VerbaNexAI at SemEval-2026 Task 7: Integrating Web Snippets and RAG for the Evaluation of Multilingual Cultural Knowledge in LLMs

Danileth Almanza Gonzalez, Jairo E. Serrano, Edwin Puertas,
Juan Carlos Martinez-Santos
Universidad Tecnologica de Bolivar
Cartagena, Colombia
{daalmanza,jserrano,epuerta,jcmartinezs}@utb.edu.co

Abstract

In multilingual and multicultural contexts, LLMs require contextualization mechanisms to generate culturally coherent responses. In this sense, this study presents a LLaMA-based approach to answer short cultural questions in different languages within Task 7 of SemEval-2026 (Track 1: SAQ), without access to official training data. The system integrates controlled synthetic data generation, evidence retrieval through web snippets, and a Retrieval-Augmented Generation (RAG) framework with Few-shot learning. BLEnD is used solely as a thematic guide, ensuring semantic independence. During development, the LLaMA-3.1-8B model achieved 38.51% global accuracy, while LLaMA-3.2-1B obtained 15.54%. In a large-scale evaluation (30,500 instances), the 1B model achieved 16.69% and maintained stability after prompt optimization. The results demonstrate that contextual retrieval improves the evaluation of multilingual cultural knowledge and highlight the importance of pipeline design and model capacity.

1 Introduction

Everyday and cultural knowledge constitutes a key element for LLMs to produce responses that are not only linguistically correct, but also semantically and contextually appropriate. Although recent advances in natural language processing (NLP) have substantially improved the fluency, coverage, and overall quality of generated responses, these models still exhibit significant limitations when reasoning about practices, social norms, traditions, and implicit knowledge that vary across cultures and territories. This issue becomes especially critical in contexts characterized by high cultural and linguistic diversity, where a single language may encompass distinct sociocultural realities, as well as in regions with low representation in training corpora, which increases the risk of bias, unwarranted generalizations, and misinterpretations.

Within this framework, SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures proposes a systematic evaluation framework to analyze the cultural awareness of language models across a wide variety of languages, countries, and regions (Myung et al., 2024). This study focuses on Track 1: Short Answer Questions (SAQ), which poses an additional challenge by not providing a training dataset, thus requiring participants to design alternative strategies for response generation and validation. To address this limitation, an information retrieval-based approach was implemented, using snippets obtained via an API to provide relevant contextual evidence during inference. These fragments were integrated into an LLaMA model, enabling the generation of concise, culturally coherent answers in multiple languages. The results show competitive performance, placing 17th, demonstrating that the combination of contextual retrieval and generative models constitutes an effective strategy for addressing everyday knowledge in multilingual environments. The repository is available at the following link: (available after review). [1]

2 Background

The dataset used in this research is from BLEnD, a manually designed benchmark for evaluating the everyday knowledge of language models across diverse cultural contexts. The construction process was carried out in four main stages: (i) collection of culturally grounded questions by native speakers, (ii) filtering and translation to avoid stereotypes and references specific to a single country, (iii) annotation of answers by multiple native annotators with lived experience in the corresponding territory, and (iv) aggregation and normalization of responses to consolidate linguistic variants and account for cultural consensus. As a result,

[1] https://github.com/VerbaNexAI/SemEval2026

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 899–904

BLEnD comprises 52.6k question–answer pairs, corresponding to 16 countries or regions and 13 languages, including low-resource languages such as Amharic, Assamese, Azerbaijani, Hausa, and Sundanese, and covers six sociocultural categories: food, sports, family, education, celebrations, and working life.

Each country or region includes 500 base questions, formulated in both the local language and English, enabling analysis of model performance in monolingual and cross-lingual scenarios. For each question, responses were collected from at least five annotators, who could provide multiple variants or explicitly indicate cases of *no answer* or *not applicable*, thereby reflecting the inherent uncertainty of everyday knowledge. The final dataset includes multiple candidate answers per instance, along with their selection frequency and versions in both the local language and English. In the context of SemEval-2026 Task 7, BLEnD is a particularly suitable setting for evaluating LLMs' ability to reason, retrieve information, and generate culturally appropriate answers under inference conditions.

