chengtang at SemEval-2026 Task 7: A Retrieval-Augmented Generation Framework for Cultural Perspective Alignment in Everyday MCQs

ZhiChao Meng, MeiZhi Jin, JinJin Cheng, LianXin Jiang, JianYu Li
Ping An Life Insurance Company of China, Ltd.
EX-MENGZHICHAO001@pingan.com.cn, JINMEIZHI005@pingan.com.cn,
CHENGJINJIN213@pingan.com.cn, JIANGLIANXIN769@pingan.com.cn,
LIJIANYU002@pingan.com.cn

Abstract

Large language models (LLMs) often exhibit significant cultural representation biases in multilingual everyday knowledge understanding, struggling to accurately capture region-specific customs and values. This paper presents our system submission for SemEval 2026 Task 7: BLEnD Challenge Track 2 (MCQ) (Ousidhoum et al., 2026). To address these challenges, we propose a training-free retrieval-augmented generation (RAG) framework. Without introducing any external data, we manually constructed a localized multicultural knowledge base for each language-region and used text-embedding-v4 for region-specific cultural background retrieval. In the generation stage, we adopted a strict zero-shot setting: prompts contain no task instance question-answer examples, only injecting locale-relevant background cultural descriptions via RAG to compensate for contextual information absence, combined with a dual-model ensemble strategy using Gemini 3 Flash (preview) (Google DeepMind, 2025) and GPT-5.2 Chat (OpenAI, 2025). Our system achieved an overall score of 96.35 on the final Evaluation dataset. Additionally, we conducted in-depth analysis of model performance on specific languages, particularly highlighting severe cultural alignment challenges faced by large models in dialectal variants like Moroccan Arabic (ar-MA) and highly localized subjective Japanese (ja-JP) everyday scenarios.

1 Introduction

With the global deployment of large language models (LLMs), evaluating and enhancing models' cultural awareness across different languages and regions has become critically important. Existing models, often trained predominantly on English-centric corpora, frequently provide answers deviating from local customs when addressing multiple-choice questions about region-specific daily life. SemEval 2026 Task 7 (BLEnD) Track 2 evaluates whether systems can select the most appropriate answer for target regions when presented with four options representing different national cultural perspectives. (Ousidhoum et al., 2026; Ghosh et al., 2026)

For this task, we avoided costly model fine-tuning and instead developed a lightweight RAG-based system. Our main contributions are:

- **Culturally-aligned RAG framework**: Without external data, we manually authored compact multicultural knowledge bases for each language-region, summarizing local practices in food, holidays, family, education, and work, combined with text-embedding-v4 to provide precise cultural background references for closed-source LLMs.
- **Dual-model ensemble with matching**: Under a unified RAG framework, we performed parallel inference with Gemini 3 Flash (preview) and GPT-5.2 Chat, extracting final answers through option matching, achieving an excellent score of 96.35 on approximately 47K Evaluation data samples.
- **In-depth negative insights**: Through detailed error analysis, we revealed cognitive blind spots in current top LLMs when handling Arabic dialects heavily influenced by other language families (e.g., Moroccan Arabic) and highly subjective localized commonsense (e.g., Japanese convenience store culture).

2 Background / Related Work

Recent research on large model cultural awareness has grown significantly. For instance, Pawar et al. (2025)'s survey (Pawar et al., 2025) points out significant deficiencies in models regarding non-textual cultural commonsense; Alam et al.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1610–1615

(2025)'s NativQA framework (Hasan et al., 2025) also emphasizes the importance of localization and everyday knowledge. Our work is inspired by these studies, attempting to dynamically inject localized knowledge during inference using RAG techniques to mitigate LLMs' inherent cultural biases. The BLEnD benchmark proposed by Myung et al. (2024) (Myung et al., 2024) systematically evaluated LLM performance in multicultural everyday knowledge scenarios, discovering significant performance gaps across different languages and regions, which directly motivated the design of SemEval-2026 Task 7.

3 Data

We evaluated our system using the official SemEval-2026 Task 7 dataset, which extends the BLEnD benchmark (Myung et al., 2024) to cover more language-culture combinations. Track 2 (MCQ) instances each contain an English question, target language-region code (Language-Region, e.g., ar-MA, ja-JP, su-JB), and four candidate options representing different regional cultural perspectives.

We used the following two official datasets:

- **Trial Data**: Used throughout the entire Development (Dev) phase, including familiarization with input format, pipeline validation, model selection, hyperparameter tuning, and ensemble strategy validation;
- **Evaluation Data**: Used for the final Evaluation (Eval) phase, with official scores obtained through CodaBench submission, containing approximately 47K samples.

