ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2026.0083

CCU-Bench: a systematic benchmark for cross-cultural understanding in large language models

Wei ZHANG[1], Ting YIN[1], Jia CHEN[1], Wenjie NING[1], Xiaoyan GU[2], Kai LIU[2], Hengnian QI[3], Wei CHEN[2]✉

[1]*School of Computer and Computing Science, Hangzhou City University, Hangzhou 310015, China*
[2]*State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China*
[3]*School of Information Engineering, Huzhou Normal University, Huzhou 313000, China*

Abstract: Large language models (LLMs) have achieved strong performance on many language tasks, but they still struggle with culturally grounded symbolic reasoning. Existing benchmarks have not systematically evaluated the progressive capability chain required for cross-cultural understanding, which involves intra-cultural symbolic understanding, cross-cultural symbolic alignment, and cross-cultural conflict identification. To address this gap, we propose CCU-Bench, a systematic benchmark for evaluating cross-cultural understanding in LLMs. Grounded in Hofstede's cultural dimensions theory, CCU-Bench focuses on three culturally distinct contexts, China, Japan, and Mexico, and is organized around three dimensions, image, connotation, and emotion. The benchmark is constructed through three stages, data collection, question generation, and quality control, resulting in a high-quality evaluation set of 3029 question–answer pairs in five formats. Experiments on 19 mainstream LLMs show that current models perform unsatisfactorily on this task, achieving an average score of 57.8%. Closed-source models consistently outperform open-source models, while cross-cultural symbolic alignment remains the most challenging sub-task. Further error analysis reveals that the dominant failures stem from deficiencies in cultural knowledge, biases in intent mapping, and weak higher-order reasoning about cross-cultural conflicts, rather than simple instruction-following issues. These findings highlight persistent limitations of current LLMs in culturally grounded reasoning and demonstrate that CCU-Bench provides a standardized benchmark for culturally aware artificial intelligence research.

Key words: Benchmark; Large language models (LLMs); Cross-cultural understanding; LLM evaluation and benchmarks

1 Introduction

Large language models (LLMs) have achieved strong performance on many language tasks, but they still struggle with cross-cultural understanding. This limitation is particularly evident when models are required to interpret cultural symbols, which often encode condensed social meaning, value, and emotion. Such knowledge is typically implicit, context-dependent, and unevenly represented in long-tailed training corpora. As LLMs are increasingly deployed in multilingual and multicultural settings, systematically evaluating their capability for cross-cultural understanding has become necessary and timely (Myung et al., 2024; Romero et al., 2024; Zhou et al., 2024; Rystrøm et al., 2025; Wang W et al., 2025).

However, existing benchmarks (Chiu et al., 2024; Kharchenko et al., 2025; Lu et al., 2025; Kabir et al., 2026) do not provide a systematic assessment of the progressive capability chain required for cross-cultural understanding. In our view, this capability chain begins with intra-cultural symbolic understanding, which refers to capturing the core connotations, cultural values, and emotional associations of symbols within a given culture. It then progresses to cross-cultural symbolic alignment, which requires establishing accurate semantic correspondences between symbols across cultures, and further involves cross-cultural conflict identification, which involves recognizing potential misunderstandings, offensive interpretations, or decision-making risks caused by

✉ Wei CHEN, chenvis@zju.edu.cn
Wei ZHANG, https://orcid.org/0000-0002-8321-4607
Ting YIN, https://orcid.org/0009-0009-6706-6622
Jia CHEN, https://orcid.org/0009-0002-6905-6005
Wenjie NING, https://orcid.org/0009-0008-4871-4586
Xiaoyan GU, https://orcid.org/0009-0009-5379-985X
Kai LIU, https://orcid.org/0009-0003-5786-6489
Hengnian QI, https://orcid.org/0000-0002-1927-7160
Wei CHEN, https://orcid.org/0000-0002-8365-4741
CLC number: TP391.1

Received: Mar. 27, 2026; Revision accepted: June 23, 2026;
Crosschecked: July 13, 2026

symbolic misalignment. These capabilities are closely connected, forming the reasoning foundation of cross-cultural understanding (Zhang et al., 2025).

To fill this gap, we propose CCU-Bench to systematically evaluate cross-cultural understanding in LLMs. CCU-Bench focuses on three countries, China, Japan, and Mexico, due to their significant cultural differences according to Hofstede's cultural dimensions theory (Hofstede, 2001; Kharchenko et al., 2025; Masoud et al., 2025; Kusnick et al., 2026). Based on a three-dimensional framework of image, connotation, and emotion, we construct three evaluation tasks corresponding to the above capability chain: intra-cultural symbolic understanding, cross-cultural symbolic alignment, and cross-cultural conflict identification. CCU-Bench is built from 3000 symbol–semantic correspondence entries collected from multiple references on cultural symbols. Through a pipeline of data collection, question generation, and quality control, we obtain a final benchmark containing 3029 high-quality question–answer (QA) pairs.

Using CCU-Bench, we conduct a systematic evaluation of 19 mainstream LLMs. The results show that current models perform poorly in cross-cultural understanding, with an average overall score of 57.8%, suggesting that existing LLMs do not yet exhibit stable and reliable capabilities for culturally grounded symbolic reasoning, though closed-source models consistently outperform open-source models. For instance, GPT-5 (OpenAI, 2025) achieves the highest overall score of 70.2%, followed by Claude Sonnet 4.5 (Anthropic, 2025) (65.8%), while open-source models such as InternLM3-8B (Bi et al., 2026) (42.9%) and Qwen3-4B (Qwen Team, 2025) (43.1%) lag significantly behind. These findings reveal persistent weaknesses in current LLMs and highlight the importance of more systematic evaluation of culturally aware artificial intelligence (AI) systems. Our contributions are summarized as follows:

1. We introduce CCU-Bench, a systematic benchmark for evaluating intra-cultural symbolic understanding, cross-cultural symbolic alignment, and cross-cultural conflict identification in LLMs.

2. We construct a benchmark of 3029 high-quality QA pairs from multiple cultural sources under a three-dimensional (image–connotation–emotion) framework.

3. We conduct a systematic evaluation of 19 mainstream LLMs, showing that current models remain far from reliable in cross-cultural understanding, especially in symbolic alignment and conflict identification.

2 Related works

2.1 Cross-cultural understanding in large language models

The limitations of LLMs in cross-cultural understanding are closely tied to cultural bias introduced during training and development. Limited cultural diversity in training data can lead to Western-centric value preferences, distort representations of non-Western input, and even produce culturally offensive outputs (Johnson et al., 2022). At the same time, cultural presuppositions embedded in model development may conflict with local norms and values, leading to cultural misunderstandings and representational bias (Prabhakaran et al., 2022). These problems are further aggravated by uneven multilingual performance, as LLMs generally perform better in English and often exhibit asymmetric cross-lingual transfer, which can cause conceptual inconsistencies across languages (Xu SY et al., 2024).

Previous work has explored several ways to improve cross-cultural adaptation in LLMs. For example, CultureLLM incorporates cultural dimensions into model design, using World Values Survey data to improve cultural understanding in single-culture settings (Li et al., 2024). Others focus on detecting and mitigating cultural bias, but are largely limited to single-culture scenarios and provide limited insight into cross-cultural reasoning (Feng et al., 2023; Santurkar et al., 2023). Meanwhile, Arora et al. (2023) and Cao et al. (2023) evaluated multilingual cultural alignment based on Hofstede's cultural dimensions theory. Masoud et al. (2025) further highlighted cross-cultural interpretive bias in LLMs. However, existing efforts do not provide a full-process evaluation covering intra-cultural symbolic understanding, cross-cultural symbolic alignment, and cross-cultural conflict identification.

In short, existing methods improve cultural adaptation to a certain extent, but do not support the complete progressive reasoning chain of cross-cultural understanding. Most focus on single-culture understanding or surface-level cultural bias detection, while ignoring fine-grained symbolic alignment and higher-order reasoning about cultural conflicts. This gap motivates us to build a systematic benchmark to reveal and evaluate such limitations.

