

SARA: Unlocking Multilingual Knowledge in Mixture-of-Experts via Semantically Anchored Routing Alignment

Tianyu Dong¹, Yangyang Liu², Jiang Zhou¹, Xinwei Wu¹, Xiaohu Zhao², Hao Wang², Heng Liu², Linlong Xu², Longyue Wang², Weihua Luo², Shaolin Zhu^{1*}, Deyi Xiong¹

¹TJUNLP Lab, School of Computer Science and Technology, Tianjin University, China

²Alibaba Group, China

{tydong, zhushaolin, dyxiong}@tju.edu.cn

Abstract

Sparse Mixture-of-Experts (MoE) architectures have emerged as an increasingly influential paradigm as they offer a strategic balance between parameter scalability and computational efficiency. However, low-resource languages—which suffer from a scarcity of high-quality training data—often have their tokens routed to different experts than those predominantly activated by high-resource inputs, which limits cross-lingual expert sharing. This cross-lingual routing divergence consequently hinders their efficacy in multilingual contexts. To address this issue, we propose **SARA** (Semantically Anchored Routing Alignment), a framework designed to transfer specialized capabilities from high-resource languages as anchors to low-resource languages. SARA explicitly aligns the routing distribution of multilingual inputs with high-resource semantic anchors using a symmetric Jensen-Shannon (JS) divergence constraint. Unlike traditional distillation methods that operate on output logits, SARA directly aligns the internal routing distributions of MoE layers, encouraging mechanistic consistency in expert selection across languages. We conduct experiments on 2 LLMs across 5 low-resource languages and 3 benchmarks. Experiment results demonstrate that SARA outperforms standard instruction tuning (e.g., +0.8% on Qwen3-30B-A3B and +1.2% on Phi-3.5-MoE-instruct on Global-MMLU benchmark). Further analyses show that SARA effectively addresses performance bottlenecks in low-resource languages, providing a scalable pathway to enhance multilingual capabilities in sparse architectures. Our code is available at <https://github.com/iMoriton/sara>.

1 Introduction

The paradigm of LLMs has shifted toward sparse MoE architectures, as evidenced by recent open-weight milestones such as Mixtral (Jiang et al.,

2024), DeepSeek-V3 (Liu et al., 2024), DeepSeek-R1 (Guo et al., 2025a) and Qwen3 (Yang et al., 2025). By decoupling model capacity from computational cost, MoE models achieve remarkable scalability and allow distinct subsets of parameters or experts to specialize in specific domains (Dai et al., 2024). However, the efficacy of MoE primarily relies on routers correctly dispatching tokens to the most competent experts. While these models show exceptional capability in their dominant training languages (e.g., English), extending their specialized prowess to a broader spectrum of low-resource languages remains a formidable challenge (Imani et al., 2023; Etxaniz et al., 2024; Zhu et al., 2024b).

Empirical analyses of the internal mechanisms of MoE reveal a fundamental bottleneck: cross-lingual routing divergence where semantically equivalent inputs in different languages trigger disparate expert activation pathways (Bandarkar et al., 2025). Recent studies have identified the existence of super experts (Su et al., 2025), defined as a sparse subset of parameters responsible for encoding sophisticated domain knowledge such as mathematical reasoning. However, these components are predominantly optimized for the data-rich patterns of high-resource languages. When the model processes semantically equivalent inputs in low-resource languages, the routing network often struggles to generalize. Due to surface-level lexical variations, the router fails to dispatch tokens to these high-competence experts and instead directs them toward generalist or irrelevant pathways (Chi et al., 2022). As a result, the model may possess the necessary parametric knowledge, but lack the routing logic required to activate it for low-resource language inputs. Such routing misalignment not only degrades reasoning performance in low-resource languages, but also disrupts the consistency of internal representations across languages. As observed by Bandarkar et al. (2025),

*Corresponding author.

this inconsistency leads to performance limitations where the hidden states of low-resource inputs diverge from their high-resource counterparts due to disparate expert composition.

Recent evaluations highlight that existing LLMs still encounter significant performance plateaus across diverse and nuanced multilingual scenarios (Zhang et al., 2026; Chen et al., 2025). Prior efforts to enhance multilingualism in LLMs have largely focused on continual pre-training or instruction tuning (Li et al., 2025c; Zhu et al., 2024a; Li et al., 2025a). While recent efforts have explored expert pruning (Zhang et al., 2025b) or load balancing optimization (Guo et al., 2025b) to improve efficiency, they primarily optimize the routing mechanism for computational throughput rather than cross-lingual semantic consistency. Recent works (Zhou et al., 2025; Dong et al., 2025; Zhu et al., 2025) regulate updates via routing priors or parameter detection, revealing that unconstrained training conversely leads to catastrophic forgetting. Consequently, the core challenge remains unaddressed: how to encourage semantically equivalent inputs to trigger similar expert activation pathways across languages?

To bridge this gap, we propose **Semantically Anchored Routing Alignment (SARA)**, a novel and statistics-driven framework designed to align expert activation patterns across languages. Our method leverages the robust routing distributions of high-resource languages as semantic anchors to rectify the routing behaviors observed in low-resource settings. Unlike traditional distillation targeting output logits, SARA explicitly minimizes the distributional discrepancy in the routing probability space. This encourages the routing logic to remain invariant to the input language, effectively transferring the model’s capabilities to low-resource languages. We implement this via a multi-stage pipeline: (1) constructing a semantically aligned parallel instruction corpus; (2) extracting reliable routing priors from dominant languages; (3) fine-tuning via a Jensen-Shannon (JS) divergence constraint that penalizes routing deviations.

Our contributions are threefold:

(i) We propose the SARA framework, which moves beyond traditional token-level distillation by treating the routing probability distribution of high-resource languages as semantic anchors.

(ii) By applying a symmetric JS divergence constraint on intermediate layers, SARA effectively leverages the model’s existing high-resource knowl-

edge to make low-resource languages attain high-resource inference capabilities by rectifying their expert activation pathways.

(iii) We conduct a comprehensive evaluation across 2 LLMs and 3 challenging benchmarks (e.g., +0.8% on Qwen3-30B-A3B and +1.2% on Phi-3.5-MoE-instruct on Global-MMLU benchmark). Our results outperform standard instruction tuning, offering a scalable pathway for enhancing multilingual capabilities in sparse architectures.

2 Related Work

Multilingual Alignment. Research on cross-lingual alignment in LLMs has predominantly focused on instruction fine-tuning and the strategic curation of training data. Recent studies demonstrate that the specific composition of supervised fine-tuning (SFT) data and the integration of parallel instruction corpora (Penedo et al., 2025) are pivotal for enhancing functional consistency across languages. Frameworks such as Lingualift (Zhang et al., 2024) introduce effective two-stage instruction tuning strategies specifically tailored to bolster performance in low-resource tasks. Aya (Üstün et al., 2024) and JetMoE (Shen et al., 2024) further emphasize that distilling low-resource language knowledge from teacher models can bridge the performance gap of student models in low-resource languages. Additionally, adapting models to severely underrepresented languages highlights the critical necessity of large-scale, language-specific data curation combined with continual pre-training (Pan et al., 2025). While these methods demonstrate substantial improvements, they are fundamentally data-centric and mainly rely on the availability of high-quality parallel instructions and expert-curated multilingual corpora.

