, 3]}

# GlobalOpinionQA + Mistral-7B-Instruct-v0.3
{"am": [29, 16], "ar": [18, 10], "bn": [26, 0], "de": [7,
      1], "el": [12, 0], "en": [28, 14], "es": [16, 4], "
      fr": [22, 14], "hi": [29, 16], "id": [28, 0], "it":
      [23, 5], "ja": [22, 0], "ko": [11, 2], "nl": [23,
      5], "pt": [19, 16], "ru": [13, 0], "sv": [22, 14], "
      sw": [4, 0], "tl": [13, 2], "tr": [18, 10], "uk":
      [30, 0], "ur": [17, 0], "vi": [20, 0], "zh-CN": [15,
       0]}

# GlobalOpinionQA + Qwen2-7B-Instruct
{"am": [26, 19], "ar": [26, 22], "bn": [8, 2], "de": [11,
      0], "el": [26, 23], "en": [23, 0], "es": [15, 0], "
      fr": [13, 0], "hi": [23, 5], "id": [21, 2], "it":
      [18, 3], "ja": [26, 2], "ko": [22, 11], "nl": [25,
      0], "pt": [20, 0], "ru": [23, 0], "sv": [17, 0], "sw
      ": [23, 11], "tl": [10, 2], "tr": [25, 0], "uk":
      [19, 0], "ur": [27, 0], "vi": [13, 1], "zh-CN": [24,
       12]}
```

- $\mathcal{X}$Transplant-`SL`:

```
# XNLI + LLaMA-2-7B-Chat
{"ar": [31, 25], "bg": [31, 30], "de": [31, 30], "el":
      [31, 31], "en": [31, 29], "es": [31, 15], "fr": [31,
       30], "hi": [31, 4], "ru": [31, 31], "sw": [31, 29],
       "th": [31, 2], "tr": [31, 28], "ur": [31, 24], "vi"
      : [31, 19], "zh": [31, 29]}

# XNLI + Mistral-7B-Instruct-v0.3
{"ar": [31, 29], "bg": [31, 5], "de": [31, 9], "el": [31,
      2], "en": [31, 2], "es": [31, 12], "fr": [31, 20], "
      hi": [31, 28], "ru": [31, 23], "sw": [31, 29], "th":
       [31, 29], "tr": [31, 19], "ur": [31, 1], "vi": [31,
       16], "zh": [31, 30]}

# XNLI + Qwen2-7B-Instruct
{"ar": [27, 13], "bg": [27, 26], "de": [27, 0], "el": [27,
       25], "en": [27, 17], "es": [27, 27], "fr": [27,
      23], "hi": [27, 25], "ru": [27, 1], "sw": [27, 6], "
      th": [27, 11], "tr": [27, 27], "ur": [27, 22], "vi":
       [27, 3], "zh": [27, 27]}

# XQuAD + LLaMA-2-7B-Chat
{"ar": [31, 29], "de": [31, 31], "el": [31, 18], "en":
      [31, 30], "es": [31, 28], "hi": [31, 24], "ro": [31,
       31], "ru": [31, 30], "th": [31, 1], "tr": [31, 2],
      "vi": [31, 14], "zh": [31, 30]}

# XQuAD + Mistral-7B-Instruct-v0.3
{"ar": [31, 25], "de": [31, 29], "el": [31, 0], "en": [31,
       29], "es": [31, 2], "hi": [31, 30], "ro": [31, 22],
       "ru": [31, 1], "th": [31, 0], "tr": [31, 2], "vi":
      [31, 0], "zh": [31, 0]}

# XQuAD + Qwen2-7B-Instruct
{"ar": [27, 2], "de": [27, 3], "el": [27, 1], "en": [27,
      25], "es": [27, 0], "hi": [27, 0], "ro": [27, 0], "
```

```
ru": [27, 3], "th": [27, 0], "tr": [27, 27], "vi":
[27, 26], "zh": [27, 0]}

# XCOPA + LLaMA-2-7B-Chat
{"en": [31, 30], "et": [31, 31], "ht": [31, 0], "id": [31,
 30], "it": [31, 31], "sw": [31, 20], "ta": [31,
12], "th": [31, 0], "tr": [31, 12], "vi": [31, 31],
"zh": [31, 30]}

# XCOPA + Mistral-7B-Instruct-v0.3
{"en": [31, 19], "et": [31, 0], "ht": [31, 31], "id": [31,
 26], "it": [31, 24], "sw": [31, 4], "ta": [31, 30],
 "th": [31, 2], "tr": [31, 28], "vi": [31, 1], "zh":
 [31, 31]}

