CultureSynth: A Hierarchical Taxonomy-Guided and Retrieval-Augmented Framework for Cultural Question-Answer Synthesis

Xinyu Zhang*, **Pei Zhang, Shuang Luo, Jialong Tang, Yu Wan, Baosong Yang, Fei Huang**
Tongyi Lab, Alibaba Group Inc
{zxy440266, xiaoyi.zp, shuangluo.ls, tangjialong.tjl}@alibaba-inc.com
{wanyu.wy, yangbaosong.ybs, f.huang}@alibaba-inc.com

Abstract

Cultural competence, defined as the ability to understand and adapt to multicultural contexts, is increasingly vital for large language models (LLMs) in global environments. While several cultural benchmarks exist to assess LLMs' cultural competence, current evaluations suffer from fragmented taxonomies, domain specificity, and heavy reliance on manual data annotation. To address these limitations, we introduce CultureSynth, a novel framework comprising (1) a comprehensive hierarchical multilingual cultural taxonomy covering 12 primary and 130 secondary topics, and (2) a Retrieval-Augmented Generation (RAG)-based methodology leveraging factual knowledge to synthesize culturally relevant question-answer pairs. The CultureSynth-7 synthetic benchmark contains 19,360 entries and 4,149 manually verified entries across 7 languages. Evaluation of 14 prevalent LLMs of different sizes reveals clear performance stratification led by ChatGPT-4o-Latest and Qwen2.5-72B-Instruct. The results demonstrate that a 3B-parameter threshold is necessary for achieving basic cultural competence, models display varying architectural biases in knowledge processing, and significant geographic disparities exist across models. We believe that CultureSynth offers a scalable framework for developing culturally aware AI systems while reducing reliance on manual annotation[1].

1 Introduction

Cultural competence is a critical component of complex human emotional intelligence (Goleman, 1998), encompassing key abilities such as empathy, cognition, and adaptability in understanding and navigating cultural differences (Earley and Ang, 2003; Earley and Mosakowski, 2004). For individuals, cultural competence is not only demonstrated through awareness and understanding of cultural knowledge but also through the ability to adapt behavior and communication styles within multicultural contexts (Bennett, 2004; Hong, 2023). As the utilization of large language models (LLMs) in interactions, communications (Achiam et al., 2023; Yang et al., 2024), and workflows (Gao et al., 2024; Zhang et al., 2024) within global environments continue to grow, it is imperative to enhance the cultural competence of these models to optimize performance (Kasneci et al., 2023). This enhancement enables LLMs to more effectively interact with individuals from diverse cultural backgrounds worldwide (Havaldar et al., 2023; Liu et al., 2024a).

Several benchmarks have been developed to evaluate the cultural competence of LLMs, primarily using multiple-choice and question-answer (QA) generation formats (Pawar et al., 2024; Minaee et al., 2024). Multiple-choice benchmarks such as BLEnD (Myung et al., 2024), CommonsenseQA (Putri et al., 2024), and MMLU (Hendrycks et al., 2021) gather culturally specific questions and manually annotate the correct and incorrect options. Generation benchmarks like CaLMQA (Arora et al., 2024), NativQA (Hasan et al., 2024), and CultureBank (Shi et al., 2024) compile naturally occurring questions from Wikipedia and community web forums (Fan et al., 2019), in addition to human-written questions annotated by native speakers. Other benchmarks, such as PRISM (Kirk et al., 2024), cluster and filter culturally specific queries from real-world inquiries sourced via an English-speaking participant crowdwork platform. Meanwhile, CulturePark (Li et al., 2024) employs a different approach, creating a cultural benchmark through dialogues between LLM-based multi-agent communication.

However, existing cultural benchmarks encounter two primary challenges. First, cultural competence spans a wide range of domains (e.g., religion, social customs, law), yet most current

*Corresponding author
[1]Benchmark is available at https://github.com/Eyr3/CultureSynth.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10448–10467
November 4-9, 2025

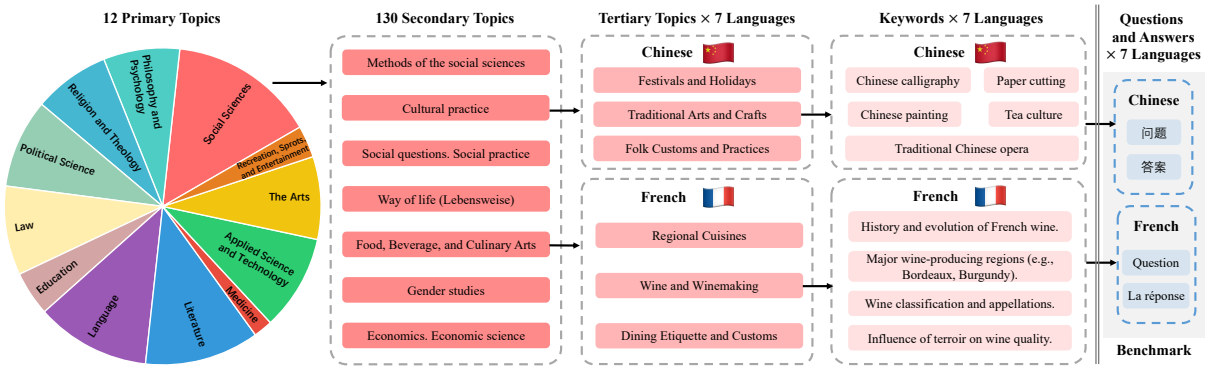


Figure 1: CultureSynth consists of a hierarchically structured multilingual cultural taxonomy (left of double line) and a RAG-based question-answer synthesis methodology (right), which together generate the benchmark.

benchmarks address only specific discrete cultural topics and lack a systematic taxonomy (Myung et al., 2024; Chiu et al., 2024a). This gap hinders comprehensive and systematic analyses of LLMs' cultural capabilities across varied domains. Second, these benchmarks mainly rely on existing forums and manual annotation (Arora et al., 2024; Shi et al., 2024). While such benchmark construction methods ensure quality, they are resource-intensive, requiring substantial human labor and material costs, and cannot guarantee comprehensiveness. Moreover, as the consumption of training data accelerates, LLMs increasingly face data shortages (Villalobos et al., 2024). Consequently, employing automated methods to synthesize high-quality data becomes essential (Long et al., 2024).

To address the aforementioned challenges, we introduce CultureSynth, as illustrated in Figure 1, a novel cultural framework featuring an automated data synthesis methodology. We propose the first hierarchically structured multilingual cultural taxonomy to tackle the issue of dispersed and unsystematic cultural topics. This framework integrates library classifications from five countries and regions to create a universal cultural taxonomy comprising 12 primary topics and 130 secondary topics (see Table 10), encompassing diverse global cultures. Additionally, we employ an expert role-playing LLM to extend over a thousand deep, country-specific cultural topics for each language. To further construct a reliable cultural benchmark while minimizing manual annotation, we propose a Retrieval-Augmented Generation (RAG)-based methodology for question-answer pair synthesis. Leveraging carefully vetted factual knowledge that ensures data authenticity, we employ expert role-playing and instruction-following prompts to generate safe, clear, and culturally relevant questions with high-quality answers.

The CultureSynth-7 benchmark encompasses 7 languages (Arabic [ar], Spanish [es], French [fr], Japanese [ja], Korean [ko], Portuguese [pt], and Chinese [zh]), totaling 19,360 QA pairs with thousands of cultural keywords per language. Native speaker annotation of a representative subset validated the benchmark quality, achieving 95.8% question clarity, 83.5% cultural relevance, and 98.8% answer quality, with no safety concerns identified.

Using CultureSynth, we conduct an extensive evaluation across 14 widely used language models, including GPT-4o (Achiam et al., 2023), Claude-3.5 (Anthropic, 2024), Qwen2.5 (Yang et al., 2024), Llama 3 (Dubey et al., 2024), and the Mistral (Jiang et al., 2024) series. The evaluation reveals a clear performance hierarchy among LLMs, i.e., ChatGPT-4o-Latest $>$ Qwen2.5-72B $>$ Claude-3.5-Sonnet $>$ others. We identify a 3B-parameter threshold for basic cultural competence, below which models default to native-language functionality. Model architecture plays a crucial role: mixture-of-experts architectures excel in cultural knowledge retrieval, while dense transformers perform better in long-context scenarios. Our evaluation also uncovers distinct geographic and domain-specific biases, particularly GPT-4o's performance gaps in East Asian contexts and Claude-3.5's limitations in Arabic and Korean language processing.

