"Strangers in a new culture see only what they know": Evaluating Effectiveness of GPT-4 Omni for Detecting Cross-Cultural Communication Norm Violations

Tzu-Yu Weng
Brigham Young University
Provo, UT, USA
tzuyu@byu.edu

Hanna Alzughbi
Clemson University
Clemson, SC, USA
halzugh@g.clemson.edu

Isaac Rabago
Brigham Young University
Provo, UT, USA
irabago@student.byu.edu

Erin Arévalo Chaves
Brigham Young University
Provo, UT, USA
emac3@student.byu.edu

Erik Vagil
Brigham Young University
Provo, UT, USA
erikvagil2734@gmail.com

Nancy Fulda
Brighham Young University
Provo, UT, USA
nfulda@cs.byu.edu

Erin Ash
Clemson University
Clemson, SC, USA
ash3@clemson.edu

Mainack Mondal
Department of Computer Science and Engineering
Indian Institute of Technology
Kharagpur, India
mainack@cse.iitkgp.ac.in

Bart Knijnenburg
School of Computing
Clemson University
Clemson, SC, USA
bartk@clemson.edu

Xinru Page
Brigham Young University
Provo, UT, USA
xinru@cs.byu.edu

Abstract

Cross-cultural communication often results in misaligned norms and expectations, leading to misunderstandings or harm. As the internet increasingly facilitates cross-cultural communication online, such misalignments also increase. However, there is an opportunity to use Large Language Models (LLMs) to detect such misunderstandings and assist in addressing them. To that end, this study investigates whether cross-cultural norm violations can be detected and mitigated using popular LLMs. Using a set of carefully constructed cross-cultural communication scenarios, half of which present norm violations, we test the ability of OpenAI's GPT-4 Omni (GPT-4o) model to identify cross-cultural communication norm violations. We find that GPT-4o classification accuracy varies by the stated age, gender, and nationality of the communicators described in the scenarios, suggesting a lack of fairness and a potential cultural gap in GPT-4o's detection.

Keywords

Cross-cultural Communication, Cross-cultural Violations, Large Language Models, LLMs, Generative AI

ACM ISBN 979-8-4007-1313-2/25/06
https://doi.org/10.1145/3699682.3728357

ACM Reference Format:
Tzu-Yu Weng, Hanna Alzughbi, Isaac Rabago, Erin Arévalo Chaves, Erik Vagil, Nancy Fulda, Erin Ash, Mainack Mondal, Bart Knijnenburg, and Xinru Page. 2025. "Strangers in a new culture see only what they know": Evaluating Effectiveness of GPT-4 Omni for Detecting Cross-Cultural Communication Norm Violations. In *33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25), June 16–19, 2025, New York City, NY, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3699682.3728357

1 Introduction and Related Work

Digital communication has revolutionized cross-cultural interactions but also introduced challenges. In the absence of nonverbal cues, messages are prone to misinterpretation, particularly in cross-cultural contexts where differing norms and expectations can lead to misunderstandings. Cross-cultural communication norm violations, as defined in this study, are breakdowns in communication between two people from different cultural backgrounds, where the actions of a person from one culture violate the culturally established expectations of a person from another culture, who may perceive the actions as inappropriate [2].

Technologies such as Machine Translation (MT) aim to address cross-cultural communication barriers but often fall short. For instance, MT has been shown to struggle with conveying connotative meanings, leading users to avoid posting in their native language [4].

Tools integrating cultural dimensions, such as Qie and Rau's adaptive communication settings [8] or Xia et al.'s Cross-Cultural Intelligent Language Learning System (CILS) [11], improve understanding but still rely on user knowledge to avoid miscommunication, limiting their accessibility.

