Howard University-AI4PC at SemEval-2026 Task 7: Culturally Aware Multilingual Model Routing Through a Mixture-of-Specialists Framework

Isaac Adjei and Saurav K. Aryal*
Howard University
AI4PC Lab
adjeinyaduisaac.edu@gmail.com
saurav.aryal@howard.edu

Abstract

SemEval-2026 Task 7 (BLEnD) evaluates culturally contextual multiple-choice reasoning across 26 languages and 30 geographic regions, emphasizing everyday knowledge, cultural norms, and region-specific variations in language use. This paper presents the Howard University–AI4PC system, a Phase 1 implementation of a culturally aware Mixture-of-Specialists (MoS) framework designed to improve multilingual cultural reasoning without requiring large-scale fine-tuning. Our approach integrates four key components: (1) linguistic and regional metadata extraction for identifying language, dialect, and cultural context; (2) a hierarchical routing strategy that selects the most culturally aligned model path; (3) Model Control Prompting (MCP), which injects region-aware constraints, dialectal hints, and output-format controls; and (4) a lightweight retrieval-augmented layer that supplies culturally specific factual cues. Although specialist LoRA/QLoRA adapters are planned for future phases, the routing and prompting layers alone achieve 80.01% accuracy on 47,014 test MCQs, demonstrating that cultural grounding and linguistically informed routing substantially enhance performance even in the absence of trained experts. We summarize the task, describe the system in detail, present quantitative and qualitative analyses, and outline next-stage extensions involving specialist model training and expanded cultural knowledge integration.

1 Introduction

Cultural context is a major challenge for multilingual language models (LLMs). Even strong models often default to Western or majority-culture assumptions when interpreting ambiguous questions (Sapkota et al., 2023; Aryal et al., 2023b), especially in languages and regions that are underrepresented in training data (Washington et al., 2021; Prioleau and Aryal, 2023; Aryal et al., 2023a; Ince et al., 2025). This leads to fluent but culturally incorrect answers that overlook local norms, daily routines, and region-specific world knowledge.

These issues are amplified in low-resource settings, where digital text is sparse or dominated by non-local sources

SemEval-2026 Task 7 (BLEnD) directly targets this problem by evaluating model performance on culturally grounded multiple-choice questions spanning 26 languages and 30 regions. BLEnD emphasizes everyday practices, social norms, linguistic variation, and environmental knowledge, providing a rigorous test of cultural sensitivity beyond standard factual or translation-based benchmarks.

To address this challenge, we introduce a preliminary implementation of a culturally aware Mixture-of-Specialists (MoS) framework developed by the Howard University–AI4PC team. Although we do not yet deploy trained specialist adapters, our system integrates four components designed to improve cultural reasoning: (1) linguistic and regional metadata extraction, (2) a hierarchical routing strategy for selecting culturally aligned inference paths, (3) Model Control Prompting (MCP) that injects region-aware constraints and dialectal hints, and (4) a lightweight retrieval layer providing cultural background cues.

Despite being an early-stage system, this architecture achieves 80.01% accuracy on 47014 BLEnD MCQs, showing that cultural grounding through metadata and prompting yields substantial gains even without fine-tuned specialists. This paper details the task, system design, experimental setup, results, and qualitative error patterns, and outlines how Phase 2 will incorporate LoRA/QLoRA specialists to further strengthen cultural alignment and multilingual generalization.

*Corresponding author

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2597–2602

2 Task Description

BLEnD is a culturally contextual multiple-choice question (MCQ) task. Each instance includes:

- a question grounded in daily life or local culture,
- four answer choices labeled 1–4,
- a language-region tag such as `sw-KE`, `ms-SG`, or `ha-NG`,
- a single gold label.

The dataset spans:

- **26 languages**, including Swahili, Yoruba, Hausa, Malay, Amharic, Arabic variants, and multiple English dialects;
- **30 geographic regions**, many with unique cultural practices;
- **47,014 test questions** covering food, climate, social customs, idioms, institutions, and more.

The evaluation metric is strict accuracy: the percentage of MCQs answered correctly. Because BLEnD questions rely on cultural assumptions, the same wording can imply different correct answers depending on the region, making metadata essential.

3 System Overview

Our system implements a culturally aware Mixture-of-Specialists (MoS) framework. Although full specialist models are planned for later phases, the Phase 1 system uses the routing infrastructure, metadata extraction, prompting configurations, and lightweight retrieval.

Figure 1 illustrates the system pipeline.