Efficiency-Oriented MoE Architectures. Recent advances in MoE have focused on improving the quality of specialized parameter subsets through diverse initialization and hybrid designs. SCoMoE (Zeng and Xiong, 2023) optimizes MoE training efficiency by restructuring communication patterns to alleviate expensive all-to-all operations. Jamba (Lieber et al., 2024) and ERNIE 4.5 (Baidu-ERNIE-Team, 2025) explore architectural fusion and heterogeneous scaling to optimize inference efficiency. Meanwhile, research into dynamic routing has sought to refine expert selection density based on task difficulty (Huang et al., 2024; Yang et al., 2024) and mitigate computational redun-

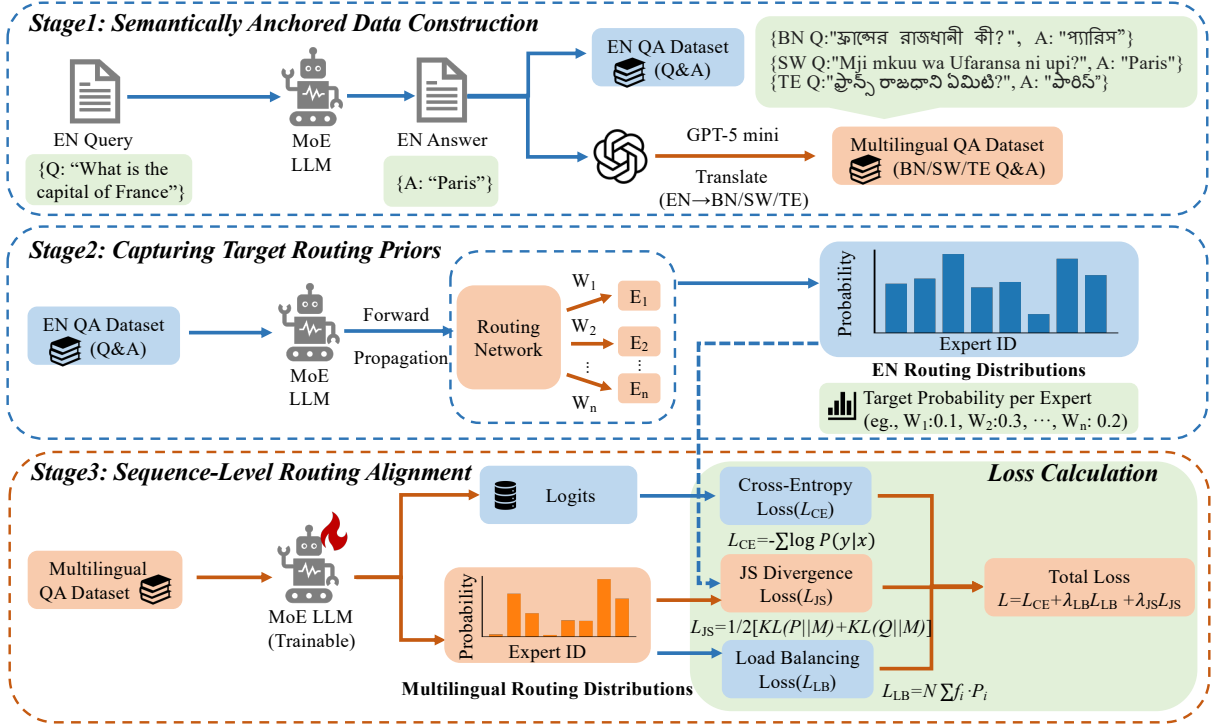


Figure 1: Illustration of the proposed SARA framework. The framework consists of three stages: (1) Generate semantically aligned parallel data via GPT-5 mini translation; (2) Perform forward propagation on high-resource inputs to extract dense routing probability distributions as target priors; (3) Fine-tune the model with a composite objective. This minimizes the Jensen-Shannon divergence between the multilingual and anchor routing distributions, along with cross-entropy and load balancing losses, to encourage cross-lingual expert consistency.

dancy via expert pruning (Lu et al., 2024). However, these optimizations are primarily designed for general-purpose efficiency and lack specialized mechanisms for multilingual tasks.

Routing Distribution Adjustments. A bottleneck in the multilingual capability of MoE is cross-lingual routing divergence (Bandarkar et al., 2025). Semantically equivalent inputs in different languages are often routed to distinct experts. This phenomenon undermines representational consistency and generalization across languages. Although existing routing-centric frameworks such as HyperMoE (Zhao et al., 2024) and RoMA (Li et al., 2025b) have explored adjustments to routing distribution, their objectives are largely confined to task-specific knowledge transfer or intra-domain generalization rather than cross-lingual semantic alignment. Consequently, they do not resolve the routing mismatch that occurs when moving from high-resource to low-resource linguistic contexts. Our SARA implements internal routing distillation specifically for multilingual consistency. By utilizing the model’s own routing distributions on high-resource languages, SARA encourages mech-

anistic consistency across the routing distribution.

3 Proposed Framework

To address the bottlenecks arising from routing misalignment in MoE, we propose the SARA framework. The key idea is to treat the expert activation patterns of high-resource languages, as the ground-truth signal for semantic processing. As illustrated in Figure 1, SARA first constructs strictly semantic-aligned parallel data to ensure content consistency. Then we perform offline forward propagation to extract the routing distributions of high-resource languages as priors. Finally, a distribution alignment objective is applied to pull the routing distributions of multilingual inputs toward those of their anchor texts. This encourages semantically equivalent inputs to induce similar expert activation distributions at the routing level, bridging the cross-lingual gap.

3.1 Semantically Anchored Data Construction

The goal of SARA is to align the routing logic of low-resource languages with that of high-resource anchors. To establish a reliable semantic foundation for this anchoring process, we construct a par-

allel instruction dataset that captures the model’s intrinsic inference logic in high-resource contexts:

- **Correctness-based Filtering:** We first perform inference on high-resource training sets (e.g., English or Chinese). As detailed in Appendix A.7 (referencing A.6), we require the model to enclose the final answer or option within a `\boxed{}`. To objectively verify the results, we utilize the regular expression `r"\boxed{(. *?)}"` to extract all formatted contents from the model’s output. We designate the final element in the resulting list as the predicted answer. Only samples where the model generates a correct answer are retained. These verified samples serve as the standard for extracting internal routing patterns.
- **Parallel Corpus Synthesis:** For each verified sample, we translate the entire interaction (including the prompt and the logical reasoning steps) into target low-resource languages. This ensures that the low-resource samples are semantically identical to the high-resource anchor samples. By maintaining strict content consistency, the discrepancy in routing can be primarily attributed to linguistic variation, providing a clean supervision signal for alignment.

3.2 Capturing Target Routing Priors

We extract routing distributions from high-resource data to obtain the optimal routing strategy for cross-lingual alignment. We perform forward propagation on the concatenated sequence of the prompt and its verified response to capture the complete inference process. We adopt the standard sparse MoE architecture. Given a token representation $\mathbf{h} \in \mathbb{R}^d$ (where d is the hidden dimension), a gating network parameterized by $\mathbf{W}_g \in \mathbb{R}^{d \times N}$ (with N denoting the number of experts) computes the routing scores u and the probability distribution \mathbf{P} :

$$\mathbf{u} = \mathbf{h} \cdot \mathbf{W}_g, \quad (1)$$

$$\mathbf{P} = \text{Softmax}(\mathbf{u}). \quad (2)$$

MoE models typically use a top- k selection strategy to improve efficiency. Let \mathcal{T} denote the set of indices for the top- k probabilities, and \mathbf{E}_j represent the j -th expert network. The sparse output $G(\mathbf{h})$ and the layer output \mathbf{y} are defined as:

$$G(\mathbf{h})_j = \begin{cases} P_j, & \text{if } j \in \mathcal{T} \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

$$\mathbf{y} = \sum_{j \in \mathcal{T}} G(\mathbf{h})_j \cdot \mathbf{E}_j(\mathbf{h}). \quad (4)$$

To obtain stable semantic routing priors, we utilize the dense routing probability distributions produced by the gating network, rather than the discrete top- k expert selections. The rationale is that for low-resource languages, the discrete expert selection during inference often diverges from high-resource anchors.

