Key Layers for Cross-Lingual Router Alignment in Multilingual MoE Models

Anonymous ACL submission

Abstract

Mixture-of-Experts (MoE) language models expand capacity through sparse expert activation, but multilingual routing is often analyzed only through aggregate expert usage. Motivated by links between non-English performance and English-routing similarity, we use answer-token sentence-level router distributions as routing representations and find depth-dependent cross-language geometry: early layers reflect script and language-family effects, while middle and mid-late layers separate low- from higher-resource languages. Because low-resource supervision makes broad multi-layer router alignment difficult, we define paired source-English overlap to select a compact set of high-overlap key layers. We propose key-layer router alignment, a lightweight router-only method that uses the model's English router as a frozen teacher and updates only selected router gates. The objective increases source-language mass on the English teacher's top-k experts while preserving original source routing. On Qwen3-30B-A3B-Instruct, this improves MGSM accuracy while preserving English MMLU performance, suggesting that router geometry is both diagnostic and useful for lightweight multilingual adaptation.

1 Introduction

Mixture-of-Experts (MoE) language models route each token to a small expert subset, increasing capacity while preserving sparse computation (Shazeer et al., 2017; Du et al., 2022). This improves compute–capacity trade-offs but makes multilingual behavior harder to interpret. We focus on routers, which determine token-level expert paths. If related inputs in different languages follow different paths, transfer may be limited even when relevant knowledge is encoded. Recent work shows that non-English MoE performance tracks similarity to English routing (Bandarkar et al., 2025), suggesting that router similarity is not only diagnostic but also a useful adaptation target.

Existing multilingual MoE analyses often rely on expert-usage statistics. These statistics identify which experts are used, but not how routing distributions are organized across languages or which layers are alignable. We instead treat sentence-level router distributions as routing representations. On multilingual mathematical reasoning data, cross-language overlap is highly non-uniform across depth: early layers mainly reflect script and language-family effects, while middle and mid-late layers separate Bengali, Swahili, and Telugu from higher-resource languages. Similar FLORES trends suggest that the pattern is not task-specific.

To connect analysis to adaptation, we define paired source-English overlap, comparing each source example with its English translation at every layer. High-overlap layers form a compact key-layer set and turn descriptive geometry into a layer-selection rule. This focus matters in low-resource settings, where scarce parallel data cannot support broad routing alignment across many layers. We then propose key-layer router alignment, a lightweight method that uses the model's own English router as a frozen target and updates selected router gates. The objective increases source-language probability mass on the English teacher's top-k experts while limiting drift from the original source routing distribution.

We contribute a geometric analysis of multilingual MoE routing, a paired source-English overlap criterion for key-layer selection, and a data- and parameter-efficient router-only alignment objective based on English teacher top-k mass alignment.

2 Related Work

MoE architectures improve compute–capacity trade-offs through sparse expert activation (Shazeer et al., 2017; Du et al., 2022). GShard (Lepikhin et al., 2021), Switch Transformer (Fedus et al.,

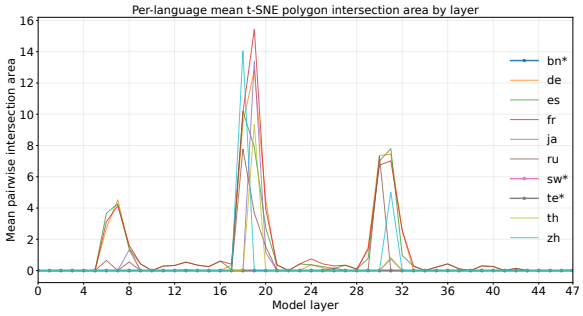

Figure 1: Layer-wise cross-language overlap on Qwen3-30B-A3B-Instruct, concentrated in a few layers.

2022), and GLaM (Du et al., 2022) established conditional computation, while recent open-weight MoE models such as Mixtral (Jiang et al., 2024), Qwen3 (Yang et al., 2025), and GPT-OSS (OpenAI et al., 2025) make router behavior important for sparse LLM analysis.