2 Related Work

2.1 Cultural Competence Benchmark

Existing cultural competence benchmarks for LLMs primarily fall into two categories: multiple-choice questions (MCQ) and question-answer generation (Pawar et al., 2024; Minaee et al., 2024). In the MCQ category, MMLU (Hendrycks et al., 2021) covers 57 subjects across humanities, so-

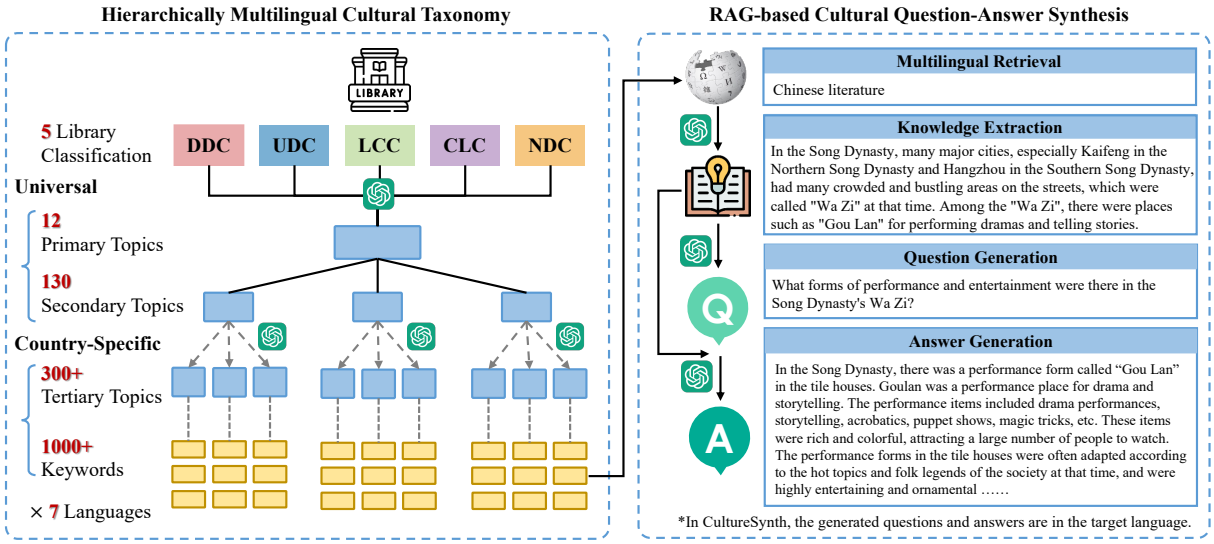


Figure 2: Overview of CultureSynth: (1) a hierarchically multilingual cultural taxonomy (see Section 3.1), and (2) an RAG-based question-answer synthesis methodology (see Section 3.2). The cultural taxonomy integrates five library classifications from different cultural backgrounds to establish primary topics, which are then hierarchically expanded using LLMs to create comprehensive and detailed cultural tertiary topics and keywords. The synthesis method leverages cultural keywords to automatically perform multilingual knowledge retrieval, extract factual knowledge, and generate question-answer pairs through LLMs.

cial sciences, and STEM, while BLEnD (Myung et al., 2024) evaluates everyday cultural knowledge across diverse languages. CulturalBench (Chiu et al., 2024b) features 1,227 human-verified questions assessing cultural knowledge across 45 regions and 17 topics. Additionally, NORMAD (Rao et al., 2024) specifically analyzes LLMs' behavior under varying socio-cultural contexts.

Question-answer generation benchmarks better evaluate cultural behavioral alignment and interaction styles. CALMQA (Arora et al., 2024) collects questions from community forums across 12 topics with native speaker validation, while NativQA (Hasan et al., 2024) employs Google's "People also ask" feature to generate queries across 18 topics efficiently. CultureBank (Shi et al., 2024) provides 23,000 structured cultural data entries from social media platforms using 11 predefined fields, enabling more flexible interpretation. PRISM (Kirk et al., 2024) contributes preference data from diverse annotators across 23 topic clusters. While these efforts advance cultural competence evaluation, our work focuses on expanding domain coverage and reducing data generation costs through automated synthesis methods.

2.2 Synthesizing Data for LLMs

The growth of LLMs has increased demands for high-quality training data, particularly in culturally sensitive domains where multilingual resources are scarce. Synthetic data generation has emerged as a promising solution for addressing scalability and domain-specific adaptability (Liu et al., 2024b). Recent studies by Gan and Liu (2024) show that synthetic data can enhance model generalization through post-training information gain. However, concerns about synthetic data reliability persist due to LLM hallucinations (Ji et al., 2023). RAG frameworks have effectively addressed this by anchoring outputs in verified knowledge bases (Huang and Huang, 2024), improving answer faithfulness by over 30% (Zhao et al., 2024). Additionally, role-playing prompts and instruction tuning have proven effective in refining synthetic outputs, with Tang et al. (2024) demonstrating improved performance through multi-agent systems. Our work builds on these advances by combining hierarchical cultural taxonomies with retrieval-augmented synthesis to produce culturally nuanced and verifiable data.

3 CultureSynth

This section presents the complete question-answer generation process of the CultureSynth framework, as illustrated in Figure 2.

3.1 Hierarchically Cultural Taxonomy

The taxonomy comprises two tiers: universal cultural topics and country-specific cultural topics. The first tier establishes cross-country analytical topics through standardized cultural dimensions,

while the second tier incorporates localized cultural elements corresponding to specific linguistic contexts. This hierarchical structure offers dual advantages: the universal tier ensures comprehensive analysis of cultural competence by preventing analytical redundancy or omissions caused by cross-cultural discrepancies, whereas the country-specific tier preserves linguistic and cultural nuances essential for granular cultural interpretation.

Universal Cultural Topics. We integrate five multinational library classification systems, i.e., Dewey Decimal Classification (DDC), Universal Decimal Classification (UDC), Library of Congress Classification (LCC), Chinese Library Classification (CLC), and Nippon Decimal Classification (NDC), to establish a cultural framework comprising primary and secondary topics. This cross-cultural synthesis ensures comprehensive coverage of universal cultural dimensions across diverse national contexts. Then through a combination of LLM-powered analysis and expert curation, we ultimately derived 12 primary topics: Social Sciences (SS), Philosophy and Psychology (PP), Religion and Theology (RT), Political Science (PS), Law (LAW), Education (EDU), Language (LAN), Literature (LIT), Medicine (MED), Applied Sciences and Technology (AST), Arts (ART), and Recreation, Sports, and Entertainment (RSE), along with 130 secondary topics, as listed in Table 10.

Country-Specific Cultural Topics. Based on primary and secondary topics, we designed a role-playing-based prompt to guide the LLM (e.g., we use GPT-4 (Achiam et al., 2023)) in generating more fine-grained cultural tertiary topics and keywords for each country (or language), as shown in Figure 10. We observe that augmenting the prompt with the instruction "*Please don't be lazy and answer this question in depth from a local's perspective*" leads to more comprehensive and culturally localized responses. Ultimately, our approach facilitates the construction of more than 300 fine-grained tertiary topics, with an average of over 1000 keywords per country/ language[2].

3.2 RAG-based Multilingual Cultural Question-Answer synthesis

The method consists of four main steps: First, we perform multilingual retrieval to obtain authentic and reliable information pages corresponding to the given keywords, followed by extracting key knowledge points. Based on these verified knowledge points, we then automatically generate high-quality questions and answers. Throughout the generation process, prompt optimization is applied to ensure the logical coherence of the questions and the comprehensiveness and depth of the answers.

Step1: Multilingual Retrieval. Given a target keyword, we first translate it into country-specific languages and retrieve corresponding pages from reliable sources (e.g., Wikipedia) in both English and the target language. We employ LLMs with prompts detailed in Figures 11 and 12 to assess the retrieved pages for keyword relevance and culture-specific content. Pages failing to meet keyword relevance or cultural content requirements are excluded from subsequent knowledge extraction.

Step2: Knowledge Extraction. For verified culturally significant pages, we employ a prompt detailed in Figures 13 and 14 that enables LLMs to systematically extract multiple knowledge points using a standardized JSON format. This stage ensures structural consistency while preserving cultural specificity in the extracted information.

Step3: Question Generation. For each knowledge point, we use LLMs with the prompt in Figures 15 and 16 to generate questions in the target language, following three key guidelines: avoiding offensive or discriminatory content, eliminating anaphoric references to ensure context clarity, and maintaining cultural and linguistic appropriateness.

Step4: Answer Generation. To construct comprehensive answers, we prompt the LLM to act as a domain expert, instructing it to incorporate cultural knowledge and provide detailed responses in the target language that address multiple aspects of each question (see Figures 17). Similar to questions, answers must avoid demonstrative pronouns to maintain contextual independence.

4 CultureSynth-7 Benchmark

4.1 Data Annotation

To validate the quality of synthetic data in CultureSynth-7, we randomly sample 120 QA pairs per language. The assessment involves two native speakers (annotators A and B) assigned to each language, totaling 14 annotators. To ensure annotation consistency and reliability, we implement a comprehensive training protocol that includes guideline alignment and trial annotation.

[2]Tertiary topics and keywords are available at https://github.com/Eyr3/CultureSynth.

Table 1: The acceptance rates (%) demonstrate inter-annotator consistency and high data quality across 7 languages.

