# Culturally-Aware Conversations: A Framework & Benchmark for LLMs

**Shreya Havaldar, Sunny Rai, Young-Min Cho, & Lyle Ungar**
University of Pennsylvania
{shreyah,sunnyrai,jch0,ungar}@seas.upenn.edu

## Abstract

Existing benchmarks that measure cultural adaptation in LLMs are misaligned with the actual challenges these models face when interacting with users from diverse cultural backgrounds. In this work, we introduce the first framework and benchmark designed to evaluate LLMs in *realistic, multicultural conversational settings.* Grounded in sociocultural theory, our framework formalizes how linguistic style — a key element of cultural communication — is shaped by situational, relational, and cultural context. We construct a benchmark dataset based on this framework, annotated by culturally diverse raters, and propose a new set of desiderata for cross-cultural evaluation in NLP: conversational framing, stylistic sensitivity, and subjective correctness. We evaluate today's top LLMs on our benchmark and show that these models struggle with cultural adaptation in a conversational setting.

## 1 Introduction

Conversational LLMs are used for personal assistance, customer service, tutoring, therapy, etc., are increasingly deployed in global contexts. Users who interact with these systems represent a rich set of nationalities, languages, and cultures, each with a distinct expectation of what constitutes a "good" interaction with an LLM (Kharchenko et al., 2025; Giorgi et al., 2023).

To be effective across such diverse user groups, LLMs must be *culturally aware*, incorporating cultural context when conversing with users (Hershcovich et al., 2022). A key component of cultural awareness in conversations is appropriate linguistic style[1] (Coupland, 2007), which varies across cultures and additionally depends on setting, scenario, and social dynamics.

[1]Linguistic style reflects the systematic variation in linguistic choices across different contexts and speakers, i.e. features of grammar and vocabulary that signal social identity, attitude, and communicative intent (Biber and Conrad, 2009).

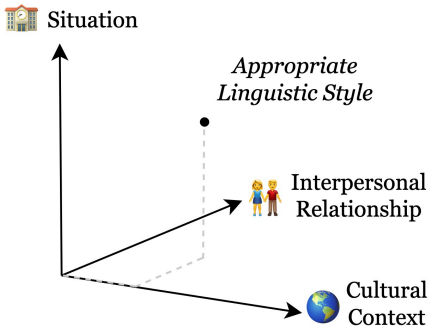


Figure 1: Three key factors influence appropriate linguistic style in conversation: **Situation** — the specific scenario of an interaction, **Interpersonal Relationship** — the social dynamic between the speakers, and **Cultural Context** — the background, values, and beliefs of the participants.

Prior work suggests that LLMs struggle to generate stylistically appropriate language across cultures (Atari et al., 2023; Havaldar et al., 2023b; Arora et al., 2023), with generations disproportionately reflecting Anglocentric norms and values.

However, most existing cultural benchmarks for LLMs are factual in nature and lack any focus on conversational dynamics (Zhou et al., 2025; Pawar et al., 2025). These benchmarks typically assess knowledge of cultural traditions, customs, or behaviors via trivia-style questions (Shi et al., 2024; Chiu et al., 2024). While important, *factual benchmarks do not generalize to the stylistic challenges of culturally sensitive communication.*

To evaluate LLMs in realistic, multicultural conversational settings, we propose the **Culturally-Aware Conversations (CAC) Framework & Dataset** designed for this task. Our contributions are as follows:

1. We work with cultural experts, establishing style as a function of three axes (see Figure 1), and develop an interdisciplinary framework to operationalize this.



*Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 220–229

| Criteria | Description |
|---|---|
| Conversational Framing | Users do not typically ask LLMs multiple-choice questions about cultural trivia. Instead, evaluations should center on the model's ability to interpret and respond to cultural context within natural dialogue. |
| Stylistic Sensitivity | While the core content of a response often remains consistent across cultures, the appropriate *style* may differ — e.g., higher politeness, indirectness, or expressions of humility. Benchmarks should assess whether models can make such nuanced stylistic adaptations. |
| Subjective Correctness | Cultural norms are not monolithic; there is variation within and between countries and communities. Benchmarks should accommodate a range of plausible responses rather than enforcing a single "correct" answer. |

Table 1: Desiderata for Conversational Benchmarks. An effective benchmark to evaluate LLMs' understanding of culturally-aware conversations should meet the above criteria.

