CULTURALLY AWARE MULTILINGUAL MODEL ROUTING THROUGH A MIXTURE-OF-SPECIALISTS FRAMEWORK

Isaac Adjei

Howard University, adjeinyaduisaac.edu@gmail.com

Saurav K. Aryal, PhD

Howard University, saurav.aryal@bison.howard.edu

Legand L. Burge III, PhD

Howard University, lburge@howard.edu

Large language models (LLMs) continue to underperform for culturally diverse and linguistically underrepresented communities, limiting their applicability in multilingual and code-switched environments. This work introduces a culturally aware Mixture of Specialists (MoS) framework coordinated by a Model Control Protocol (MCP) server to dynamically route user inputs to language- or region-specific models based on linguistic proximity, cultural relatedness, and data availability. When a dedicated specialist exists, it is used directly; otherwise, a hierarchical fallback strategy selects a linguistically related model, then a culturally proximate variant such as a West African English–tuned specialist, and finally a multilingual backbone augmented with lightweight regional adapters. As part of a multi-phase research program, this paper presents the first stage of the system, focusing on the routing architecture, cultural metadata extraction, and region-aware prompting components while specialist model training is ongoing. To support future specialization, we prepare parameter-efficient fine-tuning pipelines (LoRA and QLoRA) using openly licensed corpora rich in local context, including OSCAR, mC4, BigScience ROOTS, Tatoeba, African StoryBooks, and Global Voices, with thorough deduplication, filtering, and native-speaker validation. Evaluation on the BLEnD benchmark from SemEval 2026 Task 7 across 26 languages and 30 regions demonstrates that culturally grounded routing signals, regional metadata, and language-specific constraints yield substantial gains in contextual accuracy, robustness in low-resource settings, and cross-regional generalization. These Phase-1 results provide early empirical evidence that linguistic relatedness and cultural proximity can meaningfully enhance multilingual model performance even before full specialist integration. Overall, this work establishes a scalable foundation for developing globally adaptive and culturally grounded NLP systems.

CCS CONCEPTS Computing methodologies → Natural language processing; Information systems → Information retrieval; Computing methodologies → Machine learning algorithms

KEYWORDS: African Languages, Sentiment Analysis, Language Families, Multilingual NLP

REFERENCES

Adebara, I., Ahia, O., Orife, I., Kreutzer, J., & Muhammad, S. (2024). Cheetah: A Massively Multilingual Language Model for Over 500 African Languages. arXiv preprint. https://arxiv.org/abs/2401.06423

Dossou, B. F., & Emezue, C. C. (2022). AfroLM: A Self-Active Learning-Based Multilingual Pretrained Language Model for 23 African Languages. Proceedings of SustaiNLP (ACL Workshop). https://aclanthology.org/2022.sustainlp-1.11/

Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv preprint. https://arxiv.org/abs/2101.03961

Kuo, M., Krishna, K., Lopes, R., & OLMo Team. (2024). OLMoE: Open Mixture-of-Experts Language Model. arXiv preprint. https://arxiv.org/abs/2407.00379

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., & Wu, Y. (2021). GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. arXiv preprint. https://arxiv.org/abs/2006.16668

Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2022). DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale. arXiv preprint. https://arxiv.org/abs/2201.05596

Ruder, S., Constant, N., Firat, O., et al. (2021). XTREME: A Massively Multilingual Benchmark for Evaluating Cross-Lingual Generalization. arXiv preprint. https://arxiv.org/abs/2003.11080

SemEval Task Committee. (2026). SemEval-2026 Task 7: BLEnD Benchmark for Culturally Grounded Reasoning.