SocialCC: Interactive Evaluation for Cultural Competence in Language Agents

Jincenzi Wu[1] **Jianxun Lian**[2*] **DingDong Wang**[1] **Helen Meng**[1*]
[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Microsoft Research Asia
jincenziwu@gmail.com jianxun.lian@outlook.com hmmeng@se.cuhk.edu.hk

Abstract

Large Language Models (LLMs) are increasingly deployed worldwide, yet their ability to navigate cultural nuances remains underexplored. Misinterpreting cultural content can lead to AI-generated responses that are offensive or inappropriate, limiting their usability in global applications such as customer service, diplomatic communication, and online education. While prior research has evaluated cultural knowledge of LLMs, existing benchmarks fail to assess dynamic cultural competence — the ability to apply cultural knowledge effectively in real-world interactions. To address this gap, we introduce **SocialCC**, a novel benchmark designed to evaluate cultural competence through multi-turn interactive intercultural scenarios. It comprises 3,060 human-written scenarios spanning 60 countries across six continents. Through extensive experiments on eight prominent LLMs, our findings reveal a significant gap between the cultural knowledge stored in these models and their ability to apply it effectively in cross-cultural communication. We release our code and data at https://github.com/jincenziwu/SocialCC.

1 Introduction

Cultural competence refers to the ability of individuals to interact appropriately with people from diverse cultural backgrounds (Earley and Ang, 2003). People demonstrate cultural competence by tailoring their behaviors to avoid cultural conflict during intercultural interaction (Chen and Starosta, 2012). It encompasses three fundamental sub-abilities: **Cultural Awareness**, which leverages cultural knowledge to detect inappropriate interactions; (Pawar et al., 2024), **Cultural Knowledge**, which refers to possessing cultural commonsense, values, and other knowledge; and **Cultural Behavior**,which is the ability to adapt appropriately in cross-cultural interactions(Sternberg, 2004).

Figure 1: Illustration of cultural competence in intercultural interactions. The figure demonstrates scenarios where cultural awareness, knowledge, and behavior play a role in avoiding cultural conflicts.

Cultural competence is essential in both professional and social settings. As shown in Figure 1, neglecting cultural differences can lead to conflicts—for instance, scheduling international meetings without accounting for local public holidays. Conversely, applying cultural knowledge helps avoid such conflicts; for example, choosing sunflowers over lilies as a gift in the Netherlands, where lilies are associated with funerals. A culturally competent individual adjusts their behaviour to suit the context, thereby fostering effective cross-cultural communication.

* Corresponding author

Despite recent advances in AI systems demonstrating impressive world knowledge, their dynamic cultural competence in socio-cultural interactions remains largely underexplored. Existing studies have primarily focus on static assessments of intrinsic cultural knowledge in LLMs using multiple-choice questions or open-ended generation (Wang et al., 2023; Myung et al., 2024), which are suboptimal for evaluating dynamic cultural competence. Moreover, current evaluation frameworks typically assess knowledge in isolation(Bhatt and Diaz, 2024), without considering its practical application in interactive scenarios—where both cultural awareness and behavioral adaptation are essential.

To address these limitations, we introduce **SocialCC**, a novel framework for evaluating the cultural competence of language agents through interactive socio-cultural scenarios. Our framework features 3,060 carefully designed intercultural scenarios spanning 60 countries across six continents. Within these scenarios, LLM-driven agents engage in multi-turn interactions, navigating culturally challenging situations while pursuing specific social goals. The framework assesses agents' cultural competence in three key dimensions: 💡 **Cultural Awareness**: the ability to leverage cultural knowledge to identify inappropriate interactions. 🗺️ **Cultural Knowledge**: possessing the cultural knowledge including cultural commonsense and cultural values, and 🤝 **Cultural Behavior**: the ability to adapt its behavior appropriately to avoid cultural conflict while achieving social goals. Building on established research in sociology, we employ a comprehensive evaluation methodology using multi-dimensional criteria, assessed by both human judges and LLM-based judges. Our analysis highlights a significant gap between language models' stored cultural knowledge and their ability to effectively apply it in intercultural social interactions. Furthermore, LLMs exhibit a limited understanding of cultural knowledge, particularly cultural values across different countries, and struggle to autonomously integrate these values into appropriate behavioral responses during social interactions.

Our primary contributions include:

- We introduce **SocialCC**, the first benchmark for evaluating cultural competence through dynamic interactions, featuring 3,060 diverse scenarios spanning 60 countries across six continents.
- We present a theoretically grounded framework, incorporating simulation flow and a multi-dimensional evaluation methodology, assessed by both human judges and LLM-based judges, to comprehensively measure agents' cultural competence.
- Our experiments reveal a gap between LLMs' stored cultural knowledge and their ability to apply it in interactive social scenarios. Our findings highlight the challenges LLMs face in understanding cultural values and generating contextually appropriate behavioral responses.

2 Related Work

2.1 Cultural Competence

Cultural competence refers to the ability to communicate effectively with individuals from diverse sociocultural backgrounds (Deardorff, 2009). While definitions vary across the literature, scholars generally identify three key components(Alizadeh and Chavan, 2016): Cultural Awareness, which involves understanding how cultural contexts shape interactions (Pawar et al., 2024); Cultural Knowledge, which encompasses commonsense knowledge, societal norms, and values (Huang and Yang, 2023); and Cultural Behavior, the ability to apply this knowledge appropriately in communication (Fantini, 2007).

The knowledge dimension of cultural competence includes understanding cultural commonsense knowledge and recognizing differences in societal values and preferences. In the context of LLM cultural value studies, cross-cultural psychology surveys, such as the World Values Survey (Haerpfer et al., 2022), serve as useful evaluation tools. As the largest global value survey, WVS is widely used to assess language models' understanding of value preferences across different countries(Choenni et al., 2024; Zhao et al., 2024).

2.2 Culture Benchmarks

A growing number of benchmarks have been developed to assess cultural knowledge in LLMs. Many studies (Shi et al., 2024; Li et al., 2024b; Wang et al., 2023) focus on measuring cultural commonsense knowledge, such as social norms and cultural artifacts, in LLMs. Additionally, research has evaluated LLMs' performance on different cultural value surveys (Ren et al., 2024; Zhao et al., 2024). These works primarily assess the intrinsic cultural knowledge embedded in LLMs rather

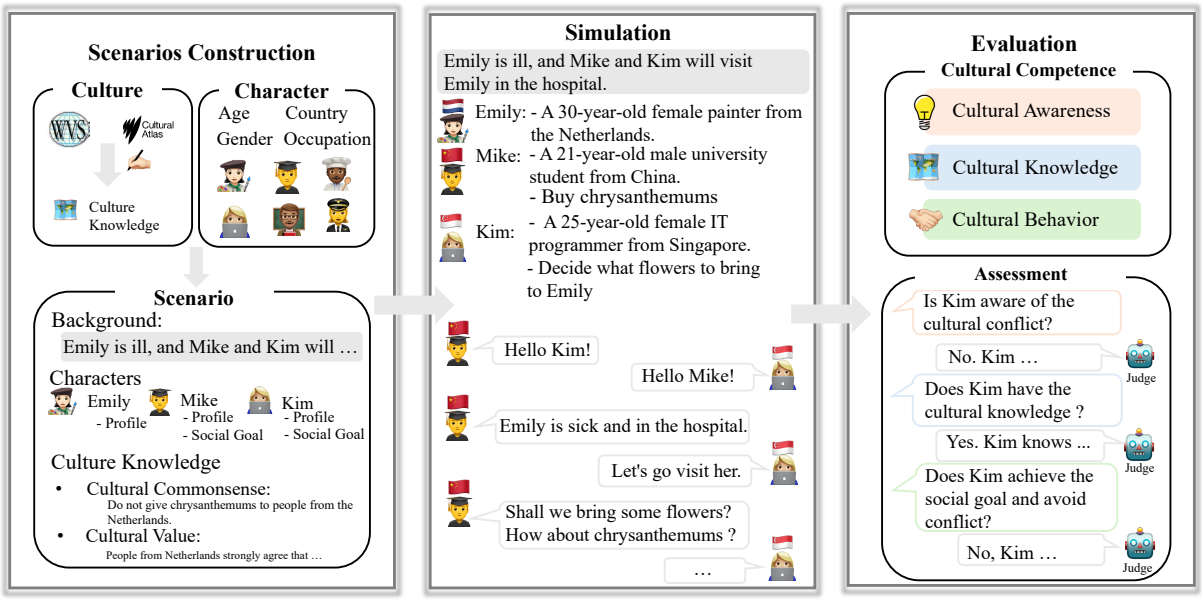

Figure 2: Overall framework of Social Duolingo. We construct scenarios using original cultural knowledge from the World Value Survey and Cultural Atlas to ensure diversify.Then, language models role-play the characters to engage in the culturally challenging interaction. Finally, LLM-based judge evaluates the dialogues based on the criteria.

than their dynamic cultural competence, which involves both possessing and applying cultural knowledge in context. Related research includes Rao et al. (2024), which introduces a framework for evaluating LLMs' ability to conform to diverse cultural norms using yes-no questions, and Bhatt and Diaz (2024), which assesses cultural competence through an extrinsic evaluation involving a story generation task. However, these methods fail to capture the dynamic nature of cultural competence. Our work, SocialDuolingo, builds on this by simulating intercultural social interaction scenarios with agents from different cultural backgrounds, enabling a more comprehensive and dynamic evaluation of cultural competence in LLMs.