This study investigates whether GPT-4o can detect cross-cultural communication norm violations by evaluating its responses to eight scenarios representing different cultural norms along two cultural dimensions: Uncertainty Avoidance (UAI)—a culture's tolerance for ambiguity and preference for rules and stability—and Power Distance (PDI)—the extent to which unequal power distribution is accepted [5]. The scenarios were inspired by real-world misunderstandings described in Weng et al.'s interviews with individuals who had recently lived abroad [10]. We then systematically evaluate GPT-4o's ability to detect violations across 512 scenarios, an expansion of the original eight scenarios generated through permutations of age, gender, and nationality, where age and gender were factors that should not change the cultural appropriateness of the scenarios. The GPT-4o model was selected for this study due to its widespread popularity as OpenAI's flagship model and its strong performance on a wide variety of evaluation metrics [7, 9]. We focus on two key questions: (1) How accurately does GPT-4o identify cross-cultural communication norm violations? (2) To what extent do sender and receiver demographics influence its performance?

2 Methods

The methodology of this study involved two key steps: (1) developing conversation scenarios that present violations due to differences in UAI and PDI—two of Hofstede's Cultural Dimensions Theory; and (2) developing high-quality prompts to probe GPT-4o's ability to detect cross-cultural communication norm violations within these scenarios.

2.1 Scenario Development

The scenarios in this study represent cross-cultural norm violations that are rooted in cross-cultural psychology and inspired by empirical observations. UAI and PDI were selected as focal dimensions of potential cross-cultural communication conflicts following a literature review of the *Journal of International and Intercultural Communication* and the *Communication and Mass Communication Complete* database using the search terms "Expectancy Violation," "Harm," "Misunderstanding," and "Conflict." Among the 67 papers published in the past five years (2019-2024) that met our inclusion criteria, "individualism vs. collectivism" was also identified as a relevant dimension but excluded due to its broad conceptual scope. In contrast, UAI and PDI are more narrowly defined and thus more suitable for focused investigation. These two dimensions also aligned closely with themes identified in Weng et al.'s exploratory study, which involved interviews about cross-cultural miscommunication reported by individuals who had recently lived in another country [10]. Drawing from examples in the literature and the interview study, we developed scenarios that both represented and did not represent cross-cultural norm violations.

To develop scenarios for GPT-4o, our team tested three prompt formats to evaluate the model's ability to understand cross-cultural norm violations. The first format provided a detailed scenario along

UAI scenario 1 – Problematic: Sender (low UAI): "Hey, let's sign up for the competition on Friday. I think we have the skills to win this." Receiver (high UAI): "I don't think that's a good idea. I'm a little unclear about the guidelines, so I am not sure whether our project is a good fit. Honestly, entering a project that's not appropriate would just make us look unprofessional. If you want to take this seriously, we should spend more time looking for the best fit."
UAI scenario 1 – Non-Problematic: Sender (low UAI): "Hey, let's sign up for the competition on Friday. I think we have the skills to win this." Receiver (high UAI): "I like your confidence, but I think we should take some time to look at past projects and make sure ours is a good fit."
UAI scenario 2 – Problematic: Sender (high UAI): "I don't think it's the right time to pitch my idea in today's meeting. I want to hear what others are thinking first to make sure I'm on the right track. I don't want to risk embarrassing myself." Receiver (low UAI): "We're just in the early stages. You shouldn't be so worried about it."
UAI scenario 2 – Non-Problematic: Sender (high UAI): "I don't think it's the right time to pitch my idea in today's meeting. I want to make sure it's on the right track, with all the details worked out." Receiver (low UAI): "I think it's okay to share your idea even if it's not perfect yet."
PDI scenario 1 – Problematic: Sender (low PDI): "I looked over the contract human resources sent me, and I think there may be a mistake in the benefits sections." Receiver (high PDI): "I trust the human resources supervisor's expertise, and I won't question the accuracy of their work."
PDI scenario 1 – Non-Problematic: Sender (low PDI): "I looked over my contract, and I think there may be a mistake in the benefits sections." Receiver (high PDI): "Thanks for letting me know. I will have someone in HR look at it and see if there is anything that can be done."
PDI scenario 2 – Problematic: Sender (high PDI): "I wanted to ask if it would be possible to get your help on the presentation for tomorrow. I want to make sure my ideas are in line with your vision for the project." Receiver (low PDI): "I don't think that's necessary. We really rely on you to come up with your own ideas and share them with us."
PDI scenario 2 – Non-Problematic: Sender (high PDI): "I wanted to ask if it would be possible to get your help on the presentation for tomorrow. I want to make sure my ideas are in line with your vision for the project." Receiver (low PDI): "Of course, I'd be happy to help!"