Recent studies have shown that, despite advances in general linguistic capabilities, LLMs continue to exhibit substantial limitations in understanding and generating culturally situated knowledge (Chen et al., 2025). In the evaluation domain, it has been demonstrated that models tend to amplify cultural biases and distortions when confronted with specific social contexts, leading to phenomena such as generative exaggeration and ideological polarization (Nudo et al., 2026). Likewise, research in psychometrics and cross-cultural adaptation has evidenced that, although LLMs can assist in translation and cultural adaptation processes, their performance requires rigorous empirical validation to avoid semantic loss, implicit biases, and reliability issues (Grobelny et al., 2025). In parallel, large-scale review and benchmarking studies have identified significant gaps between high- and low-resource cultures, revealing performance differences of more than 60% in cultural knowledge and commonsense tasks, even in state-of-the-art multilingual models (Binegde and Zhang, 2026). In response to these limitations, multiple cultural benchmarks and evaluation frameworks have been proposed, along with systematic schemes to measure cultural biases and historical inaccuracies through specific metrics and human validation in the evaluation loop (Al-Monef et al., 2026) (Martínez et al., 2025) (Mak and Luo, 2025). Complementarily, recent work presented at leading conferences has shown that explicitly incorporating local human preferences and cultural judgments significantly improves models' cultural awareness. However, these benefits are not distributed homogeneously across architectures and regions (Chiu et al., 2025) (Guo et al., 2025) (Schneider et al., 2025). In this context, BLEnD positions itself as a particularly relevant benchmark by focusing on everyday, multilingual, and multicultural knowledge, providing a realistic and challenging scenario to analyze the extent to which LLMs can generalize beyond dominant contexts and respond in a culturally coherent manner.

3 System Overview

The proposed system implements a comprehensive pipeline to solve culturally situated short-answer questions (SAQ) in BLEnD, combining controlled data generation, external evidence retrieval, and context-conditioned answer generation. The architecture operates in two clearly differentiated phases: (i) the construction of a synthetic training set and (ii) inference through Retrieval-Augmented Generation (RAG) applied to the development and test sets. As shown in Figure 1, the flow begins with BLEnD as a thematic reference, continues with question generation using an LLM, web-based evidence retrieval through text fragments (*snippets*), construction of the synthetic dataset, and finally the application of a Few-shot + RAG scheme to answer new questions under the same inference scenario.

3.1 Data Source, Feature Extraction, and Vocabulary

BLEnD serves as the system's starting point and is used exclusively as a thematic guide rather than as a direct training set. This decision ensures generalization and avoids memorization of original instances. From the questions organized by topic, keyword extraction is performed to construct a controlled vocabulary. The process includes normalization, cleaning, and lemmatization, followed by candidate identification using noun chunks and informative grammatical categories (nouns, proper nouns, and adjectives). The objective is to select terms that represent culturally relevant entities, practices, or concepts, discarding common words typical of interrogative formulations. Additionally, a semantic-similarity-based diversity criterion is applied to prevent keywords from becoming excessively cor-

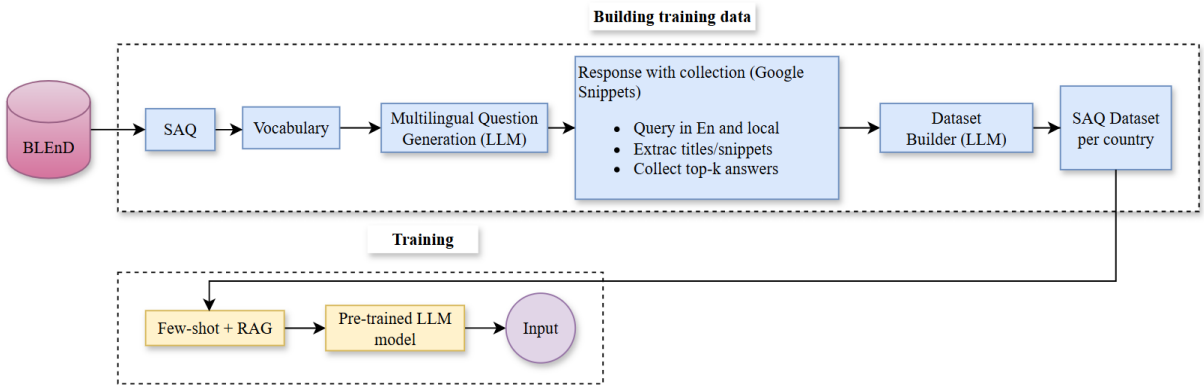

Figure 1: Architecture system.

related with a specific question, thereby reducing the risk of generating reformulations close to the base set.