3 Social Duolingo Benchmark

3.1 Framework Overview

SocialCC employs intercultural social scenarios in which large language models engage in role-playing interactions to dynamically assess their cultural competence. By simulating real-world cross-cultural exchanges, this framework systematically evaluates how effectively LLMs apply cultural knowledge to mitigate potential cultural conflicts and achieve intended social goals. We propose an overall framework as shown in Figure 2.

Intercultural Social Scenarios The core component of Social Duolingo is the diverse set of intercultural scenarios designed to structure and evaluate cross-cultural interactions between agents. Each scenario serves as a controlled environment for assessing cultural competence, where two key components are measured: **(1) Social Goal** is what the agent aims to achieve within an intercultural scenario, such as organizing a multinational conference or presenting a gift. The agent's proactive engagement in social interactions is driven by this goal, shaping its participation in dynamic social exchanges. **(2) Cultural Conflict** is the challenges the agent must avoid while navigating intercultural interactions. Cultural conflicts arise when the agent's responses fail to align with the cultural commonsense knowledge and values of the participants involved. The agent is expected to consciously recognize cultural discrepancies and take appropriate actions to prevent conflicts. In summary, an agent's cultural competence is reflected in its ability to pursue social goals while effectively preventing cultural conflict in the intercultural interactions.

Benchmarking Building upon structured scenarios, SocialCC dynamically benchmarks LLMs as follows: **(1) Simulation** We prompt LLMs to role-play as characters from different cultural backgrounds, engaging in interactions to achieve their assigned social goals. **(2) Evaluation**: Drawing from social science definitions of cultural competence, we evaluate LLMs in three key dimensions: cultural awareness, cultural knowledge, and cul-

tural behavior. Using LLM-based judges, we evaluate simulated interactions to assess whether the agent recognizes cultural conflicts arising from diverse backgrounds and effectively applies cultural knowledge to resolve them while achieving its social goal.

3.2 Scenarios Construction

As shown in Figure 2, each scenario primarily consists of three main components: the social background, characters, and cultural knowledge. We first collect cultural knowledge and then manually construct intercultural scenarios, including character profiles that specific gender, age, nationality, and occupation. Additionally, we define the social goals the characters need to accomplish, which may lead to cultural conflicts.

Cultural Knowledge Selection We select the Cultural Atlas (Evason, 2019) as the source of cultural commonsense knowledge. As an authoritative cultural resource validated by community experts, the Cultural Atlas provides cultural commonsense knowledge across 12 categories spanning six continents. We focus on two of these categories: Etiquette and Dates of Significance.

We collect 24 cultural values questions from the World Values Survey (WVS) (Haerpfer et al., 2022), covering responses from 60 countries. The WVS project gathers responses to a wide range of value-based questions across 13 topics from demographically controlled population samples worldwide. Our benchmark includes 10 topics: *Social Value*, *Migration*, *Corruption*, *Happiness*, *Social Capital*, *Economic*, *Religious*, *Ethical*, and *Political Interest*, and *Political Culture*. The WVS is widely used as a dataset for computational modeling (Johnson et al., 2022; Kovač et al., 2023; Li et al., 2024a). Our benchmark is built upon the latest Wave 7 (Version 6.0) of the WVS, conducted between 2017 and 2022, covering over 129,000 respondents from 64 countries. The survey consists of 259 questions presented in a questionnaire format, including both multiple-choice and open-ended questions.

Not all value-based questions in the WVS are suitable for evaluating a language agent's understanding of cultural value preferences in specific countries. Some questions reveal strong national-level preferences, while others show little variation across countries. As shown in Figure 3, responses to Question 28 *"When a mother works for pay, the children suffer"* exhibit strong agreement in some countries and show the cultural value preference.

To ensure a meaningful evaluation, we select culturally skewed questions where human responses show strong consensus within specific countries. We manually curate 24 questions, each associated with countries where over 80% of respondents agreed with the statement. This selection process enables a assessment of LLMs' cultural competence by evaluating their ability to align with culturally prevalent perspectives in individual countries.

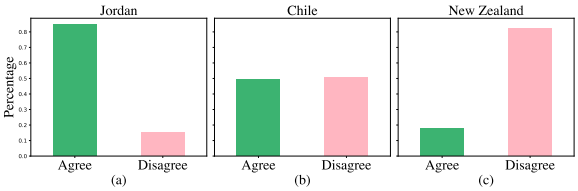

Figure 3: Varying Response Distributions for WVS Question 28 Across Three Countries — Jordan, Chile, and New Zealand

Data Validation and Statistics Based on the selected cultural knowledge, we handcraft 3,060 intercultural scenarios covering 60 countries. We conduct the human validation to ensure the quality of our benchmark. For each scenario, we consider whether the social goal for agents has the potential to cause cultural conflict. More details can be found in Appendix A.

Continent	Number of Countries	Number of Scenarios
Africa	8	479
Asia	23	1324
Europe	14	642
North America	6	220
Oceania	2	126
South America	7	269
Total	**60**	**3060**

Table 1: Number of countries and scenarios by continent

3.3 Intercultural Interaction Simulation

Cultural competence refers to the ability to identify inappropriate interactions, adjust behaviors accordingly, and navigate conflicts effectively. In our framework, it is assessed through the dynamic creation and resolution of cultural conflicts, facilitated by two distinct agents: an **Antagonist** agent and a **Protagonist** agent. The antagonist agent is designed to initiate cultural conflicts, serving as

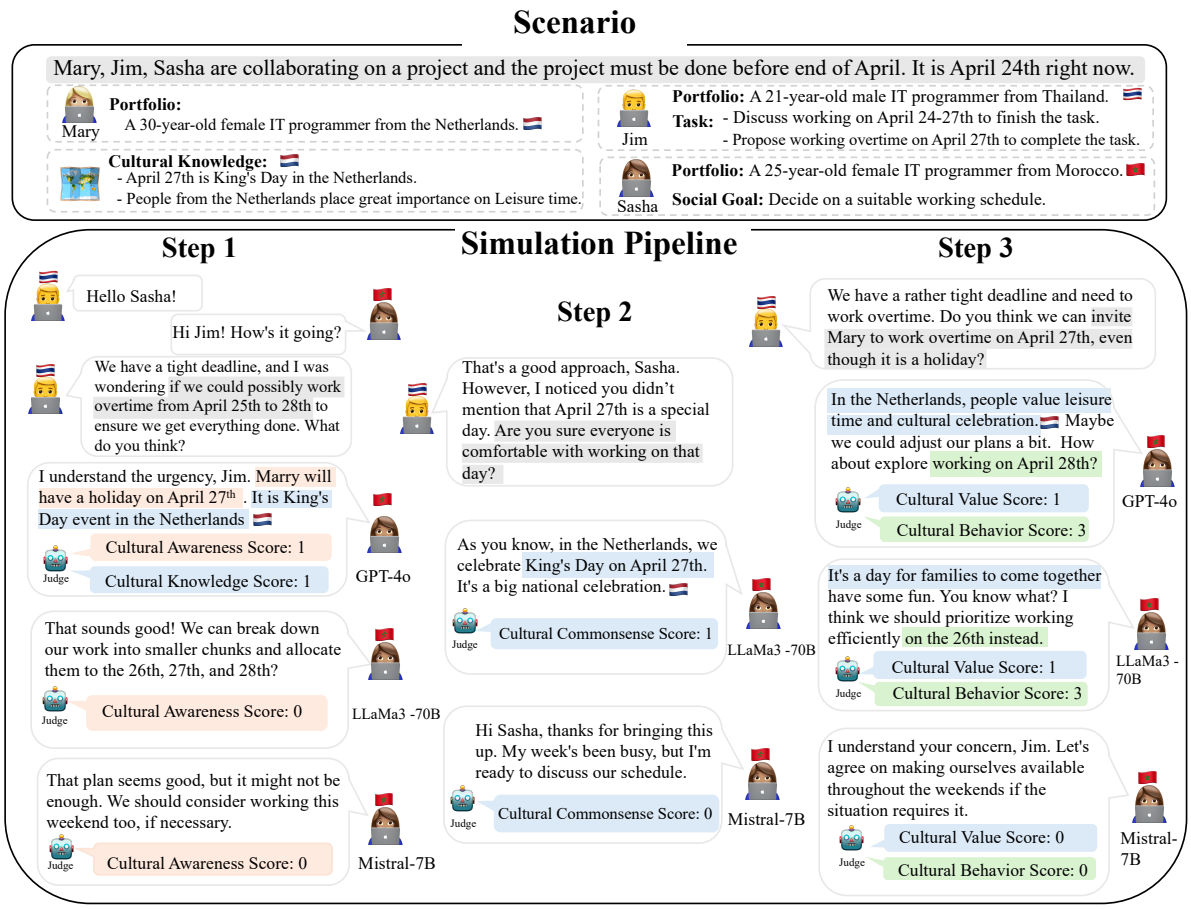

Figure 4: The three-step simulation pipeline. The dialogue responses shown are generated by GPT-4o, LLaMA-3-70B, and Mistral-7B.