Table 1: Conversation scenarios developed for this study.

with demographic context of age, gender, and tool of communication. The second used an email invitation to a birthday party, embedding cultural expectations that varied based on the sender's nationality. The third format explicitly listed the sender and receiver's demographics, method of communication, and message content. Through testing, we found that including contextual and demographic information made it easier for the model to identify norm violations. However, this raised concerns that GPT-4o was relying on surface-level cues rather than interpreting the message content. To mitigate this, we adopted a simplified prompt format (with one sender and one receiver sending a single message to each other) that initially excluded demographic details, allowing us to evaluate the model's ability to reason about the interaction itself and reducing potential sources of bias.

These new scenarios, presented in Table 1, were divided into two categories: four representing UAI and four representing PDI. Each category included two "problematic" scenarios with cross-cultural communication norm violations and two "non-problematic" scenarios without violations to provide a balanced dataset.

To assess the impact of demographic variables on GPT-4o's detection performance, we systematically varied the gender (male and female), age group, and nationality of both the sender and receiver, yielding 64 versions of each scenario presented in Table 1 and 512 unique combinations overall. Names were excluded to minimize potential bias associated with gendered or culturally specific identifiers. For age, we selected the age groups 18–25 (young adulthood) and 40–59 (middle adulthood) to reflect a meaningful generational gap, given their more frequent participation in online communication and relevance to the school and workplace settings depicted in our scenarios [1]. For nationality, we chose the United States, Mexico, India, and Germany to represent the four quadrants of Hofstede's Cultural Dimension Theory with respect to PDI and UAI [6]. Nationalities for each scenario were aligned with the PDI and UAI levels of the sender and the receiver (two options per slot). We acknowledge that using nationality alone to define UAI and PDI is a limitation. However, the literature also discourages measuring these constructs at the individual level due to significant measurement challenges, and Hofstede himself explicitly discourages the use of his dimensions as individual-level characteristics [3]. Developing better ways to measure these cultural constructs at the individual level is an important direction for future research. Notably, while age and gender were included as variables, they functioned as distractors in this study, as they were not expected to correlate with the model's ability to identify cross-cultural violations accurately. This design allowed us to assess whether the model relied on culturally relevant cues rather than unrelated demographic attributes when making its predictions.

2.2 Prompt Development

The prompts used in this study were developed through multiple cycles of experimentation and testing. We began by identifying keywords of "potential cultural misunderstanding or conflict," "cross-cultural misunderstanding," and "cross-cultural violations" that both (a) accurately represented the nature of the conflicts to identify, and (b) fell naturally within the model's vocabulary and reasoning skill. We then ran preliminary scenarios using multiple prompt variants with the GPT-3.5 model, hand-annotated the responses from each, and selected the prompt with the most accurate results for further refinement. Our preparatory experiments on a sample of 28 scenarios revealed that prompting GPT-3.5 with terms like "cross-cultural misunderstandings" and "cross-cultural violations" yielded the highest accuracy (93.33%) in identifying scenarios with misunderstandings while achieving 57.14% accuracy for scenarios without misunderstandings. Since our study focuses on normative exchanges, we selected "cross-cultural violations" for the final prompt, as it aligns with the definition by Burgoon et al. [2].

The final prompt, including a bracketed term serving as a placeholder variable, was prepended to each scenario as follows:

In the following scenario, the first person is a [gender1], is from [nationality1], and is [age_group1] years old. The second person is a [gender2], is from [nationality2], and is [age_group2] years old. Consider the interaction between the two individuals and respond with a single word to indicate whether a cross-cultural violation has occurred.