For each anchor interaction i , let $\mathbf{S}_{\text{anchor}}^{(i)} = [\mathbf{x}_i; \mathbf{y}_i]$ denote the full sequence, where \mathbf{x}_i represents the input prompt and \mathbf{y}_i represents the verified response. We perform forward propagation on this concatenated sequence to capture the routing logic of the complete inference process. We record the softmax-normalized routing distribution \mathbf{R}_l at each layer l for every token $t \in \mathbf{S}_{\text{anchor}}^{(i)}$:

$$\mathbf{R}_l(\mathbf{S}_{\text{anchor},t}^{(i)}) = \text{Softmax}(\mathbf{h}_{\text{anchor},t}^{(l)} \cdot \mathbf{W}_g) \in \mathbb{R}^N. \quad (5)$$

To obtain a stable routing distribution, we aggregate token-level routing probabilities into a sequence-level prior $\bar{\mathbf{P}}_{\text{anchor}}^{(l,i)}$ using validity masks \mathbf{m}_t :

$$\bar{\mathbf{P}}_{\text{anchor}}^{(l,i)} = \frac{1}{\sum_t \mathbf{m}_t} \sum_t \mathbf{m}_t \cdot \mathbf{R}_l(\mathbf{S}_{\text{anchor},t}^{(i)}). \quad (6)$$

The mask \mathbf{m}_t filters out [PAD] tokens to prevent non-semantic routing noise from biasing the aggregated distribution.

3.3 Sequence-Level Routing Alignment

In this stage, we fine-tune the model on multilingual inputs to align their expert activation patterns with the extracted high-resource priors. For a multilingual interaction i , we similarly concatenate the translated prompt $\mathbf{x}_{\text{lang}}^{(i)}$ and response $\mathbf{y}_{\text{lang}}^{(i)}$ to form $\mathbf{S}_{\text{lang}}^{(i)} = [\mathbf{x}_{\text{lang}}^{(i)}; \mathbf{y}_{\text{lang}}^{(i)}]$. We then compute its sequence-level routing profile $\bar{\mathbf{Q}}_{\text{lang}}^{(l,i)}$:

$$\bar{\mathbf{Q}}_{\text{lang}}^{(l,i)} = \frac{1}{\sum_t \mathbf{m}_t} \sum_t \mathbf{m}_t \cdot \mathbf{R}_l(\mathbf{S}_{\text{lang},t}^{(i)}). \quad (7)$$

Routing Alignment Loss (\mathcal{L}_{JS}). We align the multilingual profile $\bar{\mathbf{Q}}_{\text{lang}}^{(l,i)}$ with the high-resource anchor prior $\bar{\mathbf{P}}_{\text{anchor}}^{(l,i)}$ by minimizing the Jensen–Shannon (JS) divergence:

$$\mathcal{L}_{\text{JS}} = \frac{1}{|L_{\text{target}}|} \sum_{l=L_{\text{start}}}^{L_{\text{end}}} \text{JS}(\bar{\mathbf{P}}_{\text{anchor}}^{(l,i)} \parallel \bar{\mathbf{Q}}_{\text{lang}}^{(l,i)}), \quad (8)$$

where $|L_{\text{target}}|$ is the number of selected layers, and L_{start} and L_{end} denote the start and end indices of the intermediate layers (e.g., layers 7 to 34 for Qwen3). The JS divergence is defined via the Kullback-Leibler (KL) divergence as:

$$\text{JS}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{2}\text{KL}(\mathbf{P}\|\mathbf{M}) + \frac{1}{2}\text{KL}(\mathbf{Q}\|\mathbf{M}), \quad (9)$$

with $\mathbf{M} = \frac{1}{2}(\mathbf{P} + \mathbf{Q})$ being the average distribution. The choice of JS divergence over the standard KL divergence is critical due to the following 2 factors:

- **Symmetric Semantic Regularization:** Unlike KL divergence, JS divergence is symmetric, providing a balanced metric for distributional similarity. Since high-resource anchor inputs and their multilingual counterparts represent different linguistic realizations of the same semantic intent, JS divergence facilitates a more stable alignment toward a shared routing distribution.
- **Numerical Stability in Sparse Routing:** In MoE gating, the softmax function assigns non-zero routing probabilities to all experts, but in practice, the probabilities for unselected experts are often small. KL divergence is hypersensitive to low-probability tails, leading to disproportionately large or vanishing gradients.

Task Loss (\mathcal{L}_{CE}). To maintain the model’s fundamental generative and reasoning capabilities, we apply the standard cross-entropy loss for next-token prediction to the model output in the training data:

$$\mathcal{L}_{\text{CE}} = - \sum_{t=1}^T \log P(\mathbf{y}_t | \mathbf{x}_{<t}). \quad (10)$$

Load Balancing Loss (\mathcal{L}_{LB}). We retain the load balancing loss from Switch Transformers (Fedus et al., 2022) to prevent expert collapse and ensure efficient parameter utilization, formulated as

$$\mathcal{L}_{\text{LB}} = N \sum_{i=1}^N f_i \cdot \bar{\mathbf{P}}_i, \quad (11)$$

where N is the number of experts, f_i is the fraction of tokens dispatched to expert i , and $\bar{\mathbf{P}}_i$ denotes the average routing probability across the batch.

The final training process uses a composite objective that balances task performance, expert utilization and cross-lingual routing consistency:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{LB}}\mathcal{L}_{\text{LB}} + \lambda_{\text{JS}}\mathcal{L}_{\text{JS}}, \quad (12)$$

where λ_{JS} is the hyperparameter controlling the strength of the semantically anchored routing alignment.

4 Experiments

We extensively evaluated the effectiveness of the proposed SARA framework. We also conducted ablation studies and in-depth analysis to verify the specific contribution of routing alignment in low-resource languages.

4.1 Datasets

We constructed semantically anchored instruction-tuning datasets for 5 low-resource languages (hi, ne, bn, te, sw) by applying the pipeline described in Section 3.1 to the high-resource splits of MMLU-ProX (Xuan et al., 2025) and GSM8K (Cobbe et al., 2021; Yu et al.). * We established 2 independent experimental tracks using English and Chinese as high-resource anchors, respectively, to evaluate the robustness of SARA across different semantic pivots. We utilized Qwen3-30B-A3B for initial inference and capture routing distributions from samples verified for correctness. For each anchor, the filtered interactions were translated into 5 target languages using GPT-5 mini, as detailed in Appendix A.1.

To ensure a rigorous zero-shot evaluation, we removed any training samples overlapping with the test sets of our evaluation benchmarks. Following this process, the final corpus for each independent anchor track includes 7,000 parallel samples derived from MMLU-ProX and 7,000 samples from GSM8K per target language.

We evaluate performance on 3 diverse multilingual benchmarks: Global-MMLU (Singh et al., 2025), BELEBELE (Bandarkar et al., 2024) and MGSM (Shi et al.).

4.2 Baselines

We compared SARA against the following baselines: (1) **Vanilla LM:** The original instruction-tuned sparse MoE model (Qwen3-30B-A3B) without any further cross-lingual alignment. (2) **FFT:** Standard supervised full fine-tuning on the same translated dataset, but without the routing alignment objective (\mathcal{L}_{JS}). (3) **AES** (Guo et al., 2025b):

*We note that there is ongoing discussion in the community regarding whether hi(Hindi) should be categorized as a low-resource language (Holtermann et al., 2024; Hangya et al., 2022; Singh et al., 2023; Dubossarsky and Dairkee, 2024; Sharma and Bhattacharyya, 2025; Anoop et al., 2021).

It introduces an orthogonality loss to reduce representational overlap among experts and a variance loss to encourage discriminative routing decisions. (4) **ShifCon** (Zhang et al., 2025a): It enhances multilingual capabilities by aligning the internal representations of non-dominant languages with the dominant language subspace.

4.3 Settings

We implemented SARA using the PyTorch framework and fine-tune Qwen3-30B-A3B for 2 epochs with a global batch size of 256. All models were trained on $16 \times$ NVIDIA H100 80GB GPUs, and each training session took approximately 15 hours. The learning rate was set to $2e-5$ with a cosine decay scheduler. Regarding loss coefficients, the load balancing weight λ_{LB} follows the model’s default configuration, while the routing alignment weight λ_{JS} is fixed at 1.5. We selectively applied \mathcal{L}_{JS} to the intermediate layers ($l \in [7, 34]$) to focus the supervision signal on the model’s language-agnostic semantic core (see Appendix A.2 for details). The English-anchored and Chinese-anchored models were trained as independent instances to verify the framework’s robustness across different high-resource origins. A detailed sensitivity analysis of λ_{JS} and the rationale for layer selection are provided in Appendix A.3 and A.2. During evaluation, we reported the average performance over 3 independent runs (Top- $p = 1$, temperature = 0.1).

