STEERLSN: Training-Free Representation Alignment through Language-Specific Neuron Subspace Interpolation

Minjun Kim
KAIST
mjkmain@kaist.ac.kr

Dongyeon Kim
KAIST
dykim@kaist.ac.kr

Hoyun Song
Dankook University
hoyun.song@dankook.ac.kr

Jihyo Kim[†]
KAIST InnoCORE PRISM-AI Center
jihyo.kim@kaist.ac.kr

Kyungtae Lim[†]
KAIST
ktlim@kaist.ac.kr

Abstract

Recent advancements in Large Language Models (LLMs) have shown remarkable capabilities in multilingual contexts. However, previous studies suggest that multilingual LLMs primarily rely on the rich semantic knowledge of high-resource (i.e., English) representations to conduct tasks across languages. To mitigate the performance gap arising from resource imbalances, we propose *Steering via Language-Specific Neurons*, a training-free method that aligns non-English representations with the English manifold intervening on language-specific neuron activations via linear interpolation toward English activation prototypes. Experimental results show that the proposed method generally improves downstream task performance with minimal computational cost, yielding performance gains for low-resource scenarios where limited data often constrains knowledge coverage. Our findings suggest that the fundamental limitations of low-resource languages can be effectively alleviated by leveraging the robust representation space of high-resource languages, offering a promising direction for more balanced multilingual LLMs.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance on complex tasks, including reasoning, mathematics, and programming, leading to their widespread adoption (Yang et al., 2025; AI@Meta, 2024; DeepSeek-AI, 2025). As the enormous datasets utilized for training these models are sourced from the web, they naturally encompass a diverse range of languages (Luccioni and Viviano, 2021; Kreutzer et al., 2022). This induces inherent multilingual capabilities, enabling LLMs to exhibit impressive performance across various linguistic contexts (GemmaTeam, 2025; MistralAI, 2025).

Despite these inherent multilingual capabilities, multilingual LLMs exhibit performance gaps across languages, particularly underperforming in low-resource scenarios (Ahuja et al., 2023; Huang et al., 2023; Xu et al., 2025). Recent studies on the internal mechanisms of multilingual LLMs reveal that they do not operate in a language-neutral manner; instead, their processing is largely organized around an English representation space (Etxaniz et al., 2024; Wendler et al., 2024; Schut et al., 2025). In low-resource scenarios, representations tend to converge toward the English manifold as representations propagate through layers, showing increasing semantic similarities to English representation (Wendler et al., 2024; Tang et al., 2024). This observation implies that improving the alignment between low- and high-resource language representations may enhance performance. However, achieving such alignment through additional training is often computationally demanding and further constrained by limited data availability for low-resource scenarios (Guo et al., 2024; Kadyrbek et al., 2025).

To better understand the internal mechanisms of multilingual LLMs, prior studies have investigated the linguistic circuits activated during the processing of diverse languages (Resck et al., 2025; Zhang et al., 2025). One popular approach is to localize Language-Specific Neurons (LSNs) that respond selectively to particular languages. To analyze their functional roles, previous studies employ masking or ablation to intervene on targeted components and examine resulting changes in language preference and selection dynamics (Kojima et al., 2024; Zhao et al., 2024). For example, Tang et al. (2024) identify LSNs via cross-lingual comparisons of

[†] Equal corresponding authors.

In this work, we use the term "low-resource" to refer to languages with limited web-scale pretraining exposure, rather than speaker population size.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1005–1018
November 6-10, 2026 ©2026 Association for Computational Linguistics

activation patterns across a predefined set of languages, and show that modulating these neurons can steer the model's output language. Furthermore, representation-level interventions, such as adding bias or transformation vectors, have been explored to steer model behavior toward specific linguistic manifolds or task-relevant spaces (Schut et al., 2025; Etxaniz et al., 2024).

Collectively, these lines of studies indicate that multilingual behavior can be influenced through targeted neuron- and representation-level interventions. These findings motivate a lightweight approach for low-resource scenarios: aligning input language (non-high-resource in our scenario) representations toward a target language (high-resource language such as English in our scenario) manifold by manipulating a small set of language-related neurons, leading to improved downstream performance. Therefore, we propose ***Steer**ing via **L**anguage-**S**pecific **N**eurons* (STEERLSN), a training-free approach that intervenes on LSNs to shift input language representations toward a target language manifold, thereby improving alignment between them. Specifically, we employ Language Activation Probability Entropy (LAPE) (Tang et al., 2024) to identify LSNs, and linearly interpolate their subspaces defined by LSN activations toward target-language prototypes to improve alignment with the target language.

Our approach builds on recent findings suggesting that improved alignment between low- and high-resource language representations can lead to better performance (Tang et al., 2024). Since some task-relevant conceptual knowledge may be more readily accessed in high-resource representation space, improving access to that space may yield performance gains by enabling more effective reuse of such knowledge. Based on this observation, we posit the *Conceptual Access Hypothesis*: performance limitations in low-resource scenarios stem not only from a lack of encoded knowledge, but also from limited access to the English-centric conceptual space where such knowledge may be more effectively utilized. STEERLSN enhances access to this conceptual space by encouraging alignment with English representations.