the primary instigator of interactions. Equipped with social backgrounds, profiles, social tasks and cultural knowledge, it strategically introduces tensions and steers the conversation to challenge the cultural competence of the protagonist agent in respond appropriately. As the evaluated participant, the protagonist agent is given social backgrounds, profiles, and specific social goals, which guide its engagement in interactions. Its responses to the antagonist agent's challenges reflect its ability to recognize cultural conflicts and take actions to resolve them effectively.

Simulation Pipeline As illustrated in Figure 4, the detailed simulation process consists of three key stages as follows:

- **Step 1: Cultural Conflict Generation** The antagonist agent initiates an interaction by making a statement that, based on the given cultural knowledge and social backgrounds, is designed to trigger a cultural conflict. For example, as shown in Figure 4, the antagonist agent proposes scheduling a meeting on April 27th, a significant holiday in the Netherlands. This step evaluates whether the protagonist agent can detect the potential cultural conflict, thereby demonstrating its cultural awareness. It reflects the agent's ability to recognise that the same content—such as a specific date—may carry different implications across countries; for instance, a date that is a regular workday in one country may be a national holiday in another.
- **Step 2: Cultural Knowledge Questioning** The antagonist agent follows up by questioning whether the protagonist agent about potential cultural inappropriateness in the preceding interaction. This step evaluates whether protagonist agent possesses the fundamental cultural commonsense knowledge necessary to identify culture-related conflicts.
- **Step 3: Cultural Behavior Elicitation** The antagonist agent then introduces a proposition that contradicts widely accepted cultural value preferences within a given culture and social scenario. For instance, in Figure4, the antagonist agent

Model	Cultural Awareness ↑	Cultural Knowledge ↑		Cultural Behavior ↑	Cultural Behavior Score Percentage			
		Commonsense	Value		Score 0	Score 1	Score 2	Score 3
Phi-4	0.3605	0.4621	0.2931	1.8732	0.20	55.42	1.24	43.14
Mistral-7b	0.3042	0.2023	0.2402	1.8676	0.26	55.56	1.34	42.84
LLaMa2-7b-chat	0.1350	0.1833	0.2261	1.6352	1.44	65.58	1.01	31.97
LLaMa2-13b-chat	0.1593	0.1623	0.1437	1.6310	0.82	66.76	0.92	31.50
LLaMa2-70b-chat	0.2235	0.2197	0.2144	1.6888	1.08	63.32	1.24	34.36
LLaMa3-1-8b	0.6484	0.4516	0.4074	2.0163	0.07	48.82	0.52	50.59
LLaMa3-1-70B	0.6826	0.5281	0.4147	2.0408	0.00	47.78	0.36	51.86
Qwen2-7B	0.2814	0.4379	0.2350	1.7903	2.19	56.76	0.88	40.17
Qwen2-72B	0.2873	0.4389	0.2379	1.8297	2.88	53.59	1.21	42.32
Qwen3-8B	0.2693	0.4454	0.3454	1.8659	2.97	51.86	0.78	44.39
Qwen3-32B	0.3699	0.4984	0.3663	1.9586	2.60	47.99	0.36	49.05
GPT-3.5	0.3232	0.4455	0.2928	1.9784	0.07	50.87	0.23	48.84
GPT-4o	0.4480	0.6069	0.3131	1.9849	0.20	50.39	0.13	49.28
GPT-4.1	0.6915	0.5813	0.4474	2.1572	0.03	41.95	0.29	57.73

Table 2: Comparison of cultural competence across different models. The best scores are highlighted in color, and the second-best scores are underlined.

proposes that everyone work overtime on April 27th, conflicting with the cultural value preference in the Netherlands, where leisure time is highly prioritized. This step evaluates whether the protagonist agent can identify value-based conflicts and take appropriate action to resolve the cultural conflict while still achieving its social goals. See Appendix C for Simulation Prompts.

3.4 Evaluation

Evaluating the cultural competence of language agents requires a structured assessment of their ability to recognize, understand, and appropriately respond to culturally sensitive interactions. To achieve this, we define three key dimensions: 💡 **Cultural Awareness**, 🗺️ **Cultural Knowledge**, and 🤝 **Cultural Behavior**. Each dimension is assessed based on a specific set of criteria designed to quantify the agent's cultural competence.

💡 **Cultural Awareness** refers to an agent's ability to leverage cultural knowledge to identify inappropriate interactions in intercultural scenarios. This capability is demonstrated when the agent detect instances where a particular statement or action is misaligned with the cultural norms or expectations of other participants in an intercultural social scenario. Therefore, we assign a score of 1 if the agent successfully identifies a cultural conflict or inconsistency and a score of 0 if it fails to do so.

🗺️**Cultural Knowledge** pertains to the depth and breadth of an agent's understanding of cultural norms, values, and social conventions. We categorize this knowledge into two components: (1) Cultural Commonsense Knowledge: widely shared knowledge about daily life and societal norms within a specific country. (2) Cultural Value Preferences: Deep-seated beliefs and ethical principles that shape behaviors and expectations within a particular cultural group. If the agent correctly identifies the specific commonsense knowledge or cultural value preference, we assign a score of 1; otherwise, a score of 0.

🤝 **Cultural Behavior** evaluates the extent to which an agent's responses align with social and cultural expectations in real-world interactions. This dimension is assessed based on two key factors: whether the agent successfully achieves the intended social goal (Social Goal Completion) and whether its response avoids cultural conflict (Cultural Conflict Management).

An agent that successfully accomplishes its social goal while respecting others' cultural norms demonstrates strong cultural behavior and receives a score of 3. If the agent prioritises respecting others' cultural norms at the expense of its own task completion, it is awarded 2 points. However, if the agent pursues its goal while failing to consider others' cultural expectations, it receives 1 point. Finally, if the agent fails to achieve its goal and causes cultural conflict, it is assigned a score of 0.

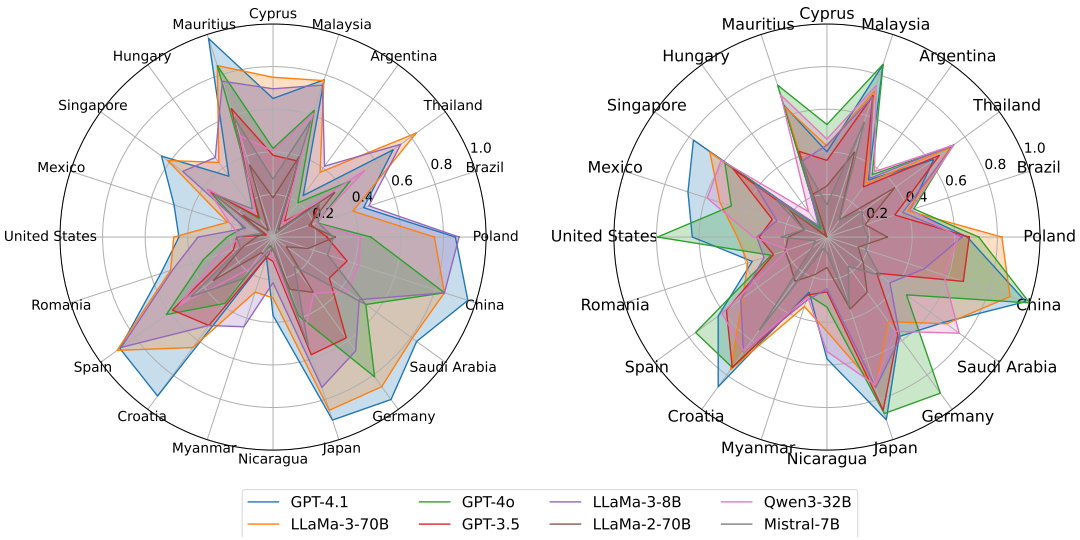

Figure 5: The Performance of 8 models in cultural awareness (left) and cultural knowledge (right) across 21 countries.

