Wei ZHANG, Ting YIN, Jia CHEN, Wenjie NING, Xiaoyan GU, Kai LIU, Hengnian QI, Wei CHEN, 2026. CCU-Bench: a systematic benchmark for cross-cultural understanding in large language models. *ENGINEERING Inform Technol Electron Eng*, 27(7):260083.
https://doi.org/10.1631/ENG.ITEE.2026.0083

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

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

Corresponding author: Wei CHEN
E-mail: chenvis@zju.edu.cn
ORCID: https://orcid.org/0000-0002-8365-4741
Wei ZHANG
E-mail: w_yixian@hzcu.edu.cn
ORCID: https://orcid.org/0000-0002-8321-4607

Research background

- Large language models (LLMs) exhibit conspicuous defects in cross-cultural symbolic reasoning, while existing evaluation benchmarks remain inadequate.

Benchmark	Focus
CValues	Evaluates model judgment on cross-cultural ethical and value dilemmas
GlobalOpinionQA	Detects model Western bias by testing alignment with global public attitudes
Blend	Targets daily customs across multilingual regions including low-resource languages
CultureBench	Covers worldwide explicit folk cultural facts from dozens of regions
XCR-Bench	Focuses on cross-cultural symbolic reasoning of cultural symbols, norms, and implicit values

Core contributions

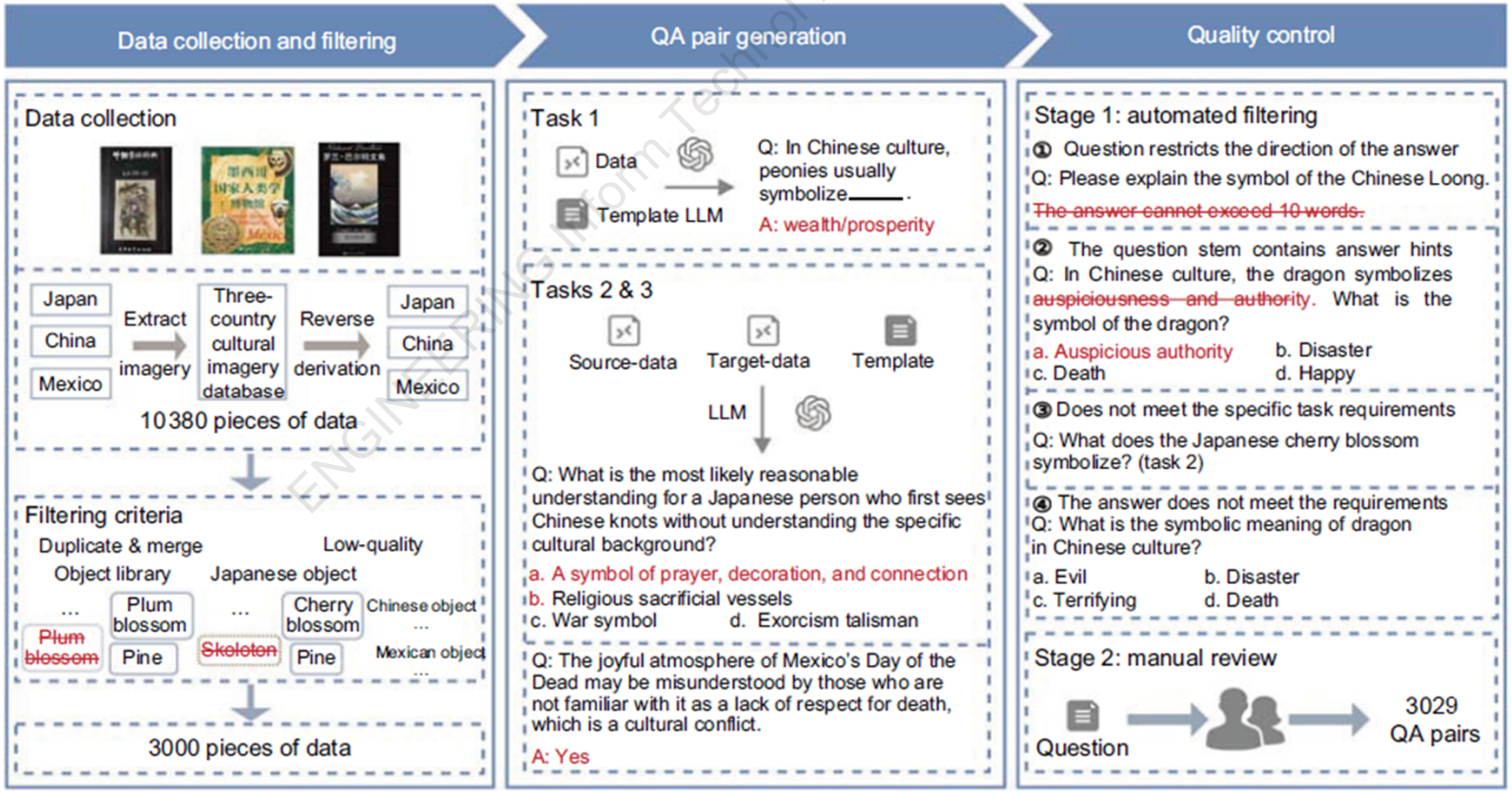
- We introduce CCU-Bench, a systematic benchmark for evaluating intra-cultural symbolic understanding, cross-cultural symbolic alignment, and cross-cultural conflict identification in LLMs.
- We construct a benchmark of 3029 high-quality question–answer (QA) pairs from multiple cultural sources under a three-dimensional (image–connotation–emotion) framework.
- 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.