We evaluate our method on widely used multilingual LLMs, including Llama 3 (AI@Meta, 2024) and Qwen 3 (Yang et al., 2025), using standard benchmarks, such as Global MMLU (Singh et al., 2025) and MGSM (Shi et al., 2023). Experimental results show consistent improvements for low-resource scenarios in knowledge-intensive evaluations spanning diverse domains without requiring additional low-resource language samples or significant computational cost. These improvements suggest that better aligning low-resource language representations with the high-resource language space enables leveraging knowledge encoded in the high-resource language representation space. We emphasize that our goal is to improve access to language-attributed knowledge, rather than to universally enhance procedural reasoning.

2 Related Work

2.1 Neuron Analysis

To elucidate the multilingual processing mechanisms of LLMs and Multimodal LLMs (MLLMs), Tang et al. (2024) proposed a quantitative methodology to identify LSNs by measuring the LAPE of neuron activation distributions across languages.

Subsequent studies have expanded beyond this probabilistic distribution analysis, introducing various approaches to enhance the precision and reliability of neuron identification. Representative advancements include the Parallel Language-specific Neuron Detection (PLND) technique utilizing label-free data (Zhao et al., 2024). Notably, to resolve polysemanticity caused by the superposition phenomenon in neural networks, attempts have been made to utilize Sparse Autoencoders (SAE) to decompose neurons into more granular features (Deng et al., 2025; Shi et al., 2025; Huben et al., 2024). Furthermore, research is actively employing gradient-based attribution scores to capture actual causal influence rather than simple activation frequency, thereby identifying key neurons that significantly contribute to model reasoning (Yamamoto et al., 2025; Chen et al., 2025; Lu et al., 2025).

However, existing research remains primarily at an analytical level, focusing on observing LSN distributions or assessing performance changes through ablation. Approaches that sophisticatedly manipulate LSNs and their surrounding vector representations layer-wise to actively transfer knowledge from high-resource to low-resource languages and improve practical reasoning capabilities remain underexplored.

2.2 Inference-time Intervention

Inference-time Intervention (ITI) has emerged as a cost-effective, parameter-preserving technique that

steers model behavior by shifting internal activations without modifying weights (Li et al., 2023). To apply such interventions effectively in multilingual contexts, it is crucial to understand the model's internal processing dynamics. Recent studies reveal that LLMs contain both core linguistic regions for general capabilities and monolingual regions specific to individual languages (Zhang et al., 2024). Furthermore, processing typically transitions from an initial input space to an English-centric concept space located in the intermediate layers, where the majority of complex reasoning occurs before generating the final output (Wendler et al., 2024).

Leveraging these layer-wise characteristics, recent frameworks attempt to bridge the performance gap between languages by focusing on these intermediate layers. For instance, extensive analysis of cross-lingual semantic alignment suggests that alignment is most prominent in middle layers, directly correlating with downstream performance (Liu and Niehues, 2025; Wu et al., 2025; Mao and Yu, 2024). Based on this, methods like INCLINE align low-resource representations to the high-performance space (Wang et al., 2025), while the MRRE framework adopts a hierarchical approach: guiding representations into the English space for reasoning and subsequently restoring them to the target language distribution (Li et al., 2025). Extending this line of research, this study proposes maximizing the reasoning capabilities of medium-to-low resource languages by explicitly operating on language-specific neurons, rather than on dense representation subspaces, to enable targeted and language-aware intervention.

3 Method

In this section, we first describe how we extract language-conditioned activations and identify LSNs for interventions in the model's language-processing mechanisms. Specifically, we employ FLORES+ (Costa-jussà et al., 2024) to obtain language-conditioned activations and apply LAPE to identify LSNs. We then provide the detailed procedures of our method. Hereafter, we refer to the input language as the *source*-language and the language used as the alignment objective as the *target*-language.

3.1 Preliminary

Extracting language-conditioned activations. To extract language-conditioned activations, we use the FLORES+ benchmark, which provides $\approx 1k$ parallel sentences per language across 225 languages (Costa-jussà et al., 2024). Following Tang et al. (2024), we define a *neuron* as a linear transformation of a single column vector in the first linear map $\mathbf{W}^i_{\text{up}} \in \mathbb{R}^{d \times 4d}$ of a feed-forward network (FFN), followed by an activation function. Let $\mathbf{h}^i \in \mathbb{R}^d$ denote the post-attention hidden state of a token at i-th layer, where d is the model hidden size. Then, a j-th neuron at i-th layer, a^i_j, is represented as $a^i_j = \sigma\left((\mathbf{h}^i)^\top \mathbf{w}^i_j\right)$, where $\sigma(\cdot)$ and $\mathbf{w}^i_j$ are an activation function and the j-th column of $\mathbf{W}^i$, respectively.

Let $L = \{l_n\}^N$ be the set of N languages l and let $\mathcal{D} = \{\mathbf{x}^{l_1}_k, \cdots, \mathbf{x}^{l_N}_k\}^K_{k=1}$ denote the K FLORES+ sentences in l_n. We extract language-conditioned activations by forwarding the model on each $\mathbf{x}^{l_n}_k$ and collecting $\{a^i_j(\mathbf{x}^{l_n}_k)\}$.