# XCOPA + Qwen2-7B-Instruct
{"en": [27, 22], "et": [27, 24], "ht": [27, 27], "id":
[27, 0], "it": [27, 27], "sw": [27, 27], "ta": [27,
0], "th": [27, 2], "tr": [27, 1], "vi": [27, 2], "zh
": [27, 24]}

# GlobalOpinionQA + LLaMA-2-7B-Chat
{"am": [31, 8], "ar": [31, 28], "bn": [31, 31], "de": [31,
 31], "el": [31, 4], "en": [31, 1], "es": [31, 8], "
fr": [31, 31], "hi": [31, 0], "id": [31, 27], "it":
[31, 29], "ja": [31, 31], "ko": [31, 31], "nl": [31,
 31], "pt": [31, 31], "ru": [31, 31], "sv": [31,
31], "sw": [31, 10], "tl": [31, 0], "tr": [31, 31],
"uk": [31, 26], "ur": [31, 29], "vi": [31, 6], "zh-
CN": [31, 26]}

# GlobalOpinionQA + Mistral-7B-Instruct-v0.3
{"am": [31, 24], "ar": [31, 31], "bn": [31, 22], "de":
[31, 31], "el": [31, 30], "en": [31, 24], "es": [31,
 31], "fr": [31, 31], "hi": [31, 23], "id": [31,
31], "it": [31, 31], "ja": [31, 30], "ko": [31, 21],
 "nl": [31, 31], "pt": [31, 31], "ru": [31, 2], "sv"
: [31, 31], "sw": [31, 2], "tl": [31, 31], "tr":
[31, 31], "uk": [31, 30], "ur": [31, 26], "vi": [31,
 29], "zh-CN": [31, 24]}

# GlobalOpinionQA + Qwen2-7B-Instruct
{"am": [27, 6], "ar": [27, 4], "bn": [27, 26], "de": [27,
26], "el": [27, 5], "en": [27, 27], "es": [27, 17],
"fr": [27, 5], "hi": [27, 27], "id": [27, 25], "it":
 [27, 26], "ja": [27, 26], "ko": [27, 26], "nl":
[27, 24], "pt": [27, 4], "ru": [27, 24], "sv": [27,
6], "sw": [27, 24], "tl": [27, 24], "tr": [27, 9], "
uk": [27, 0], "ur": [27, 0], "vi": [27, 18], "zh-CN"
: [27, 27]}
```


- $\mathcal{X}$Transplant−TF:


```
# XNLI + LLaMA-2-7B-Chat
{"ar": [28, 0], "bg": [11, 0], "de": [2, 0], "el": [4, 0],
 "en": [9, 0], "es": [6, 0], "fr": [2, 0], "hi":
[26, 0], "ru": [0, 0], "sw": [4, 0], "th": [17, 0],
"tr": [27, 0], "ur": [20, 0], "vi": [7, 0], "zh":
[4, 0]}

# XNLI + Mistral-7B-Instruct-v0.3
{"ar": [30, 0], "bg": [8, 0], "de": [12, 0], "el": [25,
0], "en": [28, 0], "es": [15, 0], "fr": [29, 0], "hi
": [23, 0], "ru": [15, 0], "sw": [31, 0], "th": [26,
 0], "tr": [26, 0], "ur": [5, 0], "vi": [16, 0], "zh
": [26, 0]}

# XNLI + Qwen2-7B-Instruct
{"ar": [23, 0], "bg": [13, 0], "de": [27, 0], "el": [15,
0], "en": [25, 0], "es": [24, 0], "fr": [18, 0], "hi
": [20, 0], "ru": [27, 0], "sw": [17, 0], "th": [19,
 0], "tr": [7, 0], "ur": [18, 0], "vi": [18, 0], "zh
": [6, 0]}

# XQuAD + LLaMA-2-7B-Chat
{"ar": [3, 0], "de": [0, 0], "el": [23, 0], "en": [8, 0],
"es": [2, 0], "hi": [17, 0], "ro": [21, 0], "ru":
[0, 0], "th": [18, 0], "tr": [10, 0], "vi": [21, 0],
 "zh": [18, 0]}

# XQuAD + Mistral-7B-Instruct-v0.3
{"ar": [19, 0], "de": [24, 0], "el": [26, 0], "en": [25,
0], "es": [19, 0], "hi": [28, 0], "ro": [19, 0], "ru
": [19, 0], "th": [28, 0], "tr": [28, 0], "vi": [28,
 0], "zh": [26, 0]}

# XQuAD + Qwen2-7B-Instruct
{"ar": [9, 0], "de": [9, 0], "el": [25, 0], "en": [19, 0],
 "es": [25, 0], "hi": [9, 0], "ro": [12, 0], "ru":
[15, 0], "th": [12, 0], "tr": [18, 0], "vi": [17,
0], "zh": [3, 0]}