4.4 Main Results

The performance of SARA on Global-MMLU, BELEBELE and MGSM is summarized in Table 1. Qwen3-30B-A3B shows limited cross-lingual generalization with a pronounced performance drop on low-resource languages such as Swahili (sw) and Telugu (te). FFT substantially improves multilingual performance and serves as a strong baseline. However, FFT does not explicitly address cross-lingual routing divergence. We perform one-tailed paired t-tests comparing SARA and FFT; the detailed statistical protocol and results are provided in Appendix A.4. We further compare SARA with AES and ShifCon under 2 anchor settings. AES relies on expert specialization and regularization, but often underperforms FFT. ShifCon applies representation-level alignment and is generally competitive with FFT. However, neither method consistently matches SARA’s performance on low-resource languages. This suggests that indirect regularization or hidden-state alignment is insufficient

to correct cross-lingual expert routing.

We also find that results with a Chinese anchor are weaker than those with an English anchor on all methods. This indicates less stable routing behavior when Chinese is used as the semantic pivot. The gap largely reflects the model’s lower baseline performance in Chinese. By directly aligning routing probability distributions between high-resource anchors and their multilingual counterparts, SARA achieves the best average performance under the English anchor. We further validate the effectiveness of SARA on Phi-3.5-MoE-instruct (Abdin et al., 2024) in Appendix A.5, demonstrating that explicit expert routing alignment is an effective mechanism for transferring semantic capabilities across languages.

4.5 Ablation Study

To analyze the effectiveness of SARA, we conducted a comprehensive ablation study on Qwen3-30B-A3B in Table 2. Respectively, $-q$ and $-g$ denote routing priors extracted from the model’s own internal inference traces and from externally generated data (via GPT-5 mini). The suffixes $-en$ and $-sw$ indicate the choice of high-resource (English) and low-resource (Swahili) anchor languages. For layer selection, $-s$ restricts routing alignment to the selected intermediate layers, while $-a$ applies alignment across all layers and $-r$ aligns a random subset of layers (see Appendix A.6 for details).

The comparison between $-q-en-s$ and $-g-en-s$ highlights the decisive advantage of self-anchoring over external distillation. While external teachers provide high-quality text, they lack mechanistic consistency with the model’s internal expert pathways. This is most evident in the MGSM benchmark, where the $-g-en-s$ variant suffers a substantial performance degradation. We observe that the average MGSM score drops to 53.00% compared to 87.40% for our proposed method. It confirms that rectifying routing logic via the model’s own high-resource inference traces is more effective than injecting external knowledge.

Layer selection strategy also proves critical for maintaining linguistic flexibility. On the MGSM benchmark, aligning a random subset of layers ($-q-en-r$) and all layers ($-q-en-a$) both lead to a decrease in average accuracy. This supports our analysis that forcing alignment on shallow and deep layers disrupts the model’s ability to process language-specific surface variations. The $-q-sw-s$ variant demonstrates that anchor quality is bounded by

Benchmark	Anchor & Method	hi	ne	bn	te	sw	en/zh	Avg.
Global-MMLU	Vanilla LM	68.15 \pm 0.67	63.20 \pm 0.26	64.86 \pm 0.63	57.46 \pm 0.71	23.66 \pm 0.69	79.65/76.97	59.28
	English Anchor							
	FFT	73.97 \pm 0.40	71.91 \pm 0.17	71.72 \pm 0.27	65.60 \pm 0.70	59.39 \pm 0.36	81.80 \pm 0.26	70.73
	AES	71.53 \pm 0.69	69.52 \pm 0.45	70.04 \pm 0.71	63.58 \pm 0.56	56.09 \pm 0.26	81.03 \pm 0.21	68.63
	ShifCon	74.32 \pm 0.61	72.25 \pm 0.78	72.18 \pm 0.56	65.68 \pm 0.40	60.15 \pm 0.77	82.12 \pm 0.41	71.12
	SARA (Ours)	74.66 \pm 0.15	72.60 \pm 0.59	72.70 \pm 0.26	65.76 \pm 0.68	61.00 \pm 0.43	82.45 \pm 0.44	71.53
	Chinese Anchor							
	FFT	72.69 \pm 0.47	70.38 \pm 0.11	70.23 \pm 0.62	64.28 \pm 0.18	57.88 \pm 0.53	77.20 \pm 0.62	68.78
	AES	69.85 \pm 0.72	68.24 \pm 0.67	68.57 \pm 0.38	62.13 \pm 0.74	55.09 \pm 0.43	76.54 \pm 0.64	66.74
	ShifCon	72.55 \pm 0.24	70.64 \pm 0.25	70.53 \pm 0.68	64.48 \pm 0.61	58.65 \pm 0.35	77.24 \pm 0.69	69.02
	SARA (Ours)	72.62 \pm 0.59	70.90 \pm 0.68	70.84 \pm 0.58	64.68 \pm 0.27	59.46 \pm 0.48	77.27 \pm 0.67	69.30
BELEBELE	Vanilla LM	82.78 \pm 0.77	79.67 \pm 0.14	84.89 \pm 0.78	73.11 \pm 0.73	56.22 \pm 0.27	94.67/91.78	78.32
	English Anchor							
	FFT	82.22 \pm 0.69	82.00 \pm 0.37	84.11 \pm 0.22	76.56 \pm 0.72	77.78 \pm 0.71	95.00 \pm 0.11	82.95
	AES	79.53 \pm 0.26	78.85 \pm 0.61	81.54 \pm 0.77	74.58 \pm 0.49	73.52 \pm 0.60	93.84 \pm 0.54	80.31
	ShifCon	82.85 \pm 0.12	80.75 \pm 0.70	83.58 \pm 0.49	77.02 \pm 0.37	77.84 \pm 0.41	95.12 \pm 0.31	82.86
	SARA (Ours)	83.44 \pm 0.64	80.89 \pm 0.22	83.67 \pm 0.17	77.44 \pm 0.35	77.89 \pm 0.28	95.22 \pm 0.49	83.09
	Chinese Anchor							
	FFT	80.89 \pm 0.43	81.44 \pm 0.18	84.44 \pm 0.23	76.78 \pm 0.66	75.67 \pm 0.77	92.00 \pm 0.44	81.87
	AES	78.56 \pm 0.67	78.03 \pm 0.46	80.58 \pm 0.20	73.54 \pm 0.24	71.09 \pm 0.39	90.55 \pm 0.21	78.73
	ShifCon	82.05 \pm 0.35	79.45 \pm 0.61	83.35 \pm 0.30	76.58 \pm 0.20	76.55 \pm 0.69	91.82 \pm 0.67	81.63
	SARA (Ours)	83.22 \pm 0.55	79.56 \pm 0.66	83.44 \pm 0.25	76.67 \pm 0.42	77.44 \pm 0.27	91.89 \pm 0.31	82.04
MGSM	Vanilla LM	–	–	84.80 \pm 0.24	77.60 \pm 0.70	48.40 \pm 0.14	96.00/89.20	79.20
	English Anchor							
	FFT	–	–	88.40 \pm 0.79	84.00 \pm 0.45	80.40 \pm 0.50	96.00 \pm 0.22	87.20
	AES	–	–	85.20 \pm 0.37	78.40 \pm 0.65	76.00 \pm 0.55	94.40 \pm 0.76	83.50
	ShifCon	–	–	87.60 \pm 0.18	80.40 \pm 0.65	82.40 \pm 0.62	97.20 \pm 0.38	86.90
	SARA (Ours)	–	–	88.00 \pm 0.51	80.80 \pm 0.50	83.20 \pm 0.77	97.60 \pm 0.60	87.40
	Chinese Anchor							
	FFT	–	–	86.00 \pm 0.58	83.60 \pm 0.38	76.80 \pm 0.25	94.80 \pm 0.26	85.30
	AES	–	–	82.40 \pm 0.64	77.60 \pm 0.30	74.00 \pm 0.69	84.80 \pm 0.52	79.70
	ShifCon	–	–	85.20 \pm 0.66	80.40 \pm 0.58	78.40 \pm 0.73	86.80 \pm 0.68	82.70
	SARA (Ours)	–	–	85.60 \pm 0.62	80.80 \pm 0.79	79.20 \pm 0.70	87.20 \pm 0.33	83.20

Table 1: Performance comparison of Qwen3-30B-A3B on Global-MMLU, BELEBELE, and MGSM. Note that the evaluation language for each benchmark is aligned with the respective semantic anchor: *en* Anchor models are evaluated on English test sets, while *zh* Anchor models are evaluated on Chinese test sets. Results for **hi** and **ne** are omitted for MGSM as they are not supported by the benchmark.