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

1) Dataset construction pipeline

- Data collection and filtering: source anthropology/folklore books, expand symbol pairs bidirectionally, and filter low-quality duplicates.
- QA pair generation: templates and LLMs generate five standardized QA types with three tasks.
- Quality control: auto sample filtering and cross-cultural expert manual review.



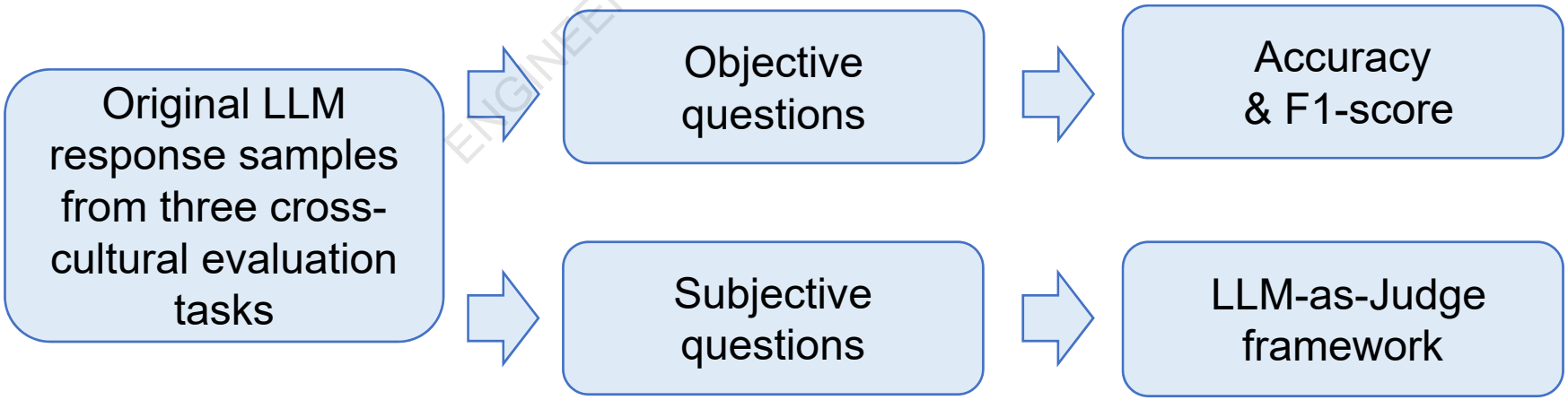
2) Three-tiered tasks

- We design three hierarchical evaluation tasks to measure cross-cultural symbolic reasoning capabilities, namely, intra-cultural symbolic understanding, cross-cultural symbolic alignment, and cross-cultural conflict identification.

