CROSS-LINGUAL ALIGNMENT FOR DECODER-ONLY MODELS USING MOE ROUTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-lingual contrastive learning has been a core component of multilingual encoder training, but the ability to explicitly align representations is not possible in decoder-only LLMs because of varying multilingual tokenization. However, a growing amount of research suggests that even in LLMs, higher cross-lingual representational alignment leads to improved cross-lingual transfer. In this paper, we propose a novel approach to reimagine cross-lingual contrastive learning given the architectural constraints of modern LLMs. Rather than applying an auxiliary alignment loss on hidden states, we propose using the outputs of the mixture-of-experts (MoE) routers as the target for alignment. Router outputs lend themselves better to pooling over many tokens, enabling more reliable cross-lingual comparisons at the sequence-level. Controlled continual pre-training experiments on four open-source MoEs show that incorporating this routing loss also aligns the underlying hidden representations across languages. Most importantly, this loss improves multilingual performance on our diverse evaluation suite, demonstrating the potential of cross-lingual MoE router alignment.

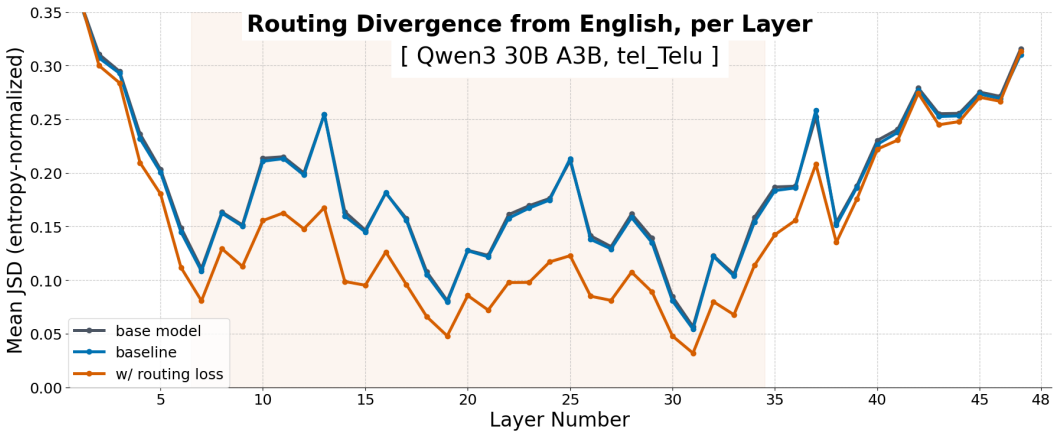

Figure 1: Impact of our auxiliary alignment loss on routing alignment during CPT of the post-trained Qwen3-30B-A3B. In this case, vanilla CPT did not alter routing behavior (the blue and gray lines overlap), whereas our method (orange) substantially reduced routing divergence. The shaded orange region indicates the middle layers to which our loss was applied.

1 INTRODUCTION

Developing LLMs that perform well in natural languages underrepresented during pretraining or post-training remains a major challenge. One longstanding idea is that *cross-lingual transfer*—the generalization of learned capabilities and knowledge across languages—is closely tied to the alignment of internal representations (Klementiev et al., 2012; Gaschi et al., 2023; Liu & Niehues, 2025). If semantically equivalent inputs in different languages are represented similarly, knowledge and capabilities learned in one language are more likely to transfer to another. In earlier encoder-decoder

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

language models, this idea could be operationalized directly because the encoder produces a natural sequence-level embedding (Zhou et al., 2016). Given the abundance of parallel (translated) text, cross-lingual contrastive learning could pull the embeddings of translated sentence pairs together while pushing those of unrelated pairs apart. Contrastive learning remains standard practice for training multilingual encoders (Wang et al., 2024a; SONAR Team et al., 2026; Zhang et al., 2025).

However, modern LLMs are decoder-only models. Consequently, they provide no natural sequence-level representation to align, and differences in tokenization and linguistic structure make it difficult to match tokens across translations. Common approaches to constructing sequence representations include using the final token's hidden state or mean-pooling hidden states (Li et al., 2024; Zheng et al., 2026), but neither reliably represents the meaning of the entire sequence. As a result, multilingual LLMs are typically trained without an inductive bias toward cross-lingual representational alignment, leaving the vast supply of translation data underutilized. Even so, these models implicitly learn language-agnostic representations, particularly in their middle layers (Wendler et al., 2024; Zhao et al., 2024; Bandarkar et al., 2026). The extent to which models use these shared representations varies significantly across languages, and a growing body of work argues that poor cross-lingual alignment is directly tied to poor performance in those languages (See Section 2).

Motivated by these limitations, we present a method for effective cross-lingual alignment using the routers in mixture-of-experts (MoE) components. MoEs have become a prominent LLM architecture (Team et al., 2023; DeepSeek-AI et al., 2025) because they enable parameter scaling at a relatively fixed inference cost (Shazeer et al., 2017). For each token, the router at each layer produces a probability distribution over the available experts, determining which experts to activate. Based on prior research, we hypothesize that mean-pooled routing weights provide strong alignment targets. We evaluate this hypothesis by developing a training method that performs continual pretraining (CPT) on parallel data with an auxiliary KL-divergence loss over the mean-pooled routing weights of middle layers. We experiment with four MoE LLMs and seven languages. First, through small-scale training runs of 200k samples, we establish that this auxiliary loss effectively increases routing agreement. We also find that its gradients increase hidden-state similarity. Across a diverse set of downstream evaluations, models trained with our auxiliary loss consistently outperform those trained with vanilla CPT. Together, these results provide causal evidence that cross-lingual alignment improves multilingual performance. In practical terms, our method offers an effective and efficient cross-lingual alignment paradigm compatible with any training stage.

This paper is organized as follows. We first present recent research that motivates our method in Section 2 and describe its design and implementation in Section 3. Next, we outline our experiments in Section 4 and present internal metrics and downstream results in Section 5. Finally, we discuss our findings in Section 6, and conclusions, and the opportunities they open for future work in Section 7.

2 Related Work and Hypotheses

2.1 Cross-Lingual Representational Alignment in Encoder Models

Because of machine translation research, there has long been a relative abundance of mined parallel data (paired multilingual texts) (Koehn, 2005; Smith et al., 2013). Parallel data was first used to align representations across languages in early word embedding models (Klementiev et al., 2012). However,languages rarely align at the word level, so later sequence-to-sequence encoder-decoder models benefitted from much more direct supervision (Zhou et al., 2016; Schwenk & Douze, 2017; Conneau et al., 2018; Ouyang et al., 2021; Chi et al., 2021). In these models, parallel data was widely used for cross-lingual contrastive learning, which unified the feature space and facilitated multilingual generalization (Wu & Dredze, 2020; Gaschi et al., 2023; Patra et al., 2023). Modern, state-of-the-art multilingual embeddings models continue to leverage such explicit contrastive signals for representation learning (Sturua et al., 2024; Wang et al., 2024a; SONAR Team et al., 2026; Zhang et al., 2025). Though less clean, data augmentation can make contrastive learning also useful for monolingual models (Wu et al., 2020; Kaushik et al., 2020; Giorgi et al., 2021)

Highly analogous to the cross-lingual setting, contrastive learning has been a major training component of multimodal embedding models, most famously for image-text representation alignment and data-efficient vision understanding (Radford et al., 2021; Zhai et al., 2023). However, as the field

has evolved toward generative multimodal language models, vision-language integration is trained primarily with image-conditioned next-token prediction (Alayrac et al., 2022; Liu et al., 2023).

2.2 Multilingual Representations in LLMs

Recent work suggests that multilingual LLMs perform substantial computation in language-shared representational spaces in the middle model layers, while retaining some language-specific structure at the beginning and end (Kojima et al., 2024; Wu et al., 2025; Zhang et al., 2026). Recent work further shows that increasing cross-lingual alignment can improve multilingual generalization through activation patching (Lim et al., 2025; Ravisankar et al., 2026), hidden-state steering (Zhao et al., 2025; Gurgurov et al., 2025; Mahmoud et al., 2025), and router steering in MoE models (Bandarkar et al., 2026). However, stronger alignment can also come at the cost of language and culture-specific information (Han et al., 2025). To improve cross-lingual alignment in decoder-only models during training, prior work has increased exposure to parallel translation data (Zhang et al., 2024; Shen et al., 2025). Similar benefits from parallel data have also been observed for programming languages (Wu et al., 2026). Closest to this work, Liu & Niehues (2025) does middle layer alignment using a cosine similarity loss with hidden states.