3.2 Multilingual SAQ Question Generation with LLM

New SAQ questions are generated using the model `meta-llama/Llama-3.2-1B-Instruct` under a controlled generation scheme. The *prompt* imposes explicit constraints: brief questions, valid interrogative structure, a single semantic intent, and coherence between the topic and the keyword. When a quantitative answer is required, a specific template is activated to restrict the expected output type and avoid ambiguities. For the multilingual component, a base set is initially generated in English and subsequently adapted to other regions through controlled translation using the same LLM. Each generated question is evaluated using embedding-based semantic similarity (cosine similarity). If similarity exceeds a 20% threshold, the question undergoes rewriting. This iterative process ensures semantic independence from the original content and enables the construction of a coherent, diverse synthetic training set. Although the synthetic data is controlled through similarity thresholds, duplicate filtering, and iterative validation to ensure diversity and avoid redundant or overly similar questions.

3.3 Evidence Retrieval with Google Snippets and Top-K Selection

For each generated question, external evidence is retrieved using the Google Custom Search API. Multiple results and their corresponding text fragments are extracted, after which the Top-K most relevant fragments ($K = 4$) are selected using *embeddings* and cosine similarity. The choice of $K = 4$ balances informational coverage and computational efficiency, reducing token cost and minimizing the risk of contradictory evidence while maintaining sufficient context to cover implicit answers (Morillo et al., 2025).

3.4 Construction of the Final Training Set

Using the questions and selected fragments, the final training set is constructed, including the identifier, topic, keyword, question in the local language, question in English, and the four associated fragments. The LLM is additionally used to generate answers consistent with the retrieved evidence, ensuring that training faithfully reproduces the inference scenario: answering short questions based exclusively on brief textual evidence. During this process, *accuracy* was used as a preliminary metric to verify the coherence between the question, the evidence, and the generated answer.

3.5 Few-shot + RAG

During evaluation on the development and test sets, the same retrieval mechanism is applied. For each target question, the Top-K most relevant fragments are selected, and semantically similar examples from the training set are incorporated using *embeddings*. These examples are inserted into the *prompt* together with the retrieved evidence under a Few-shot + RAG scheme. The model generates a single-line response in the same language as the question, under strict constraints that prevent additional explanations or lengthy outputs, prioritizing precision and cultural adequacy (Almanza-González et al., 2025).

4 Experimental Setup

The experimental configuration evaluates the system under a scheme in which training is based ex-

Table 1: Accuracy by language/region on the development data — meta-llama/Llama-3.1-8B-Instruct

Language	ar-EG	ar-MA	ar-SA	bg-BG	el-GR	en-AU	en-GB	es-EC	es-ES	es-MX	eu-ES	fa-IR
Accuracy (%)	42.86	42.86	57.14	57.14	40.00	14.29	60.00	50.00	60.00	20.00	28.57	40.00

Language	fr-FR	ga-IE	id-ID	ja-JP	ko-KR	ms-SG	ta-LK	ta-SG	tl-PH	zh-CN	zh-SG	Overall
Accuracy (%)	12.50	28.57	40.00	14.29	0.00	28.57	71.43	14.29	50.00	100.00	28.57	38.51

clusively on synthetic data, supported by external evidence. At the same time, the official development and test sets are processed using the same retrieval and generation pipeline. It allows measuring generalization capability without direct training on BLEnD.

4.1 Models and Components Used

The system integrates three main components. The central generative model is `meta-llama/Llama-3.2-1B-Instruct`, used both in synthetic dataset generation and in final inference. To analyze the impact of model size, `meta-llama/Llama-3.1-8B-Instruct` was also evaluated, allowing the study of how parametric capacity influences multilingual cultural performance. Semantic similarity tasks, novelty control, and Few-shot retrieval employ the embedding model `sentence-transformers/all-MiniLM-L6-v2`. External retrieval is performed using the Google Custom Search API, which provides textual fragments and metadata (title, link, and summary), enabling verification of whether the answer is explicitly contained in the evidence or must be generated from the available context.

4.2 Operational Parameters of the Pipeline

The pipeline prioritizes consistency and control over output format. For each question, four fragments are selected according to cosine similarity. The context sent to the model includes the question, the evidence, and, when applicable, Few-shot examples. Responses are limited to 10-15 words, with a maximum of 24 generated tokens, to avoid lengthy or explanatory outputs and ensure uniformity in evaluation.

4.3 Evidence Collection and Reproducibility

To guarantee reproducibility, API queries are stored in a local cache, avoiding repeated calls and maintaining consistency across runs. Credentials and usage limits are enforced, and a resumption mechanism is implemented to resume processing without recalculating prior instances. Additionally, version control is implemented using Git to ensure traceability and proper management of experimental development.