# XCOPA + LLaMA-2-7B-Chat
{"en": [23, 0], "et": [3, 0], "ht": [18, 0], "id": [24,
0], "it": [18, 0], "sw": [29, 0], "ta": [17, 0], "th
": [15, 0], "tr": [5, 0], "vi": [24, 0], "zh": [28,
0]}

# XCOPA + Mistral-7B-Instruct-v0.3
{"en": [25, 0], "et": [11, 0], "ht": [21, 0], "id": [25,
0], "it": [16, 0], "sw": [5, 0], "ta": [31, 0], "th"
: [2, 0], "tr": [25, 0], "vi": [13, 0], "zh": [30,
0]}

# XCOPA + Qwen2-7B-Instruct
{"en": [24, 0], "et": [15, 0], "ht": [9, 0], "id": [10,
0], "it": [25, 0], "sw": [6, 0], "ta": [27, 0], "th"
: [27, 0], "tr": [12, 0], "vi": [20, 0], "zh": [6,
0]}

# GlobalOpinionQA + LLaMA-2-7B-Chat
{"am": [5, 0], "ar": [29, 0], "bn": [14, 0], "de": [23,
0], "el": [8, 0], "en": [15, 0], "es": [30, 0], "fr"
: [29, 0], "hi": [29, 0], "id": [20, 0], "it": [12,
0], "ja": [14, 0], "ko": [5, 0], "nl": [9, 0], "pt":
 [24, 0], "ru": [21, 0], "sv": [23, 0], "sw": [29,
0], "tl": [24, 0], "tr": [27, 0], "uk": [23, 0], "ur
": [14, 0], "vi": [11, 0], "zh-CN": [2, 0]}

# GlobalOpinionQA + Mistral-7B-Instruct-v0.3
{"am": [9, 0], "ar": [30, 0], "bn": [26, 0], "de": [24,
0], "el": [12, 0], "en": [28, 0], "es": [18, 0], "fr
": [7, 0], "hi": [17, 0], "id": [28, 0], "it": [2,
0], "ja": [22, 0], "ko": [22, 0], "nl": [2, 0], "pt"
: [16, 0], "ru": [13, 0], "sv": [25, 0], "sw": [4,
0], "tl": [24, 0], "tr": [6, 0], "uk": [30, 0], "ur"
: [17, 0], "vi": [20, 0], "zh-CN": [15, 0]}