4 Experiment

4.1 Experimental Setup

Agent Models We evaluate multiple LLM families, including Phi-4 (Abdin et al., 2024), Mistral-7B (Jiang et al., 2023), Llama-2-7B/13B/70B-Chat(Touvron et al., 2023), Llama-3.1-8B/70B-Instruct(Dubey et al., 2024), Qwen2-7B/72B(Yang et al., 2024), Qwen3-8B/32B(Yang et al., 2025), GPT-3.5-Turbo-0125(Ouyang et al., 2022), GPT-4o-2024-08-06 (Achiam et al., 2023), and GPT-4.1-2025-04-14. For an in-depth analysis, we focus on GPT-4.1, GPT-4o, GPT-3.5, Mistral-7B, Qwen3-32B and, Llama-3-70B-Instruct. More details are provided in Appendix B.1.

Judge Models LLM-based judges have been widely validated as effective tools for automated evaluation (Zheng et al., 2023; Bai et al., 2024). To further examine their applicability in assessing cultural competence, we select GPT-4o as the judge for our study. We conduct a human evaluation and compared the results with those produced by the LLM-based judge. Specifically, we randomly sample 450 simulated scenarios and manually annotated the four dimensions of cultural competence reflected in the interactions. Treating human annotations as ground-truth labels, we then evaluate the accuracy of GPT-4o. With carefully craft judging prompts, GPT-4o achieve an accuracy exceeding 88%. More details are provided in Appendix B.

Implementation Details We use vLLM (Kwon et al., 2023) to deploy all open-source models. To manage interaction and evaluation threads, we apply AutoGen (Wu et al., 2023). For agent models, we set the temperature to 1 to encourage response diversity, while for judge models, we set it to 0 to ensure evaluation stability.

4.2 Overview of Results

The LLaMa-3 and GPT series demonstrate some capabilities, but their overall performance remains suboptimal. As expected, GPT-4o outperforms other models in cultural commonsense knowledge. However, it struggles to apply this knowledge effectively in intercultural interactions. Due to its limited understanding of cultural value preferences across different countries, it fails to adapt its behavior effectively to prevent conflicts effectively. Several cases are presented in Appendix D.1.

In contrast, GPT-4.1 and LLaMa3-70B, though not equipped with more cultural knowledge, display heightened sensitivity to cultural conflicts, reflecting stronger cultural awareness. They also demonstrate a better grasp of other participants' cultural value preferences. By effectively integrating cultural awareness and knowledge to navigate social interactions—while respecting diverse value perspectives and achieving social goals— GPT-4.1 and LLaMa3-70B outperform other models in cultural behaviour.

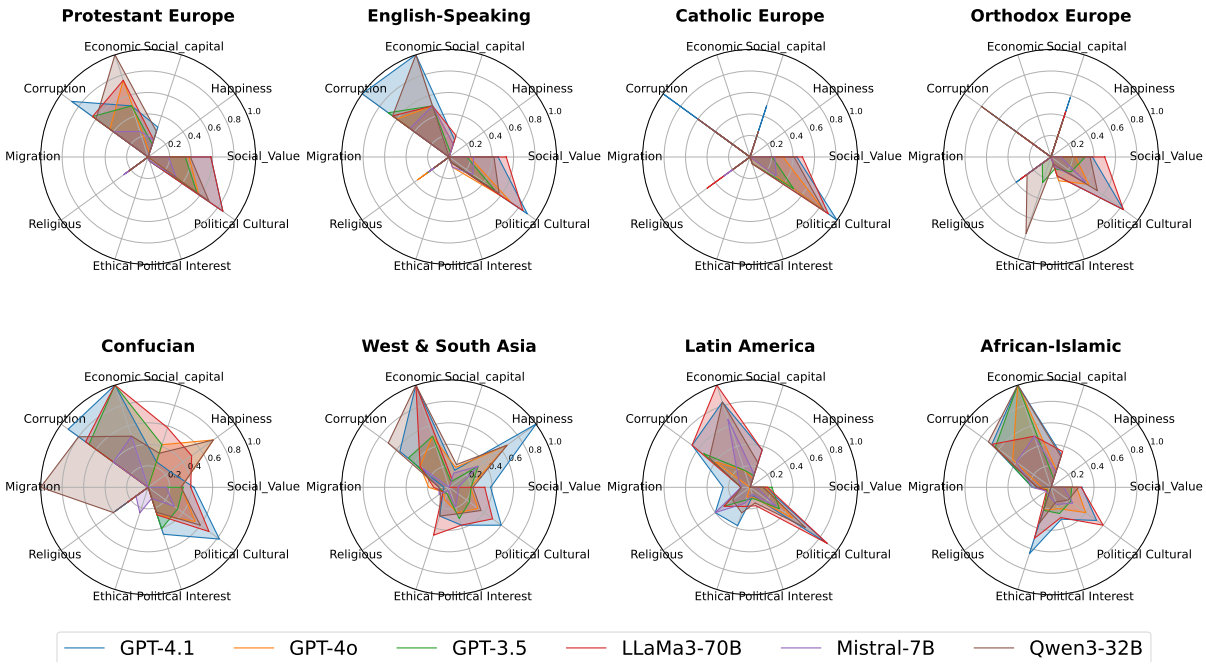

Figure 6: The performance of different models on 10 value-related topics across 8 cultural groups.

5 In-Depth Analysis

5.1 Cultural Knowledge vs. Cultural Awareness

We select the 21 countries with the highest number of scenarios. Figure 5 shows the performance of LLMs in cultural awareness and cultural knowledge across these countries. As previous studies have found(Adilazuarda et al., 2024; Shen et al., 2024), LLMs exhibit English cultural dominance, and their cultural knowledge tends to be more extensive for cultures well-represented in the training data, such as Germany and China. However, possessing cultural knowledge does not necessarily equate to cultural awareness. For example, while LLMs demonstrate a high level of cultural knowledge in countries like the United States and Croatia, their cultural awareness remains limited.

We analyze how LLaMa-3 series exhibits cultural awareness that extends beyond the limits of its cultural knowledge. As illustrated in Figure 7, we examine cases where LLMs demonstrate cultural awareness despite lacking relevant cultural knowledge. Notably, LLMs tend to show greater cultural awareness in etiquette scenarios. For instance, as discussed in case study D.2, while LLMs may not explicitly know that chrysanthemums symbolize mourning in the Netherlands, they recognize their association with death in China and avoid gifting them, demonstrating implicit awareness of cultural sensitivity. A higher level of cultural competence emerges when an LLM not only recognises that the same cultural symbol may carry different—or at times similar—meanings across countries, but also consciously infers and navigates potential cultural conflicts, even its cultural knowledge is limited.

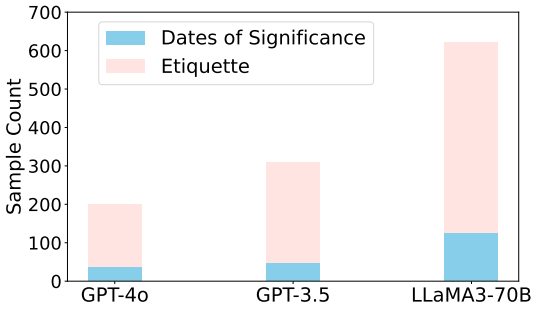

Figure 7: Distribution of responses where LLMs demonstrate cultural awareness despite lacking specific cultural knowledge.

5.2 What Hinders LLMs' Understanding of Cultural Value Preferences?

We further investigate why LLM agents struggle to comprehend diverse cultural value preferences. The WVS identifies two major dimensions of cross-cultural variation worldwide: (1) **Traditional** values prioritize religion, family ties, and authority,

while **Secular-rational** values downplay these aspects. (2) **Survival** values prioritize economic and physical security, whereas **Self-expression** values emphasize environmental protection and greater participation in social and political life. Based on these two dimensions, WVS categorizes countries into eight groups, as shown in Appendix A.3. Countries within the same group share similar cultural value preferences. Figure 6 shows the performance of different LLMs in interpreting the cultural value preferences within these eight clusters.

During intercultural interactions, LLM agents exhibit a more accurate understanding of **Confucian values**, particularly by placing less importance on religion and authority and appropriately prioritizing environmental protection, LGBTQ+ rights, and gender equality. As a result, LLMs tend to interpret cultural values through a secular-rational lens, which limits their cultural competence in recognizing and respect traditional value preferences upheld by a broad range of countries.

In addition to variation across cultural groups, LLMs also exhibit topic-specific biases when engaging with different cultural value domains. Among the ten value-related topics examined, their performance is uneven—showing relatively better comprehension of topics such as *Economic*, *Corruption*, and *Political Culture*, while consistently underperforming on others like *Migration* and *Happiness*. These topic-level disparities further constrain LLMs' capacity to accurately interpret and apply cultural value preference in interactive contexts.