2.2 Cultural understanding benchmarks

Table 1 lists a series of mainstream cross-cultural evaluation benchmarks and clarifies the unique evaluation dimensions

Table 1 Comparison of CCU-Bench with related benchmarks across key evaluation dimensions

Focus	Benchmark	Coverage	Evaluation dimensions		
			Understanding	Alignment	Conflict ID
Value alignment	CValues (Xu GH et al., 2023)	Chinese context	✓	×	×
Subjective opinions	GlobalOpinionQA (Durmus et al., 2023)	Multi-country	✓	×	×
Everyday knowledge	Blend (Myung et al., 2024)	Diverse cultures	✓	×	×
Cultural matching	CultureBench (Tam et al., 2025)	45+ regions	✓	×	×
Cultural reasoning	XCR-Bench (Kabir et al., 2026)	Multi-country	✓	×	Partial
Cross-cultural	CCU-Bench (ours)	China, Japan, Mexico	✓	✓	✓

ID: identification

on which each of them focuses. Existing benchmarks for evaluating LLMs' cultural knowledge follow three construction paradigms: large-scale cultural knowledge bases, adapted sociocultural questionnaires, and generative data expansion (Dhar and Shamir, 2021; Durmus et al., 2023; Li et al., 2024; Shi et al., 2024; Zhang et al., 2024a, 2024b). However, they suffer from several common limitations, including cultural bias, narrow topic coverage, and limited data quality. Most benchmarks focus on English and other high-resource languages, provide insufficient coverage of non-Western cultures, and restrict topics to predefined areas such as food, etiquette, and religion (Adilazuarda et al., 2024). In addition, many benchmarks directly adopt uncontrolled LLM-generated content without rigorous human verification, making them vulnerable to data leakage, factual errors, and inherited cultural bias (Nguyen et al., 2023; Peng et al., 2026; Zhang et al., 2026).

Benchmarks related to cross-cultural understanding remain incomplete. While existing studies examine preliminary cultural understanding, none cover the full capability chain targeted in this work, namely intra-cultural symbolic understanding, cross-cultural symbolic alignment, and cross-cultural conflict identification. In particular, single-culture benchmarks are often biased toward Western cultures and do not extend to cross-cultural settings (Cao et al., 2023); cross-cultural benchmarks primarily focus on surface-level symbol matching rather than semantic alignment (Chiu et al., 2024); conflict-oriented benchmarks typically identify cultural differences without analyzing their causes or downstream risks (Kabir et al., 2026). Overall, existing benchmarks lack a high-quality, multicultural, and progressive framework for evaluating cross-cultural understanding, motivating the construction of CCU-Bench.

In summary, existing research and benchmarks fail to provide a complete, progressive, and multicultural evaluation framework for cross-cultural understanding. They focus either on single-culture understanding, rely on surface-level symbol matching, or lack systematic analysis of cultural conflict risks. To bridge these gaps, we design CCU-Bench, a dedicated benchmark that covers the complete reasoning chain and supports rigorous evaluation of LLMs' cross-cultural capabilities.

3 CCU-Bench

3.1 Overview

This work introduces CCU-Bench, the new benchmark specifically designed for evaluating cultural understanding in LLMs. We adopt Hofstede's cultural dimensions theory (Hofstede, 2001) to select representative cultures for evaluating cross-cultural understanding. Since a larger cultural distance implies greater differences in symbolic connotations, values, and symbolic systems, it provides a more challenging test for LLMs. Accordingly, we choose China, Japan, and Mexico, which have clearly differentiated cultural distances (Yeh, 1988).

As shown in Fig. 1, CCU-Bench evaluates three progressive capabilities for understanding cultural symbols, intra-cultural symbolic understanding, cross-cultural symbolic alignment, and cross-cultural conflict identification, which cover the full pipeline from basic cultural cognition to advanced conflict judgment. The benchmark contains 3029 English QA pairs in

Intra-cultural symbolic understanding	Cross-cultural symbolic alignment	Cross-cultural conflict identification
Single-choice question Q: Which ancient civilization is most closely associated with the cultural symbol of "rattlesnake" in Mexico? A: c. Aztec civilization	Single-choice question Q: In Japanese culture, "lacquer" is often used to express "love." In Chinese culture, what is the most typical object used to express the same intention? A: c. Azure-winged magpie	True/false question Q: In Mexico, "butterfly" represents "butterfly of the soul," and in China, "butterfly" can also represent this meaning. A: No
Multiple-choice question Q: In Chinese culture, what are the symbolic meanings of "honeysuckle?" A: a. Beauty b. Auspiciousness c. Happiness e. Purity	Multiple-choice question Q: In Japanese culture, "horse" is often used to express "strength." In Chinese culture,which of the following objects can express a similar meaning? A: c. Ox e. Golden eagle	Open-ended question Q: When introducing the Chinese concept of pine to a Japanese audience, is it potentially misleading to state only that "pine symbolizes longevity and is commonly used in birthday celebrations?" If so, specify the misleading point and provide supporting evidence. A: Yes, it is misleading. The misleading: point is that Japanese audiences may incorrectly assume that "Japanese pine also solely symbolizes longevity," whereas in Japan, pine is also used in New Year "kadomatsu" (gate pine) decorations, symbolizing "auspiciousness and vitality." The discrepancy arises because the Chinese "pine" emphasizes "longevity and birthday blessings," while the Japanese "pine" encompasses both "longevity and New Year auspiciousness"; a single explanation obscures the multifaceted uses of pine in Japan.
Fill-in-the-blank question Q: In Chinese culture, the turtle usually symbolizes _____. A: Mystery	Fill-in-the-blank question Q: In Chinese culture, "shrike" often represents the imagery of "ominous." What connotation does "shrike" have in Japan? It typically symbolizes _____. A: Mystery	
Open-ended question Q: Interpret the symbolic meaning of "León de montaña (mountain lion)" in Mexican culture for an international audience. A: In Mexican culture, "León de montaña (mountain lion)" core symbolizes "the majesty of mountains, the loneliness of strength, and the balance of nature"…	Open-ended question Q: In Chinese culture, the "goose" often represents "love." Does Mexican culture have a similar image? Please explain. What are the connotations of the Mexican "goose?" A: 1. The core connotation of "love" for the Chinese goose is: geese maintain a rare…	

Fig. 1 CCU-Bench consists of three tasks and five question types. For the single-choice and multiple-choice questions, there are four and five options, respectively. Due to space constraints, only the answers are shown without listing all options

diverse formats, including true/false, single-choice, multiple-choice, fill-in-the-blank, and open-ended questions. Detailed statistics are provided in Table 2.

By covering three culturally distant regions and five question types, CCU-Bench provides a comprehensive and challenging testbed for evaluating basic cultural memory and advanced cross-cultural reasoning. It can help researchers identify model weaknesses, guide data optimization, and promote the development of more culturally compatible and robust LLMs.

3.2 Task definition

CCU-Bench contains three tasks to progressively evaluate the reasoning ability of LLMs in cultural symbol understanding.

Task 1: intra-cultural symbolic understanding. Accurately understanding the intended meaning of culture-specific symbols is the foundation of cross-cultural understanding (Shen et al., 2024). We therefore design this task to evaluate whether models can reliably interpret cultural symbols within a given cultural context.

Task 2: cross-cultural symbolic alignment. This task evaluates whether models can appropriately map symbols from a source culture to a target culture. It requires preserving the core connotations of source symbols while integrating them naturally into the target cultural context, thereby testing models' ability in cross-cultural knowledge transfer and adaptation (Guo et al., 2023).

Task 3: cross-cultural conflict identification. This task evaluates whether models can identify potential cultural conflicts in cross-cultural understanding and explain their causes. Since cultural symbols are often geographically and culturally specific, translation may easily introduce connotation conflicts or cognitive bias. This task further requires models to analyze the underlying reasons for such conflicts, providing evidence for improving cross-cultural understanding (Rahman and Salam, 2026).

3.3 Benchmark construction

Different from existing benchmarks, we use LLMs only for assisted QA generation based on manually curated authoritative cultural symbol data, and all QA pairs are strictly verified by human experts to ensure accuracy and cultural validity.