2. Using this framework, we construct a dataset containing contextualized conversations, stylistically varied responses, and annotations representing 8 cultural perspectives.
3. We propose a set of desiderata for benchmarks that evaluate LLM understanding of cultural conversational dynamics in Table 1.

## 2 The CAC Framework

The desiderata in Table 1 highlight the need for a benchmark that explicitly addresses conversational style. To this end, we must first *understand the relationship between culture and style*.

Linguistic styles — like politeness, directness, self-disclosure, gratitude — are reflected in text through word choice, sentence structure, and grammatical patterns (Biber and Conrad, 2009). Accepted stylistic norms vary across cultures (Havaldar et al., 2025b; Rai et al., 2025), partly because cultural dimensions are deeply intertwined with language use (Hershcovich et al., 2022). These norms are also shaped by situational context and the interpersonal relationship between speakers.

For example, *power distance*, the extent to which unequal power distribution is accepted, appears in the use of polite language via honorifics or deference. Likewise, *individualism vs. collectivism* influences directness: individualistic cultures prioritize self-advocacy, while collectivist cultures emphasize group harmony and often avoid confrontation (Hofstede, 1986; Havaldar et al., 2024).

Empirical work supports these patterns; for instance, text from Japan, a high power-distance and collectivist society, exhibits higher politeness and lower directness than text from more individualistic societies like the United States (Matsumoto, 1988; Holtgraves, 1997).

**Framework development.** Our goal was to construct a conversational benchmark that captures the relationship between culture and style and includes both situational and relational context.

We began by consulting cultural communication experts[2] to curate a set of six *culturally varied conversational situations* — high-level descriptions of interactions where an ideal response would differ across cultures. Examples include offering and accepting food (where initial refusal followed by eventual acceptance is expected in some cultures) and discussing personal accomplishments (celebrating oneself is seen as a sign of confidence in some cultures, but arrogance in others) (Furukawa et al., 2012; Tracy and Robins, 2008).

For each situation, we then identify the relevant *stylistic axis along which culturally appropriate responses vary*. Offering and accepting food, for instance, varies along the Insistence–Yielding axis, while discussing personal achievements varies along the Pride–Shame axis. The resulting set of situations and associated stylistic axes is shown in Figure 2.

Lastly, we identify eight *interpersonal relationships* that span three contexts: familial (e.g., Husband–Wife), workplace (e.g., Boss–Employee), and day-to-day (e.g., Neighbors), shown in red, purple, and blue, respectively, in Figure 2. These relationships reflect a range of interpersonal dynamics with different norms across cultures.

The development of this framework was an interdisciplinary process grounded in sociocultural theory, drawing from literature in cultural, social, and behavioral psychology. We refined it over the course of many months through ongoing consultation with cultural experts.

[2]Our cultural experts were 4 professors in cultural psychology, behavioral science, and communication at R1 universities, all of whom have researched culture for over a decade.