	Annotator	ar	zh	fr	ja	ko	pt	es	Average
Question Clarity Rate (score: 1)	Annotator A	85.8	98.3	85.8	95.0	98.3	85.0	93.3	91.6
	Annotator B	100.0	99.2	100.0	100.0	100.0	100.0	100.0	99.9
	Average	92.9	98.8	92.9	97.5	99.2	92.5	96.7	95.8
Cultural Relevance Rate (score: 1)	Annotator A	72.8	92.4	87.6	90.4	67.8	92.2	98.2	85.9
	Annotator B	70.0	92.4	75.0	87.5	66.7	90.0	91.7	81.9
	Average	71.4	92.4	81.3	89.0	67.3	91.1	95.0	83.9
High-quality Answer Rate (score: ≥ 4)	Annotator A	100.0	95.8	99.0	98.2	100.0	99.0	97.3	98.5
	Annotator B	100.0	100.0	98.3	97.5	100.0	99.2	98.3	99.0
	Average	100.0	97.9	98.7	97.9	100.0	99.1	97.8	98.8

Annotation Guideline. This guideline outlines the criteria for assessing QA pairs through three dimensions (refer to Appendix A.1 for details).

- Question Clarity & Safety: Determine if the question is self-contained and adheres to universal ethical standards (Score: 1 for yes, 0 for no).
- Cultural Relevance: Identify cultural distinctiveness based on two dimensions. Score 1 if the question shows either cultural variance (answers differ across cultures/languages) or cultural specificity (contains culture-specific elements such as regional traditions); score 0 otherwise.
- Answer Quality: Access the quality of the answers relative to the reference knowledge (i.e., Wikipedia) using the 5-point scale, with scores of ≥ 4 being high quality.

Annotation Results. Table 1 presents the acceptance rates of three assessment dimensions across 7 languages. In the cultural relevance assessment, most languages show minimal differences between annotators, with some languages achieving high agreement (e.g., zh and ja). Similarly, for the high-quality answer rate, the variations between annotators are typically less than 2%, with several languages showing strong consistency (e.g., ar and ko both at 100%). While we observe annotation variations of 14.2% to 15% in question clarity rate for ar, fr, and pt, these discrepancies can be primarily attributed to different interpretations of question self-consistency between annotators. Nevertheless, it's noteworthy that the lowest question clarity rate still achieves 85% for these three languages.

Regarding data quality, all three dimensions show strong performance across languages. The question clarity rate achieves a high overall average of 95.41%, with particularly strong results in zh and ko. The cultural relevance rate, while slightly lower, maintains a robust average of 83.91%, with es and zh performing notably well. Notably, the high-quality answer rate remains consistently strong across all languages, averaging 98.76%, with ar and ko reaching 100% and other languages scoring above 97%. These consistently high scores across all three metrics strongly validate the effectiveness of our data construction approach.

Regarding safety considerations, our question safety assessment in our annotation reveals negligible safety concerns in the generated questions, primarily due to two factors: the use of Wikipedia as the source material, which inherently limits safety risks, and the employment of LLMs with built-in safe mechanisms for question generation.

4.2 Data Release and Analysis

From over 300 tertiary topics per language in Section 3.1, we generate 19,360 QA pairs, split into two disjoint subsets with non-overlapping tertiary topics: a 4,149-example CultureSynth-7-mini[3] set for model development and resource-constrained users, and a 15,211-example CultureSynth-7-max for additional evaluation or model fine-tuning. To maintain data integrity, CultureSynth-7-max sets remain private, and can be made available upon request via email. To achieve representative sampling, we randomly selected 20 tertiary topics from each primary topic category for CultureSynth-7, including all associated QA pairs. This approach ensures zero overlaps in tertiary topics between CultureSynth-7 and CultureSynth-7-max, effectively preventing data contamination.

Table 2 presents statistics for the total of 19,360 QA pairs. Question lengths average 15 and 29 words, with language-specific variations, while answer lengths range from 38-98 words on average, reaching 560 words in Chinese. Figure 3 shows cultural topic distribution of these QA pairs, revealing distinct patterns across languages, with Arabic emphasizing social sciences and Chinese focusing

[3]We use *CultureSynth-7* to identify *CultureSynth-7-mini*, which are manually verified and released publicly.

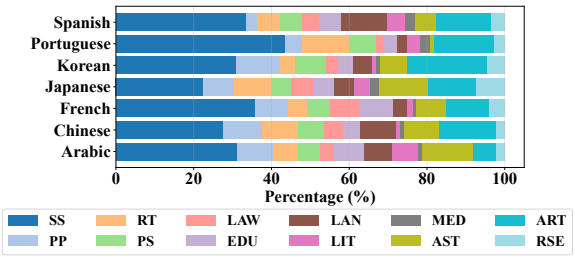


Figure 3: Topic distribution for total 19,360 QA pairs.

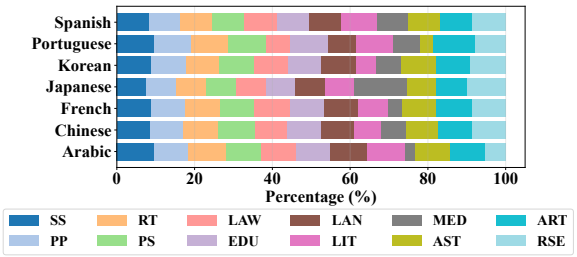


Figure 4: Balanced topic distribution in CultureSynth-7.

Table 2: Key statistics of the total 19,360 QA pairs in 7 languages from different language families.

Language	Code	Languag family	Total QA	Percentage	Question length		Answer length		
					Avg	Max	Avg	Max	Std
Spanish	es	Indo-European	2170	11.2%	26	77	90	458	110
French	fr	Indo-European	2929	15.1%	25	67	66	417	90
Portuguese	pt	Indo-European	1529	7.9%	25	77	98	451	110
Arabic	ar	Afro-Asiatic	1416	7.3%	19	52	87	310	88
Chinese	zh	Sino-Tibetan	4172	21.5%	22	86	62	560	102
Japanese	ja	Japonic	3798	19.6%	29	112	77	516	109
Korean	ko	Koreanic	3346	17.3%	15	51	38	310	56

on arts and applied sciences. For evaluation, we use the balanced and manually verified CultureSynth-7 benchmark (containing 50 to 54 samples per language and topic, see Figure 4), with all questions annotated for clarity and cultural relevance, and answers verified for high quality.

5 Experiments

5.1 Experiment Protocol

Given the subjective nature of CultureSynth and its substantial size (4,149 questions), we employ an LLM-based automatic evaluation approach (Zheng et al., 2023; Liu et al., 2023; Chiang et al., 2024). This involves a pairwise comparison using a high-performance LLM as the judge LLM. For each question in CultureSynth, we obtain responses from a moderate-performance baseline model and a target model. The judge LLM then compares both responses against the reference answer, assigning scores as follows: 1 if the target model performs better, -1 if the baseline model performs better, and 0 if performances are comparable. As referenced in (Chiang et al., 2024), the specific prompt is shown in Figure 18. After completing this process for all questions, we calculate the net win rate of the target model against the baseline.

5.2 Experimental Setup

We select ChatGPT-4o-Latest (version gpt-4o-2024-05-13) (Achiam et al., 2023) and Qwen3-32B-Think (Yang et al., 2025) as our judge LLMs, Gemma2-27B-it (Team et al., 2024) as our baseline model, and then evaluate the following target models: GPT-4o series (Achiam et al., 2023), Claude-3.5-Sonnet (Anthropic, 2024), Qwen2.5 series (Yang et al., 2024), Llama 3 series (Dubey et al., 2024), and Mixtral series (Jiang et al., 2024). Responses of proprietary models are obtained through their respective APIs.

Our evaluation metric is the net win rate:

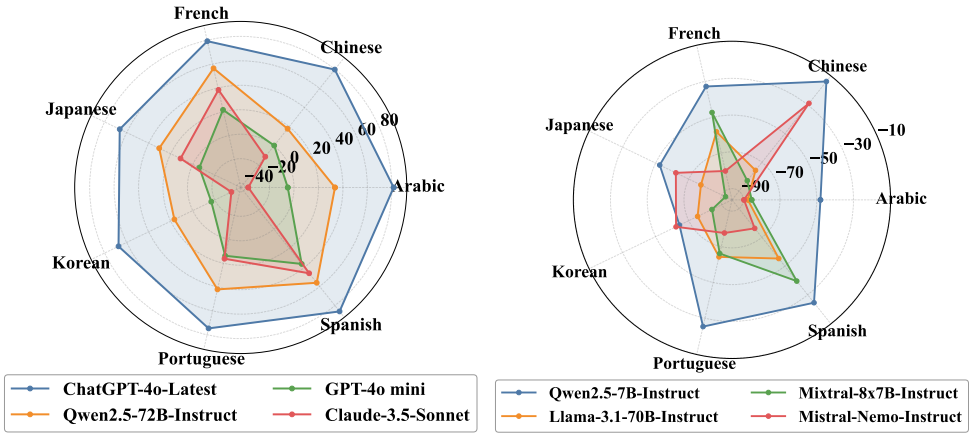
$$\text{Net win rate} = (N_{\text{target\_wins}} - N_{\text{baseline\_wins}})/N_{\text{total}},$$

where $N_{\text{target\_wins}}$ denotes the number of cases where the score is 1 (indicating the target model is superior), $N_{\text{baseline\_wins}}$ represents the number of cases with a score of -1 (indicating the baseline model is superior), and N_{total} is the total number of comparisons.

5.3 Experimental Results

Based on the evaluation results across seven languages and 12 primary topics presented in Table 3, our CultureSynth benchmark reveals clear performance stratification among different models. The overall ranking of cultural competence follows this order: ChatGPT-4o-Latest $>$ Qwen2.5-72B-Instruct $>$ GPT-4o mini $\approx$ Claude-3.5-Sonnet $>$ Qwen2.5-7B-Instruct $>$ Mistral-Nemo-Instruct-2407 $\approx$ Llama-3.1-70B-Instruct $\approx$ Mixtral-8x7B-Instruct-v0.1 $>$ Qwen2.5-3B-Instruct $>$ Llama-3.1-8B-Instruct $>$ Mistral-7B-Instruct-v0.3 $>$ Qwen2.5-0.5B-Instruc $\approx$ Llama-3.2-3B-Instruct $\approx$ Llama-3.2-1B-Instruct.

Cross-lingual Performance Analysis. As illustrated in Figure 5a), among first-tier models, ChatGPT-4o-Latest excels across most languages



(a) First-tier cultural competence models (b) Second-tier cultural competence models

Figure 5: Net win rates (%) of different models compared to the baseline model across languages.

Table 3: Net win rates (%) of models across languages, relative to the Gemma2-27B-it baseline under the ChatGPT-4o-Latest judge. Blue and yellow indicate the best and second-best performing models, respectively.

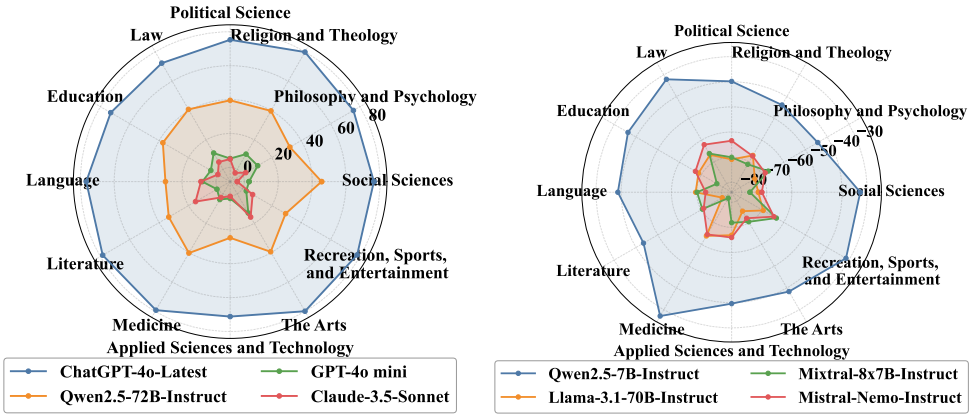
Model	Arabic	Chinese	French	Japanese	Korean	Portuguese	Spanish	Weighted Average
ChatGPT-4o-Latest	81.48	79.77	79.15	66.27	67.36	74.62	86.06	76.31
GPT-4o mini	-5.11	0.33	21.54	-6.00	-16.84	13.72	36.38	6.31
Claude-3.5-Sonnet	-37.39	-11.37	38.29	11.09	-35.07	16.17	46.15	4.58
Qwen2.5-72B-Instruct	33.51	17.89	56.41	30.43	16.67	41.92	56.09	36.13
Qwen2.5-7B-Instruct	-47.97	-13.71	-32.82	-52.47	-64.41	-25.38	-24.52	-37.48
Qwen2.5-3B-Instruct	-86.77	-32.27	-83.59	-87.41	-83.16	-74.81	-73.08	-74.48
Qwen2.5-0.5B-Instruct	-99.82	-84.62	-97.44	-97.45	-96.70	-96.43	-96.47	-95.54
Llama-3.1-70B-Instruct	-87.48	-75.59	-58.12	-77.36	-75.35	-64.29	-55.29	-70.50
Llama-3.1-8B-Instruct	-92.77	-89.13	-73.33	-90.85	-90.62	-74.81	-75.96	-84.07
Llama-3.2-3B-Instruct	-99.82	-95.32	-97.09	-98.20	-98.44	-90.41	-93.11	-96.12
Llama-3.2-1B-Instruct	-99.47	-93.48	-96.75	-98.20	-98.26	-94.36	-94.39	-96.43
Mixtral-8x7B-Instruct-v0.1	-85.36	-82.78	-47.35	-92.20	-84.03	-66.17	-39.58	-71.20
Mistral-Nemo-Instruct-2407	-89.59	-28.93	-80.00	-62.22	-62.33	-77.63	-76.28	-67.77
Mistral-7B-Instruct-v0.3	-98.77	-88.96	-86.15	-95.20	-94.27	-81.20	-83.97	-89.90

but shows limitations in East Asian cultural contexts (Japanese/ Korean). GPT-4o mini exhibits similar language performance patterns to ChatGPT-4o-Latest. Qwen2.5-72B-Instruct demonstrates strong generalization in medium-resource languages like Portuguese and Japanese. Claude-3.5-Sonnet performs well in French and Spanish but struggles with Arabic and Korean. Second-tier models, though performing below baseline, show distinct language strengths (as shown in Figure 5b): Qwen2.5-7B-Instruct excels in Chinese, while Llama-3.1-70B-Instruct and Mixtral-8x7B-Instruct perform better in French and Spanish. Mistral-Nemo-Instruct shows strength in Chinese compared to its performance in other languages.

As shown in Figure 7, model cultural competence correLatest positively with size across all languages. Models below 3B parameters experience a significant drop in language capabilities (see Figure 8), only managing to handle culture-related QA in specific native language scenarios.

Topic-wise Performance Analysis. According to Table 4, different models exhibit significant variations in their cultural topic comprehension capabilities. Cultural competence is not solely determined by the total number of parameters but is closely related to the organization of domain knowledge, the weighting of cultural data during training, and the model architectures.

For first-tier models (as shown in Figure 6a), GPT-4o-Latest demonstrates exceptional cultural literacy, particularly in topics requiring deep cultural context understanding (e.g., literature and medicine). Qwen2.5-72B-Instruct achieves 50-60% of GPT-4o-Latest's performance in practical domains like social sciences, law, and medicine, but shows relative weakness in creative fields such as language and arts. Its capability profile indicates stronger performance in structured knowledge applications, reflecting its training emphasis on professional content. Claude-3.5-Sonnet displays uneven cultural competence distribution, showing cer-



(a) First-tier cultural competence models (b) Second-tier cultural competence models

Figure 6: Net win rates (%) of different models compared to the baseline model across cultural topics.

Table 4: Net win rates (%) of models across cultural topics, relative to the Gemma2-27B-it baseline under the ChatGPT-4o-Latest judge. Blue and yellow indicate the best and second-best performing models, respectively.

Model	Social Sciences	Philosophy and Psychology	Religion and Theology	Political Science	Law	Education	Language	Literature	Medicine	Applied Sciences and Technology	Arts	Recreation, Sports, and Entertainment
ChatGPT-4o-Latest	76.39	75.56	79.84	75.27	72.30	72.96	76.50	78.42	79.17	71.21	79.89	78.10
GPT-4o mini	2.78	10.39	10.35	5.22	11.08	4.79	8.02	0.61	3.82	1.55	13.04	2.59
Claude-3.5-Sonnet	-4.17	2.25	-2.45	4.95	4.96	0.00	8.88	15.20	2.43	0.31	15.76	6.92
Qwen2.5-72B-Instruct	45.56	32.30	39.78	39.56	40.82	37.46	29.80	33.43	40.28	24.77	39.40	29.39
Qwen2.5-7B-Instruct	-32.78	-44.38	-43.87	-39.84	-32.07	-36.34	-38.68	-43.47	-27.08	-39.63	-38.32	-31.41
Qwen2.5-3B-Instruct	-68.89	-76.69	-80.11	-75.27	-67.64	-76.90	-73.64	-81.76	-75.69	-70.28	-71.74	-75.22
Qwen2.5-0.5B-Instruct	-98.33	-96.35	-97.28	-96.43	-93.59	-96.34	-95.13	-96.66	-94.10	-96.90	-94.02	-91.07
Llama-3.1-70B-Instruct	-73.33	-73.31	-67.03	-70.88	-66.76	-68.73	-69.63	-79.64	-63.89	-66.87	-75.27	-69.45
Llama-3.1-8B-Instruct	-85.56	-82.30	-88.01	-82.42	-84.55	-82.82	-85.67	-86.63	-81.25	-85.45	-80.71	-83.29
Llama-3.2-3B-Instruct	-97.22	-96.91	-96.73	-95.05	-94.17	-98.03	-96.85	-97.87	-95.83	-98.45	-92.66	-93.95
Llama-3.2-1B-Instruct	-97.50	-96.63	-96.46	-93.13	-96.79	-97.75	-97.71	-97.26	-96.53	-97.52	-95.65	-94.52
Mixtral-8x7B-Instruct-v0.1	-76.67	-67.13	-71.12	-70.05	-66.18	-77.18	-70.20	-70.52	-81.25	-71.83	-70.38	-63.40
Mistral-Nemo-Instruct-2407	-71.94	-68.54	-67.30	-63.46	-62.10	-67.32	-73.64	-71.12	-64.58	-65.94	-72.01	-64.55
Mistral-7B-Instruct-v0.3	-90.83	-92.70	-88.83	-90.93	-88.05	-93.52	-91.12	-92.10	-88.19	-88.85	-85.60	-87.90
Average	-48.04	-48.17	-47.80	-46.60	-44.48	-48.55	-47.79	-49.24	-45.91	-48.85	-44.88	-45.55

tain advantages in generative tasks (e.g., literature and arts) while demonstrating weaknesses in domains requiring objective knowledge and reasoning (e.g., social sciences and religious theology).

For second-tier models (as shown in Figure 6b), Qwen2.5-7B-Instruct and Llama-3.1-70B-Instruct perform slightly better in professional domains like medicine compared to other topics. Qwen2.5-7B-Instruct maintains baseline capabilities in highly structured fields like medicine and law. However, these models show significant deficiencies in humanities, such as philosophy, psychology, and literature, with performance declining by over 40%. Llama-3.1-70B-Instruct's superior performance in medicine reveals its strength in Western knowledge systems, while its poor performance in the humanities exposes cultural adaptation deficiencies. Comparing results between Mixtral-8x7B-Instruct and Mistral-Nemo-Instruct reveals that the mixture-of-experts architecture demonstrates effective routing for discrete knowledge points, whereas the dense transformer with a longer context window shows higher performance in political science and law domains requiring long-range textual dependencies.

For extremely small-scale models in the third and fourth-tiers (as shown in Figure 9), cultural comprehension capabilities essentially collapse when parameters are reduced to the 1B level.

Model-wise Cultural Competence Analysis. We conduct a model-wise analysis of cultural competence for two representative models, the best-performing ChatGPT-4o-Latest and its smaller counterpart GPT-4o mini, to illustrate performance variation across cultural and linguistic contexts.

Table 5 shows that ChatGPT-4o-Latest delivers strong and balanced performance across languages and topics, with notable strengths in Spanish, Arabic, and Chinese contexts. However, its relatively lower scores in Philosophy and Psychology (ja) and Applied Sciences and Technology (ko) highlight persistent challenges in certain language–domain pairs. In contrast, Table 6 indicates that GPT-4o

Table 5: Net win rates (%) of ChatGPT-4o-Latest across languages and cultural topics (red: below 60%).

	Social Sciences	Philosophy and Psychology	Religion and Theology	Political Science	Law	Education	Language	Literature	Medicine	Applied Sciences and Technology	Arts	Recreation, Sports, and Entertainment
ar	83.33	76.47	85.19	88.24	78.43	82.35	88.89	81.48	73.33	76.47	80.39	73.33
zh	88.46	86.00	79.63	84.21	64.00	76.47	74.51	90.24	76.92	72.00	82.35	82.69
fr	72.55	76.47	74.07	64.71	74.07	84.31	84.31	80.00	90.48	88.24	81.48	86.27
ja	78.00	52.94	73.58	68.63	64.71	54.00	64.71	62.75	71.11	64.00	77.36	60.61
ko	68.63	68.63	76.00	62.75	60.00	56.00	59.62	73.33	75.00	53.85	82.35	76.92
es	80.39	92.16	84.31	84.31	87.04	82.35	88.24	84.21	98.04	76.47	80.39	94.44
pt	62.75	76.47	86.27	73.08	78.79	74.51	74.36	76.47	75.00	61.11	75.44	73.81

Table 6: Net win rates (%) of GPT-4o mini across languages and cultural topics (red: below -20%).

	Social Sciences	Philosophy and Psychology	Religion and Theology	Political Science	Law	Education	Language	Literature	Medicine	Applied Sciences and Technology	Arts	Recreation, Sports, and Entertainment
ar	3.70	-15.69	1.85	15.69	13.73	-3.92	0.00	-46.30	6.67	-7.84	-3.92	-23.33
zh	7.69	16.00	-20.37	1.75	-6.00	-9.80	0.00	19.51	5.13	-6.00	7.84	-5.77
fr	19.61	45.10	29.63	7.84	18.52	29.41	15.69	4.44	33.33	35.29	27.78	-3.92
ja	-26.00	-13.73	-7.55	3.92	7.84	2.00	-17.65	-17.65	-12.22	-10.00	18.87	1.52
ko	-27.45	-11.76	0.00	-33.33	2.00	-24.00	-3.85	-6.67	-33.33	-30.77	-7.84	-25.00
pt	5.88	15.69	33.33	13.46	15.15	-9.80	-5.13	15.69	16.67	5.56	14.04	40.48
es	35.29	37.25	37.25	27.45	25.93	49.02	64.71	35.09	35.29	27.45	33.33	29.63

mini excels in Indo-European languages, especially within humanities and entertainment topics, but underperforms in Arabic literature and several Korean and Japanese domains. These disparities underscore the need for targeted improvements to enhance its cultural competence.

Consistency of Results Across Judge Models. Table 8 reports the language-specific net win rates under the Qwen3-32B-Think judge. Comparing these results with those in Table 3 (ChatGPT-4o-Latest judge) shows that, while absolute values vary slightly, the overall model ranking by weighted average net win rate is preserved: ChatGPT-4o-Latest $>$ Qwen2.5-72B-Instruct $>$ Claude-3.5-Sonnet $\approx$ GPT-4o mini $>$ Qwen2.5-7B-Instruct $>$ Llama-3.1-70B-Instruct $\approx$ Mistral-Nemo-Instruct-2407 $\approx$ Mixtral-8x7B-Instruct-v0.1 $>$ remaining models. Language-specific performance trends are also consistent. For instance, Qwen2.5-72B-Instruct excels in French and Spanish regardless of the judge model. Similarly, Table 9 reports the cultural topic-wise net win rates under the Qwen3-32B-Think judge. Comparison with Table 4 (ChatGPT-4o-Latest judge) reveals parallel topic-specific patterns. For example, Claude-3.5-Sonnet consistently ranks highest in arts and relatively lower in religion and theology under both judges.

We further compute the agreement rate (the proportion of evaluation instances where both judges yield identical pairwise preferences) between ChatGPT-4o-Latest and Qwen3-32B-Think judgments for each model. As shown in Table 7, the average agreement rates across all 14 models reaches 85.05%.

These findings indicate that LLM-based evaluations of cultural competence, when using a capable judge model and comprehensive prompts, are robust and largely judge-independent.

6 Conclusion

In this paper, we present CultureSynth, a framework for synthesizing culturally-aware QA pairs to evaluate LLMs' cultural competence. By combining a hierarchically cultural taxonomy with RAG-based synthesis, we construct CultureSynth-7, demonstrating high-quality synthetic data generation. Our extensive evaluation, spanning 14 LLMs across 7 languages and 12 cultural topics, reveals that cultural competence depends on multiple factors beyond model scale, including training data composition, architectural design, and knowledge organization. We also identify different cultural understanding gaps, particularly in geographic and domain-specific contexts.

Limitations

Our study has two main limitations. First, the uneven distribution of keywords across languages and topics creates inherent dataset imbalances in the total $19,360$ QA pairs, despite our balanced sampling approach in the evaluation benchmark (i.e., CultureSynth-7). Second, when analyzing cultural topics, we did not categorize questions based on their cognitive demands (e.g., creative thinking, factual knowledge) and difficulty levels, which provides an opportunity for future refinement through fine-grained question taxonomies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Claude 3.5 sonnet.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. Calmqa: Exploring culturally specific long-form question answering across 23 languages. *CoRR*.

MJ Bennett. 2004. Becoming interculturally competent. *Toward multiculturalism: A reader in multicultural education/Intercultural Resource Corporation*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024a. Culturalteaming: Ai-assisted interactive red-teaming for challenging llms'(lack of) multicultural knowledge. *arXiv preprint arXiv:2404.06664*.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2024b. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

P Christopher Earley and Soon Ang. 2003. Cultural intelligence: Individual interactions across cultures.

P Christopher Earley and Elaine Mosakowski. 2004. Cultural intelligence. *Harvard business review*, 82(10):139–146.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Zeyu Gan and Yong Liu. 2024. Towards a theoretical understanding of synthetic data in llm post-training: A reverse-bottleneck perspective. *arXiv preprint arXiv:2410.01720*.

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. Collabcoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–29.

Daniel Goleman. 1998. Working with emotional intelligence. *NY: Bantam Books*.

Md Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2024. Nativqa: Multilingual culturally-aligned natural query for llms. *arXiv preprint arXiv:2407.09823*.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jong Youl Hong. 2023. Multicultural society and intercultural citizens. In *Multiculturalism and Interculturalism-Managing Diversity in Cross-Cultural Environment*. IntechOpen.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment project: What participatory, representative and individualised human

feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Culturepark: Boosting cross-cultural understanding in large language models. *arXiv preprint arXiv:2405.15145*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024b. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese. *arXiv preprint arXiv:2402.17302*.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. 2024. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. 2024. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Human Annotation

A.1 Detailed Annotation Criteria

Question Clarity & Safety: Determine if the question is self-contained and adheres to universal ethical standards.

- Score 1: Question is clear, comprehensible, and self-contained

- Score 0: Question exhibits any of the following issues: requires additional context for comprehension, contains demonstrative pronouns without context, or contains unsafe elements (violence, explicit content, hate speech).

Cultural Relevance: Identify cultural distinctiveness through dual dimensions.

- Score 1: Question that exhibits either cultural variance (answers differ across cultures/languages) or cultural specificity (containing culture-specific elements such as regional traditions).
- Score 0: Question lacks cultural elements or specificity.

Answer Quality: Access the quality of the answers relative to the reference knowledge (i.e., Wikipedia) using the 5-point scale, with scores of ≥ 4 being high quality.

- Score 5: Exceptional answer (comprehensive, accurate, well-structured)
- Score 4: Strong answer (minor improvements possible)
- Score 3: Adequate answer (notable omissions or inaccuracies)
- Score 2: Insufficient answer (major gaps or errors)
- Score 1: Poor answer (significantly flawed)
- Score 0: Unacceptable answer (incorrect or inappropriate)

A.2 Characteristics Of Annotators

Our annotation process maintains rigorous professional standards throughout. For each of the seven languages (ar, es, fr, ja, ko, pt, and zh), two native speakers were recruited from a professional annotation service provider as annotators to ensure quality. The entire annotation task was completed within a two-week period. The annotation cost varied by language, ranging from approximately 0.57 to 1.71 USD for each QA pair.

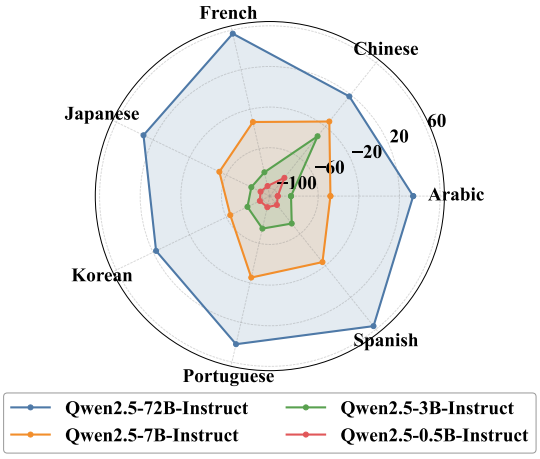
B Additional Experimental Results

Table 7 shows the average agreement rates across all 14 models. Table 8 and 9 report the language-specific and cultural topic-wise net win rates under the Qwen3-32B-Think judge, respectively.

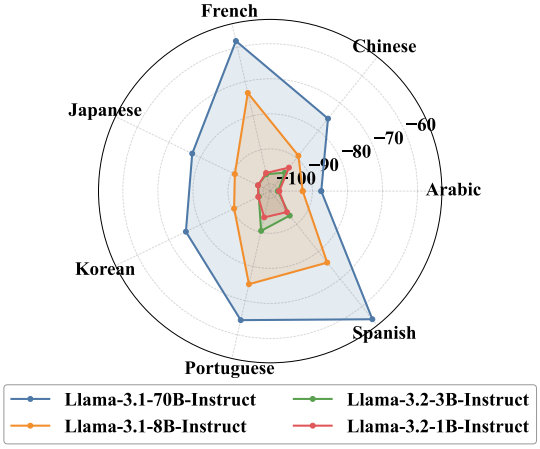
Figures 7 to Figure 9 present net win rates (%) across languages and cultural topics, comparing models of different sizes against the baseline.

Table 7: Agreement rates between ChatGPT-4o-Latest and Qwen3-32B-think judges across all models

Model	Strict Agreement
ChatGPT-4o-Latest	81.13%
GPT-4o mini	70.21%
Claude-3.5-Sonnet	75.08%
Qwen2.5-72B-Instruct	70.35%
Qwen2.5-7B-Instruct	78.38%
Qwen2.5-3B-Instruct	86.89%
Qwen2.5-0.5B-Instruct	97.06%
Llama-3.1-70B-Instruct	85.15%
Llama-3.1-8B-Instruct	90.96%
Llama-3.2-3B-Instruct	97.53%
Llama-3.2-1B-Instruct	97.08%
Mixtral-8x7B-Instruct-v0.1	84.60%
Mistral-Nemo-Instruct-2407	82.86%
Mistral-7B-Instruct-v0.3	93.42%
Average	85.05%



(a) Qwen2.5 series



(b) Llama 3 series

Figure 7: Net win rates (%) of models of different sizes compared to the baseline model across languages.

C Prompt Templates

Figure 10 shows the prompt template for cultural topic extension. Figures 11 to 17 are the step-by-step prompt templates for QA pairs generation. And Figure 18 is the prompt template for pairwise comparison of different model responses.

Table 8: Net win rates (%) of models across languages, relative to the Gemma2-27B-it baseline under the Qwen3-32B-Think judge. Blue and yellow indicate the best and second-best performing models, respectively.

Model	Arabic	Chinese	French	Japanese	Korean	Portuguese	Spanish	Weighted Average
ChatGPT-4o-Latest	68.61	79.93	80.85	65.22	71.18	80.26	88.94	76.33
GPT-4o mini	-34.04	-26.42	-9.57	-29.54	-27.26	-18.61	-1.28	-20.92
Claude-3.5-Sonnet	-51.68	-21.91	7.52	-9.75	-31.77	-14.10	19.07	-14.08
Qwen2.5-72B-Instruct	-0.35	-3.18	28.38	6.30	-4.69	16.17	32.37	10.80
Qwen2.5-7B-Instruct	-75.13	-31.44	-50.43	-65.97	-72.92	-51.13	-45.51	-56.04
Qwen2.5-3B-Instruct	-94.71	-60.37	-87.18	-85.91	-83.51	-85.53	-82.85	-82.77
Qwen2.5-0.5B-Instruct	-99.12	-93.31	-98.63	-98.20	-97.57	-98.31	-97.76	-97.54
Llama-3.1-70B-Instruct	-89.59	-78.76	-64.44	-73.01	-77.60	-73.87	-57.37	-73.29
Llama-3.1-8B-Instruct	-95.41	-89.30	-85.13	-88.89	-91.32	-85.69	-83.65	-88.45
Llama-3.2-3B-Instruct	-99.65	-96.32	-97.54	-99.10	-98.44	-93.96	-98.02	-97.64
Llama-3.2-1B-Instruct	-99.29	-95.82	-97.44	-98.65	-98.26	-96.24	-94.87	-97.23
Mixtral-8x7B-Instruct-v0.1	-92.06	-87.46	-57.95	-93.40	-82.99	-70.49	-54.17	-77.08
Mistral-Nemo-Instruct-2407	-93.47	-55.02	-81.37	-67.77	-67.19	-80.08	-81.89	-74.98
Mistral-7B-Instruct-v0.3	-99.12	-92.64	-88.55	-97.15	-95.49	-85.53	-87.02	-92.31

Table 9: Net win rates (%) of models across cultural topics, relative to the Gemma2-27B-it baseline under the Qwen3-32B-Think judge. Blue and yellow indicate the best and second-best performing models, respectively.

Model	Social Sciences	Philosophy and Psychology	Religion and Theology	Political Science	Law	Edu-cation	Lang-uage	Liter-ature	Medi-cine	Applied Sciences and Technology	Arts	Recreation, Sports, and Entertainment
ChatGPT-4o-Latest	75.00	82.87	77.93	73.08	74.05	71.27	79.66	82.98	67.01	77.40	80.98	72.33
GPT-4o mini	-25.00	-9.55	-18.53	-22.53	-14.87	-27.04	-21.78	-23.40	-22.92	-26.93	-24.18	-14.99
Claude-3.5-Sonnet	-19.72	-14.04	-26.98	-18.41	-12.24	-23.38	-8.31	-10.03	-5.56	-14.24	-2.17	-11.53
Qwen2.5-72B-Instruct	14.17	18.82	12.53	7.14	7.29	4.23	14.04	15.20	15.28	1.24	11.96	7.78
Qwen2.5-7B-Instruct	-52.50	-62.92	-55.59	-55.77	-53.64	-59.72	-57.59	-62.92	-43.06	-53.56	-59.24	-53.60
Qwen2.5-3B-Instruct	-86.11	-81.46	-83.38	-85.71	-82.51	-84.51	-84.81	-85.11	-82.64	-81.42	-79.08	-76.37
Qwen2.5-0.5B-Instruct	-98.33	-97.19	-97.28	-98.08	-96.79	-98.59	-97.99	-98.48	-97.57	-97.83	-97.01	-95.39
Llama-3.1-70B-Instruct	-80.00	-73.60	-76.29	-70.88	-66.18	-78.59	-75.07	-76.90	-70.49	-71.83	-70.92	-68.01
Llama-3.1-8B-Instruct	-93.06	-91.57	-89.10	-86.54	-88.63	-91.83	-87.97	-92.10	-86.46	-87.00	-81.79	-85.22
Llama-3.2-3B-Instruct	-98.30	-98.30	-98.62	-96.67	-96.63	-98.58	-97.98	-97.23	-97.20	-99.38	-97.27	-95.63
Llama-3.2-1B-Instruct	-98.33	-97.19	-96.19	-97.80	-96.21	-99.44	-97.13	-98.18	-98.96	-95.67	-95.65	-96.25
Mixtral-8x7B-Instruct-v0.1	-87.50	-78.93	-71.39	-74.45	-76.97	-80.56	-76.50	-77.20	-75.35	-78.02	-75.27	-72.62
Mistral-Nemo-Instruct-2407	-79.72	-75.56	-75.48	-73.08	-69.97	-74.93	-74.79	-75.08	-72.22	-72.45	-81.79	-73.49
Mistral-7B-Instruct-v0.3	-95.00	-91.29	-94.01	-92.31	-90.96	-94.93	-90.83	-95.14	-91.32	-91.64	-91.30	-88.76
Average	-58.89	-54.99	-56.60	-56.57	-54.59	-59.76	-55.50	-56.69	-54.39	-56.52	-54.48	-53.70

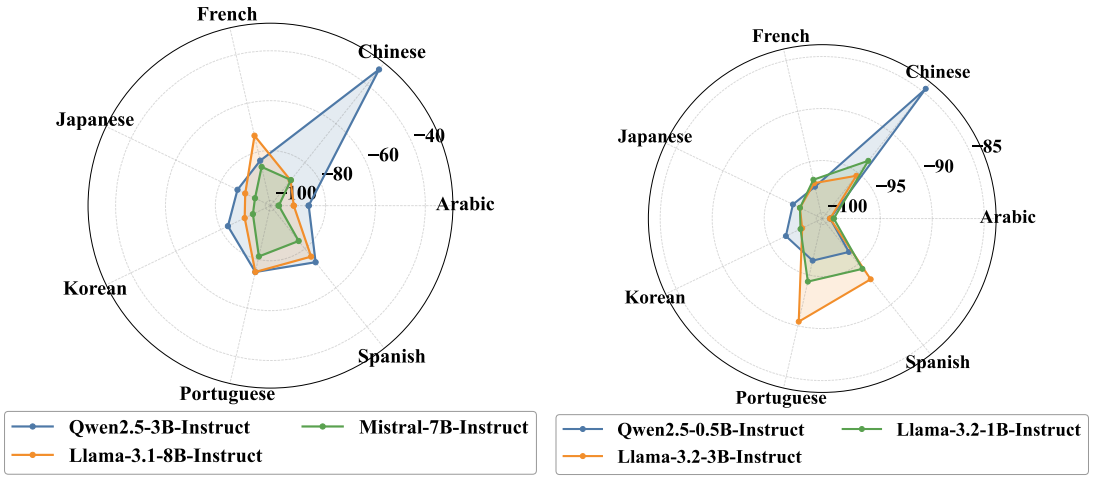
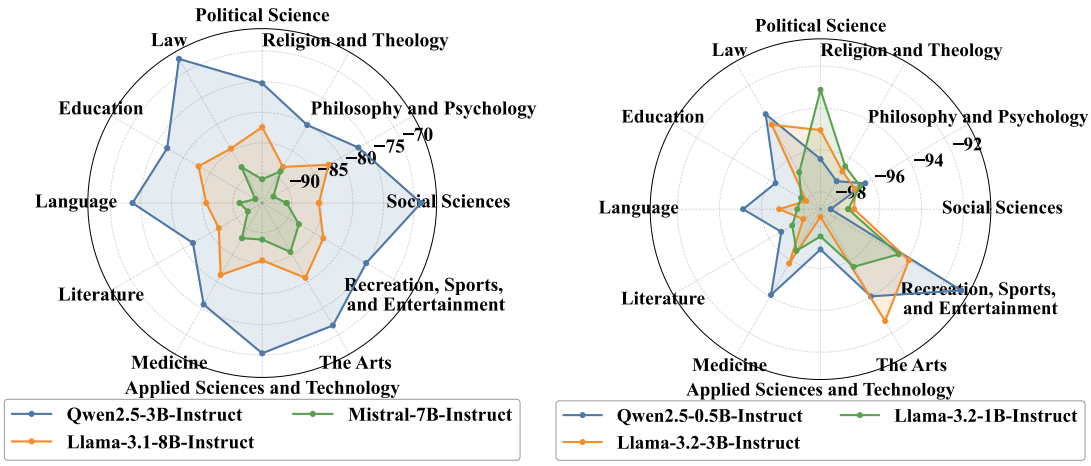


Figure 8: Net win rates (%) of models under 3B parameters compared to the baseline model across languages.

D Universal Cultural Topics

Table 10 presents the complete list of primary and secondary topics.



(a) Third-tier cultural competence models

(b) Fourth-tier cultural competence models

Figure 9: Net win rates (%) of different models compared to the baseline model across cultural topics.

Assuming you are an expert in the field of {primary_topics} in {country}, below are possible cultural differences in primary topics and secondary topics. For each secondary topic, evaluate whether there are differences between {country} and other countries worldwide. If there are no differences, output "None"; if there are differences, expand the secondary topic into five or more tertiary topics and keywords that reflect the unique characteristics of {country} and cover as much content as possible.

Primary Topics - Secondary Topics:
{primary_topics} - {secondary_topics}

Please don't be lazy and answer this question in depth from a local's perspective.

Figure 10: Prompt for extending primary and secondary topics into tertiary topics and keywords.

Can you extract knowledge points related to the "{keyword}" from the following text?
Yes or No. Do not output other content.

Text (Title: {wikipedia_title})
{wikipedia_content}

Figure 11: Prompt for determining whether the retrieved page is related to the keyword (Step 1).

Assume you are an expert in the {primary_topic}. Please determine whether the following text contains multicultural differences or is specific cultural knowledge unique to {country}. Choose A, B, or C as your output. Do not output other content.

A. Contains cultural differences from different countries
B. Is cultural knowledge specific to {country}
C. Neither of the above

Text (Title: {title})
{content}

Figure 12: Prompt for determining whether the retrieved page contains culture-specific content (Step 1).

From the following references, extract up to 3 points of knowledge in {target_language} that are important, diverse, and reflect the differences between {country} and other countries.
References (Title: {title})
{content}

Requirements
- First, summarize knowledge points for country by starting with "In {country},".
- Then, summarize the knowledge points for countries other than country and different from country accordingly, starting with "In [other countries],". If there are no knowledge points for countries other than country, explain how the USA is different based on your knowledge, starting with "In [US],".
- Do not extract historically relevant knowledge points.
- Return in JSON format: [{"knowledge1": "xxx", "differ1": "xxx"}, {"knowledge2": "xxx", "differ2": "xxx"}, {"knowledge3": "xxx", "differ3": "xxx"}]
- Use {target_language}.

Figure 13: Prompt for knowledge extraction in different cultural knowledge settings (Step 2).

Based on the following references, extract no more than 3 text fragments from the references, which contain cultural knowledge unique to {country}.

References (Title: title)
{content}

Requirements
- Do not extract historical or non-{country}-related knowledge points.
- Return in JSON format: [{"knowledge1": "xxx"}, {"knowledge2": "xxx"}, {"knowledge3": "xxx"}]
- Use {target_language}.

Figure 14: Prompt for knowledge extraction in unique cultural knowledge settings (Step 2).

Assuming you are an expert in the field of {primary_topic}, pose a situational question in {target_language} that involves the following different knowledge points from two countries.

Example
Knowledge
In Japan, 8 (八): pronounced similarly to the word for prosperity or development (はち, hachi). 7 (七): considered a lucky number, symbolizing good things come in pairs, and celebrated in many traditional festivals such as Shichi-Go-San (Children's Day). 4 (四): pronounced similar to the word for death (し, shi). 9 (九): pronounced similar to the word for suffering (く, ku), meaning pain and hardship.
Different Knowledge
In the United States, 7 (seven): widely regarded as a lucky number, often appears in gambling, lottery, and other occasions, such as 7 on Las Vegas slot machines. 3 (three): In Western culture, there is a concept of "three", such as "threesome", "lucky three", and is considered a perfectly balanced number. 13 (thirteen): considered unlucky, in many buildings, the 13th floor is even skipped and directly numbered as the 14th floor. This superstition comes from "The Last Supper", in which Judas is the thirteenth person. 666: considered to be the "devil's number", from the Book of Revelation in the Bible. In China, 8 (八), pronounced similar to "fa" (the "fa" in "facai"), symbolizes wealth and success, and is deeply loved by people. 6 (六): also represents good luck. 9 (九): symbolizes longevity and longevity. 4 (四): pronounced similar to "si" (死), is considered an unlucky number.
{{"Question": "Can you recommend a few lucky numbers?"}}

Your Turn
Knowledge
{knowledge}
Different Knowledge
{differ}

Requirements
- The question must not contain any offensive or discriminatory content, nor should it include pornography, gruesomeness, violence, or aggressive elements.
- Don't mention country names in the question.
- Do not use referential words, demonstrative pronouns, or demonstrative pronouns.
- Directly output the question, do not provide other contents.
- Use {target_language}.

Figure 15: Prompt for question generation in different cultural knowledge settings (Step 3).

Consider the following cultural knowledge and assume specific scenarios or roles to seek expert advice in {target_language}.
Cultural Knowledge
{knowledge}

Requirements
- The question must not contain any offensive or discriminatory content, nor should it include pornography, gruesomeness, violence, or aggressive elements.
- The question can be asked in different cultural contexts.
- Don't mention country names in the question.
- Do not use referential words or demonstrative pronouns.
- Directly output the question, do not provide other contents.
- Use {target_language}.

Figure 16: Prompt for question generation in unique cultural knowledge settings (Step 3).

You are an expert in the field of {primary_topic} in {country}. Please refer to the following cultural knowledge to answer the questions.
Cultural Knowledge
{knowledge}

Question
{questinon}

Requirements
- Be as detailed as possible and closely follow the cultural knowledge provided.
- Be clear, detailed, and address multiple aspects of the question comprehensively.
- Don't mention country names in the answer.
- Do not use referential words or demonstrative pronouns.
- Directly output the answer; do not provide other contents.
- Use {target_language}.

Figure 17: Prompt for answer generation in both different and unique cultural knowledge settings (Step 3).

[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer, assistant A's answer, and assistant B's answer. Your job is to evaluate which assistant's answer is better. Begin your evaluation by comparing both assistants' answers with the reference answer. Identify and correct any mistakes. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Reference Answer]
{answer_ref}
[The End of Reference Answer]

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]

Figure 18: Prompt for reference-guided pairwise comparison (Zheng et al., 2023).

Table 10: Universal Cultural Topics

Abbreviations	Primary Topics	Secondary Topics
SS	Social Sciences	Methods of the social sciences
	Social Sciences	Social questions. Social practice
	Social Sciences	Cultural practice
	Social Sciences	Way of life (Lebensweise)
	Social Sciences	Food, Beverage, and Culinary Arts
	Social Sciences	Gender studies
	Social Sciences	Sociography. Descriptive studies of society (both qualitative and quantitative)
	Social Sciences	Statistics as a science. Statistical theory
	Social Sciences	Society
	Social Sciences	Economics. Economic science
	Social Sciences	Public administration. Government. Military affairs
	Social Sciences	Safeguarding the mental and material necessities of life
	Social Sciences	Costume. Clothing. National dress. Fashion. Adornment
	Social Sciences	Customs, manners, usage in private life
	Social Sciences	Death. Treatment of corpses. Funerals. Death rites
	Social Sciences	Public life. Pageantry. Social life. Life of the people
	Social Sciences	Social ceremonial. Etiquette. Good manners. Social forms. Rank. Title
	Social Sciences	Folklore in the strict sense
	Social Sciences	Commerce
	Social Sciences	Demography
	Social Sciences	Social Interaction
PP	Philosophy and Psychology	Metaphysics
	Philosophy and Psychology	Epistemology, causation and humankind
	Philosophy and Psychology	Parapsychology and occultism
	Philosophy and Psychology	Specific philosophical schools and viewpoints
	Philosophy and Psychology	Psychology
	Philosophy and Psychology	Philosophical logic
	Philosophy and Psychology	Ethics (Moral philosophy)
	Philosophy and Psychology	Ancient, medieval and eastern philosophy
	Philosophy and Psychology	Modern western and other noneastern philosophy
	Philosophy and Psychology	Other philosophy and psychology
RT	Religion and Theology	Prehistoric religions. Religions of early societies
	Religion and Theology	Religions originating in the Far East
	Religion and Theology	Religions originating in Indian sub-continent. Hindu religion in the broad sense
	Religion and Theology	Buddhism
	Religion and Theology	Religions of antiquity. Minor cults and religions
	Religion and Theology	Judaism
	Religion and Theology	Christianity
	Religion and Theology	Islam
	Religion and Theology	Modern spiritual movements
	Religion and Theology	Other religions
PS	Political Science	Political science (General)
	Political Science	Political theory. Theory of the state
	Political Science	Political institutions and public administration
	Political Science	North America
	Political Science	United States
	Political Science	Canada, Latin America, etc.
	Political Science	Europe
	Political Science	Asia
	Political Science	Local government. Municipal government
	Political Science	Colonies and colonization. Emigration and immigration. International migration
	Political Science	International relations
	Political Science	Other Politics and Policy
LAW	Law	Law in general. Comparative and uniform law. Jurisprudence
	Law	Religious law
	Law	United Kingdom and Ireland law
	Law	Canada law
	Law	United States law
	Law	Europe law
	Law	Germany law
	Law	Asia and Eurasia law
	Law	Africa law
	Law	Latin America law

Abbreviations	Primary Topics	Secondary Topics
LAW	Law	Law of nations
	Law	Other laws
EDU	Education	Education (General)
	Education	History of education
	Education	Theory and practice of education
	Education	Special aspects of education
	Education	Other educational aspects
LAN	Language	Linguistics
	Language	English and Old English (Anglo-Saxon)
	Language	German and related languages
	Language	French and related Romance languages
	Language	Italian, Dalmatian, Romanian, Rhaetian, Sardinian, Corsican
	Language	Spanish, Portuguese, Galician
	Language	Latin and related Italic languages
	Language	Classical Greek and related Hellenic languages
	Language	Chinese, Cantonese
	Language	Arabic
	Language	Russian
	Language	Japanese
	Language	Vietnamese
	Language	Thai
	Language	Korean
	Language	Other languages
LIT	Literature	American literature in English
	Literature	English and Old English (Anglo-Saxon) literatures
	Literature	German literature and literatures of related languages
	Literature	French literature and literatures of related Romance languages
	Literature	Literatures of Italian, Dalmatian, Romanian, Rhaetian, Sardinian, Corsican languages
	Literature	Literatures of Spanish, Portuguese, Galician languages
	Literature	Latin literature and literatures of related Italic languages
	Literature	Classical Greek literature and literatures of related Hellenic languages
	Literature	Literatures of Chinese, Cantonese
	Literature	Literatures of Arabic
	Literature	Literatures of Russian
	Literature	Literatures of Japanese
	Literature	Literatures of Vietnamese
	Literature	Literatures of Thai
	Literature	Literatures of Korean
	Literature	Literatures of other specific languages and language families
MED	MEDICINE	Medical sciences
AST	Applied Sciences and Technology	Biotechnology
	Applied Sciences and Technology	Engineering. Technology in general
	Applied Sciences and Technology	Agriculture and related sciences and techniques. Forestry. Farming. Wildlife exploitation
	Applied Sciences and Technology	Home economics. Domestic science. Housekeeping
	Applied Sciences and Technology	Transportation and communications
	Applied Sciences and Technology	Accountancy
	Applied Sciences and Technology	Business management
	Applied Sciences and Technology	Public relations
	Applied Sciences and Technology	Chemical technology. Chemical and related industries
	Applied Sciences and Technology	Various industries, trades and crafts
	Applied Sciences and Technology	Industries, crafts and trades for finished or assembled articles
	Applied Sciences and Technology	Building (construction) trade. Building materials. Building practice and procedure
	Applied Sciences and Technology	Intelligent Technology
ART	Arts	Special auxiliary subdivision for the arts
	Arts	Physical planning. Regional, town and country planning. Landscapes, parks, gardens
	Arts	Architecture
	Arts	Plastic arts
	Arts	Drawing. Design. Applied arts and crafts
	Arts	Painting
	Arts	Graphic art, printmaking. Graphics
	Arts	Photography and similar processes
	Arts	Music and dance
	Arts	Drama and Movie
	Arts	Other arts

Abbreviations	Primary Topics	Secondary Topics
RSE	Recreation, Sports, and Entertainment	Games
	Recreation, Sports, and Entertainment	Sport and Exercise
	Recreation, Sports, and Entertainment	Other entertainment, leisure