Our construction is primarily divided into three stages. Fig. 2 illustrates our overall construction pipeline, including

Table 2 Summary of QA pair counts across tasks and question types

Task	Number of QA pairs				
	Single-choice	Multiple-choice	Fill-in-the-blank	Open-ended	True/false
Task 1: intra-cultural symbolic understanding	262	251	251	275	–
Task 2: cross-cultural symbolic alignment	251	251	251	251	–
Task 3: cross-cultural conflict identification	–	–	–	530	456
Total across all tasks	513	502	502	1056	456

–: not evaluated

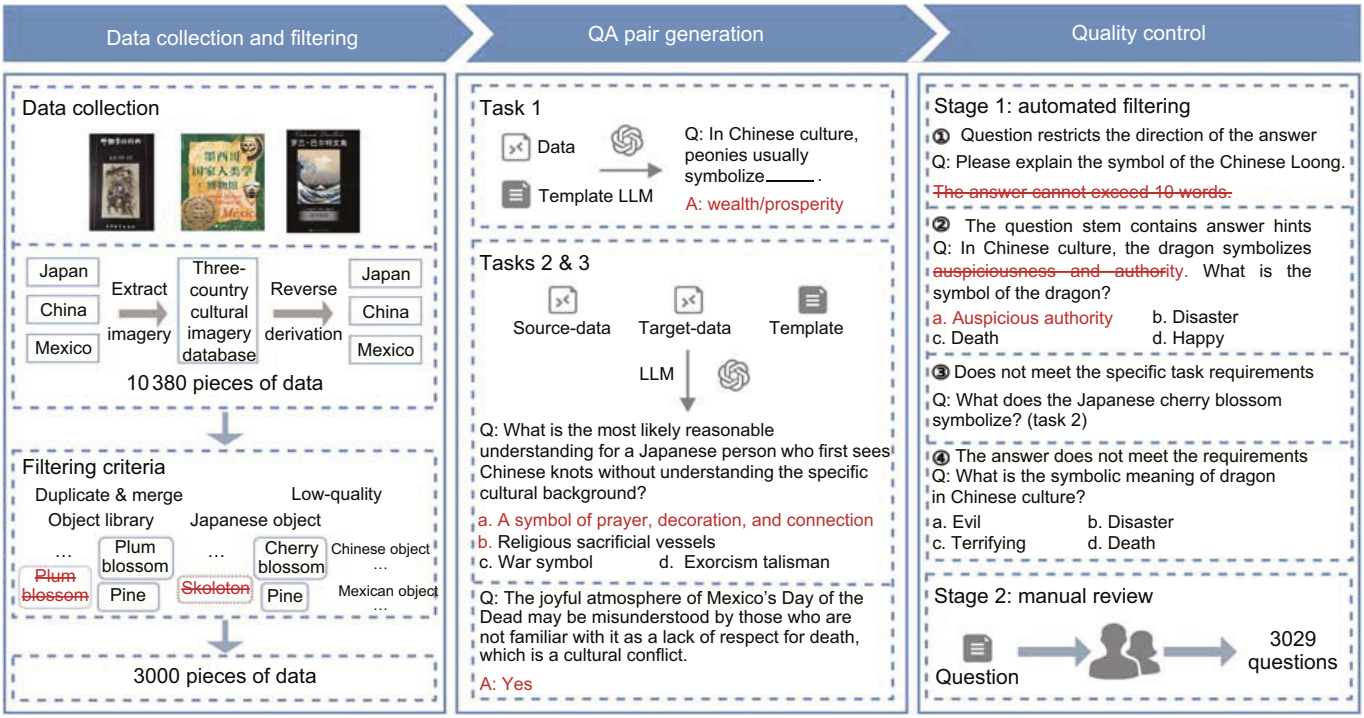

Fig. 2 Construction process of the benchmark, including data collection and filtering, QA pair generation, and quality control

data collection and filtering, QA pair generation, and quality control.

3.3.1 Data collection and filtering

We collect raw intention–object correspondences of cultural symbols from representative and authoritative books in cultural anthropology, folklore, cultural history, and intercultural communication across China, Japan, and Mexico, including *Dictionary of Chinese Symbols: Hidden Symbols in Chinese Life and Thought* (Eberhard, 1988), *National Museum of Anthropology, México* (Mei, 2023), and *The Empire of Signs: Semiotic Essays on Japanese Culture* (Ikegami, 1991). Based on these sources, we further adopt a bidirectional mapping strategy (Hogan et al., 2022) to expand coverage and diversity. Specifically, after extracting intention–object pairs within each culture, we use either symbols or intentions as anchors to retrieve semantically related counterparts across cultures. For example, the Japanese cherry blossom connotation of "transient yet brilliant life" is aligned with peach blossoms in Chinese culture and marigolds in Mexican culture. Ultimately, we establish cross-cultural correspondences among the three countries and collect 10 380 initial entries.

We then filter the data by removing duplicates and excluding symbols with weak cross-cultural relevance. For example, the Mexican skull symbol, strongly associated with Día de los Muertos, has no clear counterpart in Japanese culture. After cleaning, 3029 entries are retained.

3.3.2 Question–answer pair generation

In this stage, we distinguish the QA generation methods for task 1 and for tasks 2 and 3. We generate QA pairs with the assistance of LLMs (Wang YZ et al., 2023; Liu et al., 2024; Tang et al., 2025), along with our curated dataset and predefined rules. As exemplified in Fig. 2, we randomly select a document from the preprocessed corpus as evidence, and then use a template-based prompt to instruct the LLM to generate a question along with its corresponding answer. To adapt to the characteristics of different tasks and to examine the impact of diverse question formats on model performance, we construct five question types: true/false, single-choice, multiple-choice, fill-in-the-blank, and open-ended questions. For tasks 1 and 2, we employ four of these formats to comprehensively verify the ability of LLMs to understand and translate cultural symbols. For task 3, we set true/false and open-ended questions, which can effectively reflect the ability of LLMs to identify cultural conflicts and deeply analyze their underlying causes.

3.3.3 Quality control

To ensure data validity, we establish a pipeline consisting of automatic filtering and manual verification (Chen et al., 2021). We first use LLMs to conduct unified inspection and filtering on the generated QA pairs, focusing on removing entries where (1) the stems restrict the answering scope of LLMs, (2) the stems leak key information such as answers, (3) the content does not meet task requirements (e.g., alignment-related questions appear in task 1 for semantic recognition), and (4) the questions lack correct answers among the options. In addition, we examine the answer distribution of objective questions to avoid excessive concentration on a small subset of options.

After automatic filtering, we conduct a manual verification stage to ensure the accuracy of gold answers. Three experts in cross-cultural communication participate in the annotation process. Prior to formal annotation, all experts undergo two rounds of training and discussion to establish consistent judgment criteria. The annotation guidelines are defined as follows: (1) factual correctness—the answer must be directly supported by authoritative cultural sources; (2) cultural appropriateness—the answer must reflect the core connotation of the symbol within its original cultural context without distortion or overgeneralization; (3) task relevance—for alignment and conflict tasks, the answer must explicitly address cross-cultural correspondence or conflict reasoning rather than merely repeating intra-cultural knowledge.

4 Experiments

4.1 Setup

1. Evaluated models

Our evaluation covers 19 LLMs (7 closed-source and 12 open-source models). The selected LLMs include industry-leading closed-source models such as GPT-5 (OpenAI, 2025) and Gemini-3-Pro (Anil et al., 2023), as well as representative open-source models including the Qwen series (Qwen Team, 2025) and DeepSeek series (Wu et al., 2024). Additionally, we include other mainstream open-source models such as InternLM3 (Bi et al., 2026) and GLM-4 (Team GLM, 2024) to ensure comprehensive coverage of state-of-the-art open-source LLMs.

2. Evaluation criteria

CCU-Bench includes objective and open-ended questions. We use accuracy for single-choice and true/false questions, F1 score (the harmonic mean of precision and recall) for multiple-choice questions, and an LLM-as-judge framework for fill-in-the-blank and open-ended questions (Austin et al., 2021; An et al., 2024; Hu et al., 2025; He et al., 2026; Zhu et al., 2027). For fill-in-the-blank questions, we score each blank separately and use the average value of all blank scores as the question's final score.

Under the LLM-as-judge framework, in particular, we randomly select 100 samples to compare human judgments with GPT-5.1, Claude-4.5-Opus, and Gemini-2.5-Flash. All three models achieve strong agreement with human annotations, with Spearman's rank correlation coefficient $\rho > 0.75$ and Kendall's rank correlation coefficient $\tau > 0.6$. Although Claude-4.5-Opus achieves the best performance, GPT-5.1 offers a better balance of performance, efficiency, and cost. Consequently, we select GPT-5.1 as the evaluation model, yielding a Spearman's $\rho = 0.7688$ and a Kendall's $\tau = 0.6131$ against human annotations.