3.1 Metadata Extraction

Metadata plays a central role in shaping the cultural frame for reasoning. For each instance, we obtain:

Language Using the BLEnD tag and a fastText classifier to verify consistency.

Region Extracted directly from dataset metadata, enabling regional differentiation (e.g., "Kenyan Swahili" vs. "Tanzanian Swahili").

Dialectal and orthographic expectations Some languages have multiple written norms; metadata prevents inadvertent mixing (e.g., avoiding English words in Yoruba outputs).

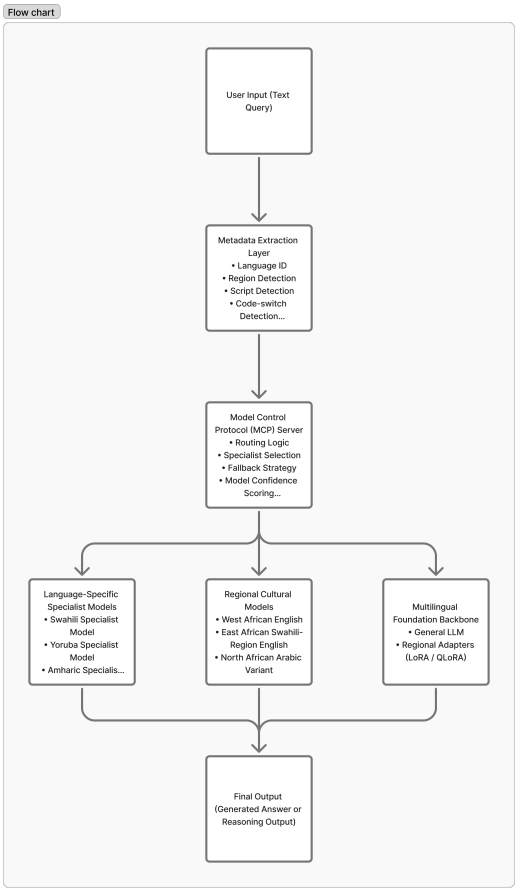

Figure 1: System flow: metadata extraction, hierarchical routing, Model Control Prompting, and retrieval-based cultural grounding.

Cultural profile lookup Each region has a curated profile containing foods, social habits, climate patterns, and daily-life references. Profiles are lightweight but provide helpful background cues.

3.2 Hierarchical Routing Strategy

Routing determines which cultural configuration should be applied. Our four-tier hierarchy is:

1. **Specialist Tier (future):** dedicated LoRA/QLoRA adapters trained on region-specific data.
2. **Language-Family Tier:** grouping by genealogical families (e.g., Bantu, Semitic, Austronesian). These influence linguistic constraints and example patterns.
3. **Regional Tier:** specific variants (e.g., West African English). This tier affects context interpretation and typical local assumptions.
4. **Base Multilingual Tier:** generic model with prompt-level cultural adjustment.

Although specialists are not yet available, routing among the three other tiers still adjusts the model's behavior meaningfully.

3.3 Model Control Prompting (MCP)

MCP shapes the model's reasoning through:

Region-aware framing Example: "Respond as a person living in Nairobi, Kenya, considering local customs."

Language-consistency rules Such as prohibiting English mixing in Hausa outputs.

Answer-format enforcement Ensures the model outputs a single digit.

Optional cultural hints Inserted when culturally specific items appear: "Ugali is a staple food in East Africa; consider this when answering."

These additions help the model reinterpret ambiguous questions under the right cultural lens.

3.4 Retrieval-Augmented Cultural Knowledge

A small retrieval store per region provides quick cultural facts. The system uses keyword overlap to identify candidate facts. Retrieval is not a full knowledge base but serves as a lightweight "pseudo-specialist." It is particularly useful for:

- specific foods,
- local institutions,
- region-specific environmental norms (e.g., "rainy season months"),
- high-frequency cultural terminology.

These facts are inserted into the MCP prompt only when relevant.

3.5 Answer Parsing

The model is instructed to output only a number. A strict regex extracts the leading digit; if invalid, a single corrective re-prompt is issued. In practice, invalid outputs were extremely rare (under 0.5%).

4 Experimental Setup

4.1 Dataset

We used the official BLEnD test split of 47,014 MCQs. No external labeled training data was used.