Benchmark	Method	hi	ne	bn	te	sw	en	Avg.
Global-MMLU	-g-en-s	73.74	72.03	72.52	69.67	58.32	81.47	71.29
	-q-en-a	69.56	68.09	67.78	61.81	55.40	81.78	67.40
	-q-en-r	70.03	67.69	68.11	61.84	55.46	80.64	67.30
	-q-sw-s	70.47	68.20	68.05	63.01	55.59	82.13	67.91
	-q-en-s (Ours)	74.66	72.60	72.70	65.76	61.00	82.45	71.53
	BELEBELE	-g-en-s	78.22	77.11	80.00	73.89	72.33	93.56
-q-en-a		79.22	77.78	80.89	72.56	74.56	94.33	79.89
-q-en-r		80.00	76.11	82.00	74.11	76.00	92.67	80.15
-q-sw-s		79.11	79.11	81.56	73.78	76.44	95.00	80.83
-q-en-s (Ours)		83.44	80.89	83.67	77.44	77.89	95.22	83.09
MGSM		-g-en-s	-	-	52.80	47.60	48.80	62.80
	-q-en-a	-	-	86.80	80.40	76.80	95.20	84.80
	-q-en-r	-	-	84.00	82.40	78.80	94.40	84.90
	-q-sw-s	-	-	85.60	79.60	77.20	95.20	84.40
	-q-en-s (Ours)	-	-	88.00	80.80	83.20	97.60	87.40

Table 2: Ablation study on Qwen3-30B-A3B investigating the impact of routing prior sources, anchor languages and layer selection strategies. Results for **hi** and **ne** are omitted for MGSM as they are not supported by the benchmark.

the model’s native reasoning stability in the pivot language. Utilizing the low-resource Swahili anchor results in an average Global-MMLU score of 67.91% lower than the 71.53% achieved with the English anchor. This result shows that stable rout-

ing patterns from high-resource anchors provide more reliable priors which can help guide expert selection for low-resource inputs.

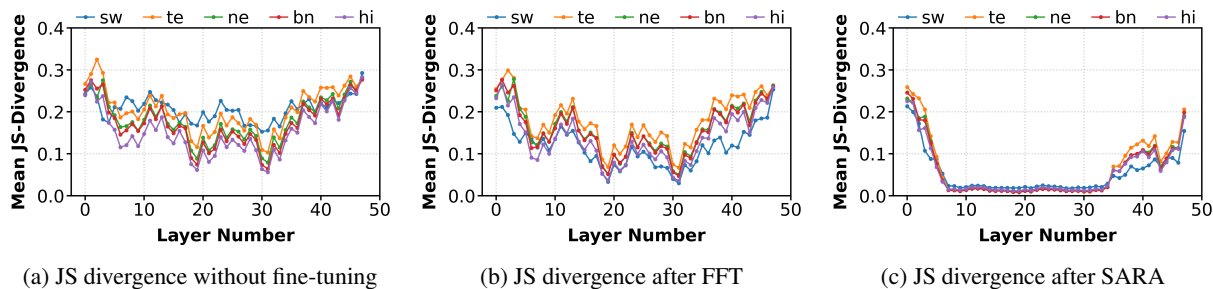


Figure 2: Comparison of layer-wise routing divergence across different fine-tuning strategies. SARA demonstrates superior alignment in intermediate layers compared to Vanilla LM and FFT.

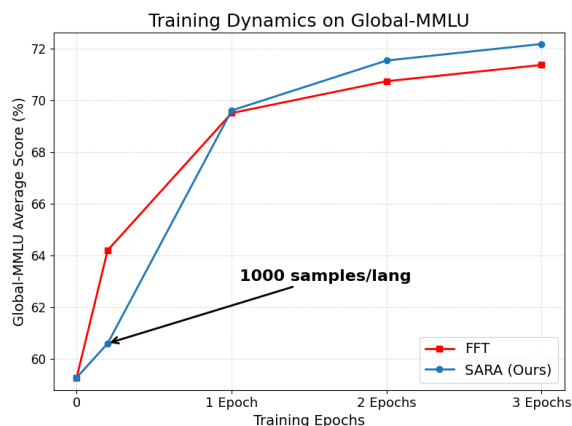


Figure 3: Training dynamics comparison between SARA and FFT on Global-MMLU.

5 Analysis

5.1 Routing Consistency

We calculated the JS divergence of the routing distributions for 5 languages relative to the English anchor on Global-MMLU, as shown in Figure 2a. We further analyzed the impact of fine-tuning on this routing behavior. As shown in Figure 2b, FFT leads to a reduction in JS divergence compared to the base model. However, this reduction is incomplete; significant divergence remains perceptible in the intermediate layers. While FFT induces some degree of alignment as a byproduct of task supervision, it is insufficient to correct the internal expert selection logic. In contrast, SARA explicitly minimizes this discrepancy. As demonstrated in Figure 2c, SARA effectively suppresses the remaining divergence, flattening the curves to near-zero levels across layers 7 to 34. This confirms that SARA effectively aligns routing distributions, encouraging semantically equivalent inputs to activate shared expert subsets (see Appendix A.2 for details).

5.2 Training Dynamics

As shown in Figure 3, we analyzed the training dynamics on Global-MMLU to understand the specific impact of data quantity and training iterations. In the initial phase, where the model processes only 1,000 samples per language, FFT outperforms SARA. This suggests that satisfying the routing alignment constraint in addition to learning the multilingual task is more challenging than the single objective of standard fine-tuning. However, this trend reverses as the training progresses. By the end of the first epoch, SARA effectively closes the performance gap with FFT. Furthermore, as training progresses into the second epoch, the performance improvement from FFT becomes insignificant. In contrast, SARA leverages these additional training iterations to consolidate mechanistic consistency, ultimately unlocking a higher performance ceiling through internal routing. By the third epoch, both methods enter a saturation regime with only marginal gains, yet SARA consistently converges to a higher plateau than FFT, indicating a superior asymptotic performance enabled by routing consistency.

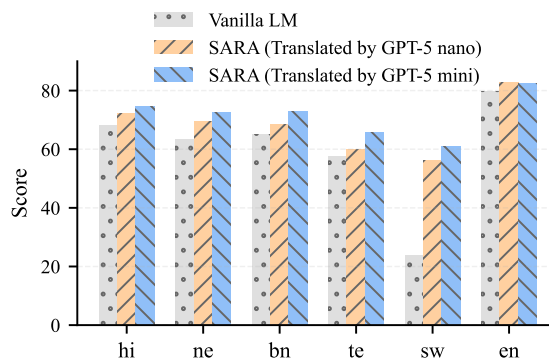


Figure 4: Comparison of SARA on Global-MMLU benchmark using translations generated by GPT-5 mini and GPT-5 nano.

5.3 Effect of Translation Quality

We analyzed the effect of translation quality on SARA by varying the models used to generate the multilingual training data. Specifically, we compared two SARA variants trained on parallel corpora translated by GPT-5 mini and GPT-5 nano, respectively.

To formally quantify the translation quality, we employed CometKiwi (Rei et al., 2022), a reference-free machine translation evaluation metric. As shown in Table 3, the GPT-5 mini model consistently achieves higher average CometKiwi scores across all evaluated target languages (hi, bn, te, sw, ne) compared to the GPT-5 nano model.