4.2 Main results

As shown in Table 3, the overall performance of 19 evaluated LLMs on cross-cultural understanding tasks is unsatisfactory, with an average overall score of 57.8%, which is below the average. Even the best-performing model, GPT-5, achieves only 70.2% in the overall score, while the weakest

Table 3 Performance comparison of the three tasks in CCU-Bench, with overall indicating the average performance across all tasks

Category	Model	Performance (%)			
		Intra-cultural symbolic understanding	Cross-cultural symbolic alignment	Cross-cultural conflict identification	Overall
Closed-source	GPT-5	**68.8**	**63.0**	**78.9**	**70.2**
	Claude Sonnet 4.5	<u>66.3</u>	57.6	<u>73.9</u>	<u>65.8</u>
	Claude Sonnet 4.6	61.4	59.0	66.0	62.1
	Gemini-3-Pro	63.6	54.9	66.5	60.9
	Gemini-2.5-Pro	66.0	<u>61.2</u>	59.2	62.2
	GPT-4o	62.8	49.7	72.8	61.7
	Qwen3.6-Plus	61.4	54.3	72.0	62.5
Open-source	GLM-4-32B	59.6	55.7	52.3	55.9
	Qwen3.5-122B-A10B	60.3	54.4	71.4	61.9
	Qwen3.5-27B	59.3	54.0	68.5	60.5
	DeepSeek-V3.2	59.8	56.1	55.2	57.1
	DeepSeek-V4-Pro	60.1	58.3	68.8	62.2
	GLM-4.7	62.4	54.0	64.0	60.1
	Qwen3-32B	58.6	51.4	41.1	50.5
	InternLM3-8B	55.2	46.2	26.5	42.9
	Qwen3-8B	54.3	48.1	40.8	47.9
	Qwen3-4B	50.9	45.8	32.0	43.1
	Qwen2.5-72B-Instruct	61.3	54.8	54.4	57.0
	Llama3.1-405B	56.4	45.8	61.8	54.6

The best results are in bold, and the second-best results are underlined

open-source model, InternLM3-8B, reaches merely 42.9%. The large gap between models reflects significant performance divergence (Masoud et al., 2025).

The distribution of model performance shows an obvious pattern of polarization. Accordingly, nearly half of the 19 models achieve an overall score below 60%, and only two closed-source models (GPT-5 and Claude Sonnet 4.5) exceed 65%. This indicates that current LLMs remain at a relatively low level in cross-cultural understanding, a knowledge-intensive task. Even the top-performing models have not reached a stable and reliable level, suggesting considerable room for optimization and improvement in this field.

Experimental results clearly demonstrate that closed-source models consistently outperform open-source models in cross-cultural understanding tasks, with advantages persisting across all core tasks. As presented in Table 3, the average overall performance of closed-source models is 63.6%, higher than the 54.5% of open-source models, representing a substantial gap. This gap varies across tasks: in the intra-cultural symbolic understanding task, the difference is relatively small (64.3% for closed-source models and 58.2% for open-source models), and some open-source models, such as Qwen2.5-72B-Instruct and GLM-4.7, can approach the performance of closed-source models. However, the gap widens significantly in cross-cultural symbolic alignment (task 2) and cross-cultural conflict identification (task 3). For task 2, closed-source models achieve an average score of 57.1% while open-source models only reach 52.1%; some open models, such as Qwen3-4B and Llama3.1-405B, even score below 46%. In task 3, closed-source models achieve 69.9% overall performance compared to 53.1% for open-source models, with most open-source models struggling to deeply explain the roots of conflicts.

In addition, closed-source models exhibit greater robustness in fill-in-the-blank questions requiring precise term extraction and open-ended questions demanding complete logical chains, without extreme weaknesses such as missing core information or broken mapping logic commonly observed in open-source models. These results suggest that although open-source models show certain potential in basic cultural knowledge memorization, they still suffer from notable deficiencies in advanced capabilities such as cross-cultural reasoning and complex context adaptation. Future research should focus on enhancing the depth of cultural knowledge integration and the architecture of multimodal reasoning.

These observations suggest that general language pretraining alone is insufficient to support stable cross-cultural understanding. Achieving reliable performance in this setting requires more explicit cultural knowledge integration, stronger cross-cultural alignment mechanisms, and higher-order reasoning capabilities beyond basic next-token prediction.

4.3 Detailed analysis

We further analyze model performance from the perspective of the three core tasks, combining overall results on all 19 evaluated models in Table 3 with fine-grained question-type performance of three representative closed-source models in Fig. 3, namely, GPT-5, Claude Sonnet 4.5, and Gemini-3-Pro. Fig. 4 displays several representative examples of question–answering variance across three different models. This analysis provides a more complete view of where current LLMs succeed and fail in cross-cultural understanding.

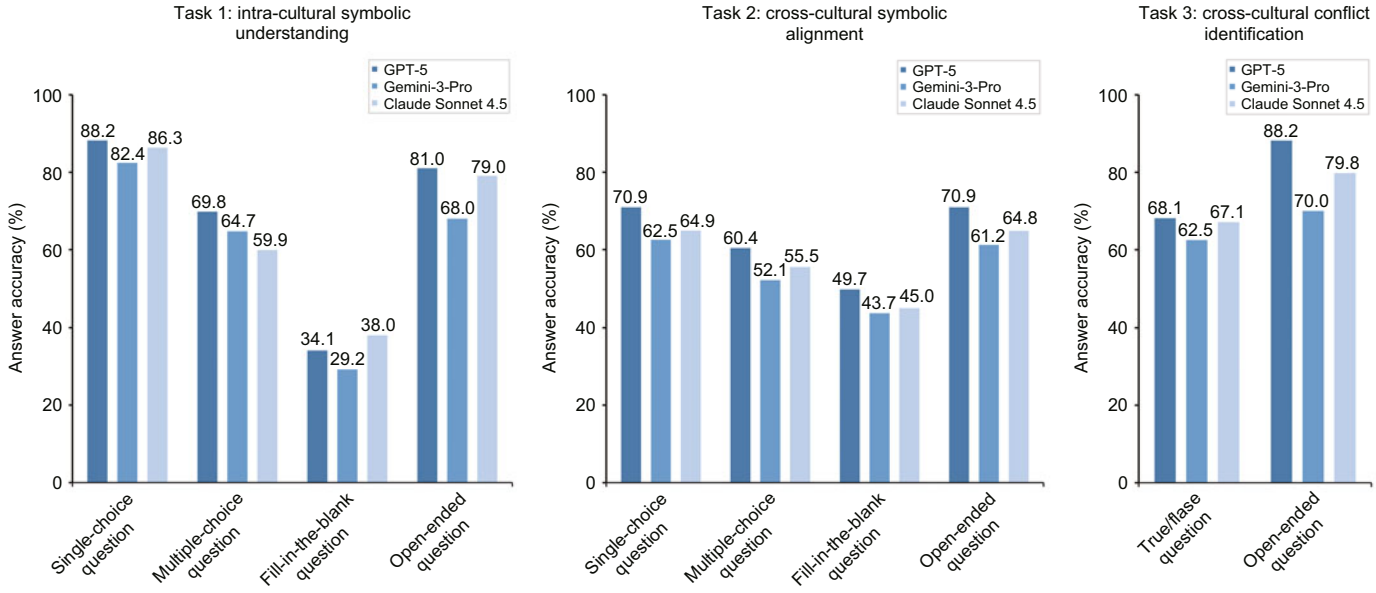

Fig. 3 Detailed performance comparison of three closed-source models on CCU-Bench

4.3.1 Task 1: intra-cultural symbolic understanding

As presented in Table 3, task 1 is the best-performing stage among the three core tasks, with closed-source models achieving an average score of 64.3% and open-source models averaging 58.2%. Among closed-source models, GPT-5 (68.8%), Claude Sonnet 4.5 (66.3%), and Gemini-2.5-Pro (66.0%) lead in performance, while advanced open-source models, such as Qwen3.5-122B-A10B (60.3%), GLM-4.7 (62.4%), and DeepSeek-V4-Pro (60.1%), achieve competitive overall results. This advantage likely stems from the large amount of single-culture common knowledge accumulated during pre-training. However, as shown in Table 4, models perform notably poorly on fill-in-the-blank questions (30.6% on average), indicating limited ability to extract precise cultural terminology.