2.3 Central Hypotheses

Based on these findings, our first hypothesis is, of course, that cross-lingual representation alignment causally improves cross-lingual transfer in modern LLMs. Our second, more novel hypothesis is that using mean-pooled MoE routing weights across parallel sequences as a target can effectively align model processing across languages. This rests on the premise that mean-pooled routing weights provide a meaningful sequence-level representation (Xu et al., 2026; Martin et al., 2026). It has a natural interpretation: the overall expert utilization through the sequence. By contrast, mean-pooling hidden states averages high-dimensional representation vectors, such that token-level features encoded in different directions can be attenuated or cancel under averaging. Li & Zhou (2025) empirically support this distinction, showing that mean-pooled routing weights perform significantly better than mean-pooled hidden states on embedding tasks. Even if underlying representations remain unaligned, Bandarkar et al. (2026) finds that promoting more consistent expert usage across languages is beneficial in its own right.

3 Methodology

3.1 Middle Layers Selection

For each model, we use cross-lingual routing divergence curves following Bandarkar et al. (2026), to identify layers that already exhibit expert sharing across languages and apply the auxiliary router loss to the corresponding layers. This provides a principled way to select these layers without exhaustively searching over the combinatorially-large space of possible layer subsets. Bandarkar et al. (2026) show that these curves identify layers where cross-lingual alignment is beneficial and layers where enforcing alignment can be detrimental. We therefore select a range of middle layers $\mathcal{M}$ for each model and hold it fixed across all experiments (see Appendix C for more details).

3.2 Loss Formulation

In order for MoE models to selectively activate a subset of experts, a *router*, or gating function, takes inputs to the MoE block of a transformer decoder layer and produces logits (one for each available expert). The input hidden state are then sent to the top-k experts only whose outputs are combined using normalized routing weights. We define our notation:

- Let E be the number of experts in each MoE layer and $k << E$ be active experts per token.
- Let $T_{\text{eng}}, T_{\text{tgt}}$ be the sequence lengths, in tokens, of English and target-language sequences.
- Let $G(\cdot)$ denote the gating function that normalizes the top-k logits and assigns zero weight to all remaining experts and is typically implemented as a softmax over the selected k logits.

- Let $\boldsymbol{p}^t_{\text{eng},l}, \boldsymbol{p}^t_{\text{tgt},l}$ be the routing weights for the tth token of the ith sample at layer l. Each $p^t_{\text{lang},l}$ is an E-dimensional probability distribution; Given router logits $\boldsymbol{z}$, we define $\boldsymbol{p}^t_{\text{lang},l} = \text{G}(\boldsymbol{z}^t_{\text{lang},l})$, yielding a sparse E-dimensional probability distribution.

Let $\mathcal{M}$ denote the set of layers in the selected middle-layer range. The cross-lingual routing loss is defined by first mean-pooling the routing weights across tokens in each sequence, $\bar{\boldsymbol{p}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{p}^t$. We then compute the KL-divergence at each layer and average these divergences across all layers in $\mathcal{M}$:

$$\mathcal{L}_{XLR} = \frac{1}{|\mathcal{M}|} \sum_{l \in \mathcal{M}} D_{\text{KL}}(\bar{\boldsymbol{p}}_{\text{eng},l} \| \bar{\boldsymbol{p}}_{\text{tgt},l}) \tag{1}$$

Appendix A provides a visual diagram of this simple loss.

For each sample, the loss is then calculated as a weighted sum with the language-modeling loss (cross-entropy) classically used in continued pre-training: $\mathcal{L} = \mathcal{L}_{LM} + \alpha \cdot \mathcal{L}_{XLR}$, where α is a hyperparameter modulating the strength of the auxiliary loss. We backpropagate gradients only through the target language sequence. The English sequence is used only to obtain routing distributions for computing $\mathcal{L}_{XLR}$, and receives no gradient updates.

3.3 Packing Optimizations

A naive implementation requires separate forward passes for the source and target sequences to compute the routing-alignment objective, followed by an additional target-language pass for the LM objective (Figure 2, left). Because English has lower token fertility on average, especially compared to the non-Latin-script languages we experiment with, the second forward pass is already less expensive. But we compute both objectives in a shared forward pass by packing source and target sequences together without padding (Figure 2, right). Variable-length attention preserves sequence boundaries, while our split-forward implementation stops processing source tokens after the last alignment layer. Together, these optimizations eliminate padding and redundant computation, resulting in only a small computational overhead relative to our baseline single-language CPT.

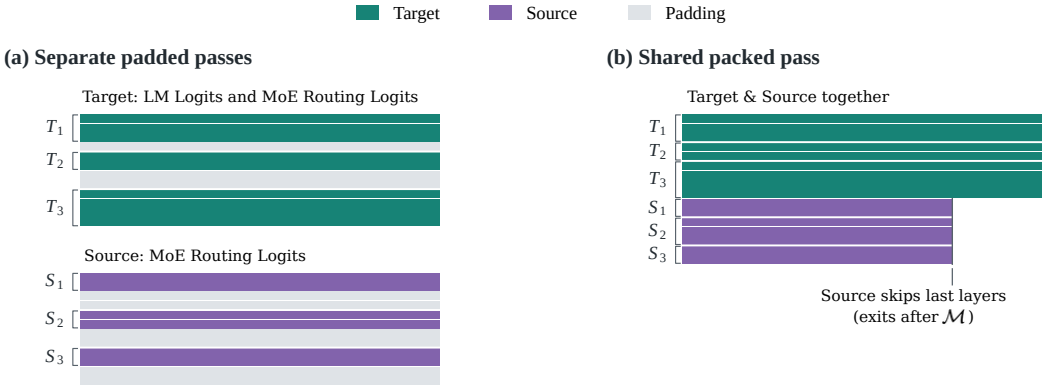

Figure 2: Comparison of the naive padded implementation (left) and our shared packed implementation (right). In the packed pass, the (typically shorter) source tokens exit early.

4 Experimental Setup

4.1 Models, Languages, and Training Data

Our goal is to experiment with a diverse set of models and languages. To this end, we work with Qwen3-30B-A3B (Yang et al., 2025), GPT-OSS-20B (OpenAI, 2025), Granite-4.0-H-Tiny (IBM Granite Team, 2025), and Marco-Nano (Jiang et al., 2026). Their MoE configurations are detailed in Appendix D. For the two smaller models, Granite and Marco, we experiment with Vietnamese, Sinhala, and Hungarian. For the two larger models, we use Telugu, Kannada, Thai, and Kyrgyz.

Because our experiments involve only a short CPT stage, we restrict evaluation to target languages in which the base models already exhibit non-trivial proficiency.

We construct a curated parallel training set of 200k samples for each language. We construct each training set from diverse, high-quality parallel sources. Whenever fewer curated examples are available, we supplement the training data with the OPUS translation dataset (Tiedemann, 2012). Indic languages are slightly over-represented (3/7) reflecting the greater availability of parallel SFT datasets for these languages. Full details are provided in Appendix E.

4.2 Experimental Comparisons

In addition to the method described in Section 3 we implement the following for comparison:

Baseline. Our baseline peforms standard CPT using only the language modeling objective $\mathcal{L}_{LM}$ applied to all tokens (not SFT). For some languages, the training data includes some parallel SFT samples, which we retain in chat-format while computing loss over the full sequence. Much of the remaining data largely consists of short, single sentence examples and therefore differs from a typical CPT distribution. Restricting CPT to these samples nevertheless provides a controlled baseline for isolating the contribution of the auxiliary loss.

Router-Only Training. We also evaluate router-only training, in which all parameters except the routers are frozen. This setting isolates the effect of directly modifying the routing function. The language modeling loss $\mathcal{L}_{LM}$ is still retained in this setting. Khandelwal et al. (2026) indicate the promise of such a parameter-efficient approach to fine-tuning.