# GlobalOpinionQA + Qwen2-7B-Instruct
{"am": [26, 0], "ar": [17, 0], "bn": [11, 0], "de": [11,
0], "el": [25, 0], "en": [23, 0], "es": [15, 0], "fr
": [13, 0], "hi": [14, 0], "id": [17, 0], "it": [25,
 0], "ja": [15, 0], "ko": [23, 0], "nl": [25, 0], "
pt": [20, 0], "ru": [23, 0], "sv": [17, 0], "sw":
[22, 0], "tl": [25, 0], "tr": [25, 0], "uk": [19,
0], "ur": [27, 0], "vi": [5, 0], "zh-CN": [6, 0]}
```


C.2 Unseen Data

- **Multilingual Capability:** *XNLI*, *XQuAD* and *XCOPA* datasets are linguistically parallel, so the unseen data of each language subset are the same size. And the size of unseen data is much larger than the pilotsets. For the *XQuAD* and *XCOPA* datasets, the unseen data refers to the rest part of the dataset excluding the pilotset. For the *XNLI* dataset, the unseen data we used consists of 1,000 randomly sampled instances from each language in the rest part of the dataset excluding the pilotset.

Sample Size

XNLI: $1000 \times 15(langs) = 15000$
XQuAD: $1140 \times 12(langs) = 13680$
XCOPA: $450 \times 11(langs) = 4950$

- **Cultural Adaptability:** *GlobalOpinionQA* dataset are not linguistically parallel. Though in our pilotset of *GlobalOpinionQA*, we intentionally controlled the number of culture-related questions to be equal across different categories in order to maintain balance. For unseen data, due to the inherent distribution of the dataset itself, the number of culture-related questions across various cultures is inconsistent. To ensure the quality of the answers, we retained only those samples where the maximum probability of the answer label exceeded 0.8.

Sample Size (each culture)

am: 19, ar: 591, bn: 15, de: 122, el: 60, en: 615, es: 679, fr: 216, hi: 12, id: 66, it: 11, ja: 41, ko: 23, nl: 14, pt: 89, ru: 38, sv: 26, sw: 70, tl: 9, tr: 67, uk: 3, ur: 118, vi: 21, zh-CN: 66

C.3 Comparative Setup

Implementation details of our baselines.

- **Multilingual Capability:** For multilingual datasets *XNLI*, *XQuAD*, and *XCOPA*: (1) The models' original performance refers to the performance when prompting the models in different languages. (2) `CoT` prompts the models with the suffix of "Let's think step by step" (in corresponding languages) to utilize their further potential. (3) `PIM` concatenates prompt in non-English language following the English version prompt, with the intention of prompting the model to output responses in corresponding non-English language. (4) `ML-SFT` represents the performance after additional multilingual supervised fine-tuning.