5 Results

The results on the development set show a clear difference between the evaluated models. With *meta-llama/Llama-3.1-8B-Instruct*, the system achieved an overall accuracy of 38.51%, whereas with *meta-llama/Llama-3.2-1B-Instruct* the accuracy dropped to 15.54%. This gap highlights the significant impact of model size on tasks that require cultural reasoning and implicit contextualization. The higher-capacity model (8B) captures complex semantic relationships more effectively and better leverages evidence retrieved from RAG. In contrast, the 1B model exhibits limitations when integrating information dispersed across multiple web snippets. The language-level analysis on the development set reinforces this structural difference. In the 8B model, outstanding performance is observed in zh-CN (100.00%) and ta-LK (71.43%), and consistent results are observed in en-GB (60.00%), es-ES (60.00%), ar-SA (57.14%), and bg-BG (57.14%). In contrast, the 1B model shows a considerable reduction across most languages, with multiple cases close to zero. This behavior suggests that the additional parametric capacity of the 8B model enables better alignment between the question, the retrieved evidence, and the final generation, especially when the answer is not explicitly contained within a single snippet.

In terms of computational efficiency, during the first submission on the development set (158 instances), the 1B model required approximately 45 minutes to generate all responses. Considering that the evaluation set consists of 30,500 instances, a prompt and inference flow optimization was performed to improve scalability without compromising performance. After this optimization, the system processed the 30,500 instances in an equivalent time (approximately 45 minutes), maintaining comparable accuracy. This result shows that although the 1B model has representational limitations, its

Table 2: Accuracy by language/region on the development data — meta-llama/Llama-3.2-1B-Instruct

Language	**ar-EG**	**ar-MA**	**ar-SA**	**bg-BG**	**el-GR**	**en-AU**	**en-GB**	**es-EC**	**es-ES**	**es-MX**	**eu-ES**	**fa-IR**
Accuracy (%)	0.00	14.29	0.00	14.29	40.00	0.00	20.00	12.50	40.00	20.00	0.00	40.00

Language	**fr-FR**	**ga-IE**	**id-ID**	**ja-JP**	**ko-KR**	**ms-SG**	**ta-LK**	**ta-SG**	**tl-PH**	**zh-CN**	**zh-SG**	**Overall**
Accuracy (%)	0.00	0.00	20.00	14.29	0.00	14.29	28.57	14.29	0.00	100.00	14.29	15.54

Table 3: Accuracy by language/region on the evaluation data — meta-llama/Llama-3.2-1B-Instruct

Language	**ar-EG**	**ar-MA**	**ar-SA**	**bg-BG**	**el-GR**	**en-AU**	**en-GB**	**es-EC**	**es-ES**	**es-MX**	**eu-ES**	**fa-IR**
Accuracy (%)	7.60	2.20	6.00	10.00	9.40	31.60	35.00	5.20	19.40	15.60	1.00	16.20

Language	**fr-FR**	**ga-IE**	**id-ID**	**ja-JP**	**ko-KR**	**ms-SG**	**ta-LK**	**ta-SG**	**tl-PH**	**zh-CN**	**zh-SG**	**Overall**
Accuracy (%)	20.40	0.20	24.20	8.20	12.80	0.00	3.20	14.60	14.60	13.80	19.40	16.69

smaller size allows for greater efficiency when the pipeline is properly optimized.

On the evaluation set, the trend observed in development is maintained. The 1B model achieves an overall accuracy of 16.69%, slightly higher than in development (15.54%), but significantly lower than that of the 8B model. This behavior indicates that although the pipeline generalizes without direct training on BLEnD, the final performance strongly depends on the generative model's capacity. At a large scale, the 1B model faces greater difficulty handling the cultural and linguistic variability of the full dataset, particularly in languages where everyday knowledge is implicit or depends on local conventions that are not easily inferred from short web evidence.

Likewise, variability across languages persists in the evaluation set. While some languages show moderate performance, others, such as ga-IE or ta-LK, register low scores. This dispersion suggests that the approach's effectiveness depends not only on model size but also on the availability and clarity of the retrieved evidence. In languages where web snippets provide direct, explicit answers, the system tends to respond correctly; in contrast, when the information is implicit or requires additional inference, the lower-capacity model exhibits limitations. Overall, the results confirm that integrating web snippets through RAG constitutes a viable strategy for addressing multilingual cultural knowledge without direct supervised training. However, they also demonstrate that model capacity plays a decisive role in the final quality of the responses. The 8B model shows greater robustness and better utilization of retrieved context. In contrast, the 1B model, although computationally more efficient and scalable after pipeline optimization, imposes constraints in high-cultural-complexity scenarios.