Multilingual models can transfer across languages despite script and resource differences (Pires et al., 2019; Conneau et al., 2020). Their representations contain language-sensitive and language-neutral structure (Chang et al., 2022), and non-English inputs can pass through English-like latent spaces (Wendler et al., 2024). We extend this view from dense hidden states to MoE router distributions, which determine expert paths.

Prior work links non-English performance to English-routing similarity through expert-frequency analysis of multilingual MoE models (Bandarkar et al., 2025), motivating router alignment. However, low-resource supervision makes broad multi-layer alignment difficult. We instead analyze layer-wise router outputs and use paired source-English overlap to identify compact key layers for targeted adaptation. We evaluate on MGSM (Shi et al., 2023) and test English-domain preservation with MMLU (Hendrycks et al., 2021).

3 Router Analysis

We analyze router distributions rather than expert-frequency counts. For each example, we average answer-token router probabilities into a sentence-level distribution per layer. We use mCoT-MATH (Lai and Nissim, 2024) and FLORES (Goyal et al., 2022), and report Qwen3-30B-A3B-Instruct (Yang et al., 2025) in the main text. For cross-language analysis, we exclude English, sample 200 examples per language, average every 10 examples into a point, project to two dimensions,

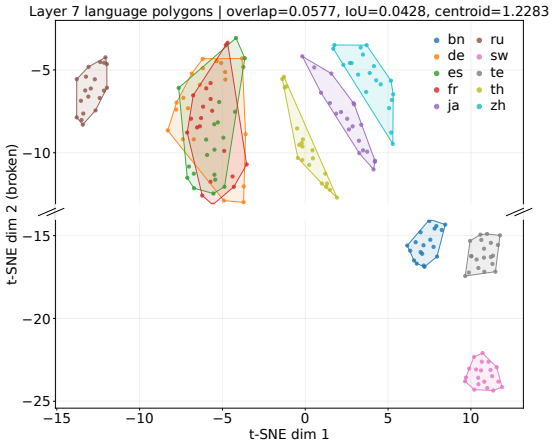

Figure 2: Layer 7 routing representations mainly reflect language-family and script-sensitive routing.

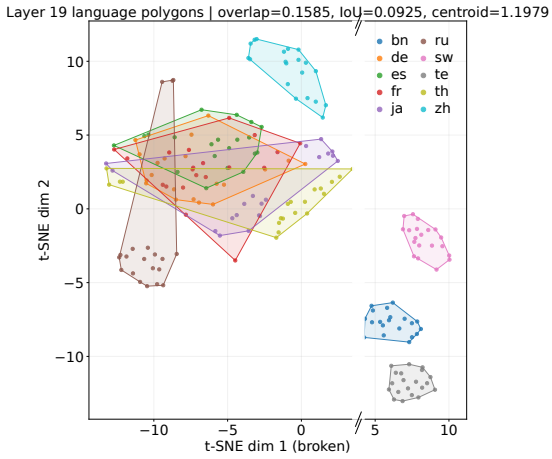

Figure 3: Layer 19 routing representations show low-resource languages separating from higher-resource languages.

and measure convex-hull overlap.

Figure 1 shows that cross-language overlap is concentrated in a few layers. The early peak reflects related scripts or typology, whereas middle and mid-late layers separate low- from higher-resource languages, suggesting divergent expert paths. Appendix A gives additional analyses.

Figure 2 illustrates the early regime: related languages overlap more than distant ones. Appendix B shows the neighboring layers. We interpret this as family- and script-sensitive routing rather than transferable cross-lingual reasoning structure.

Figure 3 shows a resource-sensitive regime that persists in surrounding layers (Appendix C). Low-resource languages such as Bengali, Swahili, and Telugu occupy regions distinct from higher-resource languages, motivating targeted adaptation in layers that differ yet remain alignable. These cross-language patterns show where routing diverges but not which layers should align to English,