Mean-Pooling Hidden States. We additionally test our hypothesis that mean-pooling router weights is more effective than hidden states. Since hidden states are embeddings and not probability distributions, we use cosine similarity as a measure of distance between two mean-pooled sequences, as in Liu & Niehues (2025). Let this variant be called $\mathcal{L}_{XLH}$. We apply it to the same layers as $\mathcal{L}_{XLR}$ and all parameters are trainable. We run this training for one model, Marco-Nano, and two languages, Hungarian and Sinhala.

4.3 Hyperparameter Selection

Under our compute budget, we tune two hyperparameters for each model-language pair while keeping all others fixed. The first is the learning rate, selected using baseline CPT runs for each model. The second is the α hyperparameter that modulates the strength of our auxiliary loss, mentioned in Section 3.2,. For all models, we intitialize α such that $\alpha \cdot \mathcal{L}_{XLR} \approx \mathcal{L}_{LM}$ during the initial training steps, and then test nearby values. The hidden-state loss $\mathcal{L}_{XLH}$ is typically much smaller in magnitude than $\mathcal{L}_{XLR}$. For fair comparison, we therefore use a larger α for $\mathcal{L}_{XLH}$, again matching the auxiliary-loss magnitude to $\approx \mathcal{L}_{LM}$. We select hyperparameters using the validation loss (just $\mathcal{L}_{LM}$). The baseline typically achieves a lower validation loss because it optimizes only (L_{LM}) Final tuned values are reported in Appendix B and all remaining hyperparameters that are fixed are listed in Appendix F.

4.4 Evaluation

We evaluate each checkpoint using the `lm-eval-harness` (Gao et al., 2024) and a wide variety of evaluation tasks, requiring diverse latent capabilities and output formatting. We use FLORES (Goyal et al., 2022; NLLB Team et al., 2022) in the eng→tgt direction to evaluate target language generation and Belebele (Bandarkar et al., 2024) for understanding. To evaluate cross-lingual knowledge transfer, we use the medicine subset of Global-MMLU (Singh et al., 2025) and MMLU-ProX (Xuan et al., 2025), and use MultiLoKo (Hupkes & Bogoychev, 2025) and INCLUDE (Romanou et al., 2025) for "local" knowledge. Global-PIQA (Chang et al., 2025) evaluates the transfer of physical reasoning. Finally, we evaluate mathematical reasoning with MGSM (Shi et al., 2023) (which includes Global-MGSM (Cohere Labs, n.d.), an extension to more languages) and PolyMath (Wang et al., 2025). Because benchmark language coverage varies, the number of available tasks differs across target languages ranging from four for Kyrgyz to nine for Vietnamese.

5 RESULTS

5.1 IMPACT ON ROUTING AND REPRESENTATIONS

We first analyze the impact of adding our routing loss to cross-lingual alignment metrics. We use FLoRES evaluation data (NLLB Team et al., 2022), out-of-distribution relative to any training sets, to measure this alignment. We first measure MoE routing alignment to English using the *routing divergence* metric proposed by Bandarkar et al. (2026). This differs from $\mathcal{L}_{XLR}$ by using two-sided KL-divergence (JS-divergence) and normalizing values by theoretical maximum entropy for better cross-layer comparison. We visualize the difference in routing alignment for one language per model in Figures 1 and 3. Inadvertently, the per-model hyperparameters induced varying magnitudes of changes to model routing than others. However, for all models and languages, the auxiliary loss training naturally produced the biggest increase in routing alignment (i.e. decrease in divergence). In constrast, baseline CPT led to only minor changes in routing relative to the original checkpoint.

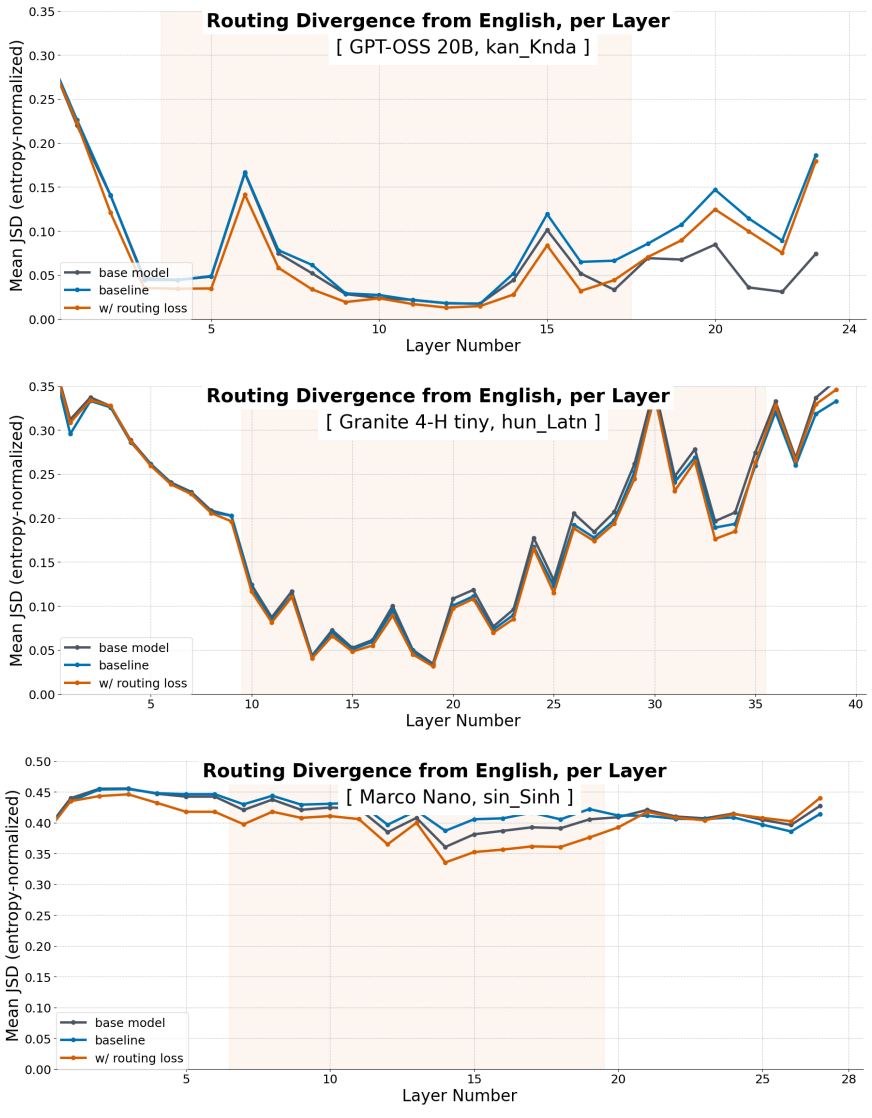

Figure 3: Layer-wise routing divergence from English for GPT-OSS-20B on Kannada (top), Granite-4.0-H-Tiny on Hungarian (middle), and Marco-Nano on Sinhala (bottom), comparing the base model, baseline training, and training with the auxiliary routing loss. Divergence is measured by mean entropy-normalized JS-divergence. The orange zone is the layers $\mathcal{M}$ where $\mathcal{L}_{XLR}$ is applied.

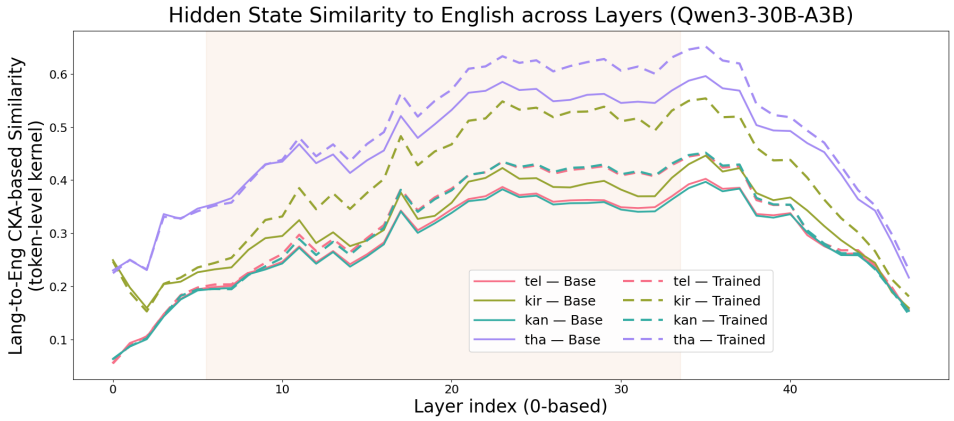

Figure 4: Layer-wise hidden-state similarity to English for Qwen3-30B-A3B. Solid lines show the base model and dashed lines show the model trained with the auxiliary cross-lingual loss for Telugu, Kyrgyz, Kannada, and Thai.