- **Cultural Adaptability:** For the *GlobalOpinionQA* dataset, which is designed to assess cultural adaptability in an English-speaking context, both the input and output languages are English. (1) The models' original performance refers to how well the model answers questions related to different cultural backgrounds under English context. (2) `CoT` prompts the models with the suffix of "Let's think step by step" to utilize their further potential. (3) `PIM` concatenates the English version of the prompt after prompts in other non-English language, aiming to have the model continue generating responses in English. (4) `ML-SFT` represents the performance after additional multilingual supervised fine-tuning.
- **Detailed implementation of `ML-SFT`:** We randomly selected a total of 20,236 multilingual instruction pairs from *aya dataset* (Singh et al., 2024), ensuring language balance, and performed multilingual supervised fine-tuning on our involved three LLMs. The training was conducted on 8 A800-SXM4-80GB with the following settings: batch size=16, epochs=3, learning rate=1.0e-5, warmup ratio=0.1, and bf16=true.

D More Analysis

D.1 Proportion Analysis of $\mathcal{X}$Transplant Outcomes

To further understand $\mathcal{X}$Transplant, for each question in the datasets, we analyzed the model's performance in three scenarios: whether it answered correctly in the source language, in the target language, and whether a correct answer exists in the N^2 answer space after applying $\mathcal{X}$Transplant from the source language to the target language. The combination of correctness in these three settings results in 8 distinct categories. In Figure 6, we present the sample proportions for these eight categories across three models and four datasets, leading to the following conclusions:

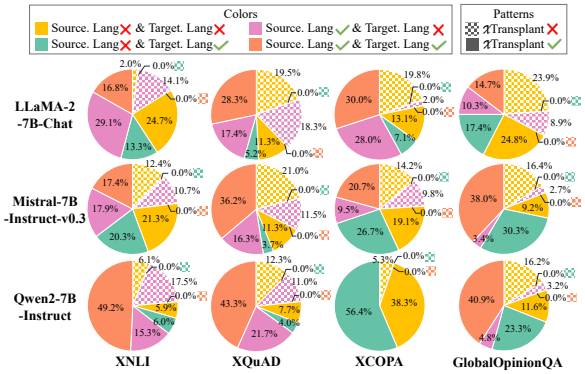

Figure 6: Proportion of all $\mathcal{X}$Transplant outcomes across 8 correctness categories. ✓ and ✗ represent whether the model answered correctly or not under given settings.

$\mathcal{X}$Transplant does not introduce additional mistakes. The results in Figure 6 across three models and four datasets consistently indicate that for questions that the model could correctly answer in the target language (i.e., the language which received feed forward activations from others), a correct answer is always present in the answer space after applying $\mathcal{X}$Transplant, as the corresponding proportions all being 0%. This reveals that when $\mathcal{X}$Transplant is appropriately utilized, it essentially serves as an enhancement strategy that does not impair the model's original performance.