LAPE. Prior studies suggest that a^i_j are often selective and sparse, and that intervening on a small subset of units (e.g., via masking or ablation) can induce detectable changes in model behavior (Dai et al., 2022; Bricken et al., 2023), making them practical control points in multilingual scenarios. Therefore, LAPE leverages a^i_j to find LSNs by estimating the activation probability $p_{i,j}$. Specifically, $p^{l_m}_{i,j}$ for a given language l_m is estimated by

$$p^{l_m}_{i,j} = \frac{\sum_k \mathbb{1}(a^i_j(\mathbf{x}^{l_m}_k) > 0)}{|\{\mathbf{x}^{l_m}_k\}|},$$

where $\mathbb{1}(\cdot)$ is an indicator function. Then, the $\tilde{\mathbf{p}}_{i,j}$, the normalized $\mathbf{p}_{i,j}$ over L, is used to compute LAPE δ score which is denoted as:

$$\delta^i_j = -\sum_{l_n} \tilde{p}^{l_n}_{i,j} \log \tilde{p}^{l_n}_{i,j}.$$

Since δ is an entropy, a subset of a^i with the lowest δ^i_j is selected as LSNs.

3.2 Steering via Language-Specific Neurons

To align a source-language, l_{src}, representation with a target-language, l_{tgt}, representation, we intervene on a^i_j of their LSNs. Our approach consists of two stages: 1) shift LSNs of l_{src}, LSNs$^{l_{src}}$, while activating those of l_{tgt}, LSNs$^{l_{tgt}}$ to steer

representations toward the target-language manifold, and 2) restore the generation capability in the source-language. For brevity, let $\mathcal{S} \subseteq \{1, \cdots, 4d\}$ be the set of LSN indices at layer i. We define the masked $\mathbf{a}^i_{\mathcal{S}}$ such that its j-th component is given by $(\mathbf{a}^i_{\mathcal{S}})_j = a^i_j \cdot \mathbb{1}(j \in \mathcal{S})$.

3.2.1 Cross-Lingual Representation Interpolation

STEERLSN focuses on the linear interpolation of a^i_j of LSNs$^{l_{src}}$ and those of LSNs$^{l_{tgt}}$. To interpret our neuron-level intervention, we simplify each Transformer block (Vaswani et al., 2017) as updating a residual representation. For clarity, let $\mathbf{W}^i$ be redefined as $\mathbf{W}^i_{\text{up}}$, since it performs as the up-projection, and $\mathbf{W}^i_{\text{down}} \in \mathbb{R}^{4d \times d}$ denote the FFN down-projection (i.e., the second linear map) that maps $\mathbf{a}^i = [a^i_1, \cdots, a^i_{4d}]^\top$ back to the residual stream. The residual stream is updated by adding $\mathbf{a}^i$ as follows:

$$\mathbf{h}^{i+1} = \mathbf{h}^i + (\mathbf{W}^i_{\text{down}})^\top \mathbf{a}^i + \cdots ,$$

with other sub-layers omitted for brevity. Therefore, intervening on a subset of neurons is analogous to injecting a controlled perturbation into the representation as follows:

$$\Delta\mathbf{h}^{i+1} = (\mathbf{W}^i_{\text{down}})^\top \Delta\mathbf{a}^i .$$

In the proposed method, Δa^i_j is non-zero only for indices of LSNs. Hence, the induced shift lies in a low-dimensional subspace spanned by the corresponding output directions:

$$\Delta\mathbf{h}^{i+1} = (\mathbf{W}^i_{\text{down}})^\top \Delta\mathbf{a}^i_{\mathcal{S}} \in \text{span}\Big(\{(\mathbf{W}^i_{\text{down}}[s,:])^\top\}_{s\in\mathcal{S}}\Big) .$$

This implies that neuron-level interventions move $\mathbf{h}^i$ along a structured subspace associated with LSNs, rather than arbitrarily perturbing the representation.

Intuitively, if LSNs capture language-dependent factors that bias representations toward a particular language manifold, then shifting their activations can steer the residual stream toward the l_{tgt} representation while preserving the remaining components of the l_{src} representation.

3.3 Steering toward Mean Target Representations

We use the language-wise mean of $\mathbf{a}^i_{\mathcal{S}}$, denoted by $\boldsymbol{\mu}_{\mathbf{a}^i_{\mathcal{S}}}$, as the steering objective (i.e., the target-language in Stage 1 and the source-language in Stage 2). Since our method intervenes at inference-time, using a mean activation pattern provides a stable reference for steering. First, $\boldsymbol{\mu}_{\mathbf{a}^i_{\mathcal{S}}}$ represents a typical activation level under the LSN$^{l_{tgt}}$ activation distribution. It can reduce the drastic distribution shifts compared to extreme statistics, such as the maximum. Second, $\boldsymbol{\mu}_{\mathbf{a}^i_{\mathcal{S}}}$ provides a prototype within the LSN$^{l_{tgt}}$ subspace. It encourages LSN$^{l_{src}}$ activations to steer towards a representative LSN$^{l_{tgt}}$ pattern, rather than towards rare activation regimes that may be related to low-density regions. Additionally, when combined with interpolation, $\boldsymbol{\mu}_{\mathbf{a}^i_{\mathcal{S}}}$ can enhance robustness by reducing sensitivity to a steering strength α relative to alternatives.

Furthermore, replacing $\mathbf{a}^i_{\mathcal{S}_{tgt}}$ with their mean introduces an additive offset along LSN$^{l_{tgt}}$ directions. As this replacement fixes the $\mathbf{a}^i_{\mathcal{S}_{tgt}}$ to a typical value under LSN$^{l_{tgt}}$ activation distribution, it reduces input-dependent variability in LSN$^{l_{tgt}}$ subspace and can make the steering effect more consistent across inputs. Consequently, l_{src} representations are encouraged to move toward a high-density region associated with the l_{tgt} mean activation pattern (restricted to the LSN subspace), facilitating alignment with the l_{tgt} representation space. In Stage 2, we instead steer LSN activations toward the source-language mean to restore source-language generation while reducing residual l_{tgt} features.