Model	hi	bn	te	sw	ne
GPT-5 mini	0.7238	0.7528	0.7356	0.6774	0.7005
GPT-5 nano	0.7179	0.7455	0.7275	0.6615	0.6982

Table 3: CometKiwi Scores (en → xx) for Training Data

This disparity in translation quality directly impacts downstream performance. As illustrated in Figure 4, while both SARA variants substantially improve performance over the vanilla MoE model across all Global-MMLU languages, the variant trained with GPT-5 mini translations achieves superior accuracy. This correlation suggests that higher-quality training translations provide cleaner and more precise semantic supervision for routing alignment, thereby enhancing the model’s cross-lingual transfer capabilities.

6 Conclusion

In this paper, we have presented SARA, a novel framework designed to address performance bottlenecks in MoE models arising from cross-lingual routing divergence. By treating the routing distributions of high-resource languages as semantic anchors, SARA encourages consistent expert activation patterns within intermediate layers via a Jensen-Shannon divergence constraint. Extensive experiments conducted across multiple benchmarks demonstrate that our method effectively unlocks latent domain knowledge for low-resource languages, outperforming standard full fine-tuning. SARA achieves these performance gains while preserving proficiency in the anchor language, thereby effectively transferring inference capabilities across linguistic boundaries.

Acknowledgments

The present research was supported by the National Key Research and Development Program of China (Grant No. 2023YFE0116400) and National Natural Science Foundation of China Youth Fund (Grant No. 62306210). We would like to thank the anonymous reviewers for their insightful comments.

Limitations

First, although SARA demonstrates strong performance, the diversity of its training datasets could be improved. The model currently adapts to the specific textual styles of its training data, which may hinder generalization to distinct stylistic characteristics. Moreover, our method relies on the semantic equivalence between anchors and target languages. Translation artifacts from the synthetic generation process could introduce noise into the routing supervision, potentially bounding the alignment performance. Then, by explicitly aligning the routing distribution of multilingual inputs with high-resource semantic anchors, there is a risk of suppressing language-specific cultural nuances. While this effectively transfers reasoning capabilities, it may introduce a high-resource language-centric bias, limiting performance on tasks requiring deep local cultural understanding. Finally, our layer selection strategy focuses on intermediate layers based on the observed U-shaped routing divergence in the tested models. While effective for Qwen3 and Phi-3.5, this heuristic may require adaptation for MoE architectures with fundamentally different gating mechanisms or depth configurations.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Chandran Savithri Anoop, AP Prathosh, and Ramakrishnan AG. 2021. Unsupervised domain adaptation schemes for building asr in low-resource languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 342–349. IEEE.
- Baidu-ERNIE-Team. 2025. [Ernie 4.5 technical report](#).

- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775.
- Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Peng. 2025. Multilingual routing in mixture-of-experts. *arXiv preprint arXiv:2510.04694*.
- Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33007–33024.
- Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, and 1 others. 2022. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Damai Dai, Chengqi Deng, Chenggang Zhao, Rx Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, and 1 others. 2024. Deepseek-moe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297.
- Tianyu Dong, Bo Li, Jinsong Liu, Shaolin Zhu, and Deyi Xiong. 2025. Mlas-lora: Language-aware parameters detection and lora-based knowledge transfer for multilingual machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15645–15660.
- Haim Dubossarsky and Farheen Dairkee. 2024. Strengthening the wic: New polysemy dataset in hindi and lack of cross lingual transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15341–15349.
- 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.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Sicong Leng, Qimei Cui, and Xudong Jiang. 2025b. Advancing expert specialization for better moe. *arXiv preprint arXiv:2505.22323*.
- Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006.
- Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4476–4494.
- Quzhe Huang, Zhenwei An, Nan Zhuang, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Liwei Chen, Songfang Huang, and Yansong Feng. 2024. Harder task needs more experts: Dynamic routing in moe models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12883–12895.
- Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, and 1 others. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Bo Li, Shaolin Zhu, and Lijie Wen. 2025a. Mit-10m: A large scale parallel corpus of multilingual image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5154–5167.

- Zhongyang Li, Ziyue Li, and Tianyi Zhou. 2025b. Routing manifold alignment improves generalization of mixture-of-experts llms. *arXiv preprint arXiv:2511.07419*.
- Zihao Li, Shaoxiong Ji, Hengyu Luo, and Jörg Tiedemann. 2025c. Rethinking multilingual continual pretraining: Data mixing for adapting llms across languages and resources. *arXiv preprint arXiv:2504.04152*.
- Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, and 1 others. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. 2024. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6159–6172.
- Leiyu Pan, Bojian Xiong, Lei Yang, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, and 1 others. 2025. Advancing large language models for tibetan with curated data and continual pre-training. *arXiv preprint arXiv:2507.09205*.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all—adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.
- Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and 1 others. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.
- Ujjwal Sharma and Pushpak Bhattacharyya. 2025. Hige: Hindi grammar error correction in low resource scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6063–6075.
- Yikang Shen, Zhen Guo, Tianle Cai, and Zengyi Qin. 2024. Jetmoe: Reaching llama2 performance with 0.1 m dollars. *arXiv preprint arXiv:2404.07413*.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 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, and 1 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.
- Shubhankar Singh, Anirudh Pupneja, Shivaansh Mittal, Cheril Shah, Manish Bawkar, Lakshman Prasad Gupta, Ajit Kumar, Yaman Kumar, Rushali Gupta, and Rajiv Ratn Shah. 2023. H-aes: Towards automated essay scoring for hindi. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15955–15963.
- Zunhai Su, Qingyuan Li, Hao Zhang, Weihao Ye, Qibo Xue, YuLei Qian, Yuchen Xie, Ngai Wong, and Kehong Yuan. 2025. Unveiling super experts in mixture-of-experts large language models. *arXiv preprint arXiv:2507.23279*.
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. [MMLU-ProX: A multilingual benchmark for advanced large language model evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Yuanhang Yang, Shiyi Qi, Wenchao Gu, Chaozheng Wang, Cuiyun Gao, and Zenglin Xu. 2024. Xmoe: Sparse models with fine-grained and adaptive expert selection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11664–11674.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo

- Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.
- Zhiyuan Zeng and Deyi Xiong. 2023. Scomeo: Efficient mixtures of experts with structured communication. In *ICLR*.
- Hengyuan Zhang, Chenming Shang, Sizhe Wang, Dongdong Zhang, Yiyao Yu, Feng Yao, Renliang Sun, Yujiu Yang, and Furu Wei. 2025a. Shifcon: Enhancing non-dominant language capabilities with a shift-based multilingual contrastive framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4818–4841.
- Hongbin Zhang, Kehai Chen, Xuefen Bai, Youcheng Pan, Yang Xiang, Jinpeng Wang, and Min Zhang. 2026. Mitigating translationese bias in multilingual llm-as-a-judge via disentangled information bottleneck. *arXiv preprint arXiv:2603.10351*.
- Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024. Lingualift: an effective two-stage instruction tuning framework for low-resource language tasks. *arXiv e-prints*, pages arXiv–2412.
- Zeliang Zhang, Xiaodong Liu, Hao Cheng, Chenliang Xu, and Jianfeng Gao. 2025b. Diversifying the expert knowledge for task-agnostic pruning in sparse mixture-of-experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 86–102.
- Hao Zhao, Zihan Qiu, Huijia Wu, Zili Wang, Zhaofeng He, and Jie Fu. 2024. Hypermo: Towards better mixture of experts via transferring among experts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10605–10618.
- Hao Zhou, Zhijun Wang, Shujian Huang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, Weihua Luo, and Jiajun Chen. 2025. Moe-lpr: Multilingual extension of large language models through mixture-of-experts with language priors routing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26092–26100.
- Shaolin Zhu, Leiyu Pan, Dong Jian, and Deyi Xiong. 2025. Overcoming language barriers via machine translation with sparse mixture-of-experts fusion of large language models. *Information Processing & Management*, 62(3):104078.
- Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024a. Landerm: Detecting and routing language-aware neurons for selectively finetuning llms to machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148.
- Shaolin Zhu, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António

A Appendix

A.1 Details of Languages

In our experiments, we designate en and zh as high-resource anchor languages. We select hi, ne, bn, te and sw as the representative low-resource target languages. The full list of languages and their corresponding ISO 639 codes is provided in Table 4.