All data was used solely for development and evaluation, strictly adhering to Task 7's evaluation-only policy, with no parameter training or few-shot example construction.

3.1 Dataset Statistics and Distribution Characteristics

To better understand Track 2's internal structure and complexity, we conducted detailed distribution statistics and analysis of both Trial Data and Evaluation Data.

1. Trial Data Distribution The Trial Data used in the Dev phase contains 146 initial samples covering 23 fine-grained Language-Region combinations. Overall, this small-scale dataset has relatively uniform distribution, with each target region containing approximately 5 to 8 samples (see table below). This small but broad-coverage dataset was effectively used to pre-build early RAG knowledge base templates and validate overall engineering pipeline robustness.

Table 1: Trial data distribution by language-region in the development phase.

Language-Region	Count	Language-Region	Count
es-EC	8	**ar-SA**	7
tl-PH	8	**ar-MA**	7
fr-FR	8	**ja-JP**	7
ms-SG	7	**ta-SG**	7
ar-EG	7	**zh-SG**	7
bg-BG	7	**el-GR**	5
ta-LK	7	**ko-KR**	5
ga-IE	7	**id-ID**	5
eu-ES	7	**es-MX**	5
en-AU	7	**es-ES**	5
zh-CN	5	**en-GB**	5
fa-IR	5	**Total**	**146**

2. Evaluation Data Language Imbalance For the Evaluation Data containing 47,014 blind test samples, we decomposed by primary language (Language). Results show extreme imbalance in language distribution (Long-tail Distribution):

Table 2: Evaluation data language distribution (by primary language).

Language	Count	Language	Count
Spanish	4,807	**Hausa**	2,008
Korean	4,697	**Indonesian**	1,995
English	4,622	**Tagalog**	1,327
Arabic	3,955	**Tamil**	1,114
Persian (Farsi)	3,699	**Basque**	1,075
Amharic	2,863	**Irish**	856
Greek	2,734	**Bulgarian**	648
Assamese	2,451	**Swedish**	447
Chinese	2,357	**Japanese**	410
Sundanese	2,345	**French**	307
Azerbaijani	2,297	**Total**	**47,014**

As shown, Spanish, Korean, and English dominate (all >4,600 samples), while Japanese (410 samples) and French (307 samples) form extremely sparse long tails. This distribution tests both high-resource language culture performance and zero-shot cultural inference in resource-scarce contexts (like our Japanese scenarios showing strong localization).

3. Core Question Stem High-Frequency Reuse Further text deduplication analysis revealed the core challenge of "cross-cultural alignment" in this task. Despite containing 47K massive samples, the Evaluation Data has only 2,706 unique

question stems (statistics shown below):

Table 3: Question stem reuse statistics in the evaluation data.

Statistic	Value
Unique question count	2,706
Mean frequency	17.37
Standard deviation	56.53
Median (50%)	3
Maximum frequency	1,144

The data shows the same English question appears on average 17.37 times, with the most frequent question repeated 1,144 times. High-frequency repetition means identical English stems are systematically assigned to dozens or hundreds of different target region codes with constantly changing option combinations. This means models cannot rely on single, static "common knowledge" to memorize answers, but must make distinctly different cultural fact reasoning based on the specific region code injected in the prompt. This extreme data skew pattern inversely validates the necessity and rationality of our "region-specific RAG injection strategy."

4 System Overview

This section details our retrieval-augmented generation (RAG) framework for cross-cultural MCQ tasks.

4.1 Multicultural Knowledge Base Construction

To enhance cross-cultural perception, we did not directly use Task 7/BLEnD question-answer pairs as retrieval documents. Instead, completely without external corpora, based on understanding of target regions' daily life, we **manually authored** an efficient, localized multicultural knowledge base. For each language-region (e.g., ar-MA, ja-JP, su-JB), we constructed short text profiles summarizing typical daily practices and social customs in food, holidays, family structure, education systems, work, and leisure.

Each profile contains at least:

- Language-Region code;
- Theme category (e.g., food, holidays, family, work);
- Several natural language descriptions reflecting common practices, noting "multiple practices exist" when necessary to avoid oversimplification into single stereotypes.

These profiles are organized by locale and vectorized for indexing as retrieval candidate documents. All locale profiles followed the same template and were manually checked for format consistency, obvious contradictions, and overly specific stereotypes. Practice shows that even without external data, relying on these compact locale cultural profiles, we can provide sufficient regional background for generation models, achieving good cultural alignment at low data cost.