Stage 1. The goal of Stage 1 is to shift the l_{src} features and amplify the strength of l_{tgt} features to align $\mathbf{h}^i$ of l_{src} into l_{tgt} representation space. Specifically, for each neuron index j, the adjusted activation $\tilde{a}^i_j$ is computed as:

$$\tilde{a}^i_j = \begin{cases} (1-\alpha)a^i_j + \alpha\boldsymbol{\mu}_{\mathbf{a}^i_{\mathcal{S}_{tgt}}} & \text{if } j \in \mathcal{S}_{src}, \\ \boldsymbol{\mu}_{\mathbf{a}^i_{\mathcal{S}_{tgt}}} & \text{if } j \in \mathcal{S}_{tgt}, \\ a^i_j & \text{otherwise,} \end{cases}$$

where $\alpha \in (0, 1]$ is the steering strength. Note that we do not zero out LSN$^{l_{src}}$ activations. Due to the *polysemantic* nature of neurons (Mondal et al., 2025), setting activations to zero can have unintended effects, may suppress task-relevant signals such as reasoning. We linearly interpolate LSN$^{l_{src}}$ activations to shift them more carefully, aiming to preserve their original contributions while enabling the model to better leverage knowledge encoded in l_{tgt}.

Stage 2. The objective of Stage 2 is to restore l_{src} generation by reactivating LSN$^{l_{src}}$. The $\tilde{a}_j^i$ in Stage 2 is computed as:

$$\tilde{a}_j^i = \begin{cases} \boldsymbol{\mu}_{\mathbf{a}^i_{\mathcal{S}_{src}}} & \text{for } j \in \mathcal{S}_{src}, \\ (1-\alpha)a_j^i + \alpha \boldsymbol{\mu}_{\mathbf{a}^i_{\mathcal{S}_{src}}} & \text{for } j \in \mathcal{S}_{tgt}, \\ a_j^i & \text{otherwise.} \end{cases}$$

Concretely, we set the mean LSN$^{l_{src}}$ activation for LSN$^{l_{src}}$ to reinstate a source-typical activation pattern for LSN$^{l_{src}}$ subspace. At the same time, for LSN$^{l_{tgt}}$, we interpolate a_j^i toward the same source prototype with strength α, gradually reducing residual l_{tgt} features.

4 Experiments

4.1 Experimental Settings

Model and evaluation benchmarks. To verify the effectiveness of SteerLSN, we evaluate it on two popular multilingual LLMs, Qwen3-4B-Instruct (Yang et al., 2025) and Llama3.1-8B-Instruct (AI@Meta, 2024), using two multilingual benchmarks: Global MMLU (Singh et al., 2025) and MGSM (Shi et al., 2023). Global MMLU covers 42 languages and includes two types of tasks, Culturally Sensitive (CS) and Culturally Agnostic (CA). In particular, CS questions can verify whether representation alignment facilitates cross-lingual transfer of language-attributed knowledge. MGSM covers 11 languages and consists of grade-school math problems. Because solving math problems requires multi-step reasoning, it can evaluate whether representation alignment enables models to leverage the more complex reasoning ability typically learned from high-resource languages (e.g., English), where training data are richer and more abundant. We do not involve Chain-of-Thought (CoT) (Wei et al., 2022) to evaluate MGSM. Consistent with prior findings that multilingual LLMs are English-centric, CoT produced by these models are often dominated by English, even when the input is in non-English. This can hinder rigorous analysis of the proposed method's effect. If the model shifts its CoT into English, performance gains may stem from CoT-induced language switching.

Language selection. We evaluate Global MMLU on 12 languages. Using the 11 language set from MGSM as a base, we add Hindi (hi) and Turkish (tr) to increase coverage of non-high-resource languages and exclude Thai (th) because it is not supported in Global MMLU. MGSM is evaluated on its full set of 11 supported languages. Based on the Distribution of Languages in Common Crawl Monthly Archives, we categorize languages with a share of at least 2.5% as high-resource, those with a share between 0.25% (inclusive) and 2.5% (exclusive) as medium-resource, and those with a share below 0.25% as low-resource. Under this criterion, seven languages–English (en), Chinese (zh), Russian (ru), German (de), Japanese (ja), French (fr), and Spanish (es)–are categorized as high-resource; Turkish (tr), Hindi (hi), and Bengali (bn) as medium-resource; and Swahili (sw) as low-resource.

LAPE. We estimate LAPE using the entire language set (i.e., 12 for Global MMLU and 13 for MGSM). Following Tang et al. (2024), we filter neurons by requiring activation probabilities above the 95-th percentile and then select LSNs as the bottom 1% by LAPE score.

SteerLSN. The steering strength α is searched based on the validation performance using a grid search over $(0, 1]$ with a step size of 0.1. To rigorously find α, we split each benchmark dataset into a validation set and a test set, with a ratio of $2:8$. Since MGSM has only 258 samples per language, we evaluate over five different splits and select α by majority vote. The validation set of each split is mutually exclusive from that of the other splits. We use the same α for all layers where the linear interpolation is applied.

4.2 Experimental Results

Layer selection. Our proposed method applies layer-wise shifts to the hidden states, so selecting appropriate layers for each stage is critical. To identify which layers to intervene on effectively, we analyze t-SNE (van der Maaten and Hinton, 2008) of language representations on different layers to visualize how they transition across layers from bottom to top.