MoE router alignment does not directly imply hidden state representation alignment. But Gradients from the routing loss propagate through preceding model parameters. We therefore evaluate whether the routing loss also changes the underlying hidden state representations, rather than only the router outputs. Moreover, activating similar experts may increase representational similarity in subsequent layers. To measure the impact on underlying representations, we perform forward passes on the same parallel data and collect the hidden states at two points each decoder layer: entering the layer and the hidden state after the attention block[1]. Then, rather than mean-pooling, we calculate within-sequence and cross-language RBF kernel matrices to soft-match tokens. This allows us to use centered kernel alignment (CKA) (Kornblith et al., 2019) without explicitly matching tokens across sequences. This cross-lingual alignment metric is displayed for Qwen3 in Figure 4. It clearly displays that across languages, our MoE alignment metric is making representations, themselves, more similar to English.

5.2 Impact on Downstream Tasks

Because benchmark coverage differs by language, Table 1 reports the average over all available tasks for each model–language pair; task-level scores and complete training configurations are provided in Appendix B. The baseline itself yields only modest gains over the original checkpoint, so differences between experimental conditions are also relatively small. At the individual task level, there is noise. However, the language-level averages show a consistent trend. The routing loss improves over this baseline for 13 of the 14 model–language pairs and ties for the remaining pair (Marco-Nano on Vietnamese). The positive gains range up to 2.1 average points, with a mean improvement of 0.9 points across all pairs. This gains are observed despite the relatively short CPT stage applied earlier to already post-trained checkpoints.

Table 2 summarizes the comparison to mean-pooling hidden states using cosine similarity ($\mathcal{L}_{XLH}$). The hidden-state objective is competitive on Sinhala, improving the baseline average from 26.1 to 26.3, but remains below the 26.8 obtained with $\mathcal{L}_{XLR}$. On Hungarian, hidden-state alignment actually decreases the baseline score from 37.4 to 36.9, whereas routing alignment increases it to 38.1. Thus, in both tested settings, mean-pooled router distributions provide a more effective alignment target than mean-pooled hidden representations. Because this ablation covers only Marco-Nano and two languages, we view it as targeted support for our design choice rather than a comprehensive comparison of representation-alignment methods.

We next analyze the router-only training condition. Despite updating fewer than $< 0.1\%$ of params, it ocasionally matches or exceeds the LM-only baseline, although performance varies substantially across settings. Its changes to the internal alignment metrics are similarly noisy. These results suggest that router-only training may provide a parameter-efficient alternative, although it is not a

[1]Concretely, this is the hidden state entering the MoE block, right after the residual connection is re-added.

Table 1: Average benchmark scores for each model, language, and training condition. Parentheses in the language headers indicate the number of tasks included in each average. Bold marks the best score within each model–language pair, including ties at the reported precision. Original checkpoint denotes the released post-trained model before our training. The baseline uses only target-language LM training; our method adds the routing-alignment loss. Router-only training applies the auxiliary loss while updating only the routers in the selected middle layers.

Model	Condition	Telugu (8)	Kyrgyz (4)	Kannada (4)	Thai (7)
Qwen3-30B-A3B	Original checkpoint	44.6	40.8	49.3	43.0
	Baseline	45.4	43.5	49.6	44.9
	+ aux routing loss (ours)	**46.3**	**45.6**	**51.4**	**45.4**
	Router-only training	45.1	42.2	49.9	43.8
GPT-OSS-20B	Original checkpoint	42.5	35.9	47.5	34.9
	Baseline	46.5	42.0	55.5	39.9
	+ aux routing loss (ours)	**47.0**	**43.8**	**56.4**	**40.8**
	Router-only training	43.2	36.4	48.8	34.7
Model	Condition	Vietnamese (9)	Sinhala (5)	Hungarian (6)	
Granite-4.0-H-Tiny	Original checkpoint	34.8	31.3	34.5	
	Baseline	36.5	32.8	38.1	
	+ aux routing loss (ours)	**36.8**	**32.9**	**38.8**	
	Router-only training	36.5	31.9	38.2	
Marco-Nano	Original checkpoint	44.2	24.0	36.3	
	Baseline	**45.6**	26.1	37.4	
	+ aux routing loss (ours)	**45.6**	**26.8**	**38.1**	
	Router-only training	44.8	25.2	37.1	

Table 2: Routing versus hidden-state alignment for Marco-Nano. Entries are the reported average benchmark scores over the benchmarks available in each language; parentheses in the headers give the number of tasks in each average. Bold marks the best score in each column. Hidden-state alignment uses $\alpha = 5000$; routing alignment uses $\alpha = 10$ for both languages. Full hyperparameters and task-level scores appear in Table 6.

Condition	Sinhala (5)	Hungarian (6)
Original checkpoint	24.0	36.3
Baseline	26.1	37.4
+ routing loss (ours)	**26.8**	**38.1**
+ hidden-state loss	26.3	36.9

reliable substitute for full-model training. Taken together, these results suggest that the gains from the routing loss arise primarily from changes to the representations supplied to the routers, rather than the updates to the routing function alone.

6 Discussion and Future Work

Experimental Limitations In order to evaluate the impact of our training method on downstream tasks, we were constrained to experiment on fully post-trained LLMs. Using base models would lead to noisy evaluations, given they would not have learned to reason, perform chain-of-thought, follow instructions, etc. On the other hand, the abundant parallel data available is largely unlabeled, requiring us to do CPT. CPT on post-trained LLMs is sensitive and inefficient, as is reflected in the very limited gains by the baseline. We elect such an experimental setting to control for the influence of the routing loss on downstream evaluations, likely at the expense of bigger experimental deltas.

Real-Life Implementation We believe this method would be much more effective during pre-training rather than post-training because that would compound all the learnings from pre-, mid-, and post-training to be tied together across languages more natively. In addition, recent studies

suggest MoE routing dynamics are established quite early on in pretraining (Xue et al., 2024; Muennighoff et al., 2025).

Dependence on Parallel Data Naturally, this method is entirely reliant on parallel data. This restricts its applicability because much of the data used for pre-training, mid-training, SFT or RL is not parallel. Existing parallel corpora are also predominantly sentence-level and cover a relatively narrow data distribution. As can be seen in our baseline, where the model only improved a small amount during a CPT of 200k training samples, this data on its own cannot substantially improve a model on downstream tasks. We therefore view MoE alignment as one component of a broader training curriculum, interleaved with stages that optimize on complementary data distributions.

No Negatives Often, the term *contrastive* learning implies the presence of positive and negative pairs. The purpose of the negative pairs is to prevent representation collapse (van den Oord et al., 2018). However, the negative pairs are unnecessary here as MoE LLMs are always trained with load-balancing (another auxiliary loss), which incentivizes equitable use of experts across different batch sizes. In our experiments, our alignment loss is only done at a small-scale, with low risk of collapsing the router distributions. As a result, we do not experiment with negative pairs nor with the interaction with load-balancing auxiliary loss. We leave this to future work, as such experiments would likely require large-scale (i.e. expensive) pretraining.

Transfer-Dependent Tasks Based on prior work (Zhu et al., 2024; Yoon et al., 2024; Hu et al., 2025), we additionally hypothesize that alignment may be more beneficial for tasks that depend less strongly on language-specific information, like reasoning, math, knowledge, etc. However, our results do not establish convincingly that router alignment benefits such tasks more than translation and reading comprehension.

Direct Early-Layer Gradients In traditional CPT, the LM loss comes from the logit distribution, which is subject to a gradient bottleneck during backpropagation through the softmax operation (Godey & Artzi, 2026). Auxiliary internal losses such as MoE load-balancing and our routing divergence loss, provide more direct supervision to early-layer parameters helping mitigate vanishing or exploding gradients (Lee et al., 2015; Kedia et al., 2024; Li et al., 2025).