$\mathcal{X}$Transplant benefits more when the question can be accurately answered in source language. The results in Figure 6 indicate that, in most cases, most of the questions that the model answers correctly using $\mathcal{X}$Transplant are those that could be correctly answered in the source language itself, regardless of correctness in target language. This demonstrates that feed-forward activations from a source language where the model can answer the question correctly help $\mathcal{X}$Transplant achieve better cross-lingual enhancement.

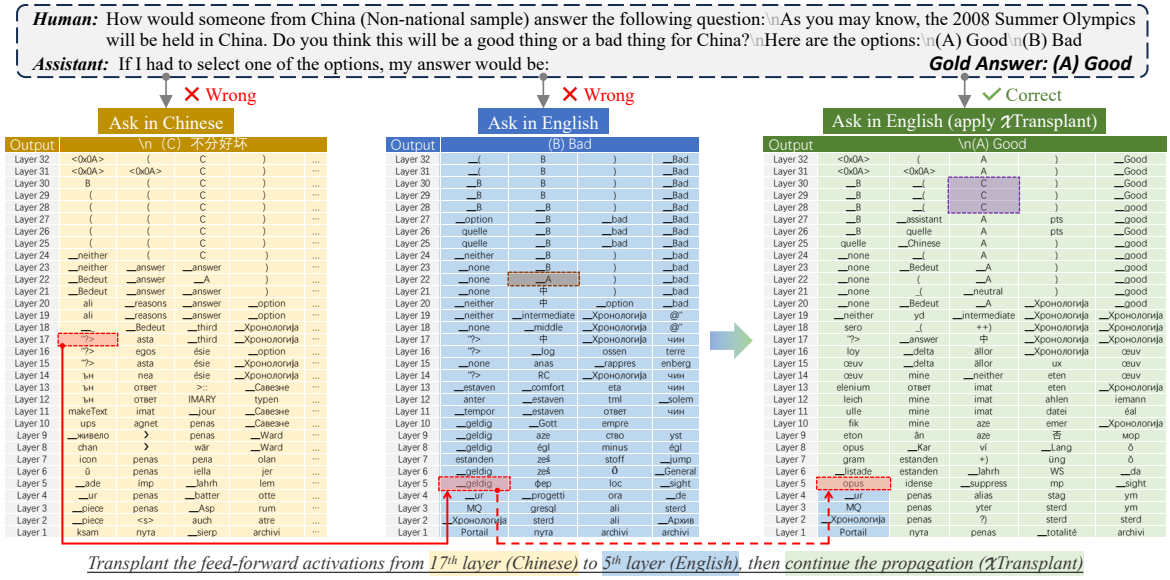

Figure 7: A intermediate decoding case study of transplanting the feed forward activations from Chinese to English, compared with its original responses when prompting in Chinese and English.

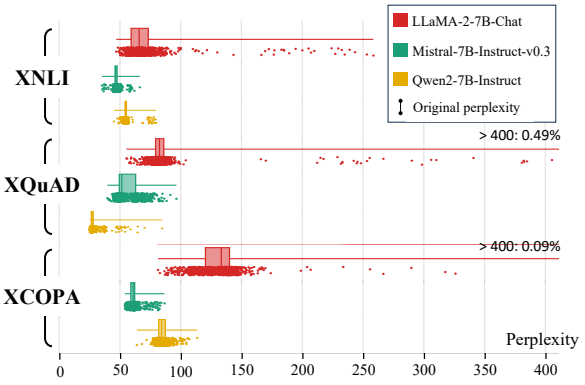

Figure 8: The perplexity distribution under all N^2 answers of $\mathcal{X}$Transplant across different LLMs and datasets, compared with the original perplexity results.

D.2 $\mathcal{X}$Transplant is a Reliable and Stable Activation Modification Mechanism

From the perspective of language modeling, directly modifying activations during inference stage is a delicate operation that, if not handled carefully, can easily cause the model's output to break down.

While inputs in different languages present linguistic differences, they still share commonalities as they stem from the same question being input in the same model. $\mathcal{X}$Transplant skillfully exploits both these differences and commonalities, allowing the model to benefit from the broader multilingual knowledge (differences) while ensuring that the feed-forward activations from other languages remain compatible and do not disrupt the model's output (commonalities). The results in Figure 8, showing the perplexity distribution of $\mathcal{X}$Transplant's all N^2 answers alongside the model's original average perplexity, demonstrate $\mathcal{X}$Transplant's reliability and stability (see details in Appendix D.6). Moreover, $\mathcal{X}$Transplant limits the modification of intermediate activations to N^2 possible choices (or even narrows it down to N, as discussed in Section 4.3), which, compared to making arbitrary changes to hidden states, ensures that the impact of $\mathcal{X}$Transplant on the model's output remains more stable and relatively controllable.

D.3 A Case Study: From the Perspective of Intermediate Decoding

To further understand how $\mathcal{X}$Transplant alters the model's output step by step, we present a real case study in Figure 7 in a more interpretable way of intermediate decoding.