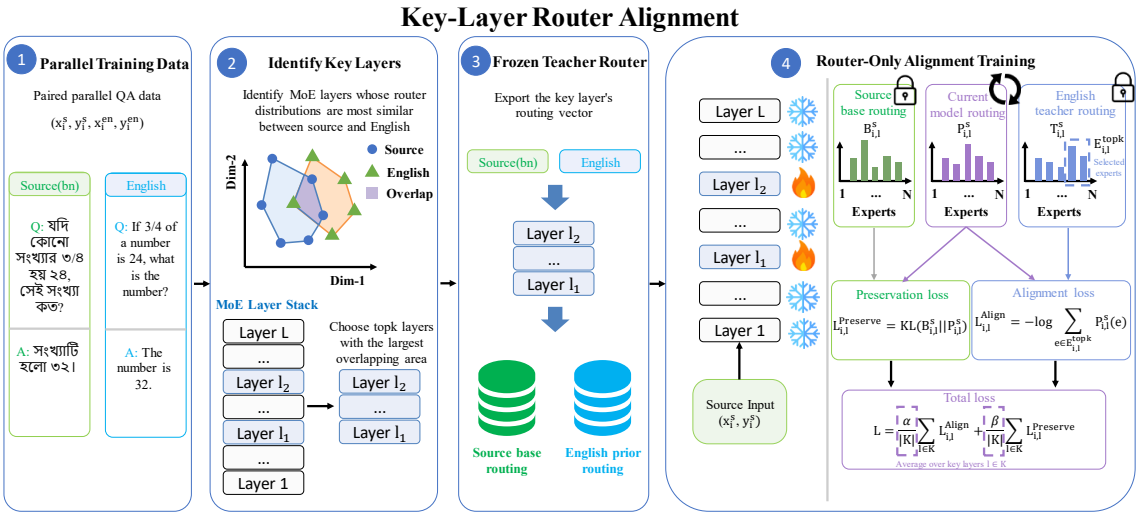

Figure 4: Overview of key-layer router alignment, which identifies key layers, caches English teacher and source base routers, and updates only selected router gates.

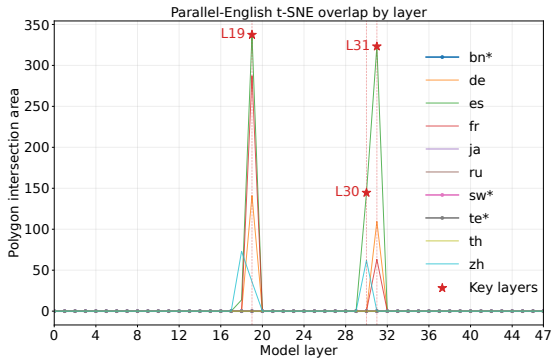

Figure 5: Layer-wise paired source-English overlap. High-overlap layers form the key-layer set for router alignment.

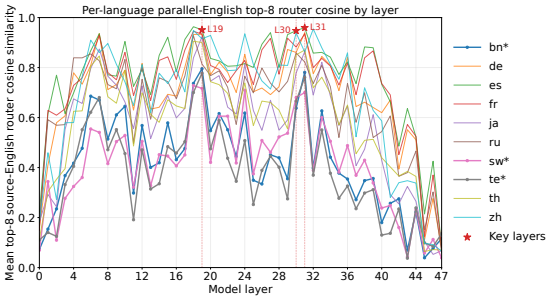

Figure 6: Original-router-space top-8 source-English router cosine similarity on Qwen3-30B-A3B-Instruct. The selected key-layer region shows high source-English routing similarity across languages.

so the method section introduces paired source-English overlap as an adaptation-targeted criterion.

4 Key-Layer Router Alignment

KLRA first selects alignable key layers with paired source-English router geometry, then adapts only their router gates. Figure 4 gives an overview: paired router geometry identifies key layers, cached English and source routers provide teacher and preservation targets, and training updates only selected gates.

We train on mCoT-MATH-Align, a parallel set constructed from mCoT-MATH (Lai and Nissim, 2024). Each example pairs a source-language question-answer with its English translation, $(x_i^s, y_i^s, x_i^{en}, y_i^{en})$. The final router-alignment setting uses Bengali, Telugu, and Swahili, with 200 examples per language; Appendix F gives details.

To identify key layers, we compute paired source-English overlap in projected router space. Figure 5 shows that this overlap is concentrated, so we rank layers by mean overlap and select the top ones. Figure 6 verifies the selected region in the original router space, where source languages maintain high cosine similarity with paired English router vectors after retaining the top-k activated experts. Unlike generic cross-language overlap, paired overlap compares matched content and directly matches the alignment objective. We select high-overlap rather than maximally divergent layers because they indicate an existing shared routing subspace, allowing limited low-resource supervision to strengthen English-like routing without forcing unrelated router states to match. Appendix D shows projection examples.