Cross-Modal Representation Learning While contrastive learning has been standard practice in encoder-based architectures (Radford et al., 2021), modern multimodal generative models do not use such supervision (Wang et al., 2024b). Recent work has further identified modality-dependent routing discrepancies in multimodal MoE models that can hinder the activation of task-relevant reasoning experts (Xu et al., 2026). Future work could investigate whether MoE-based alignment could facilitate cross-modal alignment and improve reasoning capabilities across modalities.

7 Conclusion

This work introduces a cross-lingual alignment objective for decoder-only MoE language models that minimizes the divergence between mean-pooled routing distributions for translated sequences. Grounded in recent LLM research on multilinguality, our method provides a pragmatic way to introduce sequence-level cross-lingual supervision into modern decoder-only architectures without requiring token-level correspondence. We find that its impact extends beyond the explicit routing objective by also aligning the model's hidden representations. Consistent with the hypothesis that representational alignment facilitates cross-lingual transfer, these internal changes yield positive aggregate gains across diverse multilingual evaluations, even within our limited-scale post-training setting. Together, these results provide empirical support for the role of cross-lingual alignment in multilingual transfer and demonstrate that it can be directly encouraged in MoE LLMs.

These findings establish MoE routing as a promising interface for inducing cross-lingual transfer in modern LLMs. Future work could develop objectives that reduce token-level information loss, determine when and where alignment supervision is most effective within a model's training curriculum, and reduce reliance on explicitly parallel data. Studying these questions at pretraining scale may reveal more general ways to encourage shared representations without sacrificing language- or modality-specific structure.

AI USE STATEMENT

In this work, we used generative AI tools for [propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, interpret results.]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks,clean and reformat dataset], and [Generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, support qualitative and thematic data analysis] are not applicable to this work. Additionally, we used generative AI tools for [create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references].

We have reviewed all AI-assisted work. All code, analysis, and writing has been thoroughly reviewed and edited by authors of this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This work is highly reproducible. We provide important high-level details for pre-analysis, training, and evaluations in Sections 3 and 4. Further necessary details for reproducibility are provided in the Appendix E-F, including detailed hyperparameters. But above all, we attach the repository, anonymized, to the submission providing the exact code and implementation of all discussed parts of the project.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 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)*, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Peng. Multilingual routing in mixture-of-experts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZoZR0x7tTD.

Tyler A. Chang, Catherine Arnett, et al. Global PIQA: Evaluating commonsense reasoning across 100+ languages and cultures, 2025. URL https://arxiv.org/abs/2510.24081.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL https://aclanthology.org/2021.naacl-main.280/.

Pranjal A Chitale, Varun Gumma, Sanchit Ahuja, Prashant Kodali, Manan Uppadhyay, Deepthi Sudharsan, and Sunayana Sitaram. UPDESH: Synthesizing grounded instruction tuning data for 13 Indic languages. In *Proceedings of the 64th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pp. 37997–38041, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1763. URL https://aclanthology.org/2026.acl-long.1763/.

Cohere Labs. Global-MGSM. Hugging Face dataset, n.d. URL https://huggingface.co/datasets/CohereLabs/global-mgsm. Accessed September 9, 2026.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269/.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. DeepSeek-V3 Technical Report, 2025. URL https://arxiv.org/abs/2412.19437.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, July 2024. URL https://zenodo.org/records/12608602.

Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3020–3042, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.189. URL https://aclanthology.org/2023.findings-acl.189/.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long*

Papers), pp. 879–895, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.72. URL https://aclanthology.org/2021.acl-long.72/.

Nathan Godey and Yoav Artzi. Lost in backpropagation: The lm head is a gradient bottleneck. In *Third Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2603.10145.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30/.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. Language arithmetics: Towards systematic language neuron identification and manipulation. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2911–2937, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.156. URL https://aclanthology.org/2025.ijcnlp-long.156/.

HyoJung Han, Sweta Agrawal, and Eleftheria Briakou. Rethinking cross-lingual alignment: Balancing transfer and cultural erasure in multilingual llms, 2025. URL https://arxiv.org/abs/2510.26024.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1525–1542. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.72. URL https://aclanthology.org/2025.naacl-long.72/.

Dieuwke Hupkes and Nikolay Bogoychev. MultiLoKo: a multilingual local knowledge benchmark for LLMs spanning 31 languages, 2025. URL https://arxiv.org/abs/2504.10356. Accepted for publication in Computational Linguistics.

IBM Granite Team. Granite-4.0-H-Tiny model card, 2025. URL https://huggingface.co/ibm-granite/granite-4.0-h-tiny. Hugging Face model card.

Fan Jiang, Yu Zhao, Chenyang Lyu, Tianqi Shi, Yichao Du, Feihu Jiang, Longyue Wang, and Weihua Luo. Marco-MoE: Open multilingual mixture-of-expert language models with efficient upcycling, 2026. URL https://arxiv.org/abs/2604.25578.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Akhil Kedia, Mohd Abbas Zaidi, Sushil Khyalia, Jungho Jung, Harshith Goka, and Haejun Lee. Transformers get stable: An end-to-end signal propagation theory for language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23449–23531. PMLR, 2024. URL https://proceedings.mlr.press/v235/kedia24a.html.

Aditi Khandelwal, Marius Mosbach, Verna Dankers, Siva Reddy, and Golnoosh Farnadi. Leveraging routing dynamics in mixture-of-experts models for efficient language adaptation, 2026. URL https://arxiv.org/abs/2605.29714.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. In Martin Kay and Christian Boitet (eds.), *Proceedings of COLING 2012*, pp. 1459–1474, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL https://aclanthology.org/C12-1089/.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pp. 79–86, Phuket, Thailand, September 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11/.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6919–6971, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.384. URL https://aclanthology.org/2024.naacl-long.384/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 562–570. PMLR, 2015. URL https://proceedings.mlr.press/v38/lee15a.html.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8058–8076, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.445. URL https://aclanthology.org/2024.naacl-long.445/.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation, 2023. URL https://arxiv.org/abs/2305.15011.

Pengxiang Li, Lu Yin, and Shiwei Liu. Mix-LN: Unleashing the power of deeper layers by combining Pre-LN and Post-LN. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BChpQU64RG.

Ziyue Li and Tianyi Zhou. Your mixture-of-experts LLM is secretly an embedding model for free. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eFGQ97z5Cd.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. Language-specific latent process hinders cross-lingual performance, 2025. URL https://arxiv.org/abs/2505.13141.

Danni Liu and Jan Niehues. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15979–15996, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.778. URL https://aclanthology.org/2025.acl-long.778/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6a2fe573b31fb1d0b4d6f96a9a3e6b0f-Abstract-Conference.html.

Lalita Lowphansirikul, Charin Polpanumas, Attapol T. Rutherford, and Sarana Nutanong. A large english–thai parallel corpus from the web and machine-generated text. *Language Resources and Evaluation*, 56(2):477–499, 2022. doi: 10.1007/s10579-021-09536-6. URL https://doi.org/10.1007/s10579-021-09536-6.

Omar Mahmoud, Buddhika Laknath Semage, Thommen George Karimpanal, and Santu Rana. Improving multilingual language models by aligning representations through steering, 2025. URL https://arxiv.org/abs/2505.12584.

Liu O. Martin, Lucas Bandarkar, and Nanyun Peng. Extracting small translation specialists from llms by aggressively pruning experts, 2026. URL https://arxiv.org/abs/2605.28042.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xXTkbTBmqq.

Chinh Ngo, Trieu H. Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. MTet: Multi-domain translation for english and vietnamese, 2022. URL https://arxiv.org/abs/2210.05610.

NLLB Team, 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, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. 2022.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 27–38, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.3. URL https://aclanthology.org/2021.emnlp-main.3/.

Barun Patra, Saksham Singhal, Shaohan Huang, Zewen Chi, Li Dong, Furu Wei, Vishrav Chaudhary, and Xia Song. Beyond English-centric bitexts for better multilingual language representation learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15354–15373, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.856. URL https://aclanthology.org/2023.acl-long.856/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan,

Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162, 2022. URL https://aclanthology.org/2022.tacl-1.9/.