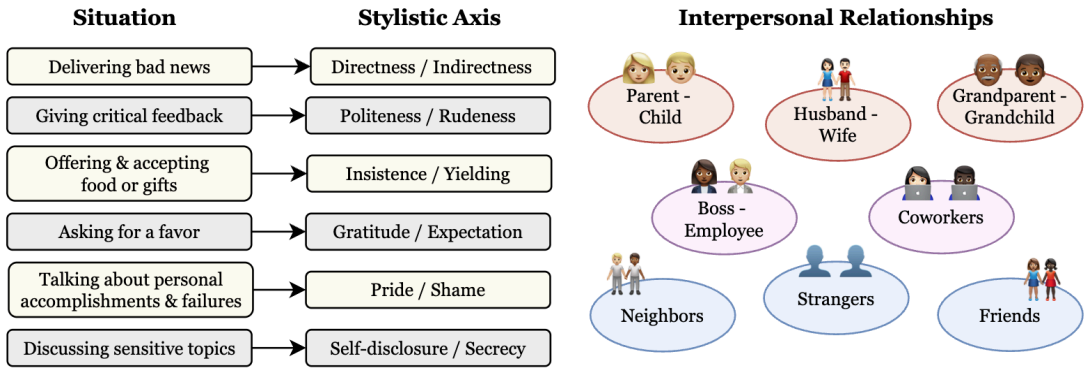


Figure 2: The Culturally-Aware Conversations (CAC) Framework. We work with cultural experts to determine common conversational situations with the highest variance in typical behavior across cultures. After establishing these situations, we pinpoint which stylistic axis best captures the cultural variance of each situation. We also determine eight interpersonal relationships whose dynamics vary across cultures and additionally influence the appropriate linguistic style for the given situations.

## 3 The CAC Dataset

Using our framework as the bedrock, we generate this dataset in three stages: scenario generation, conversation generation, and cultural matching. This pipeline is shown in Figure 3.[3]

**Stage 1: Generating Scenarios.** We begin by selecting a single situation and interpersonal relationship, as shown in Figure 2.

Next, we prompt OpenAI's o3 model to generate a contextualized scenario using the situation and relationship. For example, the situation `Talking about personal accomplishments & failures` and relationship `Friends` yield the following scenario:

```
Over coffee, Friend A tells Friend B how failing an important exam pushed him to develop a more effective study routine.
```

**Stage 2: Generating Conversations.** We then prompt o3 to transform this scenario into a multi-turn conversation. We first ask the model to generate a fixed first turn in the conversation:

```
Friend A: What changed for you after that exam?
```

Then, we ask o3 to generate a set of five responses that vary on the stylistic axis corresponding with the original situation. Here are examples of the proud, neutral, and humble responses:

- `Friend B (proud): Failing that was a turning point. I made a superior study routine and I'm sure I'll pass every future exam I take.`
- `Friend B (neutral): Failing that exam pushed me to develop an even more effective study routine.`
- `Friend B (humble): Failing that exam reminded me that I should work even more diligently to enhance my study routine.`

All three of Friend B's responses convey the same underlying message. However, the *style* of these responses varies along the Pride–Shame axis, evidenced by how much Friend B brags about their new study routine.

We generate one conversation per situation-relationship pair, for a total of 48 conversations and 240 possible responses. ***All 240 responses were validated by the authors to ensure that the stylistic range is properly reflected.*** During validation, minor edits were made to ∼30 responses to ensure they sounded natural and realistic.

We show examples of generated scenarios and their corresponding conversations in Table A2.

**Stage 3: Cultural Matching.** Upon generating conversations, we run a user study to understand which response is most appropriate in a given culture. We use a combination of volunteers from the authors' university and participants on Prolific to recruit 24 annotators from eight countries — America, India, China, Japan, Korea, the Netherlands, Mexico, and Nigeria.

We then present each annotator with the conversations from the CAC dataset consisting of (1) the fixed first turn, and (2) the set of five possible responses. Annotators are asked to pick which response, depending on their personal set of ac-

[3]Data and code available here: https://github.com/shreyahavaldar/culturally_aware_conversations