This pattern is also reflected in the three representative closed-source models. As illustrated in Fig. 3, all three models perform best on task 1, with single-choice and open-ended questions generally exceeding 65%. GPT-5 leads on single-choice questions (about 88%) and open-ended questions (about 81%), followed by Claude Sonnet 4.5 (about 86%/79%) and Gemini-3-Pro (about 82%/68%). By contrast, performance on multiple-choice questions drops to about 60%–70%, and

Table 4 Comparison of the performance of 19 LLMs on different question types in CCU-Bench

Category	Model	Performance (%)				
		True/false	Single-choice	Multiple-choice	Fill-in-the-blank	Open-ended
Closed-source	GPT-5	68.1	79.7	65.1	**41.9**	**82.2**
	Claude Sonnet 4.5	67.1	86.3	59.9	38.0	<u>79.0</u>
	Claude Sonnet 4.6	<u>70.6</u>	82.0	62.0	31.0	69.0
	Gemini-3-Pro	62.5	82.4	64.7	29.2	68.0
	Gemini-2.5-Pro	66.1	84.4	**68.4**	33.0	76.5
	GPT-4o	68.4	79.8	<u>66.6</u>	31.9	71.2
	Qwen3.6-Plus	68.6	<u>87.0</u>	63.0	29.0	65.0
Open-source	GLM-4-32B	57.9	83.6	62.8	<u>40.7</u>	51.2
	Qwen3.5-122B-A10B	64.7	86.6	62.1	29.9	61.2
	Qwen3.5-27B	67.1	84.7	66.4	25.7	59.1
	DeepSeek-V3.2	69.3	85.5	60.2	29.5	62.7
	DeepSeek-V4-Pro	64.9	77.8	57.6	39.9	63.1
	GLM-4.7	**72.8**	**87.4**	64.6	28.7	67.2
	Qwen3-32B	58.8	82.1	60.8	28.1	62.0
	InternLM3-8B	37.7	79.8	60.0	24.6	55.3
	Qwen3-8B	68.4	78.2	56.4	22.1	59.1
	Qwen3-4B	52.9	74.4	55.7	17.5	54.6
	Qwen2.5-72B-Instruct	62.9	85.0	64.0	34.0	63.0
	Llama3.1-405B	52.9	80.2	52.4	26.3	65.0

The best results are in bold, and the second-best results are underlined

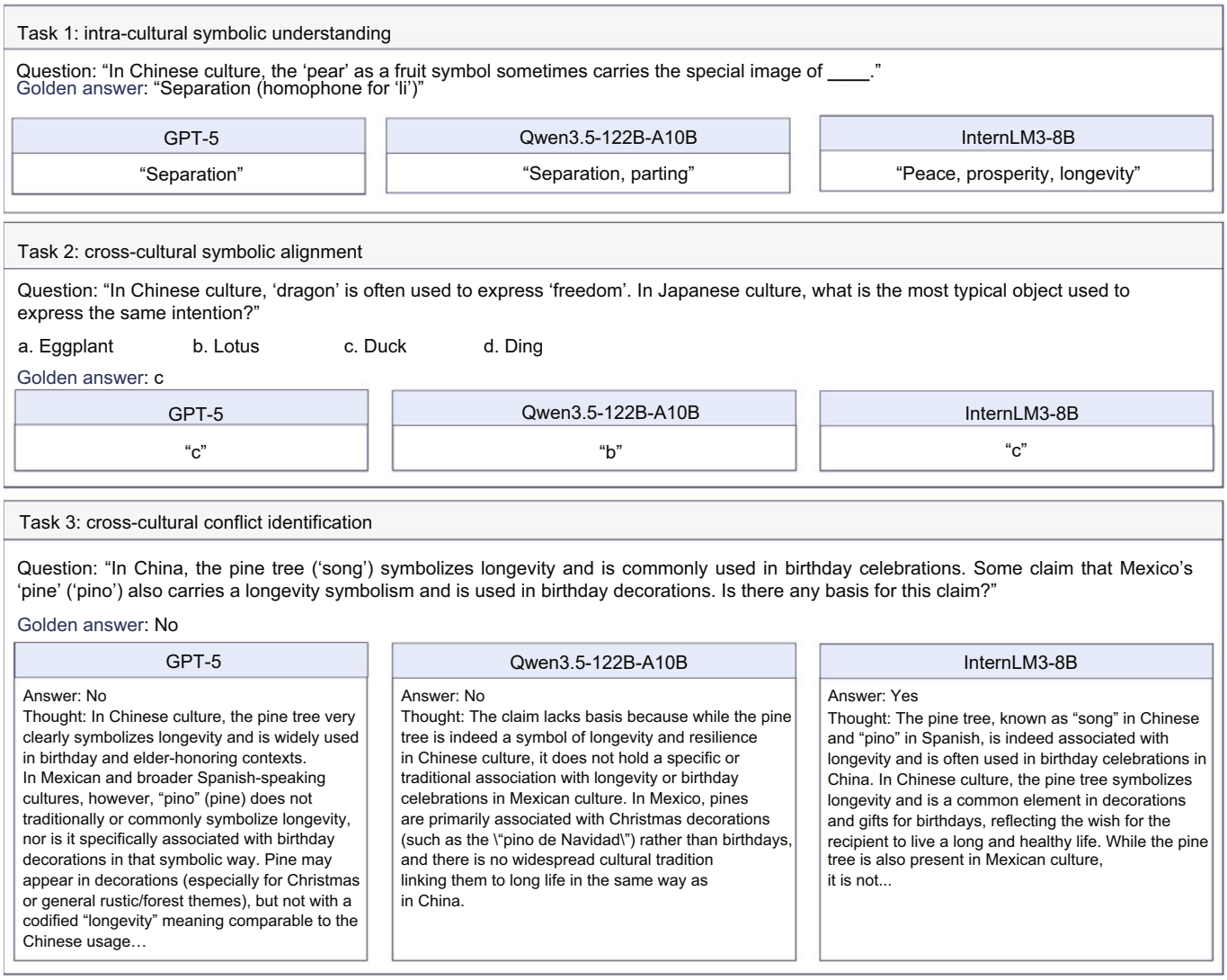

Fig. 4 QA examples of LLMs on CCU-Bench, covering three core cross-cultural tasks. Above are QA examples of the top-performing model (GPT-5), the poorest-performing model (InternLM3-8B), and the best open-source model (Qwen3.5-122B-A10B) from our experimental results on CCU-Bench

fill-in-the-blank questions remain the most difficult format, with Gemini-3-Pro scoring the lowest (about 29%). This suggests that even strong closed-source models still struggle to distinguish closely related symbolic meanings and accurately retrieve culture-specific terms.

4.3.2 Task 2: cross-cultural symbolic alignment

Task 2 is the weakest link in the overall model performance. As shown in Table 3, closed-source models obtain an average of 57.1% and open-source models obtain an average of 52.1%, resulting in an overall average of 53.9%. Among closed-source models, GPT-5 (63.0%) and Gemini-2.5-Pro (61.2%) perform relatively well, but both fall clearly behind their task 1 scores, highlighting the difficulty of cross-cultural knowledge transfer (Kong et al., 2025; Masoud et al., 2025). Open-source models generally achieve weak performance on this task. For instance, Qwen3-4B (45.8%) and Llama3.1-405B (45.8%) obtain relatively low results.

The errors of these models reveal a typical failure pattern, as demonstrated in Fig. 4. First, these models often fail to follow answer constraints in single-choice and multiple-choice settings, outputting multiple options with explanations instead of the required answer only. Second, in fill-in-the-blank and open-ended questions, they frequently produce symbolic mapping errors by relying on superficially similar meanings while ignoring deeper differences in symbolic function and cultural value. Typical examples include mapping bamboo to plum blossom, or pine tree to peony. These results show that current models, especially weaker open-source ones, lack robust cross-cultural symbolic alignment ability.

According to Fig. 3, the same bottleneck appears in the fine-grained analysis of closed-source models. Compared with task 1, the three representative models show divergent performance changes on task 2 across different question types. Their scores drop noticeably on single-choice and multiple-select formats, whereas fill-in-the-blank items yield higher results relative to task 1. Single-choice and open-ended questions remain the relative strengths among the question types, with GPT-5 achieving over 70% on both, followed by Claude Sonnet 4.5

(64.9%/64.8%) and Gemini-3-Pro (62.5%/61.2%). However, multiple-choice and fill-in-the-blank questions remain substantially weaker, with Gemini-3-Pro performing worst on both. Errors are concentrated in cross-culturally similar option interference and terminology transfer deviation, suggesting that even strong closed-source models remain limited in fine-grained cross-domain semantic differentiation. Overall, task 2 remains the core bottleneck for current LLMs.

4.3.3 Task 3: cross-cultural conflict identification

Interestingly, Table 3 implies that the performance of task 3 falls between the preceding two tasks, with closed-source models averaging 69.9% and open-source models averaging 53.1%, yielding an overall average of 59.3%. Among closed-source models, GPT-5 (78.9%) and Claude Sonnet 4.5 (73.9%) stand out, showing relatively strong capabilities in conflict judgment and causal explanation. Among open-source models, InternLM3-8B performs worst in conflict identification, with a score of 26.5% and many cases of misjudgment.