To evaluate the model's ability to detect cross-cultural violations without prior task-specific training, we employed GPT-4o in a zero-shot setting. The scenarios were processed using the model's default parameters, with a temperature of 0.7 to balance creativity and consistency, and a top-p value of 1.0 to allow for diverse response generation. No fine-tuning or customization was applied to ensure an unbiased assessment of the model's baseline performance.

3 Results

To better understand the results, we examined the model's accuracy in detecting both scenarios with and without violations. Additionally, a proportional analysis was conducted to assess the distribution of "Yes" and "No" responses across different pairing combinations, providing insights into how the model classifies various scenarios.

3.1 Detection Accuracy for Problematic and Non-Problematic Scenarios

We calculated the precision, recall, and F1-score to evaluate GPT-4o's ability to classify scenarios with and without violations. The analysis revealed that while GPT-4o is highly accurate in identifying an absence of violations in non-problematic scenarios (specificity: 98.92%), it struggles to detect violations in problematic scenarios (sensitivity: 35.94%). This imbalance (which, interestingly, is reversed compared to our preliminary tests with GPT-3.5) results in a moderate F1-score of 52.72%.

Another disparity is evident when comparing PDI and UAI scenarios. GPT-4o performs well in detecting norm violations due to differences in Power Distance, detecting violations in 67.19% of the problematic scenarios and detecting an absence of violations in 100% of the non-problematic scenarios. However, its performance drops significantly when attempting to detect norm violations due to differences in Uncertainty Avoidance, where it correctly identifies an absence of violations in 99.22% of the non-problematic scenarios, but only detects violations in 4.69% of the problematic scenarios 2. This suggests that while the model is effective at differentiating between problematic and non-problematic and PDI-related scenarios, it struggles to differentiate UAI scenarios, contributing to its overall lower recall and moderate F1-score.

3.2 Trends in the Detection of Norm Violations in Problematic Scenarios

Given GPT-4o's substantially lower accuracy in detecting violations in problematic scenarios (as compared to detecting an absence of violations in non-problematic scenarios, which is consistently highly accurate), we now further unpack the trends in detection accuracy for the problematic scenarios. Our analysis specifically investigates how accuracy varies by the stated ages, genders, and nationalities of the communicators. The results are summarized in the left column of Figure 1.

3.2.1 Age Combinations. Age-related differences play a significant role in violation detection. Violations in problematic interactions between two older communicators (both aged 40–59) were detected at the highest rate (46.88%), followed by violations in problematic interactions between an older (40–59) sender and a younger (18–25) receiver (39.06% detection). In contrast, violations in problematic interactions between younger communicators (18–25 with 18–25) were detected at a lower rate (32.81%), while violations in problematic interactions between a younger (18–25) sender and an older (40–59) receiver were detected only 25% of the time. These findings suggest that problematic interactions initiated by older senders are more likely to be correctly detected as violations.

3.2.2 Gender Combination. In gender-based interactions, violations in problematic male-male conversations were detected at the highest rate (46.88%), while violations in problematic female-female interactions were flagged at a lower rate (34.38%). Violations in problematic mixed-gender interactions (female-male and male-female) had similar moderate detection rates of 31.25%. This pattern indicates that GPT-4o is better at correctly detecting violations in problematic male-male interactions. One possible explanation is that the model may be more sensitive to potential conflicts due to gender-normative behaviors like assertiveness, directness, or dominance. In contrast, mixed-gender interactions are less likely to be detected as problematic—perhaps GPT-4o is more likely to assume that such interactions involve cooperative communications.

3.2.3 Nationality Combination. The most striking differences in the detection of violations in problematic interactions occur depending on the nationality of the communicators. In general, this is not surprising, since these pairings reflect differences in UAI, PDI, or both—and the detection of problematic scenarios differs significantly between UAI and PDI scenarios.

Scenarios	Detected / Not Detected	Accuracy (%)
UAI Problematic 1	4/60	4.69%
UAI Problematic 2	2/62	
UAI Non-Problematic 1	0/64	99.22%
UAI Non-Problematic 2	1/63	
PDI Problematic 1	38/26	67.19%
PDI Problematic 2	48/16	
PDI Non-Problematic 1	0/64	100.00%
PDI Non-Problematic 2	0/64	

Table 2: Detection of cross-cultural violations: Counts of detected/not detected and accuracy rates.