ISO 639	Language
hi	Hindi
ne	Nepali
bn	Bengali
te	Telugu
sw	Swahili
en	English
zh	Chinese

Table 4: ISO 639 language codes and names.

A.2 Detailed Analysis of Routing Divergence and Layer Selection

Figures 5 and 6 visualize the layer-wise routing divergence (measured by JS divergence) between English and various target languages for Qwen3-30B-A3B and Phi-3.5-MoE-instruct. A core finding is that both models exhibit a characteristic U-shaped distribution relative to the model depth: divergence is significantly higher in the initial layers (which handle shallow linguistic features) and the final layers (which specialize in language-specific token generation), while the intermediate layers form a stable, lower-divergence valley. This U-shaped trend suggests that the mid-range experts function as a language-agnostic semantic reasoning engine, making them the ideal targets for semantically anchored alignment.

Based on these empirical observations, we selectively apply the routing alignment loss \mathcal{L}_{JS} to the layers corresponding to these U-shaped valleys. For Qwen3-30B-A3B, the divergence reaches its nadir between layer 7 and 34, which we define as our alignment range. Phi-3.5-MoE-instruct similarly follows this U-shaped trend; despite its higher local volatility, the divergence is consistently lower within the intermediate routing logic of layers 13 to 26. By focusing SARA on these specific ranges, we ensure the supervision signal from high-resource anchors rectifies the core reasoning logic while

avoiding the noise inherent in the model’s shallow input processing and deep output generation stages.

A.3 Sensitivity Analysis of λ_{JS}

To evaluate the impact of the routing alignment strength, we conduct a sensitivity analysis on the hyperparameter λ_{JS} . This parameter controls the weight of the Jensen–Shannon divergence loss within the total training objective (\mathcal{L}_{total}). We vary λ_{JS} across $\{0.5, 1.0, 1.5, 2.0\}$ and evaluate the model’s performance on the Global-MMLU benchmark to observe how the magnitude of the alignment constraint affects multilingual reasoning.

The results are summarized in Table 5 and Figure 7. We observe that the model performance generally benefits from a moderate alignment strength. When λ_{JS} is low (0.5 or 1.0), the alignment signal is insufficient to fully transfer semantic patterns, resulting in lower average scores. Conversely, an overly strict constraint ($\lambda_{JS} = 2.0$) slightly hampers the model’s flexibility, leading to a performance drop. The setting $\lambda_{JS} = 1.5$ achieves the highest average Global-MMLU accuracy of 71.53, offering the optimal balance between semantic consistency and task-specific generation across the diverse set of languages.

A.4 Statistical Significance Testing

To rigorously evaluate the performance improvements of SARA relative to the standard full fine-tuning (FFT) baseline, we conducted one-tailed paired t-tests using performance scores across all evaluated languages as paired samples.

A.4.1 Hypothesis Formulation

Given that SARA is explicitly designed to enhance cross-lingual generalization compared to standard fine-tuning, we employed a directional hypothesis:

$$H_0 : \mu_D \leq 0$$

$$H_1 : \mu_D > 0$$

where $D_i = X_{SARA,i} - X_{FFT,i}$ represents the performance difference for a specific language i . The t-statistic is calculated as:

$$t = \frac{\bar{D}}{SE_{\bar{D}}}, \quad df = n - 1$$

The one-tailed p -value is derived from the upper tail of the t-distribution with $n - 1$ degrees of freedom, representing the probability of observing a positive mean difference at least as large as \bar{D} under the null hypothesis.

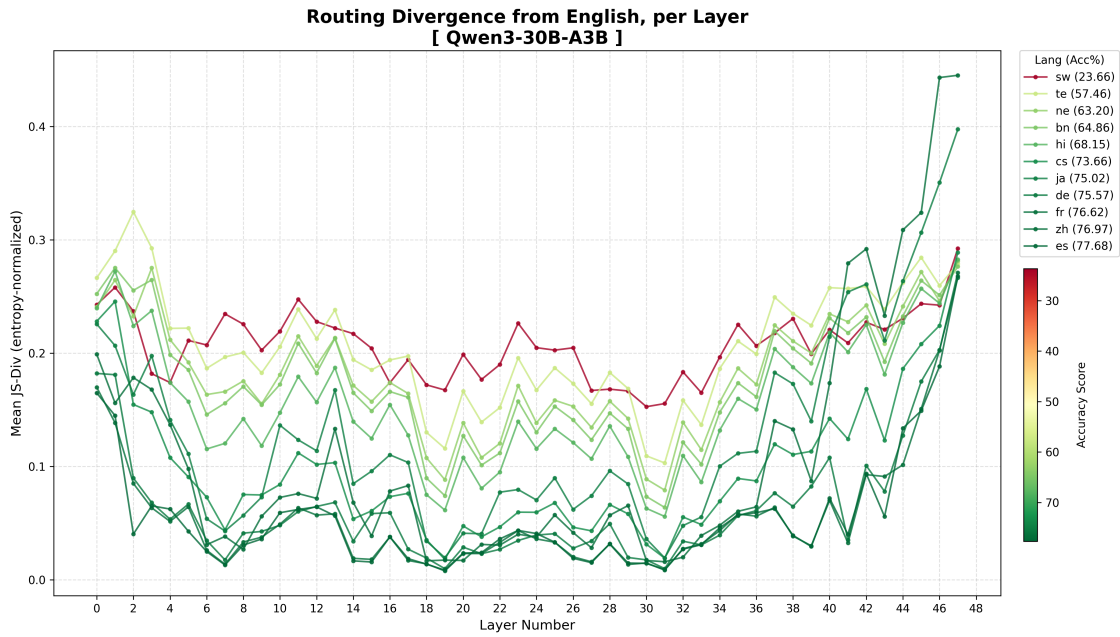


Figure 5: Visualization of routing divergence for selected languages relative to English across model layers based on Qwen3-30B-A3B on the Global-MMLU dataset.

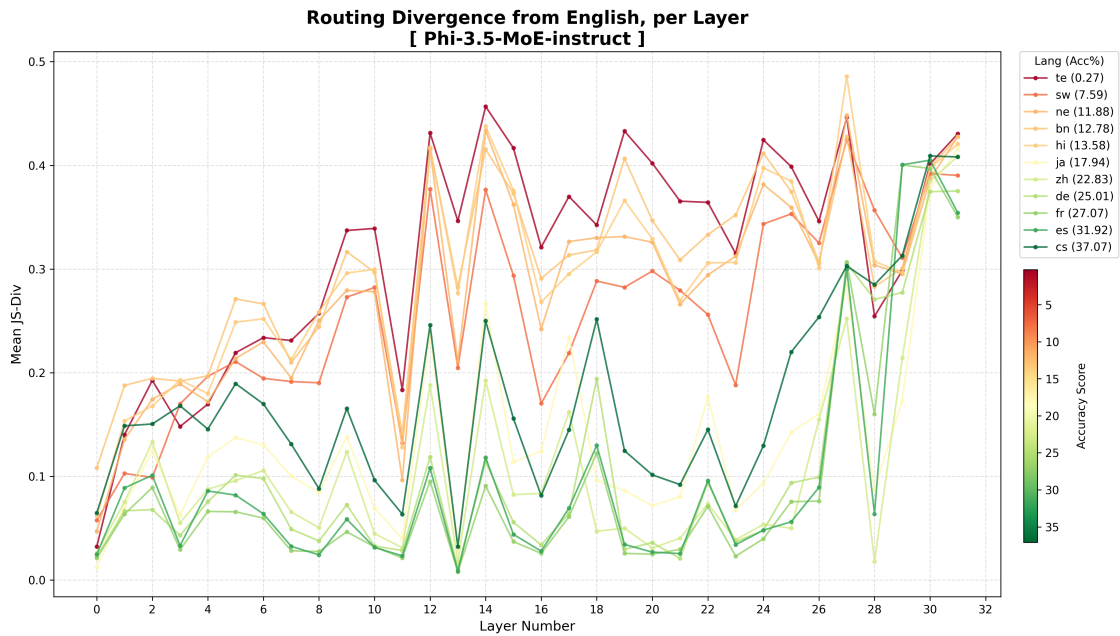


Figure 6: Visualization of routing divergence for selected languages relative to English across model layers based on Phi-3.5-MoE-instruct on the Global-MMLU dataset.