Fig. 1 visualizes t-SNE of layer-wise representations from Qwen3-4B-Instruct, computed using parallel sentences across languages. It shows that language representations form distinct clusters in lower layers, and gradually become closer to one another toward the middle layers. Toward the top layers, they diverge again, forming distinct clusters. This observation is consistent with prior studies

https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Bench.	Model	Resource Level: Highest	High						Medium			Low
		en	zh	ru	de	ja	fr	es	tr	hi	bn	sw
Global MMLU	Qwen3-4B	67.88	**62.68**	59.44	58.88	55.94	61.31	60.38	37.50	23.56	**45.50**	29.13
	+ STEERLSN	-	62.31	**59.94**	**60.94**	**58.81**	**61.75**	**61.75**	**45.75**	**30.94**	45.38	**29.81**
	Llama3.1-8B	67.19	55.19	54.50	**56.50**	50.81	**57.12**	56.88	45.75	44.06	41.56	41.12
	+ STEERLSN	-	55.19	**55.50**	56.44	**51.31**	56.44	**58.25**	**46.25**	**44.31**	**42.75**	**41.31**
MGSM	Qwen3-4B	32.80	**27.60**	**34.80**	30.40	23.60	**28.80**	**33.60**	-	-	**26.80**	8.40
	+ STEERLSN	-	24.80	34.00	**31.20**	**24.40**	28.40	33.20	-	-	26.40	**10.00**
	Llama3.1-8B	9.60	1.20	**14.00**	**14.40**	5.60	**13.60**	11.60	-	-	**11.60**	10.80
	+ STEERLSN	-	**11.20**	13.20	14.00	**8.00**	12.80	**12.40**	-	-	11.20	**11.20**

Table 1: Accuracy on Global MMLU and MGSM for Qwen3-4B-Instruct and Llama3.1-8B-Instruct. All values are percentages. **Bold** represents the best performance between the baseline and STEERLSN. "Bench." denotes the benchmark.

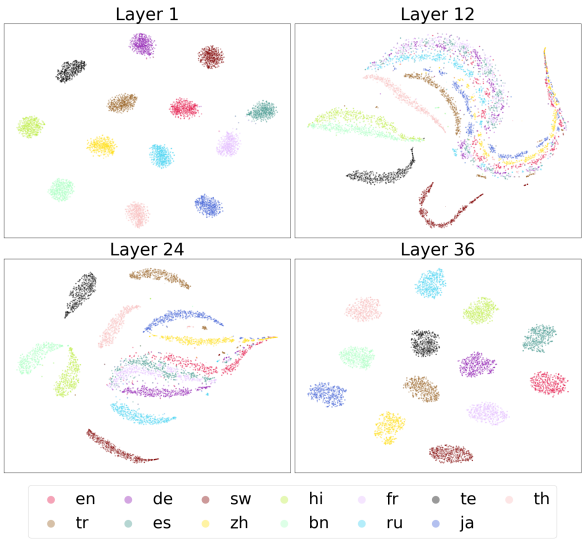


Figure 1: t-SNE of layer-wise representation from Qwen3-4B-Instruct.

suggesting that the internal mechanisms of multilingual LLMs are not inherently language-neutral, but instead predominantly English-centric, reflecting the bias inherent in high-resource training data (Etxaniz et al., 2024; Wendler et al., 2024; Schut et al., 2025; Zeng et al., 2025). Based on Fig. 1, we apply Stage 1 to the lower-middle through upper-middle layers ($25\% - 75\%$ to the model depth), where language representations become more aligned with English after being encoded with language-specific characteristics (e.g., script and syntactic structure) in lower layers. Stage 2 is applied to the upper layers, $75\% - 100\%$, where representations are transformed back to the source language for output generation. Comparison between different layer selections will be discussed in Sec. 4.3.

Overall. As shown in Tab. 1, the proposed method generally improves performance on Global MMLU, whereas MGSM exhibits a different performance improvement tendency. On Global MMLU, most languages have performance gains over the baselines. We also observe performance enhancement for five of the 6 high-resource languages, suggesting that alignment to a higher-resource target can be beneficial even when the source language is already high-resource. In contrast, improvements for sw are relatively modest; we speculate that extremely limited training data may constrain the model's ability to encode and exploit representations for this language. Medium-resource languages benefit most from STEERLSN. In particular, applying STEERLSN to *tr* and *hi* improves accuracy by 8.25% and 7.38%, respectively, over the Qwen3-4B-Instruct baseline. We infer that, unlike low-resource scenarios, medium-resource languages retain sufficient capacity to encode well-formed representations, yet may still lack language-specific knowledge coverage. This makes them especially well-suited to gain from improved access to the richer knowledge encoded in high-resource representations.

Why does STEERLSN show limited gains on MGSM? On MGSM, STEERLSN yields little to no improvement over the baselines. We hypothesize that the limited gains on MGSM stem from the nature of the benchmark itself. STEERLSN is designed to improve access to language-attributed, knowledge-intensive representations by aligning language-specific subspaces. In contrast, MGSM primarily evaluates procedural, multi-step reasoning that depends on composing intermediate states, rather than retrieving or contextualizing language-dependent knowledge. Under this hypothesis, improving conceptual access alone may be insufficient to enhance performance on MGSM, suggest-

ing that complementary mechanisms beyond representation alignment are required for such tasks.