Such errors often arise not from missing cultural knowledge, but from weak higher-order reasoning. For example, in the case of Spring Festival dumplings versus Mexican tamales, the model can correctly identify their shared festive and collective-production contexts, but incorrectly equates situational similarity with equivalence in symbolic meaning, concluding that both primarily symbolize family reunion. This suggests that weaker models struggle to distinguish between surface-level contextual similarity and deeper differences in symbolic connotation.

Fine-grained analysis in Fig. 3 further confirms this pattern. Task 3 includes only true/false and open-ended questions. In open-ended questions, GPT-5 performs best at 88.2%, followed by Claude Sonnet 4.5 (79.8%) and Gemini-3-Pro (70.0%), indicating a stronger ability in explicit conflict explanation and causal reasoning. On true/false questions, all three models are more balanced, scoring 62.5%–68.1%, with GPT-5 slightly ahead. This suggests that current closed-source models can generally detect whether conflict exists, but struggle in borderline cases where a symbolic difference does not necessarily imply an actual conflict.

4.3.4 Brief summary

Across the overall and fine-grained analyses, a consistent performance pattern emerges: task 1>task 3>task 2. In other words, current LLMs are relatively capable of understanding cultural symbols within a single-culture context, somewhat capable of identifying cross-cultural conflict, but substantially limited in cross-cultural symbolic alignment. At the model level, GPT-5 performs best overall, especially in advanced reasoning and cross-domain mapping; Claude Sonnet 4.5 shows the most stable and balanced performance; Gemini-3-Pro is comparatively weaker, particularly in terminology extraction and cross-cultural alignment. Across question types, fill-in-the-blank questions remain a common weakness, while multiple-choice and open-ended questions are generally more manageable for strong models.

4.4 Error mode analysis

To further investigate the core deficiencies of LLMs in cross-cultural understanding, we randomly sample 380 low-scoring cases from the responses of all 19 evaluated models across the three core tasks and five question types. We manually categorize model errors into four types: answer norm deviation, cultural knowledge deficit, intent mapping bias, and conflict recognition failure. The results are summarized in Table 5.

A clear pattern emerges: most model failures are not caused by superficial formatting issues, but by deficiencies along the progressive reasoning chain of cross-cultural understanding. In particular, cultural knowledge deficit (49.7%), intent mapping bias (16.6%), and conflict recognition failure (12.1%) together account for 78.4% of all errors, while answer norm deviation represents 21.6%. This suggests that the main bottleneck of current LLMs lies in culturally grounded understanding and reasoning rather than simple instruction-following.

At the first level, cultural knowledge deficit reflects

Table 5 Statistics on error type distribution across 380 sampled low-scoring questions

Primary category	Sub-error type	Count*	Total count	Proportion (%)
Answer norm deviation	Option content output	38 (46.3%)	82	21.6
	Linguistic difference in output	14 (17.1%)		
	Short-answer format	24 (29.3%)		
	Only thought chain without answer	6 (7.3%)		
Cultural knowledge deficit	Core intent omission	47 (24.9%)	189	49.7
	Unknown core intent of symbol	18 (9.5%)		
	Core intent bias	99 (52.4%)		
	Reversed intent	25 (13.2%)		
Intent mapping bias	Wrong mapping object	55 (87.3%)	63	16.6
	Core mapping bias	8 (12.7%)		
Conflict recognition failure	Cultural value orientation misjudgment	36 (78.3%)	46	12.1
	Symbol hierarchy confusion	7 (15.2%)		
	Cultural stereotype	3 (6.5%)		

* The values in the parentheses represent the proportion of each sub-error type within its corresponding primary category

insufficient understanding of the core connotations of culture-specific symbols. This type of error is mainly concentrated in open-ended and fill-in-the-blank questions, with core intent omission (24.9%) and core intent bias (52.4%) as the dominant sub-error types. For example, when asked to explain the meaning of chrysanthemum in traditional Chinese culture, a model mentions only its natural property of blooming in autumn while failing to capture its core associations with nobility and seclusion. Such errors indicate that many models lack robust intra-cultural symbolic understanding.

At the second level, intent mapping bias accounts for 16.6% of all errors, revealing that cross-cultural symbolic alignment remains a central challenge for current models. Most of these cases (87.3%) are mapping object errors, where models select an incorrect counterpart in another culture despite partial semantic similarity. For instance, one model maps the Mexican funeral counterpart of the Chinese chrysanthemum to rose rather than marigold, showing that it could not establish an appropriate cross-cultural symbolic correspondence. This type of failure suggests that models often rely on shallow semantic overlap rather than deep cultural alignment.

At the third level, conflict recognition failure accounts for 12.1% of all errors and appears exclusively in open-ended questions. All cases fall into the misjudgment of conflict existence, indicating that models often fail to determine whether symbolic differences actually constitute cultural conflict. For example, a model incorrectly interprets the differing symbolism of the crane in Chinese and Republic of Korean cultures as a hierarchical conflict, while another judges bamboo as culturally conflicting between Chinese and Japanese contexts despite their shared association with perseverance. These results suggest that current models remain weak in higher-order reasoning about symbolic incompatibility and cross-cultural risk.

By contrast, answer norm deviation accounts for 21.6% of all errors, making it a frequent but non-core category. These errors involve violations of required answer formats rather than substantive failures in cultural reasoning. Typical cases include outputting option contents instead of labels in objective questions or responding in an open-ended style when a constrained answer is required. While such issues indicate imperfect instruction-following, they are secondary compared with the deep knowledge and reasoning failures above.

Overall, the error distribution aligns closely with the three capability levels targeted by CCU-Bench. Models first fail to accurately understand the connotations of individual cultural symbols, then struggle to align symbolic meanings across cultures, and finally break down in judging whether such differences lead to actual conflict.

This progressive failure pattern highlights that the core limitation of current LLMs in cross-cultural understanding lies not in surface-level response formatting, but in insufficient culturally grounded knowledge, weak symbolic alignment, and unreliable higher-order reasoning about cross-cultural conflicts.

This progressive error pattern echoes the three-task design of CCU-Bench and validates the rationality of our benchmark construction. It indicates that future culturally aware LLMs must be equipped with hierarchical cultural understanding capabilities, rather than relying on uniform modeling across all cultural tasks (Masoud et al., 2025).

5 Discussion

From the comprehensive evaluation results on CCU-Bench, we derive several critical and generalizable insights into the inherent limitations of current LLMs in cross-cultural understanding.

First, LLMs mostly rely on shallow semantic matching rather than in-depth cultural grounding, leading to frequent misalignment in cross-cultural symbol mapping. General language pre-training alone cannot endow models with structured cultural logic or value-aware interpretation capabilities.

Second, the evident performance ranking of task 1> task 3>task 2 confirms that cross-cultural symbolic alignment is the core bottleneck, indicating that cross-cultural transfer and adaptive reasoning require dedicated design rather than emergent abilities.

Third, the considerable performance gap between closed-source and open-source models reveals that high-quality cultural annotation, targeted fine-tuning, and robust reasoning architectures are more critical than model scale alone for reliable cross-cultural understanding. These insights imply that constructing truly culture-aware LLMs requires a shift from general language modeling to culture-centric symbolic representation and reasoning.

Based on the systematic evaluation results of CCU-Bench, we distill three key directions for optimizing the cross-cultural understanding capability of current LLMs.

First, it is necessary to strengthen the precise storage and structured integration of cultural knowledge. Among the 380 sampled low-scoring questions, cultural knowledge deficit takes up the largest share of all error types, reaching 49.7%. This fully proves that existing models have prominent defects in understanding the core connotations of target cultural symbols. Inside this error category, core intent bias is the most common sub-error, accounting for 52.4% of all cultural knowledge errors. In the future, we can enhance the model's ability to accurately grasp the connotations of different cultural symbols by constructing a cross-cultural symbol knowledge base and optimizing the coverage of cultural diversity in pre-training data, with a particular focus on supplementing data on symbolic correspondence between regions with large cultural distances (e.g., China and Mexico).

Second, priority should be given to optimizing cross-cultural reasoning and semantic mapping mechanisms. Intent mapping bias accounts for 16.6% of total errors. Within this category, wrong mapping objects account for 87.3% of all mapping bias errors, which reflects severe deficiencies in the logic of cross-cultural symbol correspondence in models. This issue can be addressed by designing specialized cross-cultural symbolic alignment pre-training tasks and introducing a bidirectional cultural symbol alignment framework, enabling models to establish connections between symbolic connotations and cultural contexts and avoid confusing symbols with similar surface imagery but fundamentally different meanings.