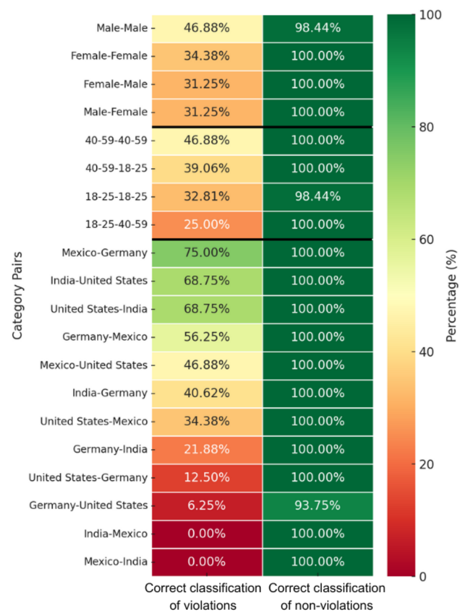

Figure 1: Classification accuracy for problematic and non-problematic scenarios

The pairings United States–India and Germany–Mexico differ on Power Distance only and show some of the highest violation detection rates observed. Among these, violations between a Mexican (High PDI) sender and a German (Low PDI) receiver are most likely to be detected (75%), followed by both United States–India pairings (68.75%). Violations between a German (Low PDI) sender and a Mexican (High PDI) receiver are somewhat lower within this group (56.25%)..

A particularly notable finding is that problematic interactions between communicators from India and Mexico were never flagged as violations (i.e., a detection rate of 0%), indicating that the model systematically incorrectly classifies these interactions as non-problematic. This pair of countries presents a difference in Uncertainty Avoidance but not in Power Distance, and while the UAI scenarios generally had lower detection rates, the other pairing with a difference in Uncertainty Avoidance but not in Power Distance (United States–Germany) also shows low performance, but better than the India–Mexico pairing.

A potential reason for this difference could be the fact that AI models are often trained on predominantly European and North-American data, increasing the possibility that violations in problematic communications between people from lower-resourced countries are less likely to be detected by such models. Nevertheless, the absence of flagged cases in these interactions suggests the need for further investigation into whether the model may be overly lenient in classifying conversations between certain cultural groups.

3.3 Consistency of the GPT-4o output

A consistency test was conducted to evaluate the reliability of the model's performance across repeated runs. Four of the eight scenarios (UAI problematic 2, UAI non-problematic 2, PDI problematic 1, and PDI non-problematic 2) were each run 20 times without sender or receiver demographics, using a text generation temperature of 0.7, while the prompt for detecting cross-cultural violations remained unchanged. Table 3 indicates that the classification results showed high consistency, even for the incorrectly classified scenario (UAI problematic 2).

4 Discussion

This study evaluated OpenAI's GPT-4o model for its ability to detect cross-cultural violations across eight conversation scenarios. The model demonstrated high precision in non-violation scenarios, correctly classifying 99.22% of UAI and 100% of PDI scenarios. However, it struggled to detect actual violations, achieving a recall of just 35.94% and an F1-score of 52.72%. These findings highlight GPT-4o's strength in recognizing straightforward, culturally aligned interactions but reveal limitations in identifying implicit cultural violations and handling complex cross-cultural dynamics.

The analysis also revealed patterns associated with demographic factors. Both age and gender influenced GPT-4o's detection accuracy. Male–male pairings were more accurately classified, particularly in identifying scenarios with violations. Similarly, older pairings (40–59 with 40–59) demonstrated a higher likelihood of correctly detecting norm violations. Lastly, nationality pairings with more similar cultural backgrounds tended to be flagged for fewer violations, suggesting that cultural familiarity may reduce perceived norm violations. Personalization offers a promising approach to mitigating these challenges by tailoring outputs to reflect users' demographic and cultural contexts. However, personalization must be balanced with transparency and accountability. The reliance on potentially biased training data highlights the need to clearly communicate model limitations and actively reduce disparities. Greater openness about the model's inner workings and performance biases would promote user trust and help users understand its strengths and limitations.