Language	$\lambda_{JS} = 2.0$	$\lambda_{JS} = 1.5$	$\lambda_{JS} = 1.0$	$\lambda_{JS} = 0.5$
hi	74.33	74.66	74.19	73.81
ne	72.60	72.60	72.40	72.52
bn	72.33	72.70	72.48	72.10
te	66.40	65.76	66.46	65.79
sw	59.76	61.00	59.71	59.41
en	82.46	82.45	82.39	82.51
Avg.	71.31	71.53	71.27	71.02

Table 5: Sensitivity analysis of the routing alignment weight λ_{JS} on the Global-MMLU benchmark. We report the performance on individual languages and the overall average. The results indicate that $\lambda_{JS} = 1.5$ yields the best balance.

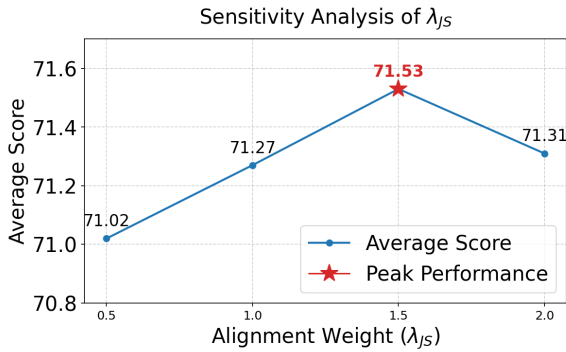


Figure 7: Average score as λ_{JS} changes.

A.4.2 Results and Analysis

The statistical results across three multilingual benchmarks—Global-MMLU, BELEBELE, and MGSM—for the Qwen3-30B-A3B and Phi-3.5-MoE-instruct models are summarized in the following table:

Benchmark	Model	t-statistic	p-value
Global-MMLU	Qwen3-30B-A3B	4.0784	0.0096
	Phi-3.5-MoE-instruct	2.4705	0.0565
BELEBELE	Qwen3-30B-A3B	0.4200	0.6904
	Phi-3.5-MoE-instruct	2.0100	0.1008
MGSM	Qwen3-30B-A3B	0.1525	0.8885
	Phi-3.5-MoE-instruct	1.4606	0.2403

Table 6: One-tailed paired t-test results comparing SARA and FFT.

- **Global-MMLU:** This benchmark provides the strongest statistical evidence due to its large evaluation set of approximately 14,000 samples per language. SARA demonstrates statistically supported improvements, particularly for the Qwen3 model ($p < 0.01$).
- **BELEBELE and MGSM:** These benchmarks consist of significantly fewer test samples—approximately 900 and 250 per language,

respectively. While SARA maintains positive performance trends, the lack of statistical significance in these instances is primarily due to reduced statistical power from smaller sample sizes rather than inconsistent model behavior.

A.5 Results on Other LLMs

We also fine-tuned the Phi-3.5-MoE-instruct model using our proposed method and compared it against baseline fine-tuning approaches. The results, presented in Table 7, demonstrate that our method achieves state-of-the-art performance across multiple datasets and languages. This finding underscores the generalizability of our approach across different model architectures. Notably, the performance gain observed on Phi-3.5-MoE-instruct is more pronounced than that on Qwen3-30B-A3B, suggesting that our method may yield greater improvements on models with relatively weaker multilingual capabilities.

A.6 Ablation Variant Definitions

To further analyze the design choices in SARA, we define several controlled ablation variants by systematically varying the source of routing priors, anchor language selection, and layer alignment strategies.

- **-q-en-s (Ours):** Uses internal inference traces generated by Qwen3-30B-A3B (-q) and English anchors (en) on the selected intermediate layers (-s).
- **-g-en-s:** Uses external data generated and translated by GPT-5 mini (-g) to obtain routing priors through forward propagation on Qwen3-30B-A3B, serving as the English anchors (en) for the selected intermediate layers (-s) to evaluate the necessity of self-anchoring.
- **-q-en-r:** Uses internal inference traces generated by Qwen3-30B-A3B (-q) and English anchors

Benchmark	Method	hi	ne	bn	te	sw	en	Avg.
Global-MMLU	Vanilla LM	13.58	11.88	12.78	0.27	7.59	49.16	15.88
	FFT	42.14	37.39	27.99	15.96	45.42	80.86	41.63
	AES	40.55	36.12	26.54	12.45	43.20	80.15	39.84
	ShifCon	43.15	37.82	28.35	17.55	45.54	80.92	42.22
	SARA (Ours)	44.14	38.25	28.70	19.09	45.66	80.98	42.80
BELEBELE	Vanilla LM	10.33	1.11	24.44	0.00	0.56	51.00	14.57
	FFT	68.89	64.56	55.56	0.00	58.22	93.78	56.84
	AES	67.25	63.15	53.25	0.00	56.15	92.55	55.39
	ShifCon	70.15	64.85	58.45	0.05	59.85	93.65	57.83
	SARA (Ours)	71.22	65.00	61.56	0.11	61.67	93.44	58.83
MGSM	Vanilla LM	-	-	9.20	0.00	11.20	65.20	21.40
	FFT	-	-	54.80	44.00	55.60	66.40	55.20
	AES	-	-	53.20	42.40	52.00	65.60	53.30
	ShifCon	-	-	56.40	45.20	53.60	68.00	55.80
	SARA (Ours)	-	-	58.00	46.00	54.00	69.20	56.80

Table 7: Performance comparison of Phi-3.5-MoE-instruct on Global-MMLU, BELEBELE, and MGSM using **English** as the semantic anchor.

(*en*) but aligns a random subset of layers (*-r*) instead of the targeted intermediate ones, keeping the total number of aligned layers consistent with the *-s* setting.

- **-q-en-a**: Uses internal inference traces generated by Qwen3-30B-A3B (*-q*) and English anchors (*en*) while applying alignment across all layers (*-a*) of the model.
- **-q-sw-s**: Uses internal inference traces generated by Qwen3-30B-A3B (*-q*) on the selected intermediate layers (*-s*) but utilizes low-resource Swahili (*sw*) as the semantic anchor to test the pivot language quality.

A.7 Prompt Templates

To evaluate the multilingual capabilities of LLMs across multiple benchmarks (Global-MMLU, BELEBELE, and MGSM), we developed standardized prompt templates in English. To ensure cross-lingual consistency and minimize instruction bias, we utilized GPT-5 mini to translate these core instructions into our target languages, maintaining the semantic structure while adapting to the natural linguistic flow of each language.

To facilitate objective and scalable evaluation, we mandated that models enclose their final answers within a `\boxed{}` command. We subsequently employed regular expressions to programmatically parse the model outputs and extract the final answers for performance scoring. The base English templates for each benchmark are detailed in Table 8, 9, and 10.

Below is a single-choice question with four options (A, B, C, D). Only one option is correct.

Question:
{question}

Options:
A) {A}
B) {B}
C) {C}
D) {D}

Analyze the question carefully and answer it. At the end of your response, clearly indicate your final answer by enclosing the correct choice in `\boxed{}`.

Table 8: Prompt template used for Global-MMLU.

Below is a passage followed by a single-choice question with four options (A, B, C, D). Only one option is correct based on the passage.

Passage:
{flores_passage}

Question:
{question}

Options:
A) {A}
B) {B}
C) {C}
D) {D}

Analyze the passage and question carefully and answer. At the end of your response, clearly indicate your final answer by enclosing the correct choice in `\boxed{}`.

Table 9: Prompt template used for BELEBELE.

Below is a question.

Question:
{question}

Please analyze the question and solve it. At the end of your response, clearly indicate your final answer by enclosing it in `\boxed{}`.

Table 10: Prompt template for MGSM.