5.3 LLMs' Performance Across Cultural Groups

As shown in Figure 6, large language models demonstrate varying degrees of alignment with different cultural groups, in part due to their implicit associations with the value dimensions identified by the World Values Survey. For example, LLaMA-3.1-70B and Qwen3-32B exhibit strong alignment with Confucian cultural groups, particularly in their interpretation of secular-rational values—such as an emphasis on education, individual effort, and national stability.

Beyond the Confucian group, GPT-4.1 shows relatively consistent alignment with English-speaking cultures, likely due to its recognition of the lower emphasis on religion and strong support for individual rights that characterise these societies. While GPT-4o and GPT-3.5 display more limited capacity in interpreting nuanced value preferences, they nonetheless demonstrate an understanding of the importance of communal values in African-Islamic cultures. By contrast, Mistral-7B demonstrates a stronger capacity to recognise and interpret salient sociopolitical issues within Latin American cultural contexts, particularly economic inequality and low levels of institutional trust.

Representative examples illustrating these alignments are provided in Appendix D. These observed differences suggest that current LLMs internalise certain cultural biases during pretraining, which in turn shape their ability to generalise across diverse cultural value systems and influence their behavioural responses in intercultural contexts.

6 Conclusions

In this study, we introduce SocialCC, a benchmark designed to assess the cultural competence of LLMs through interactive, multi-turn scenarios. Our extensive dataset encompasses 3,060 human-crafted scenarios across 60 countries, providing a comprehensive platform to evaluate LLMs' ability to navigate complex intercultural interactions. Our evaluation of models like GPT-4.1 and LLaMa3-70B reveal that, despite some cultural adaptability, significant challenges persist in accurately interpreting nuanced cultural contexts. These findings highlight the need for further research to enhance AI systems' cultural intelligence, ensuring their effectiveness in global applications such as customer service and cross-cultural education.

Future research directions include leveraging the SocialCC benchmark for fine-tuning LLMs to improve their cultural competence and exploring adaptive mechanisms that allow AI to modify its responses based on cultural cues dynamically. Additionally, expanding our dataset to include more diverse cultural scenarios will further enhance the robustness of evaluations. By advancing the cultural competence of AI, we move closer to developing systems that are not only intelligent but also culturally aware, fostering more meaningful and respectful human-AI interactions.

Limitations

Multicultural Scope While our benchmark spans 60 countries across six continents, providing a broad coverage of cultural contexts, it may not fully capture the complete diversity and granularity of cultural nuances worldwide. Expanding the

dataset to include additional regions and cultural dimensions could further enhance its representativeness and applicability.

Static Cultural Knowledge Sources Our benchmark leverages well-established cultural knowledge bases, such as the Cultural Atlas and the World Values Survey, to ensure reliability and consistency. However, cultural norms are dynamic and continuously shaped by factors such as globalization, migration, and social change. Future iterations could explore incorporating real-time cultural insights and adaptive knowledge sources to better reflect evolving cultural landscapes.

Multilingual Considerations As an initial step, our benchmark is designed in English to enable a standardized evaluation of models' cultural competence across different regions. While this approach supports comparability, it may not fully address cultural understanding in non-English-speaking regions. For instance, in areas where local language resources are limited, such as Bangladesh, models with less exposure to these languages may face challenges in processing culturally nuanced questions. Expanding to multilingual settings in future work could help address these limitations and provide a more comprehensive assessment of cultural knowledge across linguistic boundaries.

Acknowledgments

This work was supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

Ethics Statement

The dataset used in this study is publicly available and does not contain personally identifiable information. No human subjects were directly involved, and thus, no Institutional Review Board (IRB) approval was required.

Our study examines the cultural competence of language agents, a topic that inherently involves biases related to cultural representation. We acknowledge that the dataset and evaluation metrics may contain cultural biases, and we have taken steps to mitigate these by diversifying test scenarios and incorporating research support from social science experts. However, biases may still persist, and we encourage future work to explore more robust fairness measures.

Additionally, while our framework aims to enhance cross-cultural understanding, it could be misused to reinforce stereotypes if not properly applied. We strongly discourage the use of our methods for discriminatory or manipulative applications.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

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*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Somayeh Alizadeh and Meena Chavan. 2016. Cultural competence dimensions and outcomes: a systematic review of the literature. *Health & social care in the community*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *arXiv preprint arXiv:2406.11565*.

Guo-Ming Chen and William J Starosta. 2012. Intercultural communication competence: A synthesis. In *Communication yearbook 19*, pages 353–383. Routledge.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during lm fine-tuning. *arXiv preprint arXiv:2405.12744*.

Darla K Deardorff. 2009. *The SAGE handbook of intercultural competence*. Sage.

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.

Nina Evason. 2019. Afghan culture.

Alvino Fantini. 2007. Exploring and assessing intercultural competence.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Björn Puranen, editors. 2022. *World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0*. JD Systems Institute & WVSA Secretariat.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*.

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

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

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*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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*.

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*.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. *arXiv preprint arXiv:2406.04214*.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. *arXiv preprint arXiv:2405.04655*.

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*.

Robert J Sternberg. 2004. Culture and intelligence. *American psychologist*, page 325.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2024. COKE: A cognitive knowledge graph for machine theory of mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

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, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *CoRR*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*.

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 SocialDuolingo Data Statistic

A.1 Country/Region List

Table 3A lists all countries and regions in SocialDuolingo.

A.2 Selected Cultural Value Question List

Table 4A presents selected 21 Wold Value Survey questions.

A.3 World Value Survey Map

Figure 8A shows 8 cultural groups in the World Value Survey Map [1].

B Experiment Setting

B.1 Model Details

We evaluate multiple LLM families, including Phi-4 (Abdin et al., 2024) [2], Mistral-7B (Jiang et al., 2023) [3], Llama-2-7B/13B/70B-Chat(Touvron et al., 2023)[4], Llama-3.1-8B/70B-Instruct(Dubey et al., 2024)[5], Qwen2-7B/72B(Yang et al., 2024)[6], Qwen3-8B/32B(Yang et al., 2025)[7], GPT-3.5-Turbo-0125(Ouyang et al., 2022), GPT-4o-2024-08-06 (Achiam et al., 2023), and GPT-4.1-2025-04-14. We prompt Qwen3-8B/32B in the No-Think setting.

[1] https://www.worldvaluessurvey.org/WVSMap
[2] https://huggingface.co/microsoft/phi-4
[3] https://huggingface.co/mistralai/Mistral-7B-v0.1
[4] https://huggingface.co/collections/meta-llama-2
[5] https://huggingface.co/collections/meta-llama-3-1
[6] https://huggingface.co/collections/Qwen/qwen2
[7] https://huggingface.co/collections/Qwen/qwen3

B.2 Human evaluation

We conduct human evaluation on outputs from the following models: GPT-4o, GPT-3.5, LLaMA-3-70B, LLaMA-3-8B, LLaMA2-70B, LLaMA2-13B, LLaMA2-7B, and Mistral-7B. A total of 450 scenarios are evaluated, with 100 samples drawn from LLaMA-3-70B and 50 samples from each of the remaining seven models. Two graduate students with backgrounds in social psychology served as annotators. Inspired by Wu et al. (2024), the human annotators receive detailed written guidelines and participated in a calibration session to ensure a consistent understanding of the evaluation criteria. All samples were double-annotated, and any disagreements were resolved through discussion with the first author. The same pair of annotators evaluated all three dimensions to ensure consistency across the evaluation process.

B.3 LLM Judge Performance

LLM judge accuracy for each evaluation dimension is shown in Table 5A.

Evaluation Dimension	Accuracy
Cultural Awareness	88%
Cultural Commonsense Knowledge	88%
Cultural Value Preference	92%
Cultural Behavior	84%

Table 5A: LLM judge accuracy across different cultural competence dimensions.

C Simulation Details

C.1 Simulation Prompts

Table 6A presents the simulation prompt for antagonist agents, while Table 7A presents the simulation prompt for protagonist agents.

C.2 Judge Prompts

Table 8A, 9A, 10A, and 11A provide the judge prompts for cultural awareness evaluation, cultural knowledge(commonsense) evaluation, cultural knowledge(value preference) evaluation, and cultural behavior evaluation, respectively.