4.3 Analysis

Language-attributed knowledge transfer. Tab. 2 provides detailed results of Global MMLU. CS results are computed using only questions related to English-speaking countries, including "USA", "UK", and "Canada", as English is the target language. STEERLSN yields comparable improvements on CS questions and CA ones, suggesting that its benefits are not limited to broadly shared facts but are also consistent with improved access to representations underlying language-associated or culture-linked information. For example, a CS question—"Which statement best describes one of Dworkin's central arguments in *Justice for Hedgehogs*?"—requires familiarity with USA-centric jurisprudence context. For sw, the model answers this question correctly after applying STEERLSN ("Moral values are both independent and objective."), whereas it previously predicted an incorrect option ("Morality plays no role in the concept of law."), indicating that alignment can alter how language-specific representations are mapped to task-relevant semantic distinctions.

Is the improvement really due to target-language knowledge? To more rigorously assess whether CS improvements are driven by culture-linked knowledge, we vary the target language and consider a non-English high-resource target (zh). As shown by hi results in Tab. 2, even though zh is among the highest-resource languages, the CS improvement is smaller than when English is the target, whereas the CA improvement is larger. This observation supports the view that representation alignment via STEERLSN induces target-dependent alignment effects, which can differentially influence performance on CS and CA questions.

The effect of linear interpolation of each Stage. To verify whether STEERLSN operates well as intended at the representation level, we compute the cosine distance between source- and target-language (i.e., English) representations using Qwen3-4B-Instruct. We set tr as the source-language. As presented in Fig. 2, applying STEERLSN in Stage 1 ($25\% - 75\%$ to the model depth) shifts the source-language representation closer to the English representation, consistent with the intended role of Stage 1 in promoting alignment. In contrast, during Stage 2, using the mean activation of source LSNs as the interpolation endpoint gradually moves representations away from English, bringing the cosine distance closer to the baseline trajectory.

Lang.	CS-en	CA
zh	58.82	66.82
+STEERLSN	58.82 (+ 0.0)	66.82 (+ 0.00)
ru	56.86	59.64
+STEERLSN	58.82 (+ 1.96)	59.64 (+ 0.00)
de	43.14	60.27
+STEERLSN	52.94 (+ 9.80)	63.01 (+ 2.74)
ja	35.29	55.45
+STEERLSN	39.22 (+ 3.93)	59.55 (+ 4.10)
fr	54.90	65.62
+STEERLSN	54.90 (+ 0.0)	65.18 (- 0.44)
es	54.90	61.54
+STEERLSN	60.78 (+ 5.88)	64.25 (+ 2.71)
tr	27.45	40.54
+STEERLSN	37.25 (+ 9.80)	48.20 (+ 7.66)
bn	37.25	42.92
+STEERLSN	35.29 (- 1.96)	39.73 (- 3.19)
sw	27.45	27.85
+STEERLSN	27.45 (+ 0.0)	29.68 (+ 1.83)
hi	23.53	25.68
+STEERLSN	27.45 (+ 3.92)	32.88 (+ 7.20)
+STEERLSN→zh	25.49 (+ 1.96)	33.78 (+ 8.10)
Avg. +STEERLSN	45.29 (+ 3.33)	52.89 (+ 2.26)

Table 2: Performance on CS and CA tasks of Global MMLU for Qwen3-4B-Instruct per language after applying STEERLSN. Parenthesized values denote changes relative to the model without STEERLSN. All values are percentages. Red and blue means improvements and decreases, respectively.

Performance tendency in layer selection. To analyze the effect of layer selection, we compare two variants: 1) early Stage 1, which applies Stage 1 to 0-50% of the model depth and Stage 2 to 75-100%, and 2) early Stage 2, which applies Stage 1 to 25-50% and Stage 2 to 50-100%. As reported in Fig. 3, steering source representations toward the target language at the beginning of the model is less effective. Likewise, initiating the recovery of source-language generation from the middle layers is generally ineffective, likely reducing the model's ability to leverage knowledge encoded in the target-language representations. This aligns with Fig. 1, where lower layers encode language-specific structural information and mid layers show

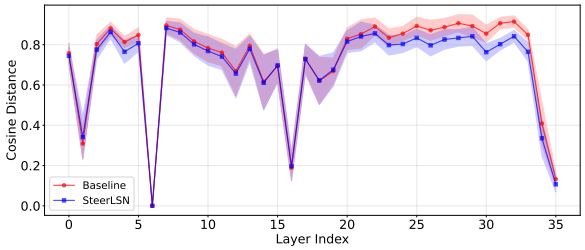


Figure 2: Layer-wise cosine distance between Turkish and English representations in Qwen3-4B-Instruct when input language is Turkish.

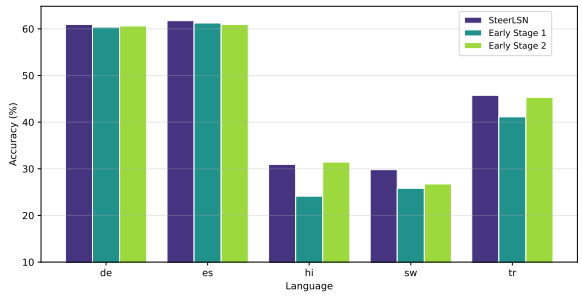


Figure 3: Ablation on layer selection for Qwen3-4B-Instruct.

the strongest cross-lingual convergence. Moreover, triggering Stage 2 too early can prematurely reintroduce source-language biases, and may weaken the target-oriented shift before it is sufficiently established.