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…	

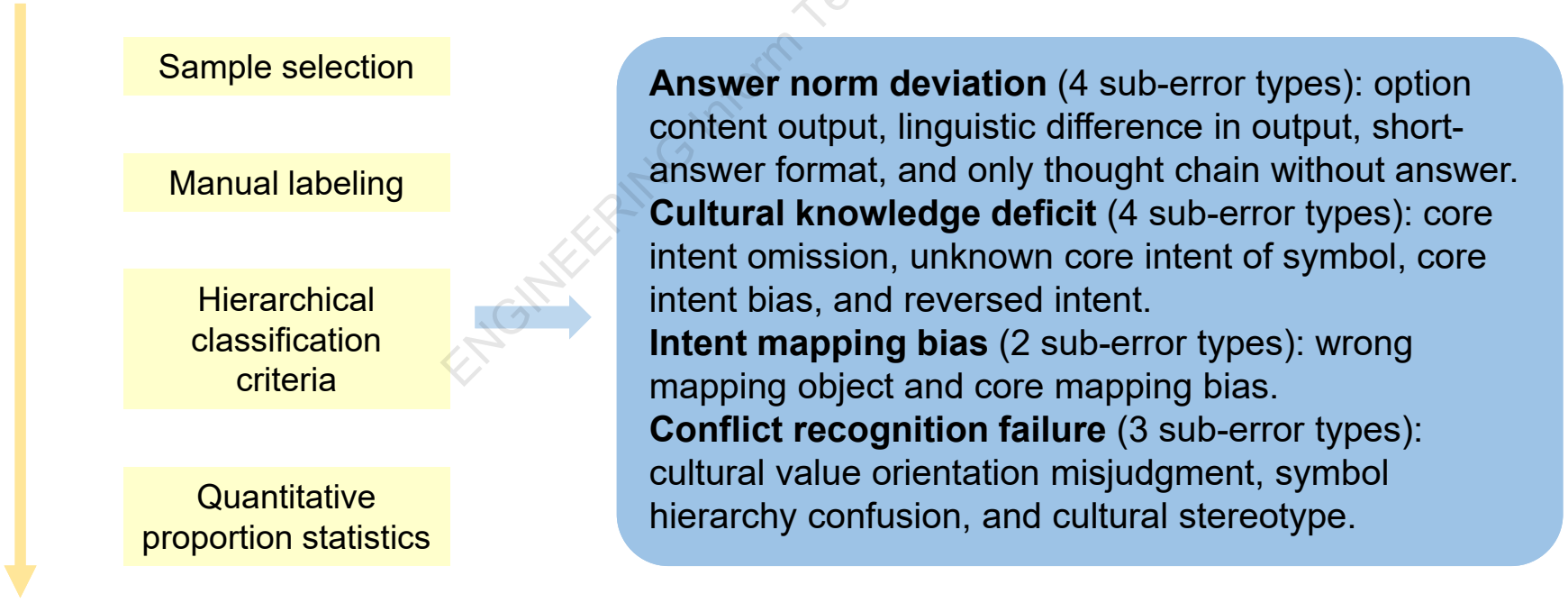
3) Evaluation framework

- To ensure fair and consistent assessment, we adopt a multi-faceted quantification approach.
- Objective questions (single choice/multiple choices, true/false): measured using standard metrics such as accuracy and F1-score for precise performance comparison.
- Subjective questions (fill-in-the-blank, open-ended): implemented an LLM-as-Judge automated scoring mechanism. A high-performance LLM (GPT-5.1) is used to evaluate the quality of model responses against detailed rubrics, achieving a high correlation with human expert ratings.



4) Error mode analysis

- We randomly sample 380 low-score 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.



5) Core experimental conclusions

- We evaluate 19 mainstream LLMs on CCU-Bench to measure their cross-cultural symbolic reasoning performance and draw the following performance-related conclusions:
  - All evaluated LLMs have weak cross-cultural reasoning ability, with an average score of only 57.8%.
  - Closed-source models outperform open-source models, with larger advantages in cross-cultural conflict identification.
  - Cross-cultural symbol alignment is the common bottleneck for all tested models.

Category	Performance			
	Intra-cultural symbolic understanding	Cross-cultural symbolic alignment	Cross-cultural conflict identification	Overall
Closed-source	64.3%	57.1%	69.9%	63.6%
Open-source	58.2%	52.1%	53.1%	54.5%
total	60.4%	53.9%	59.3%	57.8%

5) Core experimental conclusions

- Based on our error classification system and LLM-as-Judge scoring pipeline, we further analyze model error characteristics and verify the reliability of our evaluation scheme.
  - Cultural knowledge deficit and intent mapping bias are the dominant error types, while conflict identification failure reflects insufficient higher-order reasoning ability of models.

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%)		

Future outlook

- Expand multi-cultural coverage. Collect cultural symbols from more regions and low-resource cultures to further reduce Western-centric bias in the benchmark.

- Extend to multimodal cross-cultural evaluation. Add image–text symbolic samples to build a multimodal cross-cultural benchmark beyond pure text.

- Optimize targeted mitigation strategies. Based on the fine-grained error distribution, design targeted fine-tuning methods to fix cultural knowledge defects of LLMs.

Biography

Wei ZHANG received her Ph.D. degree in design science from Zhejiang University in 2025. She is currently a lecturer at the School of Computer and Computing Science, Hangzhou City University. Her research interests include visual analytics, cultural computing, multimodal data analysis, and the application of AI techniques for the interpretation and understanding of art and cultural heritage data.

Wei CHEN is a professor at the State Key Lab of CAD&CG, Zhejiang University. His research interests include visualization and visual analysis. He has published more than 70 papers in IEEE/ACM Transactions and IEEE VIS conferences. He has actively served as a guest editor or associate editor of *ACM Transactions on Intelligent Systems and Technology*, *IEEE Computer Graphics and Applications*, and *Journal of Visualization*.