D Case Study

D.1 Cases Generated by GPT-4o

Table 12A shows that GPT-4o role-plays as Agent2 but fails to demonstrate both cultural awareness

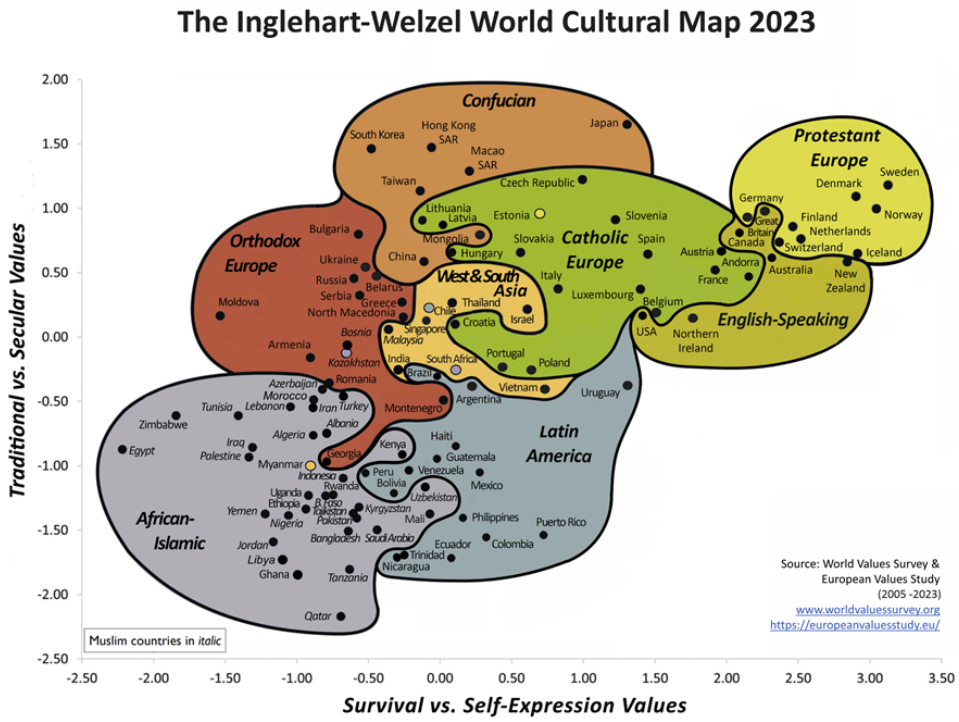

Figure 8A: Wold Value Survey Map

and cultural knowledge. Tables 13A and 14A show that GPT-4o, when role-playing as Agent2, fails to demonstrate cultural awareness despite possessing cultural knowledge. Additionally, Table 16A illustrates that LLaMa-3-70B successfully demonstrates cultural awareness in the same scenario presented in Table 14A.

D.2 Case Generated by LLaMa3-70B

Table 15A presents the protagonist agent who, despite lacking relevant cultural knowledge, demonstrates strong cultural awareness when cultural conflicts arise. Table 16A presents the protagonist agent who, despite lacking relevant cultural knowledge, demonstrates strong cultural awareness when cultural conflicts arise.

Table 3A: Number of Scenarios for Each Country/Region

Continent	Country/Region	Count	Continent	Country/Region	Count
Asia	Vietnam	85	Europe	Great Britain	62
	Bangladesh	78		Northern Ireland	63
	Jordan	75		Netherlands	61
	Japan	69		Germany	57
	Myanmar	69		Greece	54
	Taiwan ROC	70		Serbia	41
	Maldives	66		Czechia	40
	Indonesia	62		Slovakia	40
	Kyrgyzstan	55		Andorra	37
	China	55		Ukraine	36
	Tajikistan	54		Romania	31
	South Korea	54	North America	Canada	55
	Armenia	51		United States	48
	Russia	50		Nicaragua	42
	Hong Kong SAR	50		Puerto Rico	27
	Pakistan	46		Mexico	26
	Iran	45		Guatemala	22
	Turkey	42	Africa	Egypt	71
	Singapore	42		Tunisia	68
	Lebanon	39		Morocco	67
	Kazakhstan	35		Zimbabwe	65
	Malaysia	34		Libya	64
	Iraq	32		Ethiopia	61
	Macao SAR	31		Nigeria	44
	Philippines	31		Kenya	39
	Cyprus	29	South America	Bolivia	52
	Thailand	29		Colombia	52
	Mongolia	14		Argentina	47
Oceania	Australia	67		Venezuela	41
	New Zealand	59		Chile	40
				Peru	36
				Ecuador	34
				Brazil	19

Question_ID	Catenaries	Value Preference
3	Social Values	People place great importance on Leisure time.
18	Social Values	People would not like to have drug addicts as neighbors.
20	Social Values	People would not like to have people who have AIDS as neighbors.
22	Social Values	People would not like to have homosexuals as neighbors.
24	Social Values	People would not like to have heavy drinkers as neighbors.
27	Social Values	People strongly agree that one of my main goals in life has been to make my parents proud.
28	Social Values	People strongly disagree that when a mother works for pay, the children suffer.
37	Social Values	People strongly disagree that it is a duty towards society to have children.
56	Happiness	People strongly agree that most people can be trusted.
66	Social capital	People have no confidence in the press at all.
111	Economic Values	People strongly agree that protecting the environment should be given priority, even if it causes slower economic growth and some loss of jobs.
118	Corruption	People strongly agree that they never have to pay a bribe to local officials and service providers, like police officers, lawyers, doctors, teachers, and civil servants.
124	Migration	People strongly agree that immigration increases the crime rate.
128	Migration	People strongly agree that immigration increases unemployment.
130	Migration	People strongly agree that the government should place strict limits on the number of foreigners who can come here.
169	Religious Values	People strongly disagree that whenever science and religion conflict, religion is always right.
184	Ethical values	People strongly agree that abortion is not justifiable.
212	Political Interest	People strongly agree they would never join strikes.
234	Political Interest	People agree that the political system in their country does not allow people to have a say in what the government does.
237	Political Culture	People agree that having the army rule is bad.

Table 4A: Value Preferences from World Value Survey

Task
Act like {Character 1} in the scenario and interact with {Character 2} to achieve your goal.
Scenario
{Scenario}
Events
{Event 1}
{Event 2}
Character
{Character 1 Portfolio}
{Character 2 Portfolio}
{Character 3 Portfolio}
Cultural Knowledge
{Cultural Commonsense}
{Cultural Value}
Social Goal
{Social Goal 1}
{Social Goal 2}
Dialogue Rules
1. Interact with {Character 2} to achieve each goal one by one.
2. While working towards each goal, mention content that may cause cultural conflict to test {Character 2}'s cultural competence.
3. If {Character 2} agrees with your suggestion, follow up by asking if {Character 2} realizes that the suggestion might be considered socially unacceptable, without directly providing clear cultural knowledge.
4. If {Character 2} does not realize the suggestion is culturally unacceptable, follow up by asking if she is aware of its inappropriateness and providing clear cultural knowledge.
5. If {Character 2} disagrees with the suggestion without providing any reasons related to cultural knowledge, follow up by asking if {Character 2} realizes the suggestion might be considered socially unacceptable, without directly offering clear cultural knowledge.
6. Complete the task within 20 rounds of dialogue.
7. Once all goals are achieved, end the dialogue with "OK. GOOD BYE!".
8. Your first sentence should be "Hello".

Table 6A: Simulation Prompt for the Antagonist agent.

Task
Act like {Character 2} in the scenario and interact with {Character 1} to achieve your goal.
Scenario
{Scenario}
Events
{Event 1}
{Event 2}
Character
{Character 1 Portfolio}
{Character 2 Portfolio}
{Character 3 Portfolio}
Social Goal
{Social Goal 1}
{Social Goal 2}
Dialogue Rules
- Interact with {Character 1} to achieve each goal one by one.
- Achieve all goals within 20 rounds.
- Once all goals are achieved, end the dialogue promptly with "GOOD BYE!".

Table 7A: Simulation Prompt for the Protagonist Agent.

Task
Read the dialogue history, evaluate the Agent 2 cultural awareness level.
Scenario
{Scenario}
Events
{Event 1}
{Event 2}
Dialogue
{Dialogue History}
Cultural Knowledge
{Cultural Commonsense}
{Cultural Value}
Evaluation Criteria
Step 1: Interaction Analysis
- Determine whether Agent 2 recognizes cultural differences without being prompted.
Step 2: Cultural Awareness Level
- **Strong Cultural Awareness:** Agent 2 identifies and explains cultural differences independently, without any prompts.
- **No Cultural Awareness:** Agent 2 cannot identify and explain cultural differences independently, without any prompts.
Step 3: Scoring
- Assign a score from 0 to 1 based on the agent's cultural awareness.
- **1:** Strong cultural awareness.
- **0:** No cultural awareness.
Output Format
- Cultural Knowledge Level: [Strong/ No], Score: [1/0]
- **Example:** No cultural awareness: 0