Text generation after each Stage. Tab. 3 presents examples from Qwen3-4B-Instruct before and after applying SteerLSN stages sequentially. We use es as the source language and en as the target. After Stage 1, the model generates fluent English, and after Stage 2 it successfully returns to generate outputs in the source language, as intended by design of the proposed approach.

5 Conclusion

In this paper, we introduce SteerLSN, a training-free method that aligns low-resource language representations with high-resource ones by intervening on LSN activations, enabling models to better leverage knowledge encoded in high-resource languages and thereby improving downstream task performance. The experimental results demonstrate that the proposed method generally improves task performance in low-resource scenarios, suggesting that carefully controlled shifts of language representations can make representation alignment an effective mechanism for cross-lingual knowledge transfer. We expect that this LSN-based intervention framework can be extended to a wider spectrum of linguistic pairs and complex reasoning tasks, demonstrating that structural steering is a viable and computationally effective alternative to extensive finetuning for downstream performance in low-resource scenarios.

Prompt	¿Cuáles son las principales causas del cambio climático?
Baseline	Las principales causas del cambio climático son las actividades humanas que aumentan la concentración de gases de efecto invernadero (...)
SteerLSN Stage 1	The main causes of climate change are primarily human activities that increase the concentration of greenhouse gases (GHGs) in the atmosphere. These gases trap heat from the sun, leading to a warming effect known as the greenhouse effect.(...)
SteerLSN Stage 2	Las principales causas del cambio climático are: 1. Emissions de gases de efecto invernadero (GEIs)

Table 3: Generated samples of Qwen3-4B-Instruct baseline and its outputs with applying SteerLSN stage sequentially.

Limitations

Although our method improves downstream task performance, the gains are modest. We infer that this is due to the *polysemantic* nature of neuron activations: a single neuron may encode linguistic properties (e.g., script or syntactic structure) while also contributing to task-relevant ability (e.g., reasoning). As a result, modifying activations based solely on an LSN criterion may inadvertently distort signals that are essential for solving the task. Future work that better disentangles language-related and task-related contributions could enable more precise control of cross-lingual knowledge transfer. A second limitation concerns our use of mean LSN activations, which may reduce the representational richness of the original, context-dependent activation patterns. An alternative is to sample replacement values from the target-language activation distribution, potentially preserving variability while still encouraging alignment.

Use of AI Assistants

We used an AI assistant (ChatGPT) to polish the writing of this paper, limited to grammar correction

and sentence-level wording suggestions. All suggestions were reviewed and revised by the authors.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) under the Leading Generative AI Human Resources Development(No. RS-2025-25441313, Contribution: 50%) grant funded by the Korea government(MSIT), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-23524633, A Study of Adaptive Vision-Language-Action Models through Interaction in Heterogeneous Environments, Contribution: 50%). This work utilized GPU resources from the "Advanced GPU Utilization Support Program" funded by MSIT, Republic of Korea.

References

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

AI@Meta. 2024. Llama 3 model card.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Zhipeng Chen, Kun Zhou, Liang Song, Wayne Xin Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. 2025. Extracting and combining abilities for building multi-lingual ability-enhanced large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17574–17591, Suzhou, China. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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.

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

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

GemmaTeam. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697.

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

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Nurgali Kadyrbek, Zhanseit Tuimebayev, Madina Mansurova, and Vítor Viegas. 2025. The development of small-scale language models for low-resource languages, with a focus on kazakh and direct preference optimization. *Big Data and Cognitive Computing*, 9(5).

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

Qiming Li, Xiaocheng Feng, Yixuan Ma, Zekai Ye, Ruihan Chen, Xiachong Feng, and Bing Qin. 2025. Unlocking multilingual reasoning capability of llms and lvlms through representation engineering. *Preprint*, arXiv:2511.23231.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.

Cen Lu, Yung-Chen Tang, and Andrea Cavallaro. 2025. Minimal neuron ablation triggers catastrophic collapse in the language core of large vision-language models. *Preprint*, arXiv:2512.00918.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189.

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

MistralAI. 2025. Introducing mistral 3.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20465–20497.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Chuancheng Shi, Shangze Li, Shiming Guo, Simiao Xie, Wenhua Wu, Jingtong Dou, Chao Wu, Canran Xiao, Cong Wang, Zifeng Cheng, Fei Shen, and Tat-Seng Chua. 2025. Where culture fades: Revealing the cultural gap in text-to-image generation. *Preprint*, arXiv:2511.17282.

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

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

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

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Yuemei Xu, Kexin Xu, Jian Zhou, Ling Hu, and Lin Gui. 2025. Linguistic neuron overlap patterns to facilitate cross-lingual transfer on low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27646–27661.

Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. 2025. Neuron-level analysis of cultural understanding in large language models. *Preprint*, arXiv:2510.08284.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Unveiling linguistic regions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Bangkok, Thailand. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, volume 37, pages 15296–15319.

A Evaluation Prompt & Generated Samples over Languages

Fig. 4 provides the prompt used to evaluate the MGSM benchmark, with a system prompt consisting of language options.