6 Conclusion

This study demonstrates that evaluating everyday knowledge in multilingual and multicultural contexts requires architectures that explicitly integrate external evidence. The main contribution of this work lies in the systematic integration of web fragments (snippets) as contextual evidence and in reinforcing the LLM to identify, within those fragments, whether the correct answer is explicitly contained, validating its coherence before generating it. Additionally, iterative prompt adjustments were incorporated to validate the response and select the fragment with the highest semantic relevance to the question through embedding-based similarity. This scheme enabled the alignment of question, evidence, and final output within a controlled flow that reduces ambiguity.

The results confirm that parametric capacity significantly influences the integration of implicit cultural knowledge; however, they also show that, through optimization of the inference pipeline, it was possible to maintain the same performance of the lightweight model while substantially reducing execution time on the evaluation set, which contains a considerably larger volume of instances than the development set. Additional exploratory experiments using alternative configurations, including Few-shot without retrieval and standard RAG without synthetic data, showed lower performance, reinforcing the effectiveness of the proposed integrated pipeline. From a prospective standpoint, we propose designing comparative experiments using LLMs from different families and parametric scales beyond the LLaMA architecture to analyze how different pretraining.

Acknowledgments

The authors express their gratitude to the Call 933 "Training in National Doctorates with a Territorial, Ethnic and Gender Focus in the Framework of the Mission Policy — 2023" of the Ministry of Science, Technology and Innovation (Minciencia). In addition, we thank the team of the Artificial Intelligence Laboratory VerbaNex [2], affiliated with the UTB, for their contributions to this project.

References

Renad Al-Monef, Hassan Alhuzali, Nora Alturayeif, and Ashwag Alasmari. 2026. From words to proverbs: Evaluating llms' linguistic and cultural competence in saudi dialects with absher. *Alexandria Engineering Journal*, 137:25–41.

Danileth Almanza-González, Jairo Serrano, Juan Carlos Martínez-Santos, and Edwin Puertas. 2025. Prediction of human preferences and explanation generation with llm: An approach based on rag, few-shot learning, and auto-cot. In *Conference and Labs of the Evaluation Forum*.

Geleta Negasa Binegde and Huaping Zhang. 2026. Exploring cultural commonsense in multilingual large language models: A survey. *Information Systems*, 138:102649.

Cong Chen, Wei Qu, Si Su, Yukun Feng, and Tao Li. 2025. A comprehensive review of llm-based content moderation: advancements, challenges, and future directions. *Knowledge-Based Systems*, 330:114689.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *Preprint*, arXiv:2410.02677.

Jaroslaw Grobelny, Kacper Szymański, and Zuzanna Strozyk. 2025. Act as an expert in psychometry. the evaluation of large language models utility in psychological tests cross-cultural adaptations. *Acta Psychologica*, 261:105813.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. Care: Multilingual human preference learning for cultural awareness. *Preprint*, arXiv:2504.05154.

Moon-Kuen Mak and Tiejian Luo. 2025. A framework for evaluating cultural bias and historical misconceptions in llms outputs. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 5(3):100235.

[2] https://github.com/VerbaNexAI

Gonzalo Martínez, Marina Mayor-Rocher, Cris Pozo Huertas, Nina Melero, María Grandury, and Pedro Reviriego. 2025. Spanish is not just one: A dataset of spanish dialect recognition for llms. *Data in Brief*, 63:112088.

Anderson Morillo, Edwin Puertas, and Juan Carlos Martinez Santos. 2025. VerbaNexAI at SemEval-2025 task 3: Fact retrieval with Google snippets for LLM context filtering to identify hallucinations. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1534–1541, Vienna, Austria. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, V'ictor Guti'errez-Basulto, Yazm'in Ib'a nez Garc'ia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Jacopo Nudo, Mario Edoardo Pandolfo, Edoardo Loru, Mattia Samory, Matteo Cinelli, and Walter Quattrociocchi. 2026. Generative exaggeration in llm social agents: Consistency, bias, and toxicity. *Online Social Networks and Media*, 51:100344.

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. Gimmick – globally inclusive multimodal multitask cultural knowledge benchmarking. *Preprint*, arXiv:2502.13766.