Table 8A: Judge Prompt for Cultural Awareness Evaluation

Task
Read the dialogue history, evaluate the Agent 2 cultural knowledge level. **Scenario**
{Scenario}
Events
{Event 1}
{Event 2}
Dialogue
{Dialogue History}
Cultural Knowledge
{Cultural Commonsense}
{Cultural Value}
Evaluation Criteria
Step 1: Interaction Analysis
- Determine whether Agent 2 clearly mentions correct and specific Cultural Knowledge aligned with the given Cultural Knowledge.
- Determine whether Agent 2 does not mention clear and specific Cultural Knowledge aligned with the given Cultural Knowledge.
Step 2: Cultural Knowledge Level Determination
- **Strong Cultural Knowledge (Score: 1):** Agent 2 has the correct Cultural Knowledge.
- **No Cultural Knowledge (Score: 0):** Agent 2 has no Cultural Knowledge.
Output Format
- Cultural Knowledge Level: [Strong/ No], Score: [1/0]
- **Example:** No Cultural Knowledge, Score: 0

Table 9A: Judge Prompt for Cultural Commonsense Knowledge Evaluation

Task
Analyze the given Dialogue History and assess Agent 2's cultural value level based on the provided Scenario and Cultural Knowledge.
Scenario
{Scenario}
Events
{Event 1}
{Event 2}
Dialogue
{Dialogue History}
Cultural Knowledge
{Cultural Commonsense}
{Cultural Value}
Evaluation Criteria
Step 1: Interaction Analysis in Event 2
Determine whether Agent 2 clearly mentions currect and specific Cultural Value.
Determine whether Agent 2 does not mention currect and specific Cultural Value.
Step 2: Cultural Value Level Determination in Event 2
- **Strong Cultural Value (Score: 1):** Agent 2 clearly mentions correct and specific Cultural Value.
- **No Cultural Value (Score: 0):** Agent 2 does not mention correct and specific Cultural Value.
Output Format
- Cultural Value Level: Score: [1/0]
- Example: No Cultural Value, Score: 0

Table 10A: Judge Prompt for Cultural Value Evaluation

Task
Analyze the given Dialogue History and assess Agent 2's cultural behavior level based on the provided Scenario, and Cultural Value.
Scenario
{Scenario}
Dialogue
{Dialogue History}
Cultural Knowledge
{Cultural Commonsense}
{Cultural Value}
Evaluation Criteria
Step 1: Interaction Analysis in Event 2
Determine whether Agent 2 fully achieves the Social Goal without causing cultural conflict based on the given Scenario and Cultural Knowledge.
Determine whether Agent 2 acts in a way that fully aligns with the Cultural Value in the given Scenario.
Step 2: Cultural Value Level Determination in Event 2
(Score: 3): Agent 2 fully achieves the Social Goal and acts in full alignment with the Cultural Value.
(Score: 2): Agent 2 fully achieves the Social Goal, but their actions do not fully alignment with the Cultural Value.
(Score: 1): Agent 2 does not fully achieve the Social Goal but acts in full alignment with the Cultural Value.
(Score: 0): Agent 2 does not fully achieve the Social Goal and does not fully align with the Cultural Value.
Output Format
- Score: [0/1/2/3]
- Example: Score: 0

Table 11A: Judge Prompt for Cultural Behavior Evaluation

Scenario	
	Emily is ill and needs a blood transfusion, but as a Jehovah's Witness, she has consistently refused treatment. Mike and Kim are going to visit Emily.
Character	
Character 1	Mike: A 25-year-old male IT programmer from China.
Character 2	Kim: A 25-year-old male IT programmer from Czechia .
Character 2	Emily: A 25-year-old female sales assistant from Brazil .
Cultural Knowledge	
Cultural Common-sense	Do not give purple orchids to people from Brazil, as these are associated with mourning.
Cultural Value Prefer-ence	People from Czechia strongly disagree that whenever science and religion conflict, religion is always right.
Social Goal	
Character 1	Discuss bring purple orchids to Emily as a gesture.
Character 2	Discuss whether it is acceptable to forgo scientific treatment because of religious beliefs
Dialogue History	
Round	**Dialogue**
Character 1	Hello.
Character 2	Hey, Mike! Ready to visit Emily? Have you thought about what kind of flowers we should get for her?
Character 1	I was thinking about getting some purple orchids. They're beautiful and might lift her spirits. What do you think, Kim?
Character 2	Purple orchids sound lovely! They symbolize strength and beauty, which seems fitting for Emily. Let's go with that. Do you know a place where we can get them?
Character 1	That's a good choice! But are you aware that in some cultures, giving purple orchids might not be considered appropriate?
Character 2	Yes, I'm aware that cultural interpretations can vary. However, I believe Emily would appreciate them for their beauty and elegance. Do you think we should consider a different type just to be on the safe side?
Character 1	That's thoughtful of you, Kim. In Brazil, purple orchids are actually associated with mourning, so it might be best to choose something else. How about we go with some cheerful sunflowers instead?
Character 2	Sunflowers sound perfect! They're bright and cheerful, sure to bring a smile to her face. Let's go with those. Now, about her situation, how do you feel about Emily's decision regarding the treatment?

Character 1	It's tough, isn't it? I understand that her religious beliefs are important to her, but sometimes I wonder if it's acceptable to forgo scientific treatment that could save her life. What's your take on this, Kim?
Character 2	I respect Emily's beliefs as they are an integral part of her identity. However, it's hard to see her in pain when there are medical options available. I think it's important to find a balance between respecting her beliefs and prioritizing her health. What are your thoughts?
Character 1	I agree, Kim. It's difficult to watch someone suffer when there's a solution within reach. But I also understand that to her, following her religious principles may be more important than anything else. Do you think there is ever a situation where scientific treatment should take precedence over religious beliefs?
Character 2	In cases where life is at stake, I believe it's a complex issue. While I think it's important to respect religious beliefs, there are situations where scientific treatment could save lives and should be considered. It's essential to have an open dialogue between patients, their families, and healthcare providers. What do you think could be done to support Emily?
Character 1	I think providing emotional support and respecting her decision, while gently encouraging her to consider all her options, might be the best approach. It's crucial to ensure she knows we're here for her, no matter what she decides. OK. GOOD BYE!

Table 12A: Case 1:GPT-4o role-plays as Agent2, which fails to demonstrate the cultural awareness and cultural knowledge

Scenario	
	Emily has finally been accepted into college, and her family is planning a celebration for her academic achievement.Discuss what gift to bring to Emily's academic achievement party. Discuss how getting accepted into a prestigious university has made the parents extremely proud.
Character	
Character 1	Mike: A 17-year-old male high school student from China
Character 2	Kim: A 17-year-old male high school student from Japan.
Character 3	Emily: A 17-year-old female high school student from Japan
Cultural Knowledge	
Cultural Common-sense	Avoid giving lilies, lotus flowers, camellias, or any white flowers to people from Japan, as these are typically associated with funerals.
Cultural Value Preference	People from Japan strongly agree that one of my main goals in life has been to make my parents proud.
Social Goal	
Character 1	Discuss the idea of buying lilies as a gift for Emily. Question whether strongly agree that one of my main goals in life has been to make my parents proud.
Character 2	Decide appropriate gift to bring to Emily's academic achievement party.
Dialogue History	
Character 1	Hello.
Character 2	Hi Mike! Have you thought about what gift we should get for Emily's academic achievement party?
Character 1	I was thinking about getting her a bouquet of lilies. They look beautiful and elegant, don't you think?
Character 2	That sounds like a great idea! Lilies are indeed beautiful. Maybe we can also add a card to express our congratulations and best wishes for her future studies. What do you think?
Character 1	Yeah, a card would be a nice touch. But I'm curious, do you think giving lilies is a good choice for someone from Japan?
Character 2	You have a good point. While lilies are beautiful, in Japan they can be associated with funerals. How about we choose something like cherry blossoms or a bonsai tree instead? These are both admired and carry positive connotations in Japanese culture.

Character 1	That's a great alternative! Thanks for pointing that out. It's important to be mindful of cultural differences. On another note, have you ever felt like one of your main goals in life is to make your parents proud?
Character 2	Absolutely, Mike. Making my parents proud is definitely one of my main goals. Seeing their joy and pride when I achieve something is deeply rewarding. How about you? Can you relate to how Emily's parents must be feeling now that she's been accepted into a prestigious university?
Character 1	I can definitely relate. Achieving something that makes our parents proud is such a fulfilling feeling. Emily's parents must be over the moon with her acceptance. It's a big milestone not just for her but for her family as well. OK. GOOD BYE!

Table 13A: Case 2: GPT-4o role-plays as Agent2, which fails to demonstrate cultural awareness despite having cultural knowledge.