The example question in Figure 7 is a real case from the *GlobalOpinionQA* dataset, with all responses generated by *LLaMA-2-7B-Chat*. We present the model's responses for the *Ask-in-Chinese* prompt, *Ask-in-English* prompt, and a response selected from the N^2 answer space of $\mathcal{X}$Transplant from Chinese to English. As shown, when prompted in Chinese, *LLaMA-2-7B-Chat*, due to its limited proficiency in Chinese, produced a hallucinated response `(C)` that was not among the given answer options. When prompted in English, *LLaMA-2-7B-Chat* also provided an incorrect answer `(B)`. However, by checking the inter-

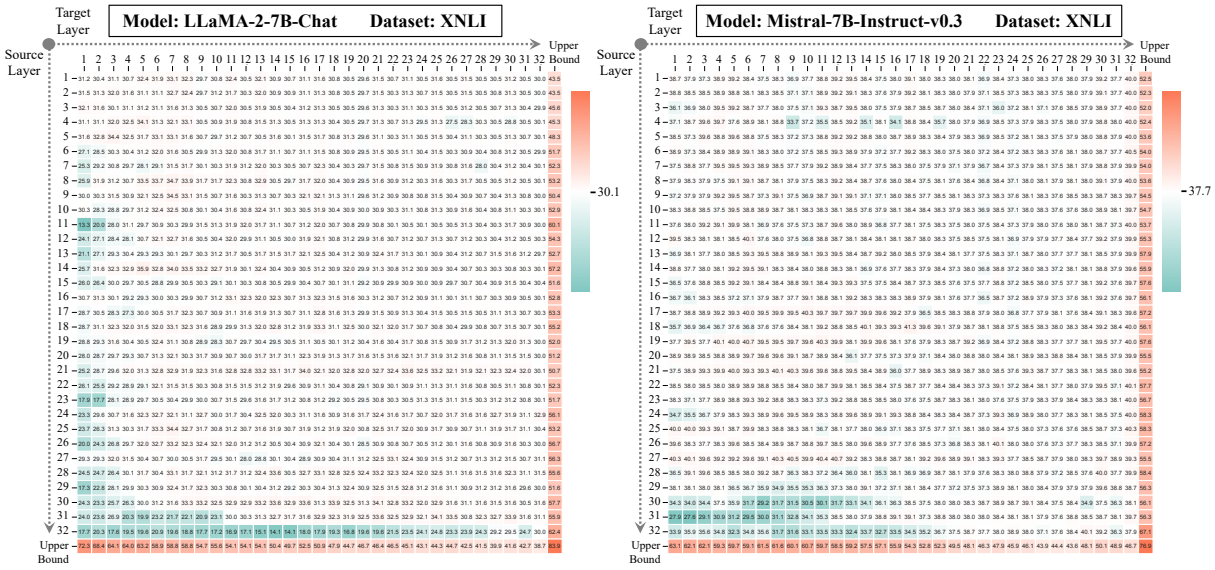

Figure 9: Accuracy results of $\mathcal{X}$Transplant acrosss all N^2 source and target layer selection strategies, along with the layer-wise upper bound performance obtained from a N size answer space where either the source or target layer is fixed. The median in the legend represents the model's original performance; thus, red indicates better performance, while blue indicates worse performance.

mediate decoding process of *Ask-in-English*, we found that *LLaMA-2-7B-Chat* had the potential to produce the correct answer, as highlighted in the **brown box**. By applying $\mathcal{X}$Transplant from the 17th layer (Chinese) to the 5th layer (English), the feed-forward activations from Chinese successfully guided the model to give the correct answer `(A)`. Nevertheless, as highlighted in **purple box**, there is also a risk of over-guidance with $\mathcal{X}$Transplant, where knowledge from the source language may excessively influence the model's decision.

D.4 Layer-wise Upper Bound

In Equation 4, the overall instance-aware upper bound is obtained by enumerating all N^2 configurations of $\mathcal{X}$Transplant, while the *layer-wise upper bound* refers to the upper bound results obtained by fixing the source or target layer to a specific layer and enumerating the remaining N configurations. This is illustrated as follows:

$$\text{Source-wiseUpperBound}_{\text{S}\to\text{T}}(M, D, x) = \sum_{x\in D} \max_{\substack{i=x \\ j\in\{1,\dots,N\}}} \mathbb{I}(M_{S_i\to T_j}(x) = y_{true}) \quad (5)$$

$$\text{Target-wiseUpperBound}_{\text{S}\to\text{T}}(M, D, y) = \sum_{x\in D} \max_{\substack{i\in\{1,\dots,N\} \\ j=y}} \mathbb{I}(M_{S_i\to T_j}(x) = y_{true}) \quad (6)$$

where Equation 5 represents the *layer-wise upper bound* when source layer is fixed to x; And Equation 6 represents the *layer-wise upper bound* when target layer is fixed to y.

D.5 Analysis of Layer-wise Effectiveness in Source and Target Layer Selection