Kartik Ravisankar, HyoJung Han, Sarah Wiegreffe, and Marine Carpuat. Can you map it to English? the role of cross-lingual alignment in the multilingual performance of LLMs. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4854–4872, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.225. URL https://aclanthology.org/2026.eacl-long.225/.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya K Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan S Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia moakhar, Ayush Tarun, Azmine Toushik Wasi, Thenuka Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ced46a50befedcb884ccf0cbe8c3ad23-Abstract-Conference.html.

Holger Schwenk and Matthijs Douze. Learning joint multilingual sentence representations with neural machine translation. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih (eds.), *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 157–167, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2619. URL https://aclanthology.org/W17-2619/.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017.

Yingli Shen, Wen Lai, Shuo Wang, Ge Gao, Kangyang Luo, Alexander Fraser, and Maosong Sun. From unaligned to aligned: Scaling multilingual LLMs with multi-way parallel corpora. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7357–7379, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.374. URL https://aclanthology.org/2025.emnlp-main.374/.

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. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

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, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 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)*, pp. 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the Common Crawl. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1374–1383, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1135/.

Omnilingual SONAR Team, João Maria Janeiro, Pere-Lluís Huguet Cabot, Ioannis Tsiamas, Yen Meng, Vivek Iyer, Guillem Ramírez, Loic Barrault, Belen Alastruey, Xiang "Tony" Cao, Yu-An Chung, Marta R. Costa-Jussa, David Dale, Kevin Heffernan, Jaehyeong Jo, Artyom Kozhevnikov, Alexandre Mourachko, Christophe Ropers, Holger Schwenk, and Paul-Ambroise Duquenne. Omnilingual sonar: Cross-lingual and cross-modal sentence embeddings bridging massively multilingual text and speech, 2026. URL https://arxiv.org/abs/2603.16606.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task lora, 2024. URL https://arxiv.org/abs/2409.10173.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models, 2023.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL https://aclanthology.org/L12-1246/.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Junyang Lin, Fei Huang, and Jingren Zhou. PolyMath: Evaluating mathematical reasoning in multilingual contexts. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1431. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3d5aa9a7ce28cdc710fbd044fd3610f3-Abstract-Datasets_and_Benchmarks_Track.html.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 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)*, pp. 15366–15394, 2024. URL https://aclanthology.org/2024.acl-long.820/.

Shijie Wu and Mark Dredze. Do explicit alignments robustly improve multilingual encoders? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4471–4482, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.362. URL https://aclanthology.org/2020.emnlp-main.362/.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FrFQpAgnGE.

Zhaofeng Wu, Shiqi Wang, Boya Peng, Anuj Kumar Goyal, Melanie Kambadur, Sebastian Ruder, Yoon Kim, and Chloe Bi. Parallel-SFT: Improving zero-shot cross-programming-language transfer for code RL. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26583–26598, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1324. URL https://aclanthology.org/2026.findings-acl.1324/.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation, 2020. URL https://arxiv.org/abs/2012.15466.

Haolei Xu, Haiwen Hong, Hongxing Li, Rui Zhou, Yang Zhang, Longtao Huang, Hui Xue, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Seeing but not thinking: Routing distraction in multimodal mixture-of-experts. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31164–31178, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1438. URL https://aclanthology.org/2026.acl-long.1438/.

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, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1513–1532, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.79. URL https://aclanthology.org/2025.emnlp-main.79/.

Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models, 2024. URL https://arxiv.org/abs/2402.01739.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7502–7522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.405. URL https://aclanthology.org/2024.acl-long.405/.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11941–11952, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL https://aclanthology.org/2020.acl-main.148/.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. Getting more from less: Large language models are good spontaneous multilingual learners. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8037–8051, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.457. URL https://aclanthology.org/2024.emnlp-main.457/.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40. AAAI Press, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/40782.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, Tat-Seng Chua, and Ting Liu. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 38608–38642. Curran Associates, Inc., 2025. doi: 10.52202/085713-1290. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/372bd0e47f2d5bceca7e300e1446849c-Paper-Conference.pdf.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ctXYOoAgRy.

Weihua Zheng, Chang Liu, Zhengyuan Liu, Xin Huang, Kui Wu, Muhammad Huzaifah, Ai Ti Aw, and Roy Ka-Wei Lee. Bridging linguistic gaps: Cross-lingual mapping in pre-training and dataset for enhanced multilingual llm performance. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 25(6), June 2026. ISSN 2375-4699. doi: 10.1145/3811819. URL https://doi.org/10.1145/3811819.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. Cross-lingual sentiment classification with bilingual document representation learning. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1403–1412, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1133. URL https://aclanthology.org/P16-1133/.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8411–8423. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.498. URL https://aclanthology.org/2024.findings-acl.498/.

A Visual Diagram of Our Methodology

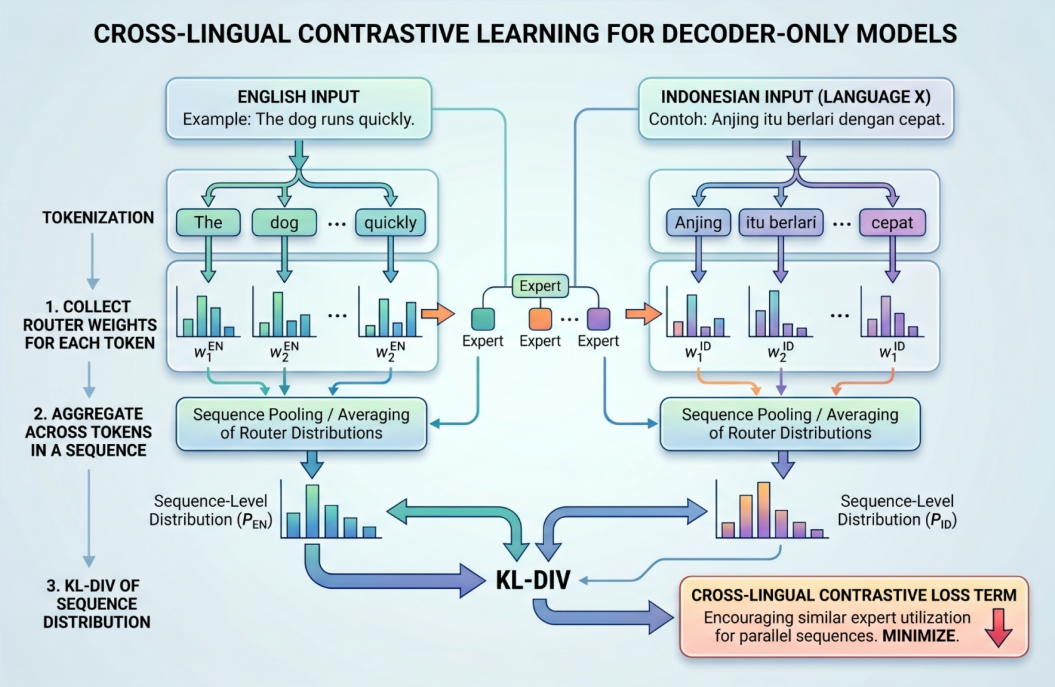

Figure 5: Overview of our cross-lingual alignment method to supplement the formal description in Section 3.2.

B Complete Experimental Results

Tables 3–6 use one row per experimental condition. *Original* denotes the released post-trained checkpoint before our training, not a base model. Baseline is target-language LM-only training; + routing adds our auxiliary loss; Router-only updates only the selected routers; + hidden aligns hidden states instead. α is the auxiliary-loss coefficient. LM loss excludes auxiliary losses; AVG averages the available benchmarks. Percentage signs are omitted, and FLORES retains its original scale. MMLU Med. and PIQA denote Global MMLU Medical and Global PIQA. Dashes indicate unavailable or inapplicable values; `scaled_alpha` is omitted.

Table 3: Complete results for Qwen3-30B-A3B. Auxiliary alignment is applied to layers 7–34. Bold AVG values mark the best result per language, including ties.

