Towards Democratizing LLMs: Investigating Multilingual Mixture-of-Experts Models

Motivation: Large Language Models (LLMs) achieve impressive performance on high-resource languages but often underperform on underrepresented ones due to data imbalances [1]. Mixture-of-Experts (MoE) architectures offer a promising solution by dynamically allocating computation to different sub-networks, potentially enabling more equitable multilingual learning [2]. In this work, we investigate whether language-specialized experts emerge naturally in decoder-only MoE models under continual multilingual pretraining. Building on OLMoE checkpoints [3], we train on a curated multilingual corpus spanning both high-resource and low-resource languages, with an emphasis on typological and script diversity. Our intrinsic analyses of expert routing patterns reveal an emergent modularity: early layers function as general-purpose experts, while later layers develop strong, language-specific specialization. Importantly, low-resource languages tend to reuse experts associated with their high-resource counterparts, especially when they share scripts or tokenization schemes. This mechanism provides an efficient pathway for knowledge transfer, suggesting that MoEs can implicitly encode cross-lingual structure. These findings highlight conditional computation as a scalable and linguistically adaptive framework for inclusive multilingual modeling.

Methodology. Step 1: Scalable Multilingual MoE Training. We extend OLMoE for continual pretraining on a large-scale multilingual corpus, ensuring coverage across diverse high- and low-resource languages. Training configurations will emphasize tokens routing strategies establishing a foundation for interpretability.
Step 2: Interpreting Expert Specialization. We analyze expert activation patterns across layers to evaluate whether language-specific experts emerge and how they interact with linguistic features such as script, morphology, or syntax. We further examine whether related languages converge on shared experts, suggesting pathways for implicit transfer.
Step 3: Transfer to Low-Resource Languages. We study how established routing patterns adapt during continual pretraining on new low-resource languages. In particular, we evaluate whether MoEs facilitate expert reuse from related high-resource languages, thereby enabling efficient adaptation under constrained data budgets.

Preliminary Results. Our initial experiments demonstrate two key findings, please refer to Figure 1:
1. Emergent Specialization. In later layers, experts become highly language-specific, while early layers remain generalist. This modularity is causally linked to performance gains.
2. Cross-Lingual Transfer. Low-resource languages predominantly reuse experts from their high-resource counterparts, especially when they share scripts (e.g., Arabic–Farsi, Hindi–Nepali). This expert reuse provides a natural mechanism for efficient transfer without explicit routing constraints.

Conclusion. Our work provides evidence that multilingual MoEs naturally develop language-specialized experts and that this emergent modularity underpins cross-lingual transfer. By leveraging conditional computation, MoEs represent a promising path toward building linguistically adaptive and equitable multilingual LLMs.

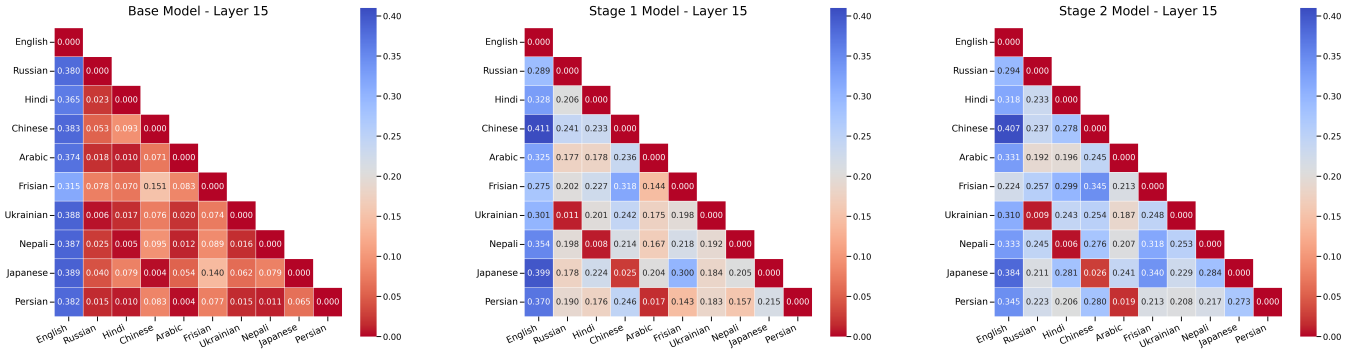

Figure 1: Jensen-Shannon Divergence for layer 15 of the base model (left), the multilingual model trained on 5 languages during Stage 1 (middle), and the model further trained on 5 low-resource languages during Stage 2 (right). Each heatmap shows the divergence in token-to-expert distributions.

References
[1] The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi et al., ACL 2020)
[2] Efficient Large Scale Language Modeling with Mixtures of Experts (Artetxe et al., EMNLP 2022)
[3] OLMoE: Open Mixture-of-Experts Language Models (Muennighoff et al., AllenAI Model 2024)