4.2 Retrieval Core

We used text-embedding-v4 as the core retrieval engine. For each multiple-choice question in Development or Evaluation phases, the system extracts [Question] and [Language-Region] (region code) as Query, performing similarity retrieval in the corresponding cultural region's knowledge base subset. To balance context richness and noise control, we set Top-K=3 based on Trial Data observations, extracting the 3 most relevant context chunks as subsequent generation references (References). Retrieval uses cosine similarity, with all profiles pre-encoded and indexed offline for online inference efficiency.

4.3 Generation and Prompting Strategy

To strictly comply with the official "prohibition of using BLEnD dataset (Myung et al., 2024) for fine-tuning or few-shot learning" evaluation rules, we adopted pure zero-shot setting in generation: besides the current question and four candidate options, prompts contain no additional question-answer examples from BLEnD/Task 7, only injecting target locale-related background cultural descriptions via RAG to compensate for contextual information absence.

We designed a strongly constrained system prompt requiring the model to act as a general Q&A assistant proficient in multiple countries' languages and cultural customs. Core prompt design includes:

- **Role setting**: Clearly informing the model of required knowledge in different countries/regions' history, culture, and daily life, guiding autonomous judgment of retrieved references' relevance and reliability.

- **Input structure**: Structured input of Language-Region (e.g., ar-MA, ja-JP, su-JB), Question, 4 candidate Options, and RAG-retrieved References, all in unified, clear format.
- **Strict output rules**: Forcing model to output only one option letter or corresponding option plain text from 4 options, prohibiting any explanations, lead-ins, or metadata to ensure perfect adaptation to automated evaluation and post-processing; simultaneously, setting LLM Temperature parameter to 0.0 uniformly for generation determinism.

5 Experimental Setup

5.1 Model Selection and Development Phase

In the Development (Dev) phase, we used Trial Data to select optimal base models for the final Evaluation phase. Under identical zero-shot prompts, we independently evaluated several mainstream large language models with and without RAG. Model accuracies in Dev phase are shown below:

Table 4: Development accuracy with and without RAG.

Model	RAG	No RAG
Gemini 3 Flash (preview)	97.97	94.59
Claude Sonnet 4.5 (20250929)	97.30	95.95
GPT-5.2 Chat	97.30	91.22
DeepSeek-V3.2	88.51	84.46

Results show that RAG consistently improved all tested models on Trial Data. Gemini 3 Flash (preview) (Google DeepMind, 2025) achieved the best RAG accuracy of 97.97. Although Claude Sonnet 4.5 (Anthropic, 2025) and GPT-5.2 Chat (OpenAI, 2025) obtained the same RAG accuracy, we selected GPT-5.2 Chat as the second ensemble model because it showed slightly more stable answer-only format compliance during manual inspection. DeepSeek-V3.2 (Liu et al., 2025) benefited from RAG but remained weaker on multilingual localized commonsense.

We used Trial Data only in Dev phase for model selection and hyperparameter tuning, making no adjustments before final Eval submission.

5.2 Final Evaluation Strategy: Option-Constrained Ensemble

Based on Dev performance, for processing 47K sample-scale Evaluation Data, we abandoned single-model inference for dual-model ensemble with option matching (Dual-Model Ensemble with Option Matching). Specifically, Gemini 3 Flash (preview) and GPT-5.2 Chat performed parallel inference on identical questions.

Since only one correct answer exists among 4 options per question, we designed strict option matching:

1. Extract two models' text outputs and match against given 4 candidate options (accepting exact matches or minor whitespace/punctuation differences);
2. If only one model's output successfully matches candidate set, adopt that prediction;
3. If both outputs match valid options but conflict, prioritize Dev-best Gemini;
4. For rare cases where neither output matches any option, use embedding similarity fallback: encode each option as vector, compare similarity with retrieved evidence average vector, select highest similarity option.

No model fine-tuning throughout, all hyperparameters fixed pre-Eval, maximizing accuracy under format compliance via ensemble and filtering.

6 Results and Analysis

6.1 Overall Performance

Our system achieved 96.35 overall in final Evaluation phase. This demonstrates region-specific cultural knowledge base combined with advanced LLM ensembling effectively handles most cross-cultural everyday knowledge MCQs. Detailed performance across 30 language-region combinations:

Most regions (e.g., as-AS, bg-BG, su-JB) achieved near/exceeding 99% accuracy, but clear performance valleys appeared in specific cultural regions.

6.2 Language-Specific Findings and Negative Insights

Combining region scores and error analysis, we identified key blind spots:

Massive Intra-Arabic Cultural and Dialectal Differences

Though excelling in Egyptian Arabic (ar-EG: 96.74) and Algerian Arabic (ar-DZ: 98.15), performance plummeted to 83.98 in Moroccan Arabic (ar-MA), with Saudi Arabic (ar-SA: 91.89)

Table 5: Accuracy by language–region on the evaluation set.