Third, it is essential to enhance higher-order reasoning about cultural conflicts and improve scenario adaptation capabilities. Open-source models lag significantly behind closed-source models in cross-cultural conflict identification tasks (53.1% vs. 69.9% on average), and among conflict recognition

failures, cultural value orientation misjudgment is the dominant error type (78.3%), which collectively exposes their shortcomings in complex reasoning. In the future, we can improve the model's ability to deeply analyze the causes of cultural conflicts by constructing a cultural conflict case library and designing hierarchical reasoning training paradigms, while optimizing the automated evaluation framework to refine the scoring dimensions for conflict judgment.

6 Conclusions and future work

This work proposes CCU-Bench, which constructs an evaluation framework consisting of three core tasks and five question types for cross-cultural understanding ability. It covers cultural symbols from China, Japan, and Mexico, with a total of 3029 QA pairs. Through systematic evaluation of 19 mainstream LLMs, we find that current models exhibit overall unsatisfactory performance on cross-cultural understanding tasks (average score 57.8%) and show a clear pattern of polarization. Closed-source models outperform open-source models in overall performance, with significant advantages especially in intra-cultural symbolic understanding and conflict identification tasks. Although models have certain potential in memorizing basic cultural knowledge, they suffer from notable deficiencies in advanced capabilities such as cross-cultural reasoning and complex context adaptation.

Error pattern analysis shows that cultural knowledge deficit and intent mapping bias are the dominant error types (accounting for 66.3% in total), while conflict identification failure reflects insufficient higher-order reasoning ability of models. These results clearly reveal the core deficiencies of current LLMs in cross-cultural symbolic alignment and point to optimization directions for future research. Follow-up studies need to deepen cultural knowledge integration, establish rigorous logical rules for cross-cultural mapping, and boost the capability of higher-order reasoning over cross-cultural conflicts. Meanwhile, researchers ought to expand evaluation dimensions and enrich data diversity for benchmarks dedicated to cross-cultural symbolic alignment tasks.

Although CCU-Bench establishes a complete evaluation paradigm for cross-cultural symbolic reasoning, there remain several unexplored dimensions and research boundaries that open promising directions for follow-up studies.

First, CCU-Bench currently focuses on three cultural regions: China, Japan, and Mexico. The selection of these three countries is theoretically grounded in Hofstede's cultural dimensions theory. According to this framework, cultural distance—the degree of difference in values, norms, and symbolic systems between cultures—is a key determinant of cross-cultural misunderstanding. Specifically, the cultural distance between China and Japan is 56.8, while the distance between China and Mexico is 96.9, providing a gradient of increasing difficulty for cross-cultural understanding tasks (Yeh, 1988). China, Japan, and Mexico represent three distinct cultural clusters, East Asian Confucian, East Asian Japonized, and Latin American, respectively, thereby offering meaningful coverage of non-Western cultural contexts. Despite this theoretical foundation, we acknowledge that three countries are insufficient to represent the full diversity of global cultures. In the future, expanding to more regions (e.g., Middle Eastern, South Asian, and Sub-Saharan African cultures) and additional cultural backgrounds will further enhance the generalizability and robustness of the benchmark.

Second, the evaluation predominantly relies on text-based questions and does not include multimodal symbolic inputs such as images and videos. Integrating multimodal data will make the benchmark closer to real cross-cultural communication scenarios.

Third, the automatic evaluation framework emphasizes correctness and consistency but cannot fully measure the naturalness, fluency, and practicality of cross-cultural understanding outputs. Future work can introduce human evaluation and multi-dimensional scoring to complement automatic metrics.

Fourth, the error analysis is based on a sample of low-score cases; a larger and more balanced sample will help reveal more detailed failure patterns and provide clearer guidance for model improvement. We will address these limitations in future versions of CCU-Bench to build a more comprehensive and practical benchmark for cross-cultural understanding.

Fifth, most open-source models evaluated in this work are smaller in scale compared to closed-source counterparts, which may limit their cross-cultural understanding capabilities and affect the comparability of results.

CCU-Bench fills the gap in progressive evaluation of cross-cultural understanding in LLMs and provides a standardized testbed for future research on culturally aware AI. Based on this benchmark, we will further embed structured cultural knowledge into model systems, refine the mechanisms for cross-cultural symbolic mapping, and continuously strengthen the models' reasoning performance when facing diverse cultural conflicts. Beyond these algorithmic improvements, we will expand the cultural coverage of CCU-Bench, introduce multimodal symbolic inputs, and polish the overall evaluation framework to support more in-depth research on cross-cultural understanding. Moreover, we will introduce comprehensive human evaluation with multidimensional scoring criteria to further enhance the reliability of the benchmark results.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62502423 and 62421003).

Author contributions

Wei ZHANG designed the research, coordinated the project, interpreted the results, and revised and finalized the paper. Ting YIN conducted the experiments, analyzed the data, and drafted the paper. Jia CHEN curated and processed the data. Wenjie NING verified the data. Xiaoyan GU and Kai LIU developed the methodology. Hengnian QI supervised the research. Wei CHEN supervised the research and provided resources.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study, including model versions, parameters, inference settings, evaluation prompts,

and scoring pipelines, are available from our anonymous repository: https://anonymous.4open.science/r/benchmark-60FF/.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Adilazuarda MF, Mukherjee S, Lavania P, et al., 2024. Towards measuring and modeling "Culture" in LLMs: a survey. Proc Conf on Empirical Methods in Natural Language Processing, p.15763-15784. https://doi.org/10.18653/v1/2024.emnlp-main.882

An CX, Gong SS, Zhong M, et al., 2024. L-Eval: instituting standardized evaluation for long context language models. Proc 62nd Annual Meeting of the Association for Computational Linguistics, p.14388-14411. https://doi.org/10.18653/v1/2024.acl-long.776

Anil R, Borgeaud S, Alayrac J, et al., 2023. Gemini: a family of highly capable multimodal models. https://doi.org/10.48550/arXiv.2312.11805

Anthropic, 2025. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5 [Accessed on Sept. 29, 2025].

Arora A, Kaffee LA, Augenstein I, 2023. Probing pre-trained language models for cross-cultural differences in values. Proc 1st Workshop on Cross-Cultural Considerations in NLP, p.114-130. https://doi.org/10.18653/v1/2023.c3nlp-1.12

Austin J, Odena A, Nye M, et al., 2021. Program synthesis with large language models. https://doi.org/10.48550/arXiv.2108.07732

Bi XJ, Li S, Xing JY, 2026. A weakly supervised preference alignment framework for robust ancient Chinese translation. *Patt Recogn*, 179:113562. https://doi.org/10.1016/j.patcog.2026.113562

Cao Y, Zhou L, Lee S, et al., 2023. Assessing cross-cultural alignment between ChatGPT and human societies: an empirical study. Proc 1st Workshop on Cross-Cultural Considerations in NLP, p.53-67. https://doi.org/10.18653/v1/2023.c3nlp-1.7

Chen M, Tworek J, Jun H, et al., 2021. Evaluating large language models trained on code. https://doi.org/10.48550/arXiv.2107.03374

Chiu YY, Jiang LW, Lin BY, et al., 2024. CulturalBench: a robust, diverse, and challenging cultural benchmark by human-AI CulturalTeaming. https://doi.org/10.48550/arXiv.2410.02677

Dhar S, Shamir L, 2021. Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks. *Vis Inform*, 5(3):92-101. https://doi.org/10.1016/j.visinf.2021.10.001

Durmus E, Nguyen K, Liao TI, et al., 2023. Towards measuring the representation of subjective global opinions in language models. https://doi.org/10.48550/arXiv.2306.16388

Eberhard W, 1988. Dictionary of Chinese Symbols: Hidden Symbols in Chinese Life and Thought. Routledge, London, UK.

Feng SB, Park CY, Liu YH, et al., 2023. From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.11737-11762. https://doi.org/10.18653/v1/2023.acl-long.656

Guo MQ, Hwa R, Kovashka A, 2023. Decoding symbolism in language models. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.3311-3324. https://doi.org/10.18653/v1/2023.acl-long.186

He C, Huang Y, Lan H, et al., 2026. AI-assisted assessment of higher education quality: a visual analytical approach. *Vis Inform*, 10:100306. https://doi.org/10.1016/j.visinf.2026.100306

Hofstede GH, 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations. Sage Publications, Thousand Oaks, USA.