MGSM Prompt

[System]

'en': "Solve the math problem step by step and end your response with 'The answer is <number>'.",
'de': "Lösen Sie das mathematische Problem Schritt für Schritt und beenden Sie Ihre Antwort mit 'Die Antwort ist <Zahl>'.",
'es': "Resuelve el problema matemático paso a paso y termina tu respuesta con 'La respuesta es <número>'.",
'fr': "Résolvez le problème mathématique étape par étape et terminez votre réponse par 'La réponse est <nombre>'.",
'ja': "数学の問題をステップバイステップで解き、'答えは<数字>'で回答を締めくくってください。",
'ru': "Решите математическую задачу пошагово и закончите свой ответ фразой 'Ответ: <число>'.",
'zh': "逐步解决这个数学问题，并在回答结束时加上'答案是<数字>'。",
'bn': "গাণিতিক সমস্যাটি ধাপে ধাপে সমাধান করুন এবং আপনার উত্তর 'উত্তর হল <সংখ্যা>' দিয়ে শেষ করুন ",
'te': "గణిత సమస్యను అంచెలంచెలుగా పరిష్కరించి, మీ సమాధానాన్ని 'సమాధానం <సంఖ్య>' తో ముగించండి.",
'th': "แก้ปัญหาคณิตศาสตร์ทีละขั้นตอนและจบคำตอบของคุณด้วย 'คำตอบคือ <ตัวเลข>'.",
'sw': "Tatua tatizo la hisabati hatua kwa hatua na umalizie jibu lako na 'Jibu ni <nambari>'."

[User]

{question}

Figure 4: MGSM evaluation prompt

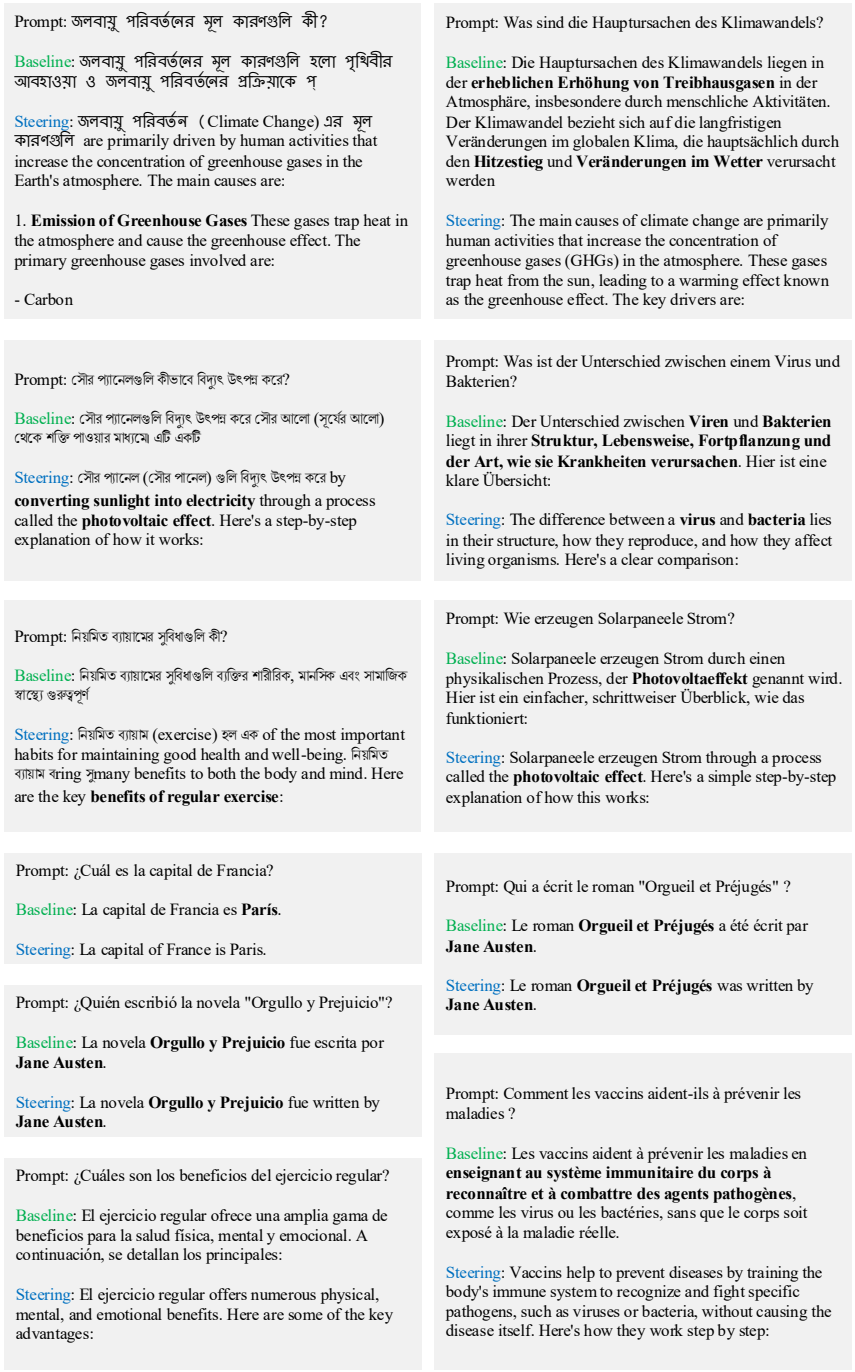


Figure 5: Samples generated by Qwen3-4B-Instruct before and after applying Stage 1 of STEERLSN. Clockwise from the top left: Bengali, German, French, and Spanish.



Figure 6: Samples generated by Qwen3-4B-Instruct before and after applying Stage 1 of STEERLSN. Clockwise from the top left: Chinese, Japanese, Hindi, and Swahili.