Before training, we cache the English teacher router $T_{i,l}^{en}$ and source base router $B_{i,l}^{s}$ for each example and key layer; both are answer-token sentence-level distributions. During adaptation,

only selected router gates are trainable, and the base model provides the teacher signal. This restriction reduces trainable parameters and concentrates low-resource supervision.

Let $P^s_{i,l}$ be the adapted source router distribution. We align it to the English teacher by maximizing mass on the teacher's top-k experts and preserve the source base router with a KL term. The two terms have complementary roles. Top-k mass alignment encourages source examples to use English-like expert choices, while KL preservation keeps the adapted router close to the model's original source-language behavior.

$$\mathcal{L}^{i,l}_{align} = -\log \textstyle\sum_{e \in \mathcal{E}^{topk}_{i,l}} P^s_{i,l}(e),$$
$$\mathcal{L}^{i,l}_{preserve} = \mathrm{KL}\left(B^s_{i,l} \parallel P^s_{i,l}\right).$$

The top-k objective avoids forcing full-distribution matching, while the KL term constrains unnecessary drift. The final objective is

$$\mathcal{L} = \frac{1}{|\mathcal{K}|} \sum_{l \in \mathcal{K}} \left(\alpha \mathcal{L}^l_{align} + \beta \mathcal{L}^l_{preserve}\right),$$

where $\mathcal{K}$ is the key-layer set, and $\mathcal{L}^l_{align}$ and $\mathcal{L}^l_{preserve}$ denote averages over training examples.

5 Results

We evaluate adaptation on Qwen3-30B-A3B-Instruct (Yang et al., 2025), with GPT-OSS-20B (OpenAI et al., 2025) results in Appendix E. We compare key-layer router alignment (KLRA) with two router-only controls that update the same number of layers: random-layer router alignment (RLRA) and neighbor key-layer router alignment (NKLRA). All methods use the same 600 Bengali, Telugu, and Swahili training examples, isolating the layer-selection criterion. Appendix F reports hyperparameters, evaluation setup, and reporting protocol.

Multilingual mathematical reasoning. Table 1 reports MGSM accuracy. KLRA achieves the best average, improving Qwen3 from 80.7% to 81.4% and raising the trained low-resource average from 64.8% to 66.4%. RLRA and NKLRA remain below the base average, so gains are not from gate updates alone; NKLRA also improves the low-resource average less than KLRA. KLRA improves eight of eleven languages over the base model, including all supervised languages and unseen German, French, Russian, Thai, and Chinese. On GPT-OSS-20B, Appendix E shows that KLRA improves MGSM from 61.1% to 62.3% and raises the trained low-resource average from 52.3% to 57.7%, suggesting stronger effects in the lower-performing setting.

Lang.	Base	RLRA	NKLRA	KLRA
bn*	78.0%	77.6%	79.6%	**80.2%**
de	85.6%	86.4%	84.8%	**87.2%**
en	94.4%	**94.8%**	92.4%	94.0%
es	**92.0%**	90.0%	91.6%	91.4%
fr	83.6%	83.6%	84.0%	**84.4%**
ja	80.8%	**81.6%**	79.2%	80.8%
ru	86.4%	86.0%	**87.2%**	87.0%
sw*	52.0%	50.0%	51.2%	**53.8%**
te*	64.4%	63.6%	**66.0%**	65.2%
th	84.4%	85.2%	85.2%	**85.4%**
zh	85.6%	**86.4%**	84.8%	86.0%
AVG	80.7%	80.5%	80.5%	**81.4%**
Low-resource*	64.8%	63.7%	65.6%	**66.4%**

Table 1: Two-shot MGSM accuracy for Qwen3-30B-A3B-Instruct. Asterisks mark training languages.