(a) Training configuration and aggregate results

Language	Condition	α	Learning rate	LM loss	AVG
Telugu (tel)	Original	–	–	0.77	44.6
	Baseline	–	3×10^{-6}	0.65	45.4
	+ routing	2	3×10^{-6}	0.7	**46.3**
	Router-only	10	3×10^{-6}	0.78	45.1
Kyrgyz (kir)	Original	–	–	3.06	40.8
	Baseline	–	4×10^{-6}	1.95	43.5
	+ routing	10	4×10^{-6}	2.07	**45.6**
	Router-only	10	4×10^{-6}	3	42.2
Kannada (kan)	Original	–	–	0.81	49.3
	Baseline	–	3.5×10^{-6}	0.69	49.6
	+ routing	4	2.5×10^{-6}	0.76	**51.4**
	Router-only	6	3×10^{-6}	0.81	49.9
Thai (tha)	Original	–	–	2.63	43.0
	Baseline	–	3×10^{-6}	2.2	44.9
	+ routing	10	3×10^{-6}	2.36	**45.4**
	Router-only	10	4×10^{-6}	2.59	43.8

(b) Benchmark scores

Lang.	Condition	Belebele	FLORES	MMLU Med.	PIQA	INCLUDE	MGSM	MMLU Pro-X	Multi LoKo	Poly Math
tel	Original	70.7	32.8	60.2	68.0	50.4	10.0	25.8	–	39.2
	Baseline	68.0	31.6	60.0	69.0	49.5	20.8	24.5	–	40.2
	+ routing	72.8	29.0	61.7	69.0	50.0	22.4	22.1	–	43.2
	Router-only	71.0	33.4	61.9	66.0	48.9	15.6	24.5	–	39.8
kir	Original	69.6	18.3	–	57.0	–	18.4	–	–	–
	Baseline	70.9	21.2	–	58.0	–	24.0	–	–	–
	+ routing	72.9	20.5	–	62.0	–	26.8	–	–	–
	Router-only	69.1	18.8	–	60.0	–	20.8	–	–	–
kan	Original	75.2	32.4	–	63.0	–	26.4	–	–	–
	Baseline	74.8	31.6	–	65.0	–	26.8	–	–	–
	+ routing	78.9	29.4	–	63.0	–	34.4	–	–	–
	Router-only	76.3	33.0	–	64.0	–	26.4	–	–	–
tha	Original	83.6	24.8	–	69.0	–	11.6	58.0	12.8	41.2
	Baseline	83.7	27.7	–	68.0	–	30.4	49.2	14.0	41.6
	+ routing	84.4	24.5	–	66.0	–	33.6	51.2	14.4	43.8
	Router-only	83.9	24.5	–	69.0	–	16.0	59.2	12.8	41.2

Table 4: Complete results for GPT-OSS-20B. Auxiliary alignment is applied to layers 4–17. Bold AVG values mark the best result per language, including ties.

(a) Training configuration and aggregate results

Language	Condition	α	Learning rate	LM loss	AVG
Telugu (tel)	Original	–	–	3.05	42.5
	Baseline	–	4×10^{-7}	2.69	46.5
	+ routing	40	4×10^{-7}	2.87	**47.0**
	Router-only	40	8×10^{-7}	3.04	43.2
Kyrgyz (kir)	Original	–	–	4.28	35.9
	Baseline	–	6×10^{-7}	3.4	42.0
	+ routing	20	4×10^{-7}	3.64	**43.8**
	Router-only	40	5×10^{-7}	4.27	36.4
Kannada (kan)	Original	–	–	2.99	47.5
	Baseline	–	6×10^{-7}	2.56	55.5
	+ routing	40	6×10^{-7}	2.76	**56.4**
	Router-only	40	8×10^{-7}	2.96	48.8
Thai (tha)	Original	–	–	3.69	34.9
	Baseline	–	6×10^{-7}	3.18	39.9
	+ routing	40	8×10^{-7}	3.47	**40.8**
	Router-only	20	1×10^{-6}	3.67	34.7

(b) Benchmark scores

Lang.	Condition	Belebele	FLORES	MMLU Med.	PIQA	INCLUDE	MGSM	MMLU Pro-X	Multi LoKo	Poly Math
tel	Original	58.3	28.3	59.8	69.0	37.4	18.8	35.2	–	32.8
	Baseline	69.4	39.2	61.7	68.0	38.7	25.6	30.2	–	39.4
	+ routing	68.9	41.3	61.0	69.0	38.5	26.0	28.4	–	42.8
	Router-only	59.3	27.9	59.5	68.0	37.0	20.4	36.9	–	36.4
kir	Original	31.2	18.1	–	62.0	–	32.4	–	–	–
	Baseline	50.2	19.1	–	62.0	–	36.8	–	–	–
	+ routing	44.2	19.1	–	67.0	–	44.8	–	–	–
	Router-only	31.4	18.2	–	60.0	–	36.0	–	–	–
kan	Original	62.2	26.2	–	63.0	–	38.6	–	–	–
	Baseline	72.0	40.0	–	63.0	–	46.8	–	–	–
	+ routing	70.9	41.8	–	65.0	–	48.0	–	–	–
	Router-only	61.8	26.6	–	64.0	–	42.8	–	–	–
tha	Original	50.9	25.9	–	65.0	–	15.2	50.2	1.6	35.6
	Baseline	70.2	37.3	–	66.0	–	14.0	47.5	8.4	36.2
	+ routing	70.6	35.2	–	67.0	–	19.6	46.0	8.2	39.2
	Router-only	50.9	27.5	–	64.0	–	14.0	48.0	2.4	36.2

Table 5: Complete results for Granite-4.0-H-Tiny. Auxiliary alignment is applied to layers 10–35. Bold AVG values mark the best result per language, including ties.

(a) Training configuration and aggregate results

Language	Condition	α	Learning rate	LM loss	AVG
Vietnamese (vie)	Original	–	–	2.93	34.8
	Baseline	–	4×10^{-7}	2.31	36.5
	+ routing	200	5×10^{-7}	2.402	**36.8**
	Router-only	200	4×10^{-7}	2.85	36.5
Sinhala (sin)	Original	–	–	1.09	31.3
	Baseline	–	8×10^{-7}	0.95	32.8
	+ routing	200	6×10^{-7}	0.99	**32.9**
	Router-only	100	4×10^{-7}	1.09	31.9
Hungarian (hun)	Original	–	–	4.77	34.5
	Baseline	–	8×10^{-7}	3.38	38.1
	+ routing	50	8×10^{-7}	3.61	**38.8**
	Router-only	100	4×10^{-7}	4.49	38.2

(b) Benchmark scores

Lang.	Condition	Belebele	FLORES	MMLU Med.	PIQA	INCLUDE	MGSM	MMLU Pro-X	Multi LoKo	Poly Math
vie	Original	62.6	37.7	47.4	59.0	46.9	23.2	9.1	8.0	18.8
	Baseline	64.2	38.3	48.6	60.0	46.4	31.6	10.2	9.6	19.6
	+ routing	65.0	38.0	48.1	62.0	46.0	32.8	10.8	9.2	19.0
	Router-only	64.1	39.2	48.8	60.0	46.7	32.4	8.8	9.6	18.6
sin	Original	47.6	14.6	42.9	48.0	–	3.2	–	–	–
	Baseline	47.0	16.2	43.6	51.0	–	6.0	–	–	–
	+ routing	47.4	16.8	44.5	50.0	–	5.6	–	–	–
	Router-only	47.7	16.8	43.6	48.0	–	3.6	–	–	–
hun	Original	61.8	25.1	–	60.0	40.5	16.0	3.7	–	–
	Baseline	60.3	25.0	–	59.0	41.8	37.6	5.0	–	–
	+ routing	60.7	25.9	–	60.0	40.5	40.0	5.8	–	–
	Router-only	59.9	26.1	–	59.0	41.6	38.0	4.8	–	–

Table 6: Complete results for Marco-Nano. Auxiliary alignment is applied to layers 7–19. Bold AVG values mark the best result per language, including ties. Hidden-state alignment is evaluated only for Sinhala and Hungarian.

