UTD-HLTRI at SemEval-2026 Task 7: Bridging Cultural Knowledge Gaps in LLMs via Web-Augmented Context

Mohammad Marufur Rahman, Rakshitha Rao Ailneni and Sanda Harabagiu
Human Language Technology Research Institute
The University of Texas at Dallas
{dal620049, rxa220074, sanda}@utdallas.edu

Abstract

Though Large Language Models (LLMs) have been serving global users through a wide range of services, concerns remain regarding their cultural bias and misalignment with people of underrepresented communities. Increasing use of LLMs presents significant implications, as they have the potential to influence people's original values toward a certain cultural perspective. Cultural alignment of LLMs with culture-specific knowledge offers a suitable solution to this concern. In our participation in the Semeval-2026 Task 7 we considered a prompt engineering-based cultural alignment strategy to address the cultural knowledge gap in LLMs. Our approach achieved promising 86.34% accuracy for Japanese culture-relevant multiple-choice questions from the BLEND (Myung et al., 2024) benchmark.

1 Introduction

Culture is a way of life that is learned by members of a society and transferred from generation to generation. It plays a vital role in shaping the way a person thinks, acts, and participates in social activities (Tao et al., 2024). Cultural differences have significant influence on fundamental perception processes, causal attribution of behavior, and human judgment (Ji et al., 2000).

Large Language Models (LLMs) (OpenAI et al., 2024; Touvron et al., 2023) have demonstrated unprecedented performance in natural language understanding and generation (Razumovskaia et al., 2024; Zhao et al., 2025; Si et al., 2024). Despite their success, LLMs struggle to identify cultural variances and give culturally aligned responses. LLMs tend to reflect the cultural values of western, educated, industrialized, rich, and demographic societies while being unable to display the cultural values of societies with less digital footprint (Masoud et al., 2024; Sukiennik et al., 2025).

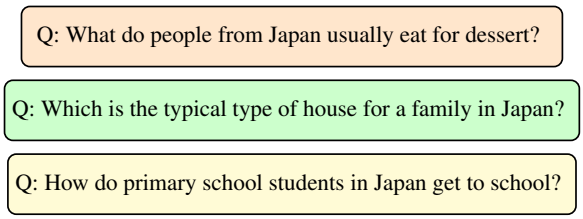

Figure 1: Example queries for Japanese culture from the BLEND benchmark.

To address this problem of cultural bias, recently BLEND (Myung et al., 2024), a hand-crafted benchmark designed to evaluate LLMs' everyday knowledge across diverse cultures and languages, was created. Initially, the BLEND benchmark covered 13 languages spoken in 16 different countries to create 500 socio-cultural question-answer pairs. These question-answer pairs addressed 6 cultural categories: (1) food; (2) sports; (3) family; (4) education; (5) holidays/celebrations/leisure; and (6) and work-life. Moreover, the set of questions and answers were created in two formats: (a) [question, short-answer] and (b) multiple-choice questions.

The Semeval-2026 Task 7 (Ghosh et al., 2026): *Everyday knowledge across different languages and cultures* expands the scope of BLEND to include 17 additional language–culture pairs (Ousidhoum et al., 2026). Furthermore, in order to reflect a model's ability to generalize to unseen, diverse cultural and linguistic contexts, the BLEND dataset was considered only for evaluation and was not used for training. SemEval-2026 Task 7 considered two possible tracks:
Track 1: Short Answer Questions (SAQ); and
Track 2: Multiple-Choice Questions (MCQ).
Our participation in the SemEval-2026 Task 7 considered queries associated with *Japanese culture* in Track 2 (MCQ). We selected the Japanese culture to perform a rigorous *stress test* of LLMs' nuanced cultural understanding, ranging from knowledge of complex Japanese social hierarchies to specific

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2657–2663

norms of everyday life in Japan. Japanese communication mostly relies on implicit contextual indicators and particular use of honorifics, which require the model to go beyond literal understandings toward proper cultural reasoning. Figure 1 shows example queries from the benchmark on Japanese culture. We were not interested in participating in Track 1, which involved searching for the correct answer to a question that could either be in English or in one of the target languages, while the answer had to be one of the target languages. This track involves (a) a retrieval component, aiming to pinpoint some candidate short answers, and (b) a verification component, akin to a reading comprehension task. We believe that the retrieval component dominates this track's processing for success, and it should rely on cross-language retrieval or multilingual retrieval. These research problems were not in our scope of interest. Instead, we were interested in Track 2, where questions are provided in English only, and the model needs to select one of four answer options, each representing a cultural perspective from a different country or region.