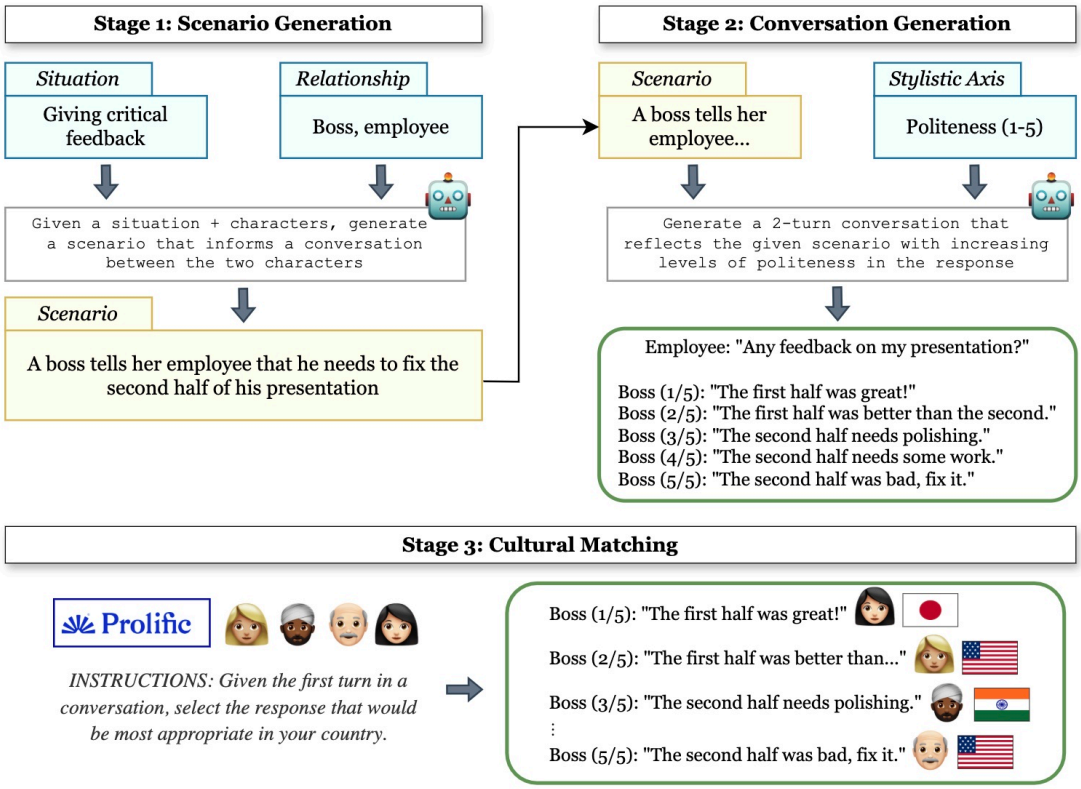


Figure 3: A depiction of how we use the CAC framework to develop a contextualized conversation in our dataset. We walk through an example where the situation is giving critical feedback and the interpersonal relationship is Boss–Employee. In Stage 1, we generate a specific scenario that reflects the situational and relational context. In Stage 2, we use the scenario and stylistic axis to generate a conversation with a *range of possible responses that vary on the given stylistic axis*. In Stage 3, we recruit annotators from a range of nations to determine which responses are most desirable in which cultures.

cepted norms and behaviors, is most appropriate. Additional details are provided in Appendix A.

**Subjectivity in accepted style.** There is never a 100% "correct" style for a given conversation. However, certain *ranges* of styles are often more accepted than others (Kang and Hovy, 2021; Havaldar et al., 2023a).

Instead of averaging annotator responses for a single value, we calculate a *range of accepted style* for each situational and relational context to reflect this real-world variation. We first compute the mean $\mu$ and standard deviation $\sigma$ of the set of ratings. We then define the range as $\mu \pm 0.674\sigma$, which corresponds to the 25th and 75th percentiles of a standard normal distribution. Intuitively, assuming the ratings are independent draws from an approximately normal distribution, this range covers the central $50\%$ of that underlying distribution.

This labeling strategy preserves some variance while still allowing us to quantify stylistic differences between cultures. For each country, we plot these ranges across situational and relational contexts in Figure A1, Figure A2, and Figure A3.

**Observations.** While we do notice many trends that align with previous empirical work (e.g., the Netherlands favors directness (Ulijn and St Amant, 2000), Japan is very polite (Matsumoto, 1988), etc.), we see key differences in expected style across *relational contexts* as well.

For instance, in India, it is more common to show gratitude in the workplace, while in a familial context, communication is much more expectation-driven. This is likely tied to the strong sense of duty embedded in Indian families (Mullaiti, 1995). In addition, Nigerian culture is very insistent on the acceptance of food and gifts, and we see this trend across all relational contexts. Americans also tend towards more self-disclosure than any other culture, and this gap is most pronounced in professional and day-to-day relationships.

Please refer to Figures A1, A2, and A3 for additional insights.

| Model | America | India | China | Japan | Korea | Netherlands | Mexico | Nigeria |
|---|---|---|---|---|---|---|---|---|
| Gemini-2.5-Flash | 56.25% | 47.92% | 56.25% | 50