Defending the Weak: Analyzing Language Differentiation in Multilingual MoEs through the Pretraining Lens

Anonymous ACL submission

Abstract

Mixture-of-Experts (MoE) architectures have become common for multilingual LLMs, but investigations into *how* routing and specialization emerge during training remain limited to English-centric models; the work on multilingual MoEs primarily analyzes, post-hoc, the patterns observed on final checkpoints. We fill the gap at the intersection: we analyze how routing and language specialization in terms of expert allocation develop *during* multilingual MoE pretraining. Across a controlled sweep of pretraining mixtures, we find that (i) the *expert allocation profiles* (EPs) of languages ossify quickly (with last layers converging the fastest), limiting early in training which experts a language can access; (ii) *language differentiation* in terms of expert allocation correlates with substantially better multilingual language-modeling performance: training with mixed multilingual batches encourages such language differentiation; (iii) by clustering languages based on their EPs, we find that stronger language differentiation yields typologically coherent clusters that protect the performance of their low-resource members under high capacity pressure from English-dominated corpora.

1 Introduction

Mixture-of-Experts (MoE) architectures have become a standard recipe for scaling large language models (LLMs), replacing each dense feed-forward block with a set of smaller expert networks and a learned router that activates only the top-k experts for each token (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022). More recently, the focus shifted towards multilinguality with several open massively multilingual MoEs released, including Qwen3-MoE (Yang et al., 2025), Qwen3.5-MoE (Qwen Team, 2026), and Gemma4-MoE (Gemma Team, 2026). Despite this, investigation of how routing and specialization in MoEs emerge *during* training remains limited to English-centric models and data (Xue et al., 2024; Muennighoff et al., 2025; Zoph et al., 2022; Qiu et al., 2025). In contrast, the body of work that analyzes multilingual MoEs probes for specialization patterns *post-hoc*, on final checkpoints (Bandarkar et al., 2025b; Chen et al., 2026; Fan et al., 2024; Zheng et al., 2026), revealing that routing in later layers reflects typological similarities between languages (Jiang et al., 2026; NLLB Team et al., 2024).

Our work targets the gap at the intersection of these two research streams, focusing on two orthogonal questions. ① We trace the development of languages' expert allocation profiles (EPs)—the subsets of actual experts that tokens of a language route to, and the relative frequency with which they do so—aiming to find out (a) how early these EPs of languages stabilize, and (b) when the typologically-driven routing structure, reported by related work (Jiang et al., 2026; NLLB Team et al., 2024), emerges. Analyzing English-centric models, Muennighoff et al. (2025) find that the routing ossifies rather early during English-centric pretraining (after ca. 20B tokens) and that per-token routing decisions at that point overlap with those of the final model checkpoint by 40% to 60%, depending on the layer (Xue et al., 2024; Xu et al., 2026). This high degree of early per-token routing saturation, however, does not necessarily mean that the broader EPs of individual languages also converge early in pretraining. ② We investigate how the structure of training batches, monolingual vs. multilingual, affects the language differentiation in multilingual MoE models. Yang et al. (2025) showed that monolingual batches limit domain differentiation in English MoEs. Analogously, we examine whether multilingual batches lead languages to develop more distinct and less overlapping EPs than monolingual ones. Crucially, we look at how the degree of language differentiation relates to multilingual language modeling performance.

Across a controlled sweep of pretraining runs,

we find that the EPs saturate already after 20B tokens of training, with late layers stabilizing first. This locks in language-specific routing long before the pretraining ends, suggesting that the remaining training budget is spent refining what individual experts learn, rather than changing which experts a language routes to. Furthermore, and in line with Yang et al. (2025), we find that mixed multilingual batches yield stronger language differentiation. Crucially, we show the link between stronger language differentiation and substantially better multilingual language-modeling performance. Moreover, we find that differentiation, especially in late layers, softens the impact of the size of languages' pretraining data on their respective performance.

Together, these findings warrant further investigation of the fast-ossifying (i.e., locked-in) routing structure. We thus provide a qualitative analysis of the late-layer language clusters based on the languages' EPs formed early in training. We observe that stronger language differentiation, as provided by multilingual batches, manifests in typologically more coherent clusters (compared to monolingual batches). This particularly benefits low-resource languages under rising English pressure. We thus believe that late-layer language differentiation protects low-resource languages by shielding them from competition for the finite capacity inside shared experts/parameters—a competition in which they do not fare well due to their small relative portions of the pretraining corpus (Conneau et al., 2020; Wang et al., 2020; Chang et al., 2024).

2 Preliminaries

MoE Block. We consider MoE models that replace the dense feed-forward block of a transformer with a sparse mixture of size k over m experts $\{E_i\}_{i=1}^{m}$. At any given layer, a token representation $x \in \mathbb{R}^d$ is fed through a learned router $g : \mathbb{R}^d \to \mathbb{R}^m$, followed by a softmax over all m experts that assigns a gate weight $w_i(x)$ to each. The top-k experts per token $\mathcal{S}_k(x)$ are selected. The MoE block output is computed as the gate-weighted sum over the selected experts: $\mathrm{MoE}(x) = \sum_{i \in \mathcal{S}_k(x)} w_i(x) \cdot E_i(x)$.

Load Balancing. Without an auxiliary signal, the router tends to concentrate updates on a small subset of experts, leading to a severe imbalance in expert utilization (Fedus et al., 2022). The Load-Balancing Loss (LBL) regularizes this behavior by penalizing the router whenever it routes a disproportionate share of tokens to a few experts. For a batch of tokens $\mathcal{B}$, let f_i be the fraction of tokens routed to expert E_i and P_i its mean routing probability. The LBL is then the sum of the products $f_i \cdot P_i$ across all m experts, scaled by m: $\mathrm{LBL} = m \sum_{i=1}^{m} f_i \cdot P_i$.

3 Analysis

3.1 Measuring Expert Allocation

Expert Allocation Profile (EP). We define the EP of a language L at layer ℓ, $\mathbf{f}_\ell(L) \in \mathbb{R}^m$, as a vector of relative expert activation frequencies; $f_{i,\ell}(L)$ is the share of L's tokens routed to E_i.

Expert Profile Alignment (EPA). We quantify the alignment between languages' EPs at layer ℓ via Pearson correlation, i.e., $\mathrm{EPA}_\ell(L_1, L_2) = \rho(\mathbf{f}_\ell(L_1), \mathbf{f}_\ell(L_2))$. We use EPA to measure (i) routing convergence across training checkpoints and (ii) language differentiation across languages at a fixed checkpoint.

3.2 Experimental Details

Model. We train a 12-layer decoder-only MoE Transformer with 64 experts per layer and top-8 routing (1B total / 378M active parameters). Our design follows the fine-grained expert recipe of recent decoder-only MoE (Dai et al., 2024; Krajewski et al., 2024; Muennighoff et al., 2025).

Training and Evaluation. We source English training data from FineWeb (Penedo et al., 2024) and the multilingual data from FineWeb-2 (Penedo et al., 2025), covering a diverse set of 30 languages. We run controlled experiments: we fix the training budget to 100B tokens and vary the proportion of English data from 20% (EN20) to 80% (EN80). Within the multilingual portion, per-language sampling weights are proportional to FineWeb-2 data sizes with temperature upsampling at τ=3.3. The overall setup follows Foroutan et al. (2025); detailed information about the exact data mixture is provided in Appendix A. We evaluate the language modeling performance on FLORES+ (NLLB Team et al., 2024) due to its fully multi-parallel nature, and report bits per byte (BPB), i.e., a byte-normalized negative log-likelihood (per-language) as a metric. Since language-modeling performance is not directly comparable between languages in absolute terms, we report either (1) the average across languages or (2) the relative change of BPB for each language, across English proportions. We compute EPAs across the non-English languages.

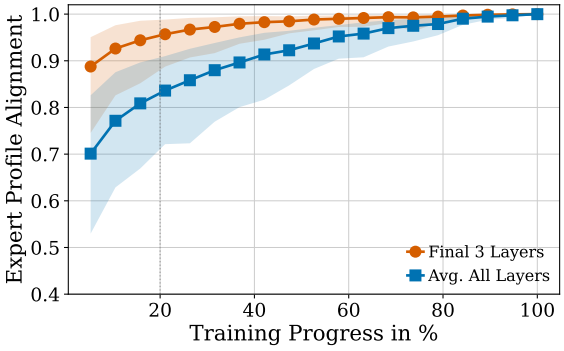

Figure 1: Average expert profile convergence across training measured by EPA to the final checkpoint (EN40). Shaded area: range between the minimal and maximal per-language EPA.

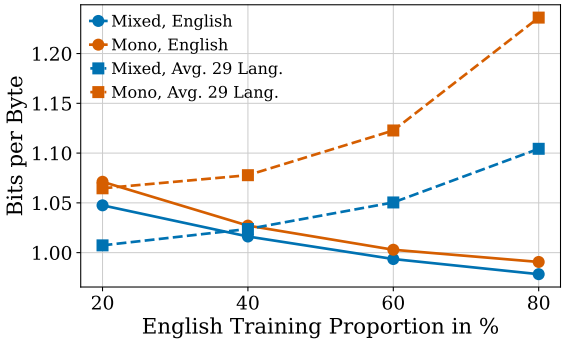

Figure 2: Language modeling performance in bits per byte (BPB; lower is better) across different shares of English pretraining data for MONO and MIXED batches.

3.3 Routing Convergence

We first explore EP convergence in multilingual MoE training, measuring the per-language EPA between each checkpoint t and the final one.

Late Layer Convergence. Figure 1 averages the resulting EPAs across non-English languages and reveals a twofold pattern. Averaged across all layers, the routing structure shows early ossification: we observe strong convergence already after 20B tokens (20% of training) with an average EPA of 0.84. But the routing in the last three layers ossifies even faster, reaching an average EPA of 0.96 at 20B tokens. The result is surprisingly consistent across languages, with per-language EPAs ranging between 0.89 and 0.99. Since late layers comprise the language-specific output mapping (Zheng et al., 2025; Bandarkar et al., 2025b,a; Tang et al., 2024), locking in their EPs early in training fixes which experts a language will ever access. This leaves only the experts' internal representations to be refined by the bulk of the (remaining) pretraining.

3.4 Language Differentiation

We next explore how batch composition shapes language differentiation in multilingual MoE training and how differentiation drives language modeling performance. The LBL directly penalizes uneven expert allocation within a batch. The language mix of each batch thus directly affects the set of experts a language is encouraged to use (Qiu et al., 2025).

Setup. We compare two training regimes that differ only in micro-batch composition. In the monolingual (MONO) regime, each micro batch contains instances from a single language: batch languages are sampled proportionally to the pretraining corpus portions. In the MIXED regime, each micro batch contains instances from multiple languages, with instance languages sampled independently from the same distribution of the pretraining corpus.

Performance Gap. Figure 2 compares the language modeling performance for training with MONO vs. MIXED batches, across training mixes with increasing English share. Mixed-batch training outperforms monolingual across the board. For English, more in-language data expectedly improves performance, but the gaps are rather small both between batch compositions and pretraining mixes. Averaged across all other languages, the trends are more prominent: MIXED not only outperforms MONO for each pretraining mix; the gap also widens under increasing English pressure (i.e., larger English share in pretraining data). At EN60, MONO (1.12) is 6.7% worse than MIXED (1.05); this almost doubles to 11.9% at EN80.

Differentiation Gap. We theorize that the observed gap widening between MIXED and MONO for languages other than English stems from better language differentiation in terms of EPs under MIXED, which protects languages under increasing English pressure. To measure the language differentiation, we compute the EPA averaged across all non-English language pairs. A lower EPA indicates a higher degree of language differentiation. At high English pressure (EN80), we observe lower EPA scores (i.e., larger language differentiation) with MIXED than with MONO training. But the differences vary starkly across layers: from modest 0.45 EPA for MIXED vs. 0.54 for MONO in middle layers, to pronounced 0.35 (MIXED) vs. 0.72 (MONO) in late layers. We provide EPA scores for all training mixes in Table 5 in Appendix D.

Figure 3 shows relative performance drops under increased English pressure (EN20 vs. EN80),

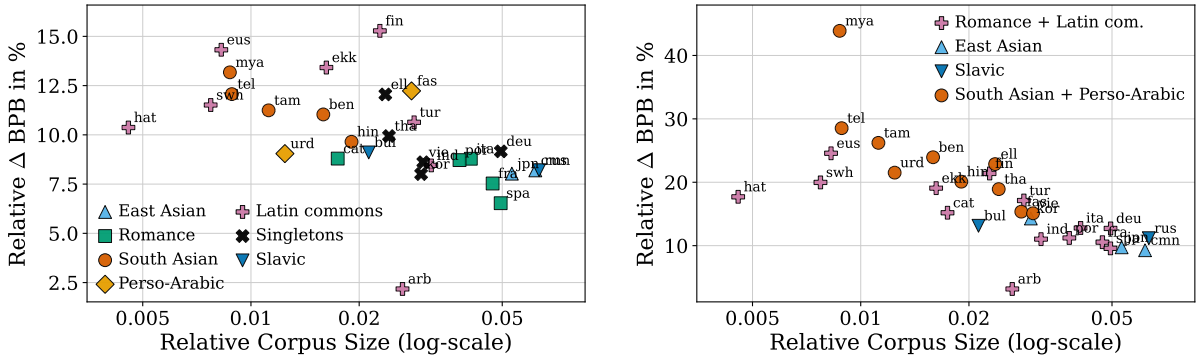

Figure 3: Relative performance decrease (BPB increase) of languages under increased English pressure (EN20 vs. EN80) against their relative size in the training corpus, for MIXED (*left*) and MONO (*right*). Colors denote language clusters induced from pairwise EPA scores, computed from concatenated EPs of the last three layers.

for both MIXED (left) and MONO (right). MONO makes a language's performance more determined by its relative corpus size than MIXED (see Appendix D). For most languages, MIXED leads to more graceful performance drops under increased English pressure than MONO, e.g., +12.1% (MIXED) vs. +28.5% (MONO) for Telugu.

3.5 Expert Competition

Seeing that late-layer routing converges early and shapes language modeling performance, we next analyze the emerging language clusters. We run hierarchical clustering based on the pairwise EPA scores over the last three layers, computed under EN20 for MIXED and MONO at 20% of training. We pick the setting with the lowest EN pressure, as it allows the other languages to route most freely. We provide full details on the clustering including the dendrograms in Appendix E.

Cluster Analysis Results. Looking at Figure 3, we see that the clusters (depicted as colors) follow typological similarity and script overlap, but to different degrees for the two batching strategies. While MIXED at EN20 induces fine-grained and coherent clusters (e.g., Romance, East Asian, etc.), MONO induces less coherent, coarser-grained supersets, because it drives a more uniform allocation of experts across languages.

Comparing the relative change in performance between MIXED and MONO under rising EN pressure (at EN20 vs. EN80), we find that the performance gap (in pp) between high- and low-resource languages *within* clusters widens under MONO. Where clusters remain coherent, the gap stays small: in the Slavic cluster, higher-resource Russian goes from 8.2% (MIXED) to 11.2% (MONO, +3.0pp) and lower-resource Bulgarian from 9.1% (MIXED) to 13.1% (MONO, +4.0pp), differing only by 1pp. However, where clusters under MONO merge and thus become less coherent, the gap grows: in the Latin commons cluster, higher-resource Indonesian (+2.5pp) vs. lower-resource Basque (+10.3pp) and Swahili (+8.5pp); or in the merged Perso-Arabic-South Asian cluster: higher-resource Persian (+3.2pp) vs. lower-resource Urdu (+12.5pp). Overall, we observe that low-resource languages in typologically incoherent clusters are most affected (see Table 7 in App. E for details).

Combined, our results show that MIXED strengthens language differentiation in EPs, which, especially in the final layers, shields the performance of low-resource languages. This is particularly important when training under high English pressure, which is common for modern LLMs. Without late-layer language differentiation w.r.t. expert allocation (like in MONO), the MoE falls back to language competition *within* experts, much like in dense LLMs, weakening its modular benefits.

4 Conclusion

We analyzed the dynamics of language specialization in terms of expert allocation profiles in multilingual MoE pretraining. We find that: (1) the languages' expert allocation profiles lock in early in training, with late layers ossifying the fastest; (2) increased language specialization w.r.t. expert allocation in late layers (a) correlates with better multilingual language-modeling performance and (b) reduces the impact of a language's share of the pretraining corpus on performance. Our analysis of the language clusters revealed that stronger language differentiation yields typologically coherent clusters that protect their low-resource members under strong capacity pressure from English.

5 Limitations

Model Scale. We evaluate a single model size. Our analysis focuses on EPs as shaped by the LBL, and we argue that, under the same number of total experts and the same batch composition, the language differentiation patterns we observe likely remain similar at larger scales. We acknowledge that scaling the model under these constraints (i.e., same total number of experts) grows per-expert capacity and would therefore dampen the within-expert competition we observe; however, relying on representational capacity to absorb suboptimal routing decisions remains a suboptimal use of the modularization advantage that MoEs are designed to provide.

MoE Module Design. We evaluate a single MoE architecture (64 experts, top-8). This choice reflects the design that recent open MoE work has converged on. We emphasize that the purpose of our work is not a broad pretraining sweep over hyperparameters but rather an analysis of how language differentiation emerges—an analysis that can only be conducted through the lens of pretraining. We argue that the clustering diagnostic we contribute can serve as an additional signal in such a sweep: for instance, in determining whether shared experts or more fine-grained expert designs better support language differentiation.

Number of Languages. We train on 30 languages. While a larger language set would strengthen the statistical power of our analysis, we note that 30 is already a comparatively small set for a multilingual MoE. Even at this scale we observe substantial expert competition, suggesting that the effect will only grow with broader language coverage.

References

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2025a. Layer swapping for zero-shot cross-lingual transfer in large language models. In *International Conference on Learning Representations*, volume 2025, pages 51315–51338.

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

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Yuxin Chen, Zhengzhou Cai, Xiangtian Ji, Weixiang Zhao, An Zhang, Xiang Wang, and Tat-Seng Chua. 2026. Understanding multilingualism in mixture-of-experts llms: Routing mechanism, expert specialization, and layerwise steering. *arXiv preprint arXiv:2601.14050*.

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

Damai Dai, Chengqi Deng, Chenggang Zhao, R.x. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.k. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297, Bangkok, Thailand. Association for Computational Linguistics.

Dongyang Fan, Bettina Messmer, and Martin Jaggi. 2024. Towards an empirical understanding of moe design choices. *arXiv preprint arXiv:2402.13089*.

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.

Negar Foroutan, Paul Teiletche, Ayush Kumar Tarun, and Antoine Bosselut. 2025. Revisiting multilingual data mixtures in language model pretraining. *arXiv preprint arXiv:2510.25947*.

Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. 2023. Megablocks: Efficient sparse training with mixture-of-experts. *Proceedings of Machine Learning and Systems*, 5:288–304.

Gemma Team. 2026. Gemma 4: Open models built from gemini 3 research and technology. Accessed: 2026-05-13.

Fan Jiang, Yu Zhao, Chenyang Lyu, Tianqi Shi, Yichao Du, Feihu Jiang, Longyue Wang, and Weihua Luo. 2026. Marco-moe: Open multilingual mixture-of-expert language models with efficient upcycling. *arXiv preprint arXiv:2604.25578*.

Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, and 1 others. 2024. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, and 1 others. 2025. Olmoe: Open mixture-of-experts language models. In *International Conference on Learning Representations*, volume 2025, pages 62061–62121.

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, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Demons in the detail: On implementing load balancing loss for training specialized mixture-of-expert models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5005–5018, Vienna, Austria. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

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*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

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

Yuqi Xu, Rizhen Hu, Zihan Liu, Mou Sun, and Kun Yuan. 2026. Grouter: Decoupling routing from representation for accelerated moe training. *arXiv preprint arXiv:2603.06626*.

Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. 2024. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, and Wang Benyou. 2025. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts. In *International Conference on Learning Representations*, volume 2025, pages 7151–7179.

Kening Zheng, Wei-Chieh Huang, Jiahao Huo, Zhonghao Li, Henry Peng Zou, Yibo Yan, Xin Zou, Jungang Li, Junzhuo Li, Hanrong Zhang, and 1 others. 2026. Unveiling language routing isolation in multilingual moe models for interpretable subnetwork adaptation. *arXiv preprint arXiv:2604.03592*.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Experimental Details

We report the architecture and training details in Tables 1 and 2. Our model is implemented in a custom fork of the TinyLlama repository (Apache-2.0) and LLaMmlein repository (Apache-2.0). We follow their architecture closely (Zhang et al., 2024;

Property	Value
Architecture	
Total parameters	1B
Active parameters	378M
Hidden Dim	768
Layers	12
Attention heads	16
Attention variant	GQA
KV Heads	4
Query Heads	16
Positional encoding	RoPE
LayerNorm	Pre-Norm
Type	FusedRMSNorm
ε	10^{-6}
Bias	False
Tied embeddings	False
Tokenizer	Qwen3
Vocabulary size	151 669
MoE	
Number of experts	64
Top-k	8
Activation / MLP	SwiGLU
Expert FFN dim	528
Shared expert	False

Table 1: Architecture hyperparameters.

Property	Value
Training	
Training budget	100B tokens
Optimization steps	190 735
Global batch size	512
Micro batch size	32
Sequence length	1024
Optimizer	AdamW
β_1	0.9
β_2	0.95
Weight decay	0.1
Dropout	0.0
Peak LR	6×10^{-4}
LR schedule	
Warmup steps	8 000
Constant phase	until 80% of training
Decay target	4×10^{-6}
Gradient clipping	1.0
Precision	bf16-mixed
Distributed training	FSDP
Strategy	`SHARD_GRAD_OP`
Auxiliary losses	
Load-balancing weight	10^{-2}
Router z-loss weight	10^{-3}

Table 2: Training hyperparameters.

Pfister et al., 2025). We implement the MoE module using dropless MoE (Gale et al., 2023) from the megablocks repository (Apache-2.0). We report the relative corpus size of each language in Table 3. We use the devtest split of FLORES+ (NLLB Team et al., 2024) (CC-BY-SA-4.0) with 1012 sentences per language for both BPB computation and our routing analyses. We run all training and evaluation on 4 nodes with 4 H100 GPUs each (16 GPUs total). We estimate the total training and evaluation cost at approximately 10 000 GPU hours.

B Load Balancing and Language Differentiation

The LBL penalizes uneven expert utilization within each micro batch, which raises the question of whether MONO batches trivially prevent language differentiation. The LBL is a soft constraint: weighted at only 10^{-2} relative to the main language modeling loss, it discourages but does not prevent uneven allocation. The router can still specialize if doing so improves the language modeling loss, and our analysis (§3.4 and §3.5) confirms that MONO still exhibits language differentiation, particularly in the late layers—it is simply weaker than under MIXED. We point out that it is specifically the degree of language differentiation and its impact on language modeling performance that we study.

C Routing Convergence Details

We provide further details for §3.3. We measure routing convergence by computing the per-language EPA between each training checkpoint and the final one, separately for each MoE layer. We average EPA scores within layer groups (all layers, or the last three) and across the 29 non-English languages; the shaded region in Figure 1 shows the per-language range. Additionally, we report the per-layer scores at 20% of training for both MIXED and MONO across all English proportions in Table 4. The pattern is consistent: late layers saturate before early layers, and the overall convergence across the last three layers is strong for both MIXED and MONO. We provide further evidence for the converging routing in Figure 4 and Figure 7. The figures show the clustering dendrograms for EN20 (MIXED and MONO) at 20% and 100% of training, respectively. For MIXED, the cluster assignments largely agree between 20% of training and the final checkpoint, and, for MONO, we observe marginal differences in the assignment of Singletons (Korean, Thai, and Vietnamese).

D Language Differentiation Details

We provide further details for §3.4: (i) the layer analysis of language differentiation and (ii) the analysis of how language differentiation mediates

Language	Code	EN20	EN40	EN60	EN80
English	eng	0.2000	0.4000	0.6000	0.8000
Russian	rus	0.0633	0.0475	0.0317	0.0158
Chinese	cmn	0.0618	0.0463	0.0309	0.0154
Japanese	jpn	0.0532	0.0399	0.0266	0.0133
German	deu	0.0496	0.0372	0.0248	0.0124
Spanish	spa	0.0495	0.0371	0.0248	0.0124
French	fra	0.0470	0.0353	0.0235	0.0118
Italian	ita	0.0409	0.0307	0.0204	0.0102
Portuguese	por	0.0380	0.0285	0.0190	0.0095
Indonesian	ind	0.0317	0.0238	0.0159	0.0079
Vietnamese	vie	0.0301	0.0226	0.0151	0.0075
Korean	kor	0.0297	0.0223	0.0149	0.0074
Turkish	tur	0.0284	0.0213	0.0142	0.0071
Persian	fas	0.0280	0.0210	0.0140	0.0070
Arabic	arb	0.0264	0.0198	0.0132	0.0066
Thai	tha	0.0242	0.0182	0.0121	0.0061
Greek	ell	0.0236	0.0177	0.0118	0.0059
Finnish	fin	0.0228	0.0171	0.0114	0.0057
Bulgarian	bul	0.0213	0.0159	0.0106	0.0053
Hindi	hin	0.0190	0.0143	0.0095	0.0048
Catalan	cat	0.0174	0.0131	0.0087	0.0044
Estonian	ekk	0.0162	0.0122	0.0081	0.0041
Bengali	ben	0.0159	0.0119	0.0079	0.0040
Urdu	urd	0.0124	0.0093	0.0062	0.0031
Tamil	tam	0.0112	0.0084	0.0056	0.0028
Telugu	tel	0.0089	0.0066	0.0044	0.0022
Burmese	mya	0.0087	0.0066	0.0044	0.0022
Basque	eus	0.0083	0.0062	0.0041	0.0021
Swahili	swh	0.0077	0.0058	0.0039	0.0019
Haitian Creole	hat	0.0046	0.0034	0.0023	0.0011

Table 3: Per-language relative corpus sizes across the four English-share configurations with 100B tokens total.

the relationship between corpus size and relative changes in BPB.

Average Expert Profile Alignment. Table 5 reports the average EPA across non-English language pairs at four layer granularities: all layers, early (L0–L2), mid (L3–L8), and late (L9–L11). For each layer, we compute the pairwise EPA of all non-English pairs. Next, we average the EPAs to get a single alignment score per layer. Finally, we average over the respective layer granularity. The layer-wise pattern is consistent with related work (Bandarkar et al., 2025b; Chen et al., 2026): mid layers show the strongest alignment, as they perform the most language-agnostic processing. Early layers follow, primarily routing based on token id (Muennighoff et al., 2025; Zoph et al., 2022). Finally, late layers exhibit the weakest alignment, reflecting the strongest language differentiation. Both MONO and MIXED show increasing EPA with rising English share in the pretraining mixture, but the gap between the two is most pronounced in the late layers. Since lower EPA indicates stronger language differentiation, the late-layer gap between MIXED and MONO suggests that late-layer routing differentiation is a key factor in the performance difference between the two regimes.

	MIXED				MONO			
Layer	EN20	EN40	EN60	EN80	EN20	EN40	EN60	EN80
L0	0.500	0.459	0.495	0.456	0.376	0.483	0.427	0.384
L1	0.733	0.728	0.712	0.765	0.179	0.234	0.291	0.272
L2	0.697	0.745	0.776	0.795	0.198	0.288	0.587	0.619
L3	0.841	0.852	0.835	0.788	0.393	0.599	0.743	0.657
L4	0.820	0.834	0.847	0.827	0.548	0.596	0.592	0.644
L5	0.888	0.879	0.848	0.826	0.618	0.675	0.658	0.710
L6	0.872	0.868	0.861	0.837	0.612	0.683	0.779	0.644
L7	0.850	0.880	0.921	0.912	0.749	0.801	0.799	0.787
L8	0.898	0.918	0.916	0.941	0.834	0.832	0.772	0.852
L9	0.934	0.932	0.937	0.936	0.796	0.792	0.789	0.831
L10	0.963	0.962	0.961	0.956	0.817	0.860	0.843	0.781
L11	0.980	0.977	0.965	0.953	0.844	0.859	0.829	0.752
Avg	0.831	0.836	0.840	0.833	0.580	0.642	0.676	0.661
Last 3	0.959	0.957	0.955	0.948	0.819	0.837	0.820	0.788

Table 4: Per-layer EPA between 20% of training and the final checkpoint, averaged across 29 non-English languages for MIXED and MONO.

Config	All	Early	Middle	Late
		MIXED		
EN20	0.103	0.106	0.127	0.051
EN40	0.149	0.159	0.180	0.076
EN60	0.224	0.202	0.273	0.149
EN80	0.378	0.271	0.447	0.346
		MONO		
EN20	0.397	0.264	0.507	0.309
EN40	0.370	0.131	0.514	0.320
EN60	0.497	0.375	0.545	0.525
EN80	0.588	0.557	0.538	0.721

Table 5: Average expert profile alignment (EPA) across non-English languages over all, early (L0-L2), mid (L3–L8), and late (L9–L11) layers.

Relative Corpus Size vs relative ΔBPB. Table 6 reports the Pearson and Spearman correlation between a language's relative corpus size (log-scale) and its relative ΔBPB. We compute the ΔBPB as the change between EN20 and each of EN40, EN60, and EN80 across the 29 non-English languages. MIXED consistently shows less weight-dependence than MONO across all transitions. The gap is largest at EN20 → EN60: high-resource languages are still relatively unaffected by English in both regimes, while low-resource languages benefit from stronger language differentiation under MIXED. By EN20 → EN80, even high-resource languages begin to degrade, which narrows the gap between the two regimes.

Transition	MIXED		MONO	
	Pears.	Spear.	Pears.	Spear.
EN20 → EN40	−0.41	−0.55	−0.53	−0.64
EN20 → EN60	−0.49	−0.64	−0.71	−0.78
EN20 → EN80	−0.53	−0.70	−0.70	−0.82

Table 6: Pearson and Spearman correlation between a language's relative corpus size (log-scale) and its relative ΔBPB across the 29 non-English languages, under MIXED and MONO batch training.

E Expert Competition Details

Detailed Results. Table 7 shows the detailed scores that are used for Figure 3 and are discussed in §3.5.

Cluster Analysis Setup. We provide further details on the clustering presented in §3.5. We run agglomerative hierarchical clustering with average linkage over the 29 non-English languages. Each language is represented by its layer-ℓ EP for $\ell \in \{9, 10, 11\}$, concatenated and renormalized into a 192-dimensional probability vector. We use the EPA distance $1 - EPA$ as the metric, and cut the dendrogram at $d = 0.7$: i.e., we are merging clusters with an average pairwise EPA larger than 0.3. Figure 4 shows the dendrograms extracted at 20% of training.

Robustness of Cluster Analysis: Threshold. We point out that there are two ways for a fair comparison between the clustering induced by MIXED and MONO: (i) fixing the threshold d for both batching regimes and analyzing the emerging clusters (i.e., as presented in the main body) or (ii) fixing the number of clusters and analyzing the resulting threshold d.

We present the dendrograms of the former in Figure 4 and the latter in Figure 5. We find that both views largely align. The figures show the same typological patterns for both MONO and MIXED: i.e., MONO still shows similar clusters as observed under MIXED. However, under MONO the cluster boundaries erode, making coherent clusters more similar to incoherent neighbors than under MIXED: e.g., South Asian and Perso-Arabic show an average EPA distance of 0.79 under MIXED, but only 0.55 under MONO (see Figure 5). Consequently, we see larger clusters for MONO under a fixed threshold d (see Figure 4).

Robustness of Cluster Analysis: English Proportions. We observe a similar erosion of cluster boundaries in the dendrogram at EN80 (see Figure 6). The pattern is the same: increasing the English share during training compresses a range of linguistically diverse languages into ever more aligned EPs (i.e., smaller EPA distance). Similarly, the effect is stronger for MONO than for MIXED.

We conclude that both views support the same mechanism: late-layer routing is driven by typological similarity, but the eroding cluster boundaries result in less language differentiation for MONO compared to MIXED. The degree of language differentiation in EPs is thus closely linked to the protection of low-resource languages under capacity pressure from English.

F Expert Competition for EN20 vs. EN60

We provide further results for the robustness of our findings in §3.5. Figure 8 replicates the results from Figure 3 comparing EN20 vs. EN60. The overall degradation pattern mirrors the pattern in the main body with the absolute differences in pp being smaller. We show the detailed scores in Table 8 and corresponding dendrogram in Figure 9. Revisiting the same examples from the main body: in the Slavic cluster, higher-resource Russian (−0.4pp) and lower-resource Bulgarian (+0.4pp) differ by less than 1pp. In the Latin commons cluster, higher-resource Indonesian (+0.5pp) degrades less than lower-resource Basque (+0.9pp) and Swahili (+2.7pp). The same asymmetry appears in the Perso-Arabic cluster: higher-resource Persian (+0.9pp) vs. lower-resource Urdu (+3.1pp).

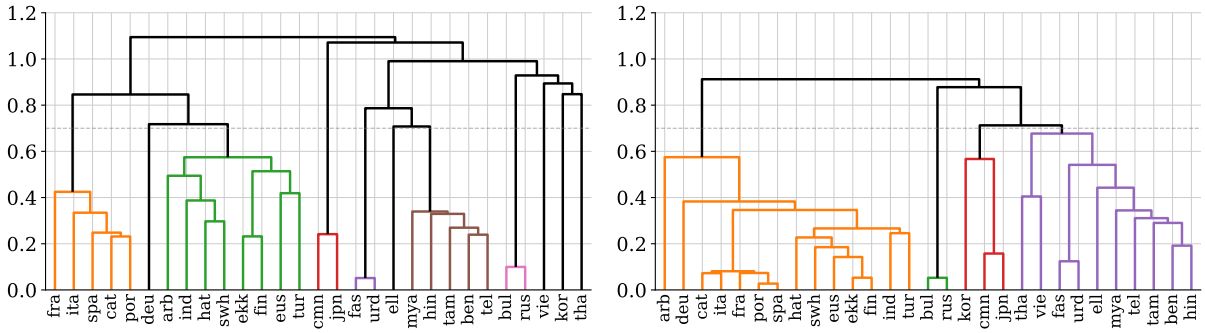

Figure 4: Clustering dendrograms of 29 non-English languages at EN20, computed at 20% of training (average linkage, $1-EPA$ distance) with threshold $d = 0.7$. Left: MIXED. Right: MONO.

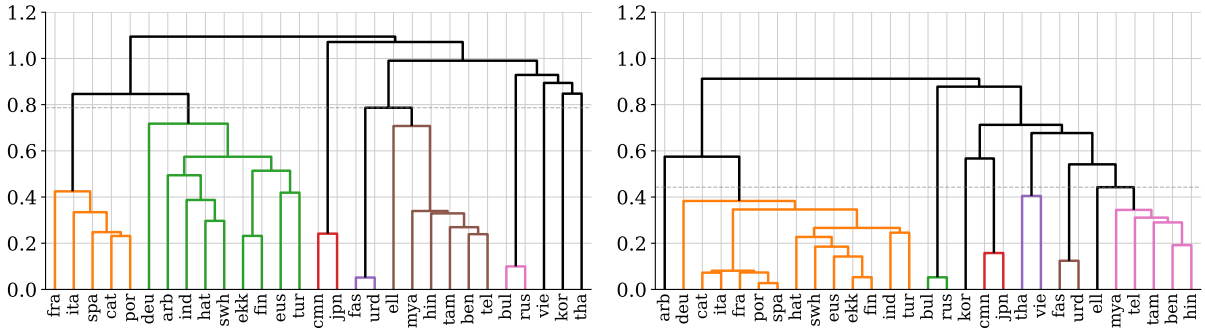

Figure 5: Clustering dendrograms of 29 non-English languages at EN20, computed at 20% of training (average linkage, $1-EPA$ distance) with the number of clusters fixed to 9. Left: MIXED. Right: MONO.

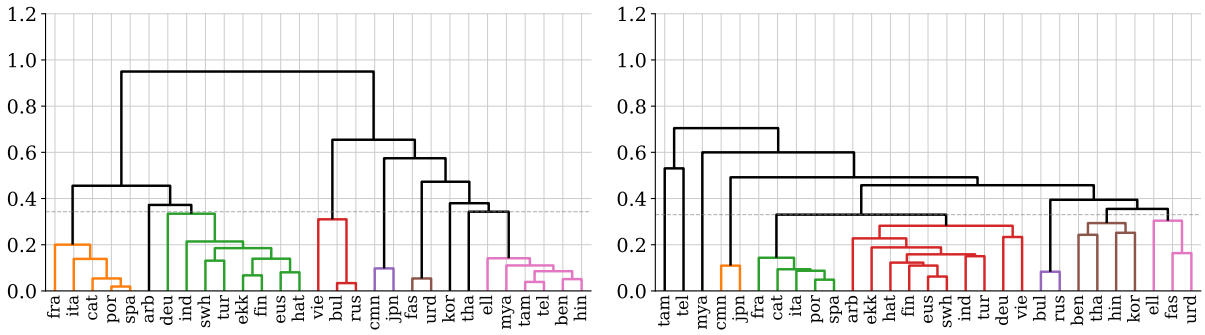

Figure 6: Clustering dendrograms at EN80 of 29 non-English languages, computed at 20% of training (average linkage, $1-EPA$ distance) with the number of clusters fixed to 9. Left: MIXED. Right: MONO.

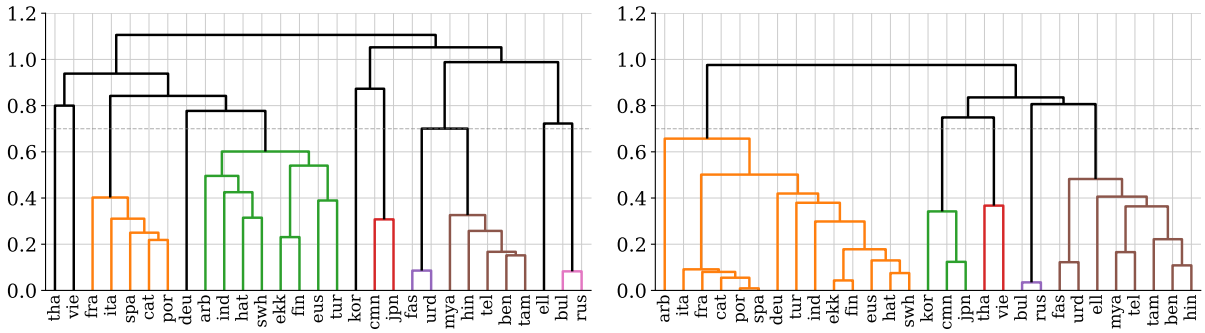

Figure 7: Clustering dendrograms at EN20 of 29 non-English languages, computed at the final checkpoint (average linkage, $1-EPA$ distance) with threshold $d = 0.7$. Left: MIXED. Right: MONO.

Lang	Wgt.	EN20			EN80			EN20 →EN80 relative		
		Mix	Mono	|Δ|	Mix	Mono	|Δ|	Mix Δ%	Mono Δ%	|Δ|
East Asian										
cmn	0.0618	1.424	1.464	+0.040	1.541	1.600	+0.059	+8.2	+9.3	+1.1
jpn	0.0532	0.985	1.020	+0.035	1.065	1.120	+0.055	+8.0	+9.7	+1.7
Slavic										
rus	0.0633	0.620	0.646	+0.026	0.671	0.718	+0.047	+8.2	+11.2	+3.0
bul	0.0213	0.663	0.696	+0.033	0.724	0.788	+0.064	+9.1	+13.1	+4.0
Romance										
spa	0.0495	1.011	1.045	+0.033	1.077	1.145	+0.067	+6.5	+9.6	+3.0
fra	0.0470	0.947	0.982	+0.034	1.018	1.085	+0.067	+7.5	+10.6	+3.0
ita	0.0409	1.018	1.058	+0.040	1.107	1.193	+0.086	+8.8	+12.8	+4.0
por	0.0380	1.061	1.096	+0.035	1.153	1.218	+0.065	+8.7	+11.2	+2.5
cat	0.0174	1.137	1.175	+0.039	1.237	1.354	+0.117	+8.8	+15.2	+6.4
Singletons										
deu	0.0496	1.044	1.090	+0.046	1.140	1.229	+0.089	+9.2	+12.7	+3.6
vie	0.0301	0.904	0.940	+0.036	0.982	1.083	+0.101	+8.6	+15.1	+6.5
kor	0.0297	1.139	1.190	+0.051	1.230	1.360	+0.130	+8.0	+14.3	+6.3
tha	0.0242	0.489	0.516	+0.026	0.538	0.613	+0.075	+9.9	+18.9	+9.0
ell	0.0236	0.648	0.701	+0.053	0.726	0.861	+0.135	+12.1	+22.9	+10.8
Perso-Arabic										
fas	0.0280	1.024	1.083	+0.058	1.150	1.249	+0.099	+12.2	+15.4	+3.1
urd	0.0124	0.920	0.986	+0.066	1.003	1.199	+0.195	+9.0	+21.5	+12.5
Latin commons										
ind	0.0317	1.141	1.189	+0.048	1.237	1.320	+0.083	+8.5	+11.0	+2.6
tur	0.0284	1.185	1.237	+0.053	1.311	1.449	+0.139	+10.6	+17.1	+6.5
arb	0.0264	2.002	2.070	+0.068	2.046	2.136	+0.090	+2.2	+3.2	+1.0
fin	0.0228	1.205	1.295	+0.090	1.389	1.572	+0.183	+15.3	+21.4	+6.1
ekk	0.0162	1.353	1.454	+0.100	1.535	1.731	+0.196	+13.4	+19.1	+5.7
eus	0.0083	1.335	1.448	+0.112	1.526	1.803	+0.277	+14.3	+24.6	+10.3
swh	0.0077	1.376	1.475	+0.099	1.535	1.769	+0.235	+11.5	+20.0	+8.5
hat	0.0046	1.683	1.805	+0.123	1.857	2.125	+0.268	+10.4	+17.7	+7.3
South Asian										
hin	0.0190	0.590	0.633	+0.042	0.647	0.760	+0.112	+9.6	+20.1	+10.4
ben	0.0159	0.609	0.659	+0.050	0.676	0.817	+0.141	+11.0	+23.9	+12.9
tam	0.0112	0.512	0.569	+0.057	0.570	0.718	+0.149	+11.2	+26.2	+15.0
tel	0.0089	0.656	0.745	+0.089	0.735	0.958	+0.223	+12.1	+28.5	+16.5
mya	0.0087	0.530	0.609	+0.079	0.600	0.876	+0.276	+13.2	+43.9	+30.7

Table 7: Per-language absolute BPB at EN20 and EN80 and relative ΔBPB (EN20 → EN80) under MIXED and MONO batch training. Each group includes a diff column: |Δ| for absolute BPB and pp for relative ΔBPB. Grouped by MIXED EN20 cluster.

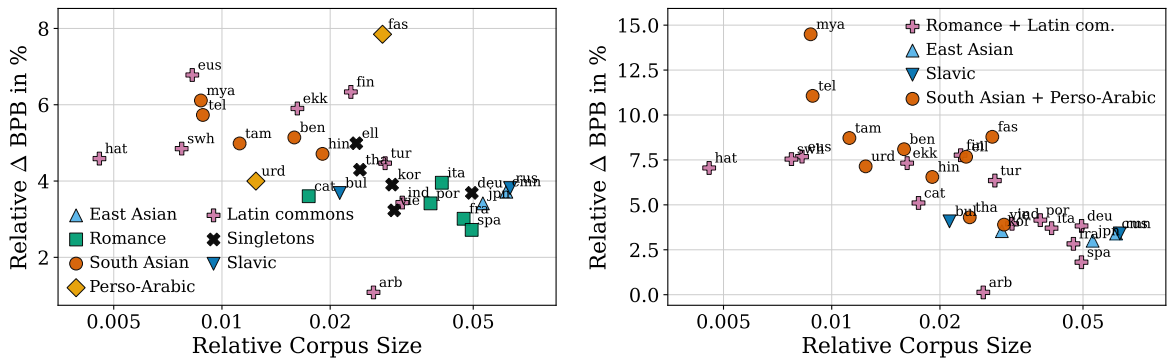

Figure 8: Relative performance decrease (BPB increase) of languages under increased English pressure (EN20 vs. EN60) against their relative corpus size (log-scale), for MIXED (*left*) and MONO (*right*). Colors denote language clusters induced from pairwise EPA scores, computed from concatenated EPs of the last three layers.

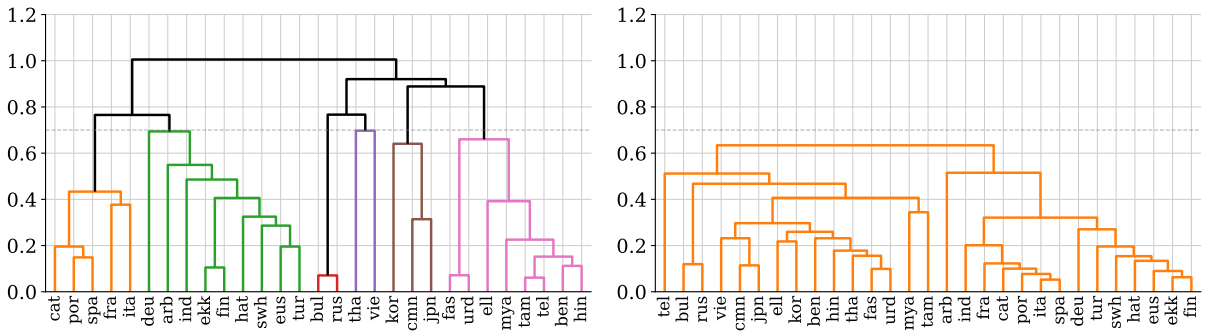

Figure 9: Clustering dendrograms at EN60 of 29 non-English languages, computed at 20% of training (average linkage, $1-EPA$ distance) with threshold $d = 0.7$. Left: MIXED. Right: MONO.

		EN20			EN60			EN20 →EN60 relative		
Lang	Wgt.	Mix	Mono	$\lvert\Delta\rvert$	Mix	Mono	$\lvert\Delta\rvert$	Mix Δ%	Mono Δ%	$\lvert\Delta\rvert$
East Asian										
cmn	0.0618	1.424	1.464	+0.040	1.477	1.513	+0.036	+3.7	+3.4	−0.3
jpn	0.0532	0.985	1.020	+0.035	1.019	1.051	+0.032	+3.4	+3.0	−0.4
Slavic										
rus	0.0633	0.620	0.646	+0.026	0.644	0.668	+0.024	+3.8	+3.4	−0.4
bul	0.0213	0.663	0.696	+0.033	0.688	0.725	+0.037	+3.7	+4.1	+0.4
Romance										
spa	0.0495	1.011	1.045	+0.033	1.039	1.064	+0.025	+2.7	+1.8	−0.9
fra	0.0470	0.947	0.982	+0.034	0.976	1.009	+0.034	+3.0	+2.8	−0.2
ita	0.0409	1.018	1.058	+0.040	1.058	1.097	+0.039	+4.0	+3.7	−0.3
por	0.0380	1.061	1.096	+0.035	1.097	1.141	+0.044	+3.4	+4.2	+0.7
cat	0.0174	1.137	1.175	+0.039	1.178	1.236	+0.058	+3.6	+5.1	+1.5
Singletons										
deu	0.0496	1.044	1.090	+0.046	1.082	1.132	+0.049	+3.7	+3.8	+0.1
vie	0.0301	0.904	0.940	+0.036	0.933	0.977	+0.044	+3.2	+3.9	+0.7
kor	0.0297	1.139	1.190	+0.051	1.183	1.232	+0.048	+3.9	+3.5	−0.4
tha	0.0242	0.489	0.516	+0.026	0.510	0.538	+0.028	+4.3	+4.3	+0.0
ell	0.0236	0.648	0.701	+0.053	0.680	0.755	+0.075	+5.0	+7.7	+2.7
Perso-Arabic										
fas	0.0280	1.024	1.083	+0.058	1.105	1.178	+0.073	+7.8	+8.8	+0.9
urd	0.0124	0.920	0.986	+0.066	0.957	1.057	+0.100	+4.0	+7.1	+3.1
Latin commons										
ind	0.0317	1.141	1.189	+0.048	1.180	1.235	+0.056	+3.4	+3.9	+0.5
tur	0.0284	1.185	1.237	+0.053	1.237	1.316	+0.079	+4.5	+6.4	+1.9
arb	0.0264	2.002	2.070	+0.068	2.024	2.073	+0.049	+1.1	+0.1	−1.0
fin	0.0228	1.205	1.295	+0.090	1.281	1.395	+0.114	+6.3	+7.8	+1.4
ekk	0.0162	1.353	1.454	+0.100	1.433	1.560	+0.127	+5.9	+7.3	+1.4
eus	0.0083	1.335	1.448	+0.112	1.426	1.559	+0.133	+6.8	+7.7	+0.9
swh	0.0077	1.376	1.475	+0.099	1.443	1.586	+0.143	+4.8	+7.5	+2.7
hat	0.0046	1.683	1.805	+0.123	1.760	1.933	+0.173	+4.6	+7.1	+2.5
South Asian										
hin	0.0190	0.590	0.633	+0.042	0.618	0.674	+0.056	+4.7	+6.6	+1.8
ben	0.0159	0.609	0.659	+0.050	0.640	0.712	+0.072	+5.1	+8.1	+3.0
tam	0.0112	0.512	0.569	+0.057	0.538	0.619	+0.081	+5.0	+8.7	+3.7
tel	0.0089	0.656	0.745	+0.089	0.694	0.828	+0.134	+5.7	+11.1	+5.3
mya	0.0087	0.530	0.609	+0.079	0.562	0.697	+0.134	+6.1	+14.5	+8.4

Table 8: Per-language absolute BPB at EN20 and EN60 and relative ΔBPB (EN20 → EN60) under MIXED and MONO batch training. Each group includes a diff column: $|\Delta|$ for absolute BPB and pp for relative ΔBPB. Grouped by MIXED EN20 cluster.