To address Track 2, we utilize a unique, simple prompting strategy where web-curated context augments information in the prompt to improve LLMs' cultural understanding. We experimented with prompting open-source and commercial LLMs using the testing BLEnD benchmark dataset, which consists of potentially unseen cultural and language-related knowledge from a wide range of sources (Myung et al., 2024). The performance reflected in the evaluation results was surprisingly good, given the simplicity of the method, as we relied on Retrieval Augmented Generation (RAG) (Lewis and et. al., 2020), which is well known to improve the reasoning of LLMs.

2 Related Work

Current research has revealed that LLMs show similar cultural biases and discrimination present in human society (Li et al., 2024). The ideal solution to this pressing issue is to refine LLMs' cultural awareness. Two main categories of approaches, prompt engineering and pre-tuning, are being adopted for this. Pre-training LLMs for different cultures requires large-scale datasets reflecting values of target cultures, followed by a fine-tuning process for better alignment (Chan et al., 2024; Abbasi and et. al., 2023). CultureLLM, proposed by Li et al. (2024), uses samples from the World Value Survey (WVS) and generates semantically equivalent data for fine-tuning culture-aware language models.

Prompt engineering-based approaches are cheap and require limited resources while being significantly effective (Wang et al., 2024). The study by Tao et al. (2024) evaluated cultural bias and alignment in popular LLMs by comparing responses from the WVS benchmark. The finding indicated that all models mirror cultural values of English-speaking and Protestant European countries. This study introduced *Cultural Prompting*, which indicates the cultural identity within the prompt. Another study proposed *Anthropological Prompting*, which enhances the cultural alignment of LLMs by injecting anthropological reasoning into the prompt (AlKhamissi et al., 2024). This method provides the model a structured persona indicating specific demographic details.

Obtaining culture-aware LLMs requires diverse and rich data containing culturally salient topics. To this endeavor Ziems et al. (2025) proposed *Culture Cartography*, a process where an LLM generates an annotation tree containing low-confidence queries for human annotators to provide culturally unique and relevant responses to fill the gap.

3 The BLEND Dataset

BLEND (Myung et al., 2024), is a hand-crafted benchmark containing 52.6K question-answer pairs from 16 regions and 13 different languages to address the lack of day-to-day cultural knowledge in LLMs. It aims to address the major gap in LLMs' understanding of everyday cultural knowledge, which is most of the time absent in common data sources like Wikipedia. Using both short-answer and multiple-choice questions, it evaluates models' understanding of mundane cultural details, such as typical birthday foods, common sports, local spices, etc.

4 The Method

Our method integrates structured prompting with external knowledge as context to improve cultural relevance and accuracy of LLMs' responses. Given a user query, the system constructs a composite prompt consisting of three elements: (i) a system-level instruction to guide the model to know its purpose, (ii) web-sourced contextual information to provide culturally unique background knowledge, and (iii) the user query. This prompt is then

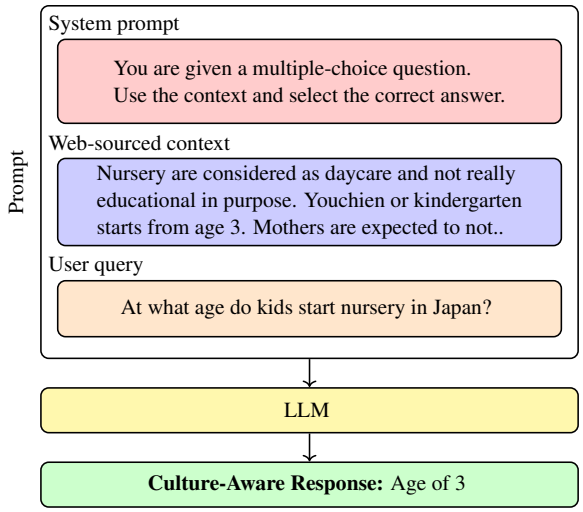

Figure 2: Overview of the UTD-HLTRI system that generates culturally grounded responses guided by Language Language Models (LLMs) when answering multiple-choice questions.

passed through the LLM, which utilizes both its pretrained knowledge and the injected new knowledge to generate a suitable response.