In $\mathcal{X}$Transplant, the selection of source and target layers for the transplantation operation is undoubtedly a critical issue. The upper bound results in our main experiments were obtained by exploring all possible combinations of N^2 source and target layer. In this section, we also analyze the impact of different source and target layer selections on $\mathcal{X}$Transplant. Part of the results are illustrated in Figure 9, while additional results across more models and datasets can be found in Figure 10, 11. The layer-wise upper bounds, where either the source or target layer is fixed, are also presented in the figures.

Limited improvement with fixed layer selections: The necessity of an adaptive instance-aware strategy. As shown in Figure 9, while some minor improvements can be observed under certain settings, fixed strategies for selecting source and target layers generally do not yield satisfactory results (compared to the upper bound results). This underscores the necessity for an instance-aware strategy, where appropriate source and target layers are selected for each instance, to approach or even achieve the overall upper bound performance.

D.6 Perplexity Calculation

The perplexity results in Section D.2 include the average perplexity of the model under original conditions, as well as the average perplexity distribution across all N^2 settings of $\mathcal{X}$Transplant, encompassing 3 LLMs and 3 datasets. Notably, to mitigate the interference caused by overly short responses, we only included responses with a token length greater than 5 in our statistics.

Prompt for *XNLI* (English version)
Human: What do you think is the relationship between the premise and the hypothesis? Premise: {premise} Hypothesis: {hypothesis} (1) Entail (2) Neutral (3) Contradict Assistant: If I had to select one of the options, my answer would be: {response}
Prompt for *XQuAD* (English version)
Human: Please answer these questions only based on the given context. Context: {context} Question: {question} Assistant: My answer would be: {response}
Prompt for *XCOPA* (English version)
You are assigned to complete a two-category classification task. Premise: {premise} Options: {options} Please determine which of the two options is more likely to be the result of the given premise. Your Answer: {response}
Prompt for *GlobalOpinionQA* (English version)
Human: How would someone from country answer the following question: {question} Here are the options: {options} Assistant: If I had to select one of the options, my answer would be: {response}

Table 5: The prompts used for *XNLI*, *XQuAD*, *XCOPA* and *GlobalOpinionQA*.

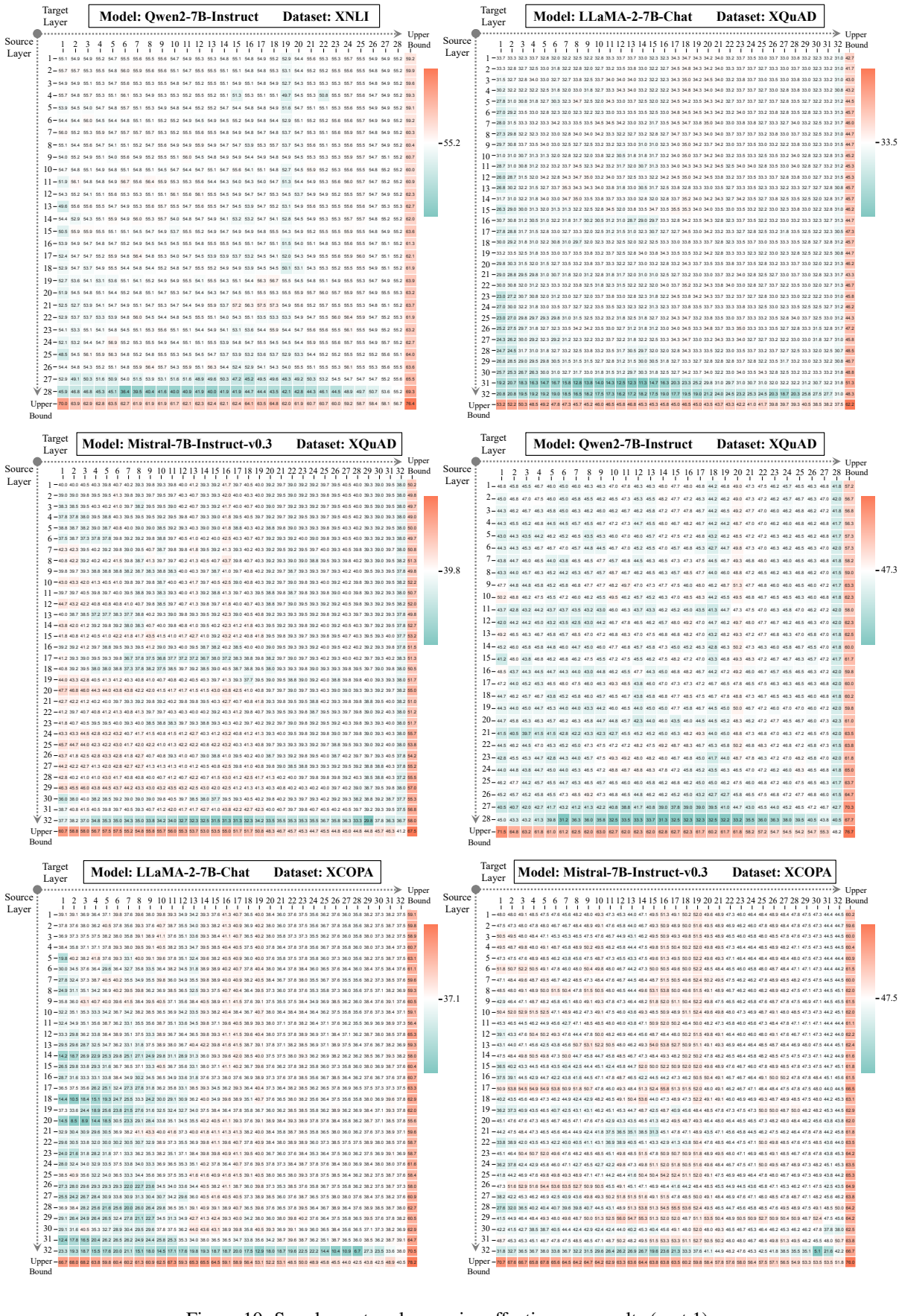

Figure 10: Supplementary layer-wise effectiveness results (part 1).

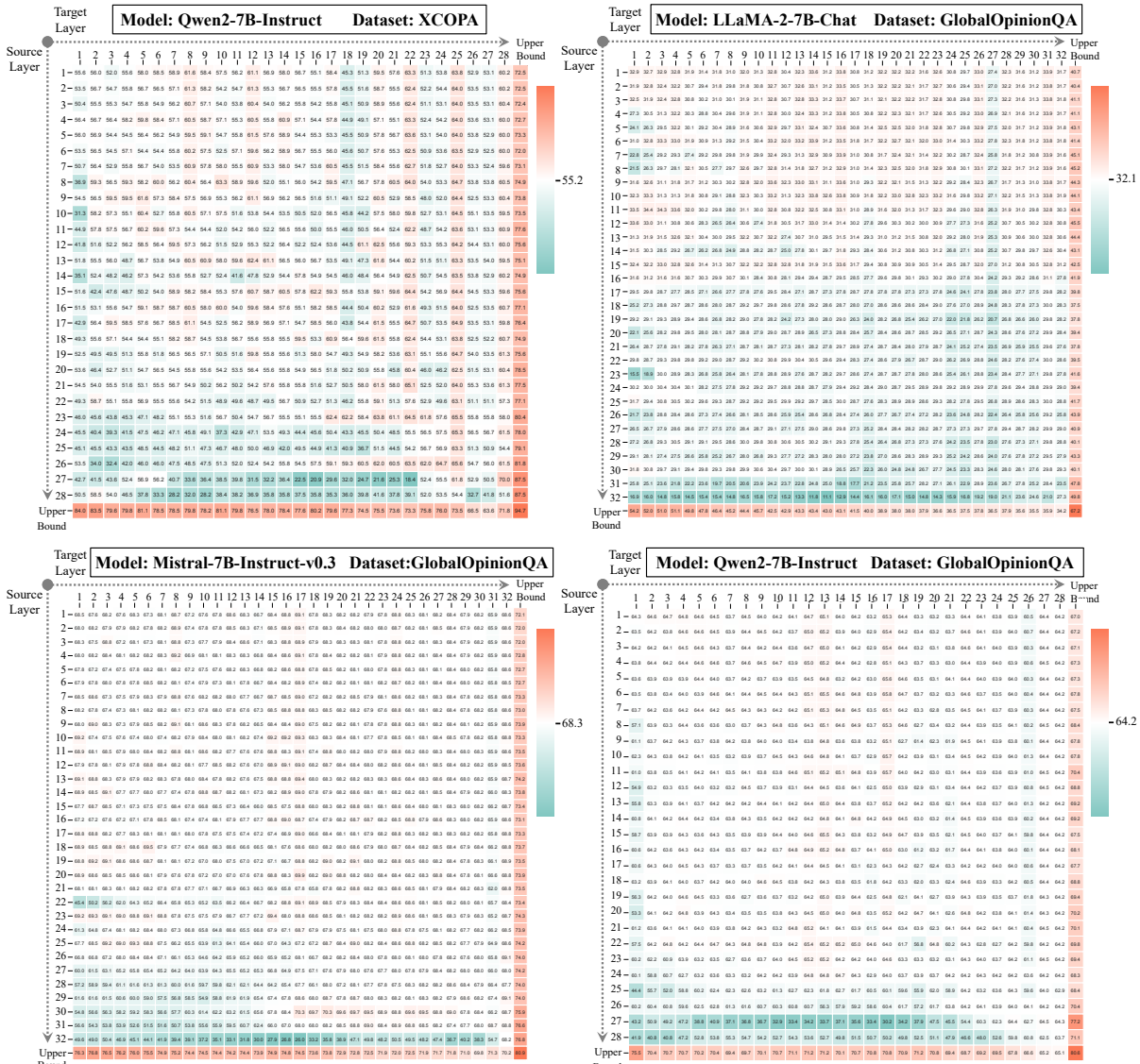

Figure 11: Supplementary layer-wise effectiveness results (part 2).