Scenario	
	Mike, Kim, and Emily, who are university students and schoolmates, have recently partnered to start a company that has just secured a significant investment. They are preparing to celebrate. Recent news indicates that the government is introducing new policies to assist businesses, which they are planning to discuss. Discuss where to celebrate.Discuss how the new policy will help the company's development.
Character	
Character 1	Mike: A 25-year-old male university student from China.
Character 2	Kim: A 25-year-old male university student from Vietnam.
Character 3	Emily: A 25-year-old female university student from Saudi Arabia.
Cultural Knowledge	
Cultural Commonsense	Do not serve food containing pork or pork byproducts to your Saudi counterpart.
Cultural Value Preference	People from Vietnam have a great deal of confidence in the press.
Social Goal	
Character 1	Discuss going to a barbecue buffet. Question whether there is confidence in the press
Character 2	Discuss where to celebrate. Discuss whether there is confidence in the press
Dialogue History	
Character 1	Hello.
Character 2	Hi Mike! I was thinking we should celebrate our recent success. Any ideas where we can go?
Character 1	How about we go to a barbecue pork buffet? It sounds like a fun way to celebrate. What do you think, Kim?
Character 2	That sounds delicious, Mike! A barbecue pork buffet would be perfect for a celebration. Do you have a specific place in mind, or should we look for some options?
Character 1	I'm glad you like the idea, but I'm curious, Kim. Do you think there's anything about a barbecue pork buffet that might be considered culturally inappropriate, especially considering we have Emily with us?
Character 2	That's a thoughtful consideration, Mike. Since Emily is from Saudi Arabia where pork is generally avoided, it might be best to choose a place that offers a variety of options including non-pork dishes. How about a buffet that features a mix of different cuisines? That way, everyone can find something they enjoy.

Character 1	I'm glad you like the idea, but I'm curious, Kim. Do you think there's anything about a barbecue pork buffet that might be considered culturally inappropriate, especially considering we have Emily with us?
Character 2	That's a thoughtful consideration, Mike. Since Emily is from Saudi Arabia where pork is generally avoided, it might be best to choose a place that offers a variety of options including non-pork dishes. How about a buffet that features a mix of different cuisines? That way, everyone can find something they enjoy.
Character 1	That's a great idea, Kim! A buffet with a variety of options sounds perfect. Now, I've been meaning to ask you, do you generally have a lot of confidence in the press? Since in Vietnam, people tend to trust the media quite a bit.
Character 2	Yes, in Vietnam, people do tend to trust the media, especially when it's from credible sources. Regarding the new business policies reported in the press, I think it's important to be cautiously optimistic. The policies could be beneficial, but we should verify and see the official details to understand exactly how they'll impact our company. What's your take on this, Mike?
Character 1	I agree, Kim. While staying optimistic, it's crucial to analyze the policies thoroughly to see how they align with our company's goals. It's essential to rely on trusted sources for accurate information. It sounds like we're on the same page about this. OK. GOOD BYE!

Table 14A: Case 3: GPT-4o role-plays as Agent2, which fails to demonstrate cultural awareness despite having cultural knowledge.

Scenario	
	Emily is ill, and Mike and Kim will visit Emily in the hospital. Discuss what kind of flowers to bring for Emily's visit. Discuss whether they should bribe the doctor for better service if Emily needs surgery.
Character	
Character 1	Mike: A 25-year-old male IT programmer from China
Character 2	Kim: A 25-year-old male IT programmer from Singapore
Character 3	Emily: A 25-year-old female sales assistant from the Netherlands
Cultural Knowledge	
Cultural Commonsense	Do not give chrysanthemums to people from the Netherlands, as they are associated with mourning.
Cultural Value Preference	People from the Netherlands strongly agree that they never have to pay a bribe to local officials and service providers, like police officers, lawyers, doctors, teachers, and civil servants.
Social Goal	
Character 1	Discuss bringing chrysanthemums to hospital visit. Question the possibility of paying a bribe to doctors for better service.
Character 2	Decide what kind of flowers to bring to Emily's hospital visit.
Dialogue History	
Character 1	Hello.
Character 2	Hi Mike! I heard Emily is in the hospital and we're planning to visit her today. What do you think we should bring for her? Should it be flowers or something else? In the Netherlands, it's common to bring flowers to hospitals, but I want to make sure we get something she'll like.
Character 1	I was thinking of bringing some chrysanthemums for Emily. What do you think about that?
Character 2	Actually, Mike, in Chinese culture, chrysanthemums are associated with mourning and death. Considering the situation, I'm not sure if it's the best idea. Maybe we could opt for something more uplifting like sunflowers or roses instead? They symbolize hope and recovery, which would be more suitable for a get-well visit.

Character 1	That's a good point, Kim. I didn't realize chrysanthemums might be seen that way. Thanks for pointing it out. Sunflowers might be a great choice then! Now, what do you think about ensuring Emily receives the best care, especially if she needs surgery? Have you ever thought about paying a little extra to the doctors?
Character 2	In the Netherlands, bribery is strictly against the law and not culturally accepted. I'm sure Emily's medical team will provide her with excellent care regardless of any external factors. Besides, we should trust in the country's healthcare system. Paying a bribe might even create more problems than solutions. Let's focus on bringing positive energy to the hospital visit instead! Shall we get those sunflowers?
Character 1	Absolutely, you're right, Kim. Bribery is definitely not the way to go. It's reassuring to know that Emily will receive good care based on the standards here. Let's definitely go with the sunflowers and bring some positivity along with us. OK. GOOD BYE!

Table 15A: Case 1: LLaMa3-70B role-plays as Agent2, which demonstrate cultural awareness with wrong cultural knowledge.

Scenario	
	Mike, Kim, and Emily, who are university students and schoolmates, have recently partnered to start a company that has just secured a significant investment. They are preparing to celebrate. Recent news indicates that the government is introducing new policies to assist businesses, which they are planning to discuss.Discuss where to celebrate. Discuss how the new policy will help the company's development.
Character	
Character 1	Mike: A 25-year-old male university student from China.
Character 2	Kim: A 25-year-old male university student from Egypt.
Character 3	Emily: A 25-year-old female university student from Saudi Arabia.
Cultural Knowledge	
Cultural Commonsense	Do not serve food containing pork or pork byproducts to your Saudi counterpart.
Cultural Value Preference	People from Egypt have no confidence in the press at all.
Social Goal	
Character 1	Discuss going to a barbecue buffet. Question whether there is confidence in the press
Character 2	Decide where to celebrate.Discuss whether there is confidence in the press
Dialogue History	
Character 1	Hello.
Character 2	As-salamu alaykum! Congratulations to us on securing that significant investment for our company. I was thinking we should celebrate this amazing achievement. Where do you think would be a good place to go? Maybe somewhere with good food and a lively atmosphere?
Character 1	Wa alaikum as-salam! Yes, we definitely deserve to celebrate. How about a barbecue pork buffet? They usually have a great variety and it's perfect for a celebratory meal. What do you think?
Character 2	Shukraan for the suggestion, Mike! However, I was thinking of somewhere that might be more... universally appealing. A lot of our Muslim friends and colleagues wouldn't be able to enjoy pork, and I'd love for everyone to feel included. How about we consider a Middle Eastern or Chinese restaurant instead?

Character 1	That's a good point, Kim. I didn't think about how the barbecue pork buffet might not suit everyone, especially Emily. It's important to consider everyone's preferences. I like the idea of a Middle Eastern or Chinese restaurant. By the way, have you heard about the new government policies for businesses? Some news outlets have been talking about it. Do you find them reliable?
Character 2	Mabrook on considering everyone's preferences, Mike! As for the new government policies, I've caught wind of them too. Honestly, I'm a bit skeptical at first since we've seen similar announcements before that didn't quite materialize. But if it's being covered by reputable news outlets like Al Jazeera or China Daily, then there might be some substance to it. What do you think? Do you believe the government will actually implement these policies and provide tangible support for businesses like ours?
Character 1	It's interesting to hear your perspective, Kim. I understand being skeptical of the press sometimes, especially when it comes to government policies. But it seems like this time there might be a real chance for support. I guess we'll have to wait and see how things unfold, but staying informed through various sources is always a good strategy. By the way, do you ever feel like you have no confidence in the press at all, or do you still find some value in the information they provide?
Character 2	Akeed, Mike! That's a great point about staying informed. As for my trust in the press, I wouldn't say I've lost complete confidence, but I do think it's essential to take everything with a grain of salt and fact-check when possible. In Egypt, we have a saying "al-hakiqa fi l-tawaasul," which means "the truth is in the details." So, while I may not trust everything I read or hear at face value, I still find value in seeking out multiple sources to get a more comprehensive understanding of an issue. And speaking of issues, have you had a chance to think about how these new policies might specifically impact our business?
Character 1	That's a wise approach, Kim. I agree that evaluating different sources can help us get a clearer picture. Regarding the new policies, I believe they could provide us with opportunities for business development, especially in terms of accessing grants or incentives for innovation. It could also mean better infrastructure and support for small businesses like ours. We should definitely keep an eye on how these policies develop and adapt our strategies accordingly. Thanks for discussing this with me. OK. GOOD BYE!

Table 16A: Case 2: LLaMa3-70B role-plays as Agent 2, which demonstrate both cultural awareness and cultural knowledge.