Language-Region	Accuracy (%)
am-ET	98.67
ar-DZ	98.15
ar-EG	96.74
ar-MA	83.98
ar-SA	91.89
as-AS	99.35
az-AZ	99.22
bg-BG	99.54
el-GR	98.10
en-AU	90.06
en-GB	98.25
en-US	98.66
es-EC	98.36
es-ES	98.55
es-MX	98.00
eu-PV	97.12
fa-IR	97.95
fr-FR	97.72
ga-IE	96.26
ha-NG	96.07
id-ID	97.59
ja-JP	87.32
ko-KP	94.60
ko-KR	99.76
su-JB	99.87
sv-SE	94.63
ta-LK	96.86
tl-PH	94.35
zh-CN	97.82
zh-SG	95.09
Overall	96.35

also notably lower. Analysis suggests Moroccan Arabic (Darija) heavily influenced by Berber and French in daily life creates vast vocabulary and custom gaps from standard Middle Eastern or Egyptian Arabic. LLMs clearly lack sufficient alignment with North African informal daily language and unique customs in pretraining corpora.

Japanese Cultural Specificity and Commonsense Misalignment

Japanese (ja-JP: 87.32) ranks second-lowest overall. Beyond Japanese's complex writing system (hiragana, katakana, kanji mix) comprehension challenges, we found many daily questions exhibit extreme subjectivity and localization.

Error Case Analysis

For question " コンビニで売ってないもの" (things not sold at convenience stores), gold answer is " おにぎり" (rice ball), with distractors including " 電車の切符" (train tickets). In experiments, both Gemini and GPT violated gold annotation, erroneously selecting " 電車の切符" based on common real-world experience.

Analysis

Such questions often stem from specific group subjective cognition or contexts, while LLMs tend toward statistical inference from broad fact distributions (train tickets aren't convenience store staples like rice balls). Models systematically fail on highly localized or counterintuitive subjective cultural questions, exposing "local consensus vs universal fact" weighing blind spots. Australian English (en-AU: 90.06) relative underperformance further confirms region-specific daily commonsense remains LLM weakness even within same language family.

7 Conclusion

This paper presents an RAG Q&A framework for cross-cultural everyday knowledge MCQs. Experiments prove region-specific cultural knowledge bases combined with advanced LLM dual-model ensembling significantly enhance multicultural performance (final Eval phase: 96.35 overall accuracy). However, detailed score distributions and error analysis reveal systematic deficiencies in handling multi-language influenced complex dialects (e.g., Moroccan Arabic) and highly subjective local commonsense (e.g., specific Japanese scenarios). These results also suggest that RAG cannot fully compensate for the underlying model's limited understanding of low-resource dialectal variants and highly subjective local commonsense. Future work will explore better cross-cultural representation learning, finer-grained local cultural knowledge graphs, and local expert collaboration to address these gaps.

8 Ethical Considerations

Though using no external data, we recognize original annotation data for evaluation and knowledge base referencing may contain specific group subjective preferences. LLMs outputting culturally-attributed answers may inadvertently amplify regional stereotypes or ignore minority voices. Thus, deploying such cultural perception models in real-world cross-cultural communication or education requires stricter human review and bias mitigation; future work must incorporate more diverse annotator groups and stakeholders in evaluation and feedback.

References

Anthropic. Claude sonnet 4.5 system card, September 2025. URL https://www.anthropic.com/

system-cards.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States, July 2026. Association for Computational Linguistics.

Google DeepMind. Gemini 3 flash model card, December 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. NativQA: Multilingual culturally-aligned natural query for LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.770. URL https://aclanthology.org/2025.findings-acl.770/.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, 2024.

OpenAI. Introducing GPT-5.2. Technical report, OpenAI, 2025. URL https://openai.com/index/introducing-gpt-5-2/.

Nedjma Ousidhoum, Junho Myung, Carla Perez-Almendros, Jiho Jin, Amr Keleg, Meriem Beloucif, Yi Zhou, Rodrigo Agerri, Vladimir Araujo, Naomi Baes, James Barry, Joanne Boisson, Nancy F. Chen, Christine de Kock, Aleksandra Edwards, Joseba Fernandez de Landa, Mohamed Fazli Imam, Huda Hakami, Shu-Kai Hsieh, Joseph Marvin Imperial, Roy Ka-Wei Lee, Chenyang Lyu, Younes Samih, Johan Sjons, Bryan Tan, Asahi Ushio, Weihua Zheng, Zhengyuan Liu, Alice Oh, and Jose Camacho-Collados. SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, 2026.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004, 2025. doi: 10.1162/COLI.a.14. URL https://doi.org/10.1162/COLI.a.14.