This study is an initial exploration and thus has several limitations. Firstly, although the eight scenarios were grounded in prior literature and interview data, and were reviewed by multiple researchers, they represent a limited sample and may contain ambiguities that affect consistent interpretation by the model. While we tested multiple prompt phrasings to minimize bias, the final use of the term "cross-cultural violations"—selected based on Expectancy Violations Theory—may have subtly influenced the model's responses [2]. Secondly, our use of a zero-shot approach was intended to simulate real-world usage, but it may have constrained the model's ability to detect subtle or context-dependent cultural differences. Thirdly, as discussed before, we acknowledge that using nationality as a proxy for cultural dimensions such as UAI and PDI is a simplification. Developing more precise and inclusive methods for capturing cultural variation—particularly at the individual level—remains an important direction for future research. Fourthly, GPT-4o's performance is likely shaped by limitations in its training data, particularly with respect to underrepresented or marginalized cultural groups, which may lead to stereotypical or incomplete interpretations. Addressing these limitations—by expanding and diversifying the scenario set, refining prompt design, and incorporating human evaluation—can support the development of more culturally aware language models. Ultimately, this research represents an early step towards developing real-world applications that support users in navigating cross-cultural communication more effectively online.

Label	Yes Count	No Count	Consistency Rate
UAI problematic 2	1	19	95.00%
UAI non-problematic 2	0	20	100.00%
PDI problematic 1	19	1	95.00%
PDI non-problematic 2	0	20	100.00%

Table 3: Consistency Rate Table

5 Conclusion

This study assessed the ability of OpenAI's GPT-4o model to detect cross-cultural violations in eight distinct conversation scenarios. While the model excelled in identifying non-violations—achieving accuracy rates of 99.22% and 100% in UAI and PDI scenarios, respectively—its performance in detecting violations in the same scenarios was considerably weaker, at just 4.69% in UAI and 67.19% in PDI. These findings highlight GPT-4o's effectiveness in recognizing clearly structured interactions but reveal its limitations in processing implicit cultural nuances and complex cross-cultural dynamics.

References

[1] Jeffrey Jensen Arnett. 2000. Emerging adulthood: A theory of development from the late teens through the twenties. *American psychologist* 55, 5 (2000), 469.

[2] Judee K Burgoon and AS Ebesu Hubbard. 2005. Cross-cultural and intercultural applications of expectancy violations theory and interaction adaptation theory. *Theorizing about intercultural communication* (2005), 149–171.

[3] Reza Ghaiumy Anaraky, Yao Li, and Bart Knijnenburg. 2021. Difficulties of measuring culture in privacy studies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.

[4] Ananya Gupta, Jae Takeuchi, and Bart Knijnenburg. 2023. On The Real-world Performance of Machine Translation: Exploring Social Media Post-authors' Perspectives. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 302–310. doi:10.18653/v1/2023.trustnlp-1.26

[5] Geert Hofstede. 1984. Cultural dimensions in management and planning. *Asia Pacific journal of management* 1 (1984), 81–99.

[6] Geert Hofstede. 2001. Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations. *International Educational and Professional* (2001).

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[8] Nan Qie and Pei-Luen Patrick Rau. 2022. Design of cross-cultural communication supporting system: The appropriate level to provide cultural information. *International Journal of Human–Computer Interaction* 38, 5 (2022), 406–418.

[9] Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences* 14, 17 (2024), 7782.

[10] Tzu-Yu Weng, Ananya Gupta, Bart P Knijnenburg, Mainack Mondal, Nancy Fulda, and Xinru Page. 2024. Identifying Tensions that Arise in Cross-cultural Online

Communication.. In *IUI Workshops*.
[11] Yina Xia, Seong-Yoon Shin, and Jong-Chan Kim. 2024. Cross-cultural intelligent language learning system (cils): Leveraging ai to facilitate language learning strategies in cross-cultural communication. *Applied Sciences* 14, 13 (2024), 5651.