We propose a Retrieval Augmented Generation (RAG) framework, presented in Figure 2, that relies on external knowledge retrieved from the Web. To retrieve this knowledge and make it available to an LLM, we relied on the Tavily Search API[1] for dynamic context retrieval instead of index-based retrieval results.

The question q, is passed to the Tavily API, which returns a set of top-k relevant documents $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$, where each document contains textual content and metadata. Since retrieved web documents often contain irrelevant information, we apply a cleaning step to produce a refined set $\mathcal{D}' \subseteq \mathcal{D}$ by removing artifacts and enforcing minimum length constraints of each document. Each document $\mathcal{D}'$ is then segmented into smaller chunks (containing 200 tokens) to improve granularity and ensure compatibility with LLM token limits, resulting in a set of *document chunks*.

To further enhance relevance, we perform document chunk ranking by computing the cosine similarity score $s(q, c)$ between the question and each chunk using embeddings obtained from Sentence-BERT (Reimers and Gurevych, 2019) and select the top-3 most relevant chunks. These selected chunks are then concatenated to form a *structured context block* under a predefined token budget. Finally, the context is incorporated into a prompt template together with the original question q, and the resulting augmented prompt is passed to the LLM to generate the final response.

By explicitly grounding the most relevant knowledge for the question, as retrieved from the Internet, into the answer inference process, we mitigate the knowledge gap of the LLM and attempt to align it with cultural and social values. The system's ability to integrate external knowledge into the answer decision-making process for culturally sensitive questions is demonstrated by the final output, which is a succinct, context-aware response.

5 Experimental Setup

All experiments are conducted on a cloud-based service provided by Google Colab [2] with access to a single NVIDIA T4 GPU (16 GB VRAM), which is used for efficient model inference. For external knowledge retrieval, the Tavily Search API (Tavily AI, 2024) is used to obtain culturally relevant web-sourced contextual information.

Our evaluation includes open-source LLMs (*DeepSeek-R1- Distill-Qwen-32B* (DeepSeek-AI, 2025), *GPT-OSS-20B* (OpenAI, 2025)) from the Hugging Face ecosystem and a proprietary model, *GPT-5 mini*, accessed via the OpenAI API [3].

The LLMs are prompted by a unified template that integrates system instructions, contextual information, and user questions to maintain consistency across experiments. For the *DeepSeek-R1-Distill-Qwen-32B* model, temperature and top-p are set to 0.6 and 0.95, respectively, while *GPT-OSS-20B* and *GPT-5-mini* use the default parameter settings (temperature 1.0, top-p 1.0).

6 Results

For our participation in the SemEval-2026 Task 6, we have experimented with using both open-source and commercial LLMs. We have also relied on two different prompting strategies, namely with web-sourced context augmentation and without context. Our experiments targeted only Japanese culture relevant MCQ task in Track 2.

Overall, the inclusion of external context consistently improves performance of the HLTRI-UTD system across all LLMs used in the experiments, as shown in Table 1. Among the open-source models, the *DeepSeek-R1-Distill-Qwen-32B* model exhibits

[1] https://www.tavily.com/

[2] https://colab.research.google.com/

[3] https://developers.openai.com/api/docs/models/

Model	Without Context	With Context
DeepSeek-R1-Distill-Qwen-32B	60.97%	70.73%
GPT-OSS-20B	54.14%	56.09%
GPT5-mini	85.12%	86.34%*

Table 1: Performance of the HLTRI-UTD system in the Semeval-2026 Task 6, Track 2 (MCQ) (*indicates reported results).

a noticeable increase from 60.97% to 70.73%, indicating a substantial gain from context injection, while *GPT-OSS-20B* showed minor improvement. The commercial model *GPT-5 mini* achieves the highest performance in both settings. Its performance improves from 85.12% to 86.34%, which indicates that stronger proprietary models already possess rich and diverse internal knowledge.

Notably, the model with strong reasoning capability (*DeepSeek-R1-Distill-Qwen-32B*) benefits more significantly from context augmentation, suggesting external knowledge is more effective when combined with superior reasoning ability.

To showcase the impact of context injection on the *DeepSeek-R1-Distill-Qwen-32B* LLM performance in culturally grounded question answering we present two examples. In Example 1 (*without context*), a generic prompt with a question and possible options is passed to the LLM, leading to an incorrect response (Option B).

On the other hand, Example 2 (*with context*) contains additional web-sourced, culturally relevant information with the prompt, enabling the model to identify the most appropriate answer, "Chinese" (Option A). The comparison of Example 1 with Example 2 highlights the role of additional information of the specific culture in helping the LLM to better align its response toward that culture.

Example 1: Without Context

```
Prompt: You are an expert in
Japanese culture. You are given a
multiple-choice question.
Question: What is a popular
second language for secondary
school students in Japan?
Options: A. Chinese B. French
C. German D. Italian
Output only the option letter (A,
B, C, or D).
Response: B
Correct Answer: A
```

Example 2: With Context

```
Prompt: You are an expert in
Japanese culture. You are given a
multiple-choice question.
Use the context and select the
correct answer.
Question: What is a popular
second language for secondary
school students in Japan?
Context: There is a notable
history of the use of Kanbun
(Classical Chinese) as a language
of literature and diplomacy in
Japan, ...
Options: A. Chinese B. French
C. German D. Italian
Output only the option letter (A,
B, C, or D).
Response: A
Correct Answer: A
```

7 Discussion

The Impact of Context Quality: While the context used by RAG in the prompting helped many times to pinpoint the correct answer, we were interested to quantify the impact of the *quality* of the context. For this reason we have considered three possible scenarios for the context:
∘ *Scenario 1*∘ the context contains an explicit mention of the correct answer of the question – therefore it receives a relevance score of 2;
∘ *Scenario 2*∘ the context contains information that is relevant to the question, but it does not explicitly mention of the correct answer of the question – therefore it receives a relevance score of 1;
∘ *Scenario 3* ∘ the context contains information that is irrelevant to the question – therefore it receives a relevance score of 0.

Table 2 lists the relevance scores of the contexts

Category (No.)	Context Relevance Quality		
	0	1	2
Food (172)	43 (25%)	52 (30%)	77 (45%)
Sports (83)	39 (47%)	17 (20%)	27 (33%)
School (38)	0 (0%)	15 (39%)	23 (61%)
Holiday (55)	4 (7%)	40 (73%)	11 (20%)
City (29)	0 (0%)	19 (66%)	10 (34%)
Language (7)	1 (14%)	6 (86%)	0 (0%)
Lifestyle (7)	0 (0%)	1 (14%)	6 (86%)
Family (7)	0 (0%)	4 (57%)	3 (43%)
Work (5)	0 (0%)	0 (0%)	5 (100%)
Transportation (3)	0 (0%)	2 (67%)	1 (33%)
House (2)	0 (0%)	0 (0%)	2 (100%)
Pet (2)	0 (0%)	0 (0%)	2 (100%)

Table 2: Evaluation of the quality of retrieved context for each category of answers.

retrieved for questions that have various answer categories. The Table shows that for the *Food* answers, a large proportion (43%) of contexts were irrelevant. The second answer category has a significant proportion of irrelevant contexts pertaining to answers about *Sports*. These irrelevant contexts represent almost 50% of all irrelevant contexts that were added to the RAG prompts.

In general, the number of irrelevant contexts (= 167) is equal to the number of relevant contexts mentioning the answers (having relevance scores = 2), while the number of contexts of relevance score = 1 is 156. This indicates that in our experiments, the LLMs were successful with "reasoning" about the requests of the questions, given that 66.73% of the contexts did not mention the answer.

In addition, the information listed in Table 2, reveals that answers about *School*, and *Food* were selected correctly due to predominant contexts that mentioned the answer, in a proportion of 61%, and 45%, respectively, of all used contexts. Notably, answers about *Work*, *House*, and *Pet* benefited from contexts 100% of the times when the answer was mentioned. Conversely, answers about *Sports* faced the most challenges in pinpointing the correct answer, as about 47% of contexts were irrelevant. Moreover, answers about *Holiday* (73%), *City* (66%), and *Language* (86%) predominantly used relevant contexts that did not explicitly mention the answers.

Error Analysis: Table 3 details the number of erroneous answers across all the answer categories. The Table also lists the number of erroneous answers when using RAG by including the context or when ignoring the context (without RAG). Clearly, the largest number of erroneous answers was observed for questions addressing the culture of *Food*. For this specific question category, the context added to the prompt reduced the incorrect answers only by a count of 2, which indicates that the knowledge available in the Web-retrieved context is not sufficient. Considering cultural knowledge focusing on food, available from (Li and Zhang et. al., 2024; Winata and Hudi et. al., 2025) could have alleviated this problem.

Table 3 shows that the second category of erroneous answers concerned questions focusing on *Sports*. Reducing these errors would be possible by accessing cultural knowledge about sports, as considered in (Singh and Kumar et. al., 2025).

Interestingly, erroneous answers concerning the concept of *Holiday* were more numerous when

Category	Sample Count	Incorrect Answer w/ RAG	Incorrect Answer w/o RAG
Food	172	28	30
Sports	83	8	10
School	38	5	8
Holiday	55	8	6
City	29	3	3
Language	7	1	1
Lifestyle	7	0	0
Family	7	1	1
Work	5	1	1
Transportation	3	0	0
House	2	0	0
Pet	2	1	1

Table 3: Distribution of erroneous answers across answer categories with both with-RAG and without-RAG prompting.

context was considered than when it was ignored. This can be explained by the fact that, as shown in Table 2, most of the added contexts did not mention the answer, thus requiring the LLM to "reason" and most probably making the answer selection more difficult than when no context was used.

Table 3 also shows that for answers concerning *City*, *Family*, *Work* and *Pet* the same number of erroneous answers were obtained when the context was provided or ignored. While for the answers focusing on *Family* and *City* this can be explained by the predominance of contexts that do not mention the answer (as illustrated in Table 2), for the answers concerning *Work* and*Pet*, these errors are surprising, given that, according to the analysis illustrated in Table 2, the prompts for their questions were provided with contexts that mentioned the correct answer.

Performance Analysis of the HLTRI-UTD system: The analysis detailed in Table 4 highlights the fact that we have 4 performance cases:
◇ *Case 1*◇ representing the highest number of cases (81%), where the HLTRI-UTD system, both with RAG and without RAG, accurately pinpointed the correct answers.
◇ *Case 2*◇ where the HLTRI-UTD system successfully pinpointed 20 answers when using contexts, while the system that ignored the contexts missed those answers. Notably, 15 of these successful answers were due to contexts that mentioned the answers.
◇ *Case 3*◇ represents the performance of the HLTRI-UTD system when ignoring the contexts

Case	Context Quality			Total
	0	1	2	
1	79	134	121	334 (81%)
2	0	5	15	20 (5%)
3	2	10	3	15 (4%)
4	6	8	27	41 (10%)

Table 4: Performance cases of the HLTRI-UTD system when using RAG or ignoring it, considering the quality of the retrieved context.

for 15 questions proved to be more beneficial than using it, as in this case the contexts, while relevant, did not mention the answer, confusing the reasoning of the LLMs.

◇ *Case 4*◇ represents the performance of the HLTRI-UTD system in the 10% of the processed questions when incorrect answers were pinpointed, both when considering contexts (RAG) or when ignoring them. When further analyzing this case, we found that in a significant number of cases the contexts mentioned the answer, but the LLMs ignored it, leading us to believe that LLMs suffer from memorization.

To illustrate the memorization of LLMs, we present Example 3. Despite providing a context when the correct answer (*tea*) is mentioned and the answer is pinpointed by the HLTRI-UTD system, *green tea* is not even mentioned, but the LLM selects it. While the context mentions several specific varieties of tea, e.g., milk tea and lemon tea, it does not mention the selected answer, which was incorrect. This suggests that sometimes the context may confuse the LLM, guiding it to rely on its memorized answers.

Example 3: Performance Case 4

Question: Which is the most popular hot drink in Japan?
Options: A. green tea B. herbal tea C. hot chocolate D. tea
Context: While Coca-Cola is a global giant, its Japanese division is a leading manufacturer of tea,... featuring a blend of Darjeeling and Assam, alongside specialized varieties like milk tea and lemon tea. These diverse offerings, ranging from quirky convenience store finds to refined, limited-edition tea blends, ...
Response: A. green tea
Correct Answer: D. tea

8 Conclusion

Our participation in the SemEval-2026 Task 7 allowed us to analyze the quality of contexts used by RAG for bridging cultural knowledge gaps in LLMs. We found that one third of contexts mentioned the answers, thus greatly facilitating answer pinpointing for Multiple-Choice Questions (MCQ). We also found that another third of the contexts were irrelevant to the questions, while the last third of contexts were relevant but did not mention the answer. Therefore, augmenting prompts with web-curated contextual information has limited impact on the quality of selected answers, given the significant number of irrelevant contexts that were used.

However, we found that our naive RAG-based approach provided some cultural enhancement on the task performance, achieving an accuracy of 86.34% on Japanese culture-related MCQs from the BLEnD benchmark, a small improvement from the accuracy of 85.12% performed when no RAG was used. We also found that even when the context mentioned the answers, sometimes LLMs would rely on memorization, preventing them from pinpointing the correct answer. Finally, we argue that additional cultural knowledge for specific types of questions would probably be more helpful than some mere contexts retrieved from the Web.

9 Ethical Consideration

This study completely adheres to the ACL Code of Ethics[4]. We used trial data provided for the SemEval-2026 task 7 during the development and evaluation of the system performance. Since the proposed system fully relies on prompt engineering techniques, no dataset was used for further fine-tuning the model. While gathering contextual information, we retrieve only from publicly available open sources and carefully process it before adding it to the prompt.

References

Mohammad Amin Abbasi and Arash Ghafouri et. al. 2023. Persianllama: Towards building first persian large language model. *Preprint*, arXiv:2312.15713.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *Preprint*, arXiv:2402.13231.

[4]https://www.aclweb.org/portal/content/acl-code-ethics

Alex J. Chan, José Luis Redondo García, Fabrizio Silvestri, Colm O'Donnell, and Konstantina Palla. 2024. Enhancing content moderation with culturally-aware models. *Preprint*, arXiv:2312.02401.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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

Li-Jun Ji, Kaiping Peng, and Richard E Nisbett. 2000. Culture, control, and perception of relationships in the environment. *Journal of personality and social psychology*, 78(5):943.

Patrick Lewis and Ethan Perez et. al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Wenyan Li and Crystina Zhang et. al. 2024. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *Preprint*, arXiv:2309.12342.

Junho Myung, Nayeon Lee, and Yi et al. Zhou. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. volume 37, pages 78104–78146. Curran Associates, Inc.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et. al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nedjma Ousidhoum, Junho Myung, and Carla Perez-Almendros et al. 2026. 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.

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2024. Analyzing and adapting large language models for few-shot multilingual nlu: Are we there yet? *Preprint*, arXiv:2403.01929.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Preprint*, arXiv:2409.04109.

Punit Kumar Singh and Nishant Kumar et. al. 2025. Let's play across cultures: A large multilingual, multicultural benchmark for assessing language models' understanding of sports. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15194–15241, Suzhou, China. Association for Computational Linguistics.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *Preprint*, arXiv:2504.08863.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Tavily AI. 2024. Tavily: The search engine for ai agents.

Hugo Touvron, Louis Martin, and Kevin Stone et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *Preprint*, arXiv:2310.12481.

Genta Indra Winata and Frederikus Hudi et. al. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Caleb Ziems, William Barr Held, Jane Yu, Amir Goldberg, David Grusky, and Diyi Yang. 2025. Culture cartography: Mapping the landscape of cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1757, Suzhou, China. Association for Computational Linguistics.