MMLU Group	Base	RLRA	NKLRA	KLRA
Humanities	82.1%	82.2%	**83.3%**	83.2%
Social Sciences	83.3%	**83.7%**	82.8%	83.1%
STEM	86.3%	**86.8%**	86.4%	86.7%
Other	82.9%	82.9%	**83.0%**	**83.0%**
Average	83.7%	83.9%	83.9%	**84.0%**

Table 2: Five-shot English MMLU accuracy for Qwen3-30B-A3B-Instruct.

English-domain preservation. Because the target is English routing, English-domain degradation is a risk. Table 2 shows that MMLU performance remains close to the base model. KLRA slightly improves the average from 83.7% to 84.0% and keeps all categories close to the base model, suggesting that key-layer restriction and source-router preservation limit harmful drift.

6 Conclusion

We studied multilingual MoE routing through sentence-level distributions and found cross-language routing structure concentrated in a few layers. Early layers mainly capture language-family and script effects, while middle and mid-late layers separate low- from higher-resource languages, showing that multilingual behavior depends on how routers allocate computation across depth. We defined paired source-English overlap for key-layer selection and proposed KLRA, which uses the model's English router as a frozen target and updates selected router gates. KLRA improves MGSM while maintaining English MMLU performance, offering a data- and parameter-efficient way to leverage English-like routing for low-resource languages.

7 Limitations

Our analysis uses low-dimensional projections and convex-hull overlap, which approximate high-dimensional router geometry. The adaptation setting is narrow, covering parallel mathematical reasoning data and a small set of low-resource languages. Because router-only alignment changes gating while leaving experts and dense representations fixed, its ceiling is lower than broader adaptation methods. English may also be a suboptimal target for languages whose reasoning traces or cultural priors differ substantially from English.

References

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Peng. 2025. Multilingual routing in mixture-of-experts. *Preprint*, arXiv:2510.04694.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136. Association for Computational Linguistics.

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

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, and 8 others. 2022. GLaM: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR.

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.

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, and 5 others. 2023. A framework for few-shot language model evaluation.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 106 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

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

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

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

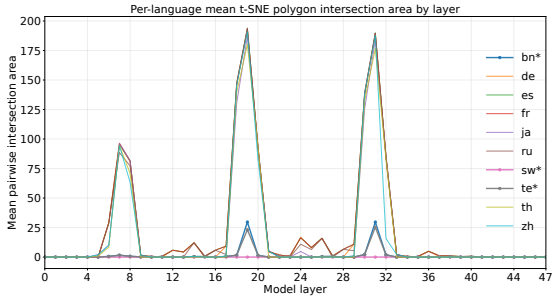

Figure 7: Layer-wise cross-language overlap of routing representations on FLORES. The same concentration pattern appears outside mathematical reasoning data.

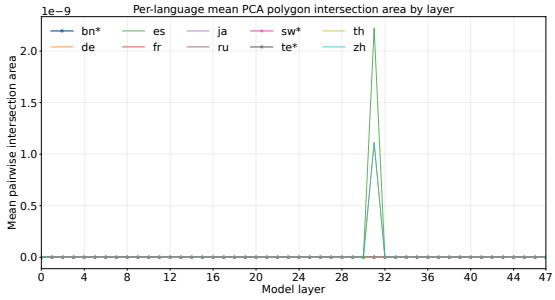

Figure 8: Layer-wise cross-language overlap under PCA. The middle-layer concentration remains visible under a different projection method.

A Additional Router Overlap Analyses

Figure 7 shows that the same concentration of cross-language overlap appears on FLORES, while Figure 8 shows that the pattern remains visible under PCA. Figure 9 shows a similar non-uniform depth pattern in GPT-OSS-20B.

B Layers Around the Early Overlap Peak

Figure 10 shows that the early region consistently exhibits language-family- and script-sensitive routing.

C Layers Around the Resource-Sensitive Region

Figure 11 shows that the resource-sensitive pattern extends across a compact middle-to-late region rather than a single isolated layer.

D Qwen Key-Layer Router Geometry

Figures 12 and 13 show projected examples for two selected Qwen key layers from the high-overlap region. We treat t-SNE overlap as a diagnostic rather than a strict high-dimensional criterion, and Figure 6 complements it in the original router space.