(a) Training configuration and aggregate results

Language	Condition	α	Learning rate	LM loss	AVG
Vietnamese (vie)	Original	–	–	3.46	44.2
	Baseline	–	1×10^{-6}	3.16	**45.6**
	+ routing	10	1.2×10^{-6}	3.23	**45.6**
	Router-only	40	2×10^{-6}	3.45	44.8
Sinhala (sin)	Original	–	–	1.79	24.0
	Baseline	–	1×10^{-6}	1.57	26.1
	+ routing	10	1.3×10^{-6}	1.59	**26.8**
	Router-only	20	1.2×10^{-6}	1.79	25.2
	+ hidden	5000	1.2×10^{-6}	1.74	26.3
Hungarian (hun)	Original	–	–	3.35	36.3
	Baseline	–	1.5×10^{-6}	2.76	37.4
	+ routing	10	1.5×10^{-6}	2.76	**38.1**
	Router-only	40	2×10^{-6}	3.34	37.1
	+ hidden	5000	8×10^{-7}	3.16	36.9

(b) Benchmark scores

Lang.	Condition	Belebele	FLORES	MMLU Med.	PIQA	INCLUDE	MGSM	MMLU Pro-X	Multi LoKo	Poly Math
vie	Original	73.7	30.4	52.1	68.0	52.0	50.4	35.1	10.8	25.2
	Baseline	73.3	38.7	52.4	72.0	49.6	54.0	33.2	9.6	27.2
	+ routing	72.9	38.0	52.6	72.0	51.8	50.8	33.1	10.4	29.0
	Router-only	73.6	32.1	51.9	66.0	51.3	55.2	35.8	11.2	26.4
sin	Original	34.0	8.5	32.4	44.0	–	1.2	–	–	–
	Baseline	35.0	9.2	31.2	53.0	–	2.0	–	–	–
	+ routing	33.4	10.6	31.2	58.0	–	0.8	–	–	–
	Router-only	33.8	9.5	31.4	50.0	–	1.2	–	–	–
	+ hidden	34.4	9.8	31.6	55.0	–	0.8	–	–	–
hun	Original	70.6	11.7	–	54.0	43.3	26.4	12.0	–	–
	Baseline	69.3	14.8	–	52.0	44.0	35.2	9.2	–	–
	+ routing	69.0	14.0	–	53.0	42.5	36.8	13.2	–	–
	Router-only	70.3	12.3	–	54.0	43.3	30.0	12.4	–	–
	+ hidden	70.6	11.9	–	53.0	43.6	30.4	12.1	–	–

C VISUALIZATIONS OF MOE ROUTING FOR LAYER SELECTION

Below, we provide the MoE routing emphdivergence graphs using the normalized JS-divergence metric defined in Bandarkar et al. (2026). For Qwen3-30B-A3B and GPT-OSS-20B, that work already selects layers and establishes they work well. Notably, that work identifies that router steering is very sensitive to these exact layers. For Granite-4-H-Tiny and Marco-Nano, we subjectively pick layers based on the below visualizations. The final layer ranges are in the last column of Table 7 in the next section.

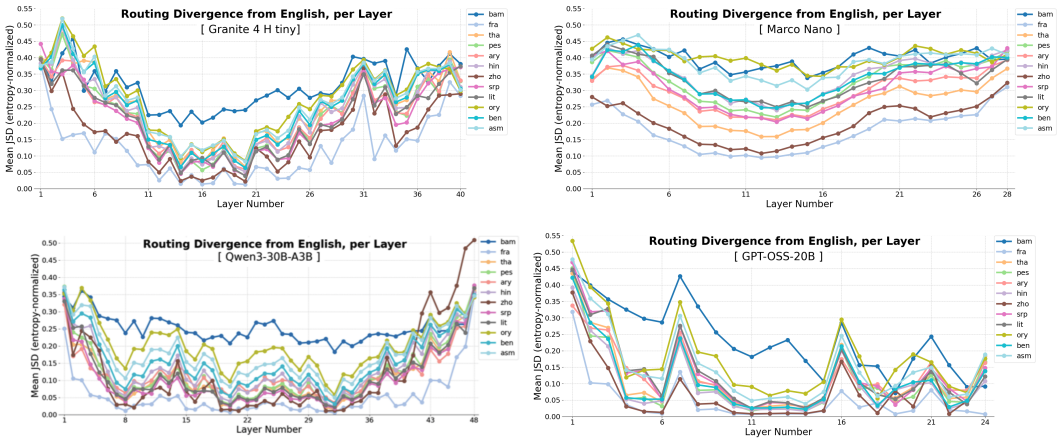

Figure 6: MoE router divergence for the four models used in experiments. Lower values represent higher alignment, and corresponds to higher-resource languages.

D Model MoE Details

Table 7: MoE architecture and router-layer details for each model. Active parameter counts are shown in parentheses.

Model	Params (Active)	Num. Layers	Active Experts / Total	Layers for $\mathcal{L}_{XLR}$
Qwen3-30B-A3B	31B (A3B)	48	8 / 128	7–34
GPT-OSS-20B	22B (A4B)	24	4 / 32	4–17
Marco-Nano	8B (A0.6B)	28	8 / 238	7–19
Granite-4.0-H-Tiny	7B (1B)	40	6 / 64	10–35

E Parallel Training Data

We construct a separate English–target-language parallel dataset for each of the seven languages in our experiments. Each dataset targets 200,000 training examples, with 1,000 additional examples reserved for validation. The examples comprise both sentence-aligned bitext and parallel instruction–response conversations. For instruction-tuning sources, we align the English and target-language versions of the same example and preserve their multi-turn chat structure. We remove empty pairs, pairs whose two sides are identical, and duplicate pairs from the same source.

Table 8 lists the data sources used for each language. The source weights in our data-construction pipeline specify requested allocations rather than the final composition: some sources do not contain enough examples that pass filtering, and their shortfalls are refilled from the remaining sources with available data. Thus, the planned weights do not reliably describe the realized number of examples from each source.

Language	Training-data sources
Vietnamese	MTet (Ngo et al., 2022); Bactrian-X (Li et al., 2023); NLLB English bitext (NLLB Team et al., 2022); OPUS-100 (Zhang et al., 2020).
Sinhala	Bactrian-X (Li et al., 2023); NLLB English bitext (NLLB Team et al., 2022); OPUS-100 (Zhang et al., 2020).
Hungarian	NLLB English bitext (NLLB Team et al., 2022); OPUS-100 (Zhang et al., 2020).
Telugu	Updesh (Chitale et al., 2026); Bactrian-X (Li et al., 2023); Samanantar (Ramesh et al., 2022); NLLB English bitext (NLLB Team et al., 2022); OPUS-100 (Zhang et al., 2020).
Kannada	Updesh (Chitale et al., 2026); Samanantar (Ramesh et al., 2022); NLLB English bitext (NLLB Team et al., 2022); OPUS-100 (Zhang et al., 2020).
Thai	Bactrian-X (Li et al., 2023); SCB-MT-EN-TH-2020 (Lowphansirikul et al., 2022); OPUS-100 (Zhang et al., 2020).
Kyrgyz	NLLB English bitext (NLLB Team et al., 2022); OPUS-100 (Zhang et al., 2020).

Table 8: Parallel training-data sources for each target language. The table records source inclusion, rather than nominal or realized mixture ratios.

Length filtering. We tokenize both the English and target-language sides before admitting an example and discard examples that exceed the applicable sequence-length limit. The default limit is 768 tokens per side. Because the model tokenizers represent some scripts inefficiently, we use stricter language-specific caps for Sinhala (512 tokens), Kannada (640 tokens), and Telugu (640 tokens). This prevents poorly tokenized examples from producing disproportionately long sequences while retaining the same filtering procedure across sources.

F Training Details

Section 4.2 describes the training setup and some hyperparameters. All remaining training settings are fixed across experiments. Each run consists of one epoch over 200,000 training examples. We use a warmup–stable–decay (WSD) learning-rate schedule, with a linear warmup over the first 10,000 examples and a decay over the final 40,000 examples. The intervening 150,000 examples therefore use a constant learning rate. We optimize with the fused PyTorch implementation of AdamW and do not apply NEFTune. Although the per-device batch size varies with the memory requirements of each model, we adjust the number of gradient-accumulation steps to maintain an effective batch size of 32 examples.

Compute Requirements The two larger models require fully sharded data parallelism (FSDP) across multiple GPUs with 80 GB of memory: GPT-OSS-20B uses three or four GPUs, depending on the run, while Qwen3-30B-A3B uses four. In contrast, Granite-4.0-H-Tiny and Marco-Nano each fit on a single GPU. Training Granite-4.0-H-Tiny and Marco-Nano on one GPU is comparatively slow, with a run taking approximately 15 hours on H100 or A100.