Hogan A, Blomqvist E, Cochez M, et al., 2022. Knowledge graphs. *ACM Comput Surv*, 54(4):71. https://doi.org/10.1145/3447772

Hu Y, Song Z, Yu J, et al., 2025. TimeJudge: empowering video-LLMs as zero-shot judges for temporal consistency in video captions. *Front Inform Technol Electron Eng*, 26(11):2204-2214. https://doi.org/10.1631/FITEE.2500412

Ikegami Y, 1991. The Empire of Signs: Semiotic Essays on Japanese Culture. John Benjamins Publishing, Amsterdam, the Netherlands.

Johnson RL, Pistilli G, Menédez-González N, et al., 2022. The ghost in the machine has an American accent: value conflict in GPT-3. https://doi.org/10.48550/arXiv.2203.07785

Kabir M, Ahmed T, Rahman MM, et al., 2026. XCR-Bench: a multi-task benchmark for evaluating cultural reasoning in LLMs. https://doi.org/10.48550/arXiv.2601.14063

Kharchenko J, Roosta T, Chadha A, et al., 2025. How well do LLMs represent values across cultures? Empirical analysis of LLM responses based on Hofstede cultural dimensions. https://doi.org/10.48550/arXiv.2406.14805

Kong L, Zhong X, Chen J, et al., 2025. Multi-perspective consistency checking for large language model hallucination detection: a black-box zero-resource approach. *Front Inform Technol Electron Eng*, 26(11):2298-2309. https://doi.org/10.1631/FITEE.2500180

Kusnick J, Andersen NS, Liem J, et al., 2026. A survey on visualization-based storytelling in digital humanities and cultural heritage. *Vis Inform*, 10(2):100312. https://doi.org/10.1016/j.visinf.2026.100312

Li C, Chen MZ, Wang JD, et al., 2024. CultureLLM: incorporating cultural differences into large language models. Proc 38th Int Conf on Neural Information Processing Systems, Article 2693. https://doi.org/10.52202/079017-2693

Liu RB, Wei J, Liu FY, et al., 2024. Best practices and lessons learned on synthetic data. https://doi.org/10.48550/arXiv.2404.07503

Lu JG, Song LL, Zhang LD, 2025. Cultural tendencies in generative AI. *Nat Hum Behav*, p.2360-2369. https://doi.org/10.1038/s41562-025-02242-1

Masoud RI, Liu ZQ, Ferianc M, et al., 2025. Cultural alignment in large language models: an explanatory analysis based on Hofstede's cultural dimensions. Proc 31st Int Conf on Computational Linguistics, p.8474-8503.

Mei C, 2023. National Museum of Anthropology, México. Guangxi Normal University Press, Guilin, China (in Chinese).

Myung J, Lee N, Zhou Y, et al., 2024. BLEND: a benchmark for LLMs on everyday knowledge in diverse cultures and languages. Proc 38th Int Conf on Neural Information Processing Systems, Article 2483. https://doi.org/10.52202/079017-2483

Nguyen TP, Razniewski S, Varde A, et al., 2023. Extracting cultural commonsense knowledge at scale. Proc ACM Web Conf, p.1907-1917. https://doi.org/10.1145/3543507.3583535

OpenAI, 2025. OpenAI GPT-5 system card. https://doi.org/10.48550/arXiv.2601.03267

Peng YZ, Zhang GR, Zhang MS, et al., 2026. LMM-R1: empowering 3B LMMs with strong reasoning abilities with two-stage rule-based RL. *Front Comput Sci*, 20:1-28.

Prabhakaran V, Qadri R, Hutchinson B, 2022. Cultural incongruencies in artificial intelligence. https://doi.org/10.48550/arXiv.2211.13069

Qwen Team, 2025. Qwen3 technical report. https://doi.org/10.48550/arXiv.2505.09388

Rahman H, Salam H, 2026. CCD-Bench: probing cultural conflict in large language model decision-making. Proc 40th AAAI Conf on Artificial Intelligence, p.39125-39133. https://doi.org/10.1609/aaai.v40i46.41260

Romero D, Lyu CY, Wibowo HA, et al., 2024. CVQA: culturally-diverse multilingual visual question answering benchmark. 38th Conf on Neural Information Processing Systems, p.11479-11505. https://doi.org/10.52202/079017-0366

Rystrøm J, Kirk HR, Hale SA, 2025. Multilingual != multicultural: evaluating gaps between multilingual capabilities and cultural alignment in LLMs. Proc 1st Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models, p.74-85. https://doi.org/10.26615/978-954-452-101-1-009

Santurkar S, Durmus E, Ladhak F, et al., 2023. Whose opinions do language models reflect? Proc 40th Int Conf on Machine Learning, p.29971-30004.

Shen SQ, Logeswaran L, Lee M, et al., 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. Proc Conf on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.5668-5680. https://doi.org/10.18653/v1/2024.naacl-long.316

Shi WY, Li R, Zhang YT, et al., 2024. CultureBank: an online community-driven knowledge base towards culturally aware language technologies. Findings of the Association for Computational Linguistics, p.4996-5025. https://doi.org/10.18653/v1/2024.findings-emnlp.288

Tam ZR, Wu CK, Chiu YY, et al., 2025. Language matters: how do multilingual input and reasoning paths affect large reasoning models? https://doi.org/10.48550/arXiv.2505.17407

Tang Y, Qiao L, Yin L, et al., 2025. Training large-scale language models with limited GPU memory: a survey. *Front Inform Technol Electron Eng*, 26(3):309-331. https://doi.org/10.1631/FITEE.2300710

Team GLM, 2024. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. https://doi.org/10.48550/arXiv.2406.12793

Wang W, Yang Y, Pan Y, 2025. Visual knowledge in the big model era: retrospect and prospect. *Front Inform Technol Electron Eng*, 26(1):1-19. https://doi.org/10.1631/FITEE.2400250

Wang YZ, Kordi Y, Mishra S, et al., 2023. Self-instruct: aligning language models with self-generated instructions. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.13484-13508. https://doi.org/10.18653/v1/2023.acl-long.754

Wu ZY, Chen XK, Pan ZZ, et al., 2024. DeepSeek-VL2: mixture-of-experts vision-language models for advanced multimodal understanding. https://doi.org/10.48550/arXiv.2412.10302

Xu GH, Liu JY, Yan M, et al., 2023. CValues: measuring the values of Chinese large language models from safety to responsibility. https://doi.org/10.48550/arXiv.2307.09705

Xu SY, Dong WL, Guo ZS, et al., 2024. Exploring multilingual concepts of human values in large language models: is value alignment consistent, transferable and controllable across languages? Findings of the Association for Computational Linguistics, p.1771-1793. https://doi.org/10.18653/v1/2024.findings-emnlp.96

Yeh RS, 1988. On Hofstede's treatment of Chinese and Japanese values. *Asia Pac J Manag*, 6(1):149-160. https://doi.org/10.1007/BF01732256

Zhang W, Zhang J, Wong K, et al., 2024a. Computational approaches for traditional Chinese painting: from the "six principles of painting" perspective. *J Comput Sci Technol*, 39(2):269-285. https://doi.org/10.1007/s11390-024-3408-x

Zhang W, Kam-Kwai W, Chen Y, et al., 2024b. ScrollTimes: tracing the provenance of paintings as a window into history. *IEEE Trans Vis Comput Graph*, 30(6):2981-2994. https://doi.org/10.1109/TVCG.2024.3388523

Zhang W, Kam-Kwai W, Xu BY, et al., 2025. CultiVerse: towards cross-cultural understanding for paintings with large language model. Proc 33rd ACM Int Conf on Multimedia, p.6710-6719. https://doi.org/10.1145/3746027.3755698

Zhang W, Gu X, Liu H, et al., 2026. DAVA: decoding art with visual analytics through feature modeling and multi-agent collaboration. *IEEE Trans Vis Comput Graph*, 32(3):2773-2786. https://doi.org/10.1109/TVCG.2026.3653892

Zhou J, Ke P, Qiu XP, et al., 2024. ChatGPT: potential, prospects, and limitations. *Front Inform Technol Electron Eng*, 25(1):6-11. https://doi.org/10.1631/FITEE.2300089

Zhu JC, Zhu MH, Rui RT, et al., 2027. Evolutionary perspectives on the evaluation of LLM-based AI agents: a comprehensive survey. *Front Comput Sci*, 21:2101341.