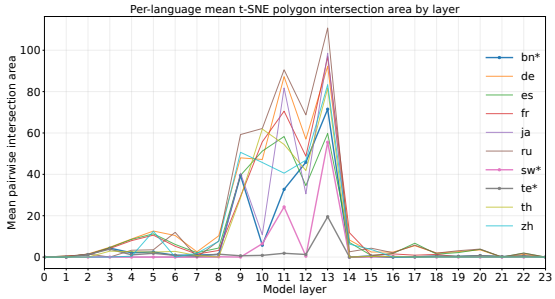

Figure 9: Layer-wise cross-language overlap of routing representations on GPT-OSS-20B. The same non-uniform pattern across depth appears in a second MoE model.

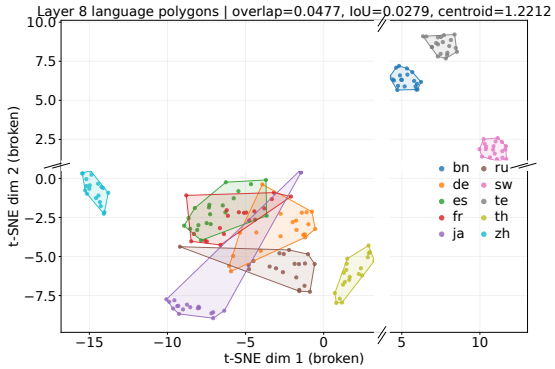

Figure 10: Routing representations at layer 8 on Qwen3-30B-A3B-Instruct. The early overlap pattern remains consistent in the neighboring layer.

E GPT-OSS-20B Results

Figure 14 shows the paired source-English overlap ranking used to select key layers in GPT-OSS-20B.

Table 3 reports GPT-OSS-20B MGSM results. KLRA improves the average from 61.1% to 62.3% and raises the low-resource average from 52.3% to 57.7%, mainly through gains on Bengali and Telugu. This provides additional support for router-only adaptation while showing model-dependent effect sizes. Figure 15 reports the corresponding per-language standard errors, which range from 2.71% to 3.17%.

F Data and Experimental Details

mCoT-MATH-Align construction. mCoT-MATH-Align is constructed from mCoT-MATH (Lai and Nissim, 2024) by sampling 200 examples from each of ten non-English languages, including Bengali (bn), German (de), Spanish (es), French (fr), Japanese (ja), Russian (ru), Swahili (sw), Telugu (te), Thai (th), and Chinese (zh). For each source-language question-answer pair, the base model produces an English translation, yielding $(x_i^s, y_i^s, x_i^{en}, y_i^{en})$, and the resulting pool

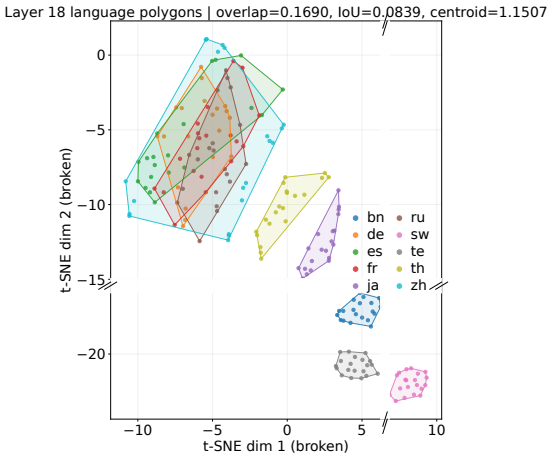

Figure 11: Routing representations at layer 18 on Qwen3-30B-A3B-Instruct. The transition toward resource-sensitive separation is already visible.

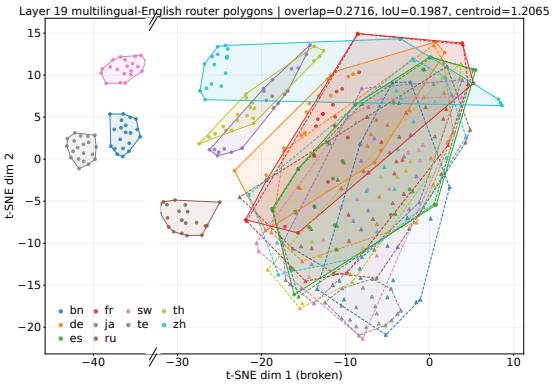

Figure 12: Paired source-English router geometry at selected key layer 19 on Qwen3-30B-A3B-Instruct. Same-color triangle markers denote projected samples after translating the source language into English.

contains 2,000 parallel examples. Router-overlap analyses use the non-English side, while paired source-English analyses use the source-English pairs. Router-alignment training uses only Bengali, Swahili, and Telugu, with 200 examples per language and 600 examples total, for training and frozen router-cache construction. Because we create no development or test split, all evaluation uses external benchmarks. Table 4 summarizes these data statistics.

Layer-selection variants. The Qwen experiments compare the base model with three router-only variants. KLRA uses the top three layers ranked by paired source-English router overlap. NKLRA and RLRA update the same number of router layers but use control layers, with NKLRA using layers 18, 29, and 32 and RLRA using layers 22, 25, and 35. These variants correspond to Table 1.

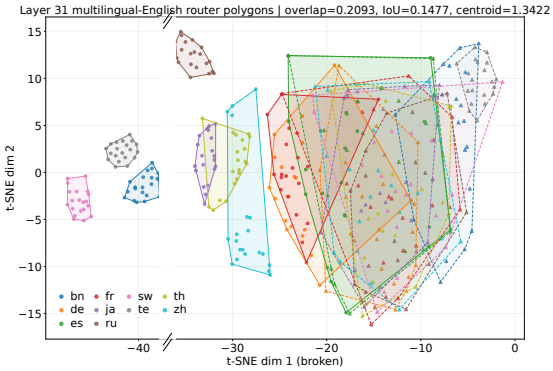

Figure 13: Paired source-English router geometry at selected key layer 31 on Qwen3-30B-A3B-Instruct. Same-color triangle markers denote projected samples after translating the source language into English.

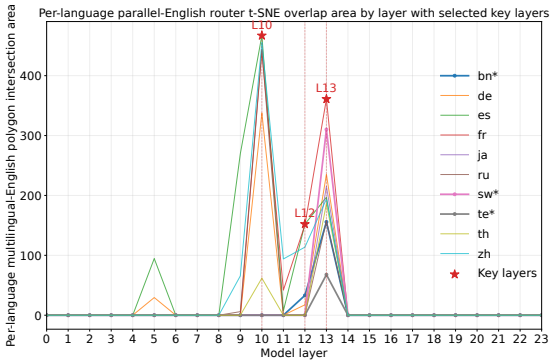

Figure 14: Layer-wise paired source-English overlap on GPT-OSS-20B. The top-ranked layers are used as key layers for router alignment.

Training setup and hyperparameters. All router-only Qwen variants use the same training setup to isolate the effect of layer selection. We train Qwen3-30B-A3B-Instruct, a 32B-parameter model with 128 experts, for 3 epochs with learning rate 5×10^{-5}, per-device batch size 2, gradient accumulation 64, cosine learning-rate schedule, warmup ratio 0.08, weight decay 0.01, gradient clipping 1.0, maximum sequence length 512, and random seed 42. Training uses 600 sentence-level alignment examples and runs on a single NVIDIA A800 GPU. Because only selected router gates are trainable and the dataset is small, each router-alignment run takes about 1.5 minutes at batch size 2. The router alignment objective uses teacher top-$k = 8$ expert mass, $\alpha = 0.15$, $\beta = 0.05$, $\gamma = 0$, and router-loss clipping at 5. We freeze the model except for the selected router gates, and we do not tune hyperparameters separately for KLRA, NKLRA, or RLRA.

Evaluation and reporting. We use the `lm-eval` library (Gao et al., 2023) for benchmark evaluation. MGSM is evaluated with two-shot prompting,

Lang.	Base	RLRA	NKLRA	KLRA
bn*	58.8%	55.2%	56.0%	**62.0%**
de	62.0%	60.0%	62.0%	**62.4%**
en	75.6%	74.8%	**76.0%**	75.6%
es	**62.0%**	60.0%	60.8%	61.6%
fr	**61.2%**	58.8%	**61.2%**	60.0%
ja	54.0%	54.0%	**56.8%**	54.0%
ru	**60.0%**	58.4%	**60.0%**	**60.0%**
sw*	55.2%	56.4%	56.8%	**58.0%**
te*	42.8%	46.0%	50.0%	**53.2%**
th	**70.0%**	64.4%	**70.0%**	**70.0%**
zh	**70.0%**	64.4%	65.6%	68.0%
AVG	61.1%	59.3%	61.4%	**62.3%**
Low-resource*	52.3%	52.5%	54.3%	**57.7%**

Table 3: Two-shot MGSM accuracy for GPT-OSS-20B. Asterisks mark training languages.

	bn	de	en	es	fr	ja	ru	sw	te	th	zh
Base	3.12	3.08	2.72	3.09	3.16	3.10	3.08	3.14	3.15	3.15	3.01
RLRA	3.15	3.10	2.75	3.12	3.16	3.12	3.07	3.16	3.14	3.16	3.03
NKLRA	3.08	3.07	2.71	3.13	3.14	3.10	3.09	3.13	3.14	3.16	3.01
KLRA	3.15	3.03	2.74	3.08	3.13	3.11	3.07	3.13	3.13	3.17	2.90

Figure 15: Per-language MGSM standard errors (%) for GPT-OSS-20B, reported by `lm-eval` from 250 test examples per language. Darker cells indicate larger standard errors.

greedy decoding, and maximum generation length 512. Each MGSM language has 250 test examples. MMLU is evaluated with five-shot prompting. For Qwen3-30B-A3B-Instruct, Base, KLRA, NKLRA, and RLRA denote the methods compared in Table 1. Figure 16 visualizes the per-language standard errors of MGSM exact-match accuracy across methods and languages, which range from 1.41% to 3.17%. In MGSM tables, AVG is the macro-average over all reported languages, and LOW-RESOURCE is the mean over the three training languages, bn, sw, and te.

Artifact licenses and terms. We use publicly available datasets, models, and evaluation tools, including mCoT-MATH, FLORES, MGSM, MMLU, Qwen3-30B-A3B-Instruct, GPT-OSS-20B, and `lm-eval`. We cite the original creators and use these artifacts for research purposes under their associated terms of use. The derived mCoT-MATH-Align data is used only for router-overlap analysis, frozen router-cache construction, and router-alignment training in this research context.

Data privacy and content. We do not collect private user data or personal records. The datasets used in this work are publicly available research datasets and benchmarks. Our constructed mCoT-MATH-Align data is derived from mCoT-MATH by sampling source-language math examples and generating English translations for research use.

Language	Constructed	Alignment train	Dev/Test
Bengali (bn)	200	200	0
German (de)	200	0	0
Spanish (es)	200	0	0
French (fr)	200	0	0
Japanese (ja)	200	0	0
Russian (ru)	200	0	0
Swahili (sw)	200	200	0
Telugu (te)	200	200	0
Thai (th)	200	0	0
Chinese (zh)	200	0	0
Total	2,000	600	0

Table 4: Statistics for mCoT-MATH-Align. The constructed examples are used for router-overlap analysis, while router-alignment training uses only Bengali, Swahili, and Telugu.

	bn	de	en	es	fr	ja	ru	sw	te	th	zh
Base	2.63	2.22	1.46	1.72	2.35	2.50	2.17	3.17	3.03	2.30	2.22
RLRA	2.64	2.17	1.41	1.90	2.35	2.46	2.20	3.17	3.05	2.25	2.17
NKLRA	2.55	2.28	1.68	1.76	2.32	2.57	2.12	3.17	3.00	2.25	2.28
KLRA	2.53	2.12	1.51	1.78	2.30	2.50	2.13	3.16	3.02	2.24	2.20

Figure 16: Per-language MGSM standard errors (%) for Qwen3-30B-A3B-Instruct, reported by `lm-eval` from 250 test examples per language. Darker cells indicate larger standard errors.

We inspected the constructed examples for obvious personally identifying information and offensive content during data construction, and did not use any fields intended to identify individuals. No additional anonymization was required.

G Use of AI Assistants

AI assistants were used in a limited manner to support code programming and routine editing tasks, such as drafting helper scripts, checking implementation details, and maintaining consistency in code and LaTeX files. The research questions, experimental design, analyses, reported results, and final manuscript decisions were made and verified by the authors.