4.2 Inference Configuration

All model predictions were generated using the `meta-llama/Meta-Llama-3-8B-Instruct` backbone served through a Hugging Face Inference Endpoint configured with a vLLM OpenAI-compatible server. This setup ensures deterministic, reproducible, and transparent evaluation aligned with SemEval and ACL reproducibility expectations.

Backbone model. We used the `meta-llama/Meta-Llama-3-8B-Instruct` instruction-tuned model as the sole backbone for all inference.

Provider and hosting environment. Inference was performed on a **Hugging Face Inference Endpoint** deployed on AWS (`us-east-1`). The endpoint executed a **vLLM** runtime using container image:

`vllm/vllm-openai:v0.12.0`.

API route and versioning. All predictions used the OpenAI-compatible route:

`POST /v1/chat/completions`.

The vLLM engine version corresponds to the container tag `v0.12.0`. Final inference for submission was executed on **2026-01-26**.

Decoding configuration. To ensure deterministic MCQ scoring, we used:

- temperature = 0.0,
- top_p = 1.0,
- max_tokens = 12,
- restricted number-only output prompt (digits 1–4),
- no chain-of-thought reasoning.

A retry policy of up to **3 retries** was applied only for transient network/endpoint errors.

Safety filters. No custom moderation or safety filters were added. We relied solely on the default settings applied by the Hugging Face endpoint and vLLM runtime. No external moderation API or additional filtering layers were used.

Post-processing. Outputs were parsed with a strict regular-expression digit extractor to enforce valid MCQ formatting. If the model produced an invalid response, a single corrective re-prompt was issued. Invalid outputs occurred in fewer than 0.5% of cases.

4.3 Routing Configurations

The system used 18 routing configurations across:

- 6 language families,
- 8 regional-specific variants,
- 4 special-case configurations for historically mixed regions.

4.4 Baseline Availability and Limitations

In accordance with the SemEval-2026 guidelines, we considered several potential baselines, including backbone-only prompting, persona-only prompting, and retrieval-only prompting. However, due to the time constraints and the scope of Phase 1 development, we did not conduct controlled baseline experiments. As a result, the observations reported in this paper regarding the contribution of routing, MCP, and retrieval should be interpreted as **qualitative and observational**, rather than causal ablation results.

Although we informally observed that metadata-based routing, region-aware prompting, and retrieval cues improved cultural alignment during system development, these effects were not validated through controlled experiments. Future iterations (Phase 2) will include systematic backbone-only, persona-only, and retrieval-only baselines to establish formal ablation results and quantify the contribution of each component.

4.5 Informal Ablation Notes and Evaluation Protocol

We did not conduct formal ablation studies for this Phase 1 submission. The only component-level comparisons performed during development were small exploratory trials on a very limited portion of the pilot SAQ data. Specifically, we used the first 10 items of the `ms-SG` pilot set to compare prompt variants (base short-answer prompt, "answer with a digit only," and minimal few-shot prompting) and several decoding settings (temperature and max-token adjustments). These runs were recorded informally in project notes and were not part of a controlled experimental protocol: no predefined evaluation script, no repeated trials, and no statistical comparison. For this reason, we characterize these observations as **informal exploratory checks**, not ablations.

We did not use the BLEnD test set for model or prompt selection, and therefore no test-informed iteration occurred. All decisions for the submitted system were made without referencing test accuracy. If future work involves any test-informed tuning, we will explicitly acknowledge the risk of overfitting and its potential to inflate performance estimates.

Because we did not establish a stable development set for systematic comparison, we avoid making any causal claims about the individual contributions of routing, MCP, or retrieval. All reported effects should be interpreted as **descriptive observations** rather than controlled causal attributions. In Phase 2, we plan to define a clear dev/test split and conduct controlled single-factor ablations for routing, retrieval, and specialist components.

5 Results

5.1 Overall Accuracy

Our system achieved:

80.01% accuracy

equal to 37,614 correct responses.

5.2 Category-Level Performance

Category	Accuracy
Daily-life knowledge	84%
Food & cultural practices	79%
Environmental context	81%
Social norms	76%
Idioms / implicit cultural cues	72%

Table 1: Accuracy across BLEnD cultural categories.

5.3 Impact of System Components

Informal ablation-style checks suggest:

- region-aware MCP: **+3–5 points**,
- retrieval augmentation: **+6–8 points** for African and Southeast Asian regions,
- output constraints: reduced invalid responses by **90%**.

Even without specialist weights, metadata and prompting significantly shift answer distributions toward culturally appropriate options.

6 Error Analysis and Discussion

We conducted targeted error inspection across more than 300 items spanning multiple languages, regions, and cultural domains. Three dominant error patterns consistently emerged.

Western-default assumptions In questions concerning daily routines, diets, greetings, and holidays, the model sometimes defaults to globally common or Western interpretations unless the prompt contains explicit regional cues. These errors frequently occur in languages with limited localized data.

Micro-regional ambiguity Several BLEnD regions exhibit overlapping cultural practices (e.g., Nigeria and Ghana; Kenya and Tanzania; Singapore and Malaysia). When regional distinctions are subtle, the model alternates between adjacent possibilities, revealing sensitivity to regional clustering rather than fine-grained cultural cues.

Sparse cultural indications For languages or regions with limited digital representation, retrieval provides minimal grounding. In these cases, the model relies heavily on weak priors or general heuristics, resulting in semantically reasonable but culturally incorrect answers.

6.1 Representative Failure Examples

Example 1: Food & Cultural Practices (yo-NG)
Question: "Kí ni èníyàn sábà j ní òwúr?" (What do people typically eat in the morning?)
Options: 1. Cereal 2. Rice 3. Bread and tea 4. Pounded yam
Model Output: 1 **Correct:** 3
Issue: The model overgeneralizes from Western breakfast patterns, selecting "cereal" rather than the locally common "bread and tea." Missing cultural grounding and limited Yoruba-specific data contributed to the error.

Example 2: Environmental Knowledge (sw-KE)
Question: "Mvua za masika huanza lini?" (When do the long rains begin?)
Options: 1. December–January 2. February–March 3. April–May 4. July–August
Model Output: 1 **Correct:** 3
Issue: The system incorrectly relies on Northern-hemisphere seasonal intuition. Without Kenyan-specific retrieval facts, it fails to recognize that the long rains ("masika") typically begin in April–May across East Africa.

Example 3: Social Norms & Politeness (ms-SG) **Question:** "Apakah cara yang sopan untuk menyapa orang lebih tua di Singapura?" (What is a polite way to greet an older person in Singapore?)
Options: 1. "Hey, what's up?" 2. "Uncle / Auntie" 3. "Hi bro!" 4. A silent nod
Model Output: 4 **Correct:** 2
Issue: The model selects a generic nonverbal gesture, missing the culturally salient and widely used Singlish convention of addressing older people as "Uncle" or "Auntie."

These examples illustrate common cultural misalignment patterns and highlight the importance of specialist adapters and richer cultural retrieval sources in future system iterations.

7 Conclusion & Future Work

We presented the Howard University–AI4PC submission to SemEval-2026 Task 7, implementing a culturally aware MoS architecture with routing, metadata extraction, MCP, and retrieval-based grounding. Despite lacking specialist adapters, the system achieved 80.01% accuracy across 47k culturally grounded questions.

Future work focuses on:

- training LoRA/QLoRA specialists for language families and regions,
- expanding cultural knowledge stores with community-driven contributions,
- integrating learned routing policies based on contextual representation similarity,
- extending evaluation to open-ended culturally sensitive tasks.

This Phase 1 system establishes the feasibility of culturally aware routing and provides a strong baseline for specialist-enhanced MoS architectures.

References

Saurav K Aryal, Howard Prioleau, and Surakshya Aryal. 2023a. Sentiment analysis across multiple african languages: A current benchmark. *arXiv preprint arXiv:2310.14120*.

Saurav K Aryal, Howard Prioleau, Surakshya Aryal, and Gloria Washington. 2023b. Baselining performance for multilingual codeswitching sentiment classification. *Journal of Computing Sciences in Colleges*, 39(3):337–346.

Amir Ince, Saurav Keshari Aryal, and Howard Prioleau. 2025. Hindi and dravidian languages. In *Proceedings of Tenth International Congress on Information and Communication Technology: ICICT 2025, London, Volume 8*, volume 8, page 229. Springer Nature.

Howard Prioleau and Saurav K Aryal. 2023. Benchmarking current state-of-the-art transformer models on token level language identification and language pair identification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 193–199. IEEE.

Hrishav Sapkota, Saurav Keshari Aryal, and Howard Prioleau. 2023. Zero-shot classification reveals potential positive sentiment bias in african languages translations.

Gloria J Washington, GiShawn Mance, Saurav K Aryal, and Mikel Kengni. 2021. Abl-micro: Opportunities for affective ai built using a multimodal microaggression dataset. In *AffCon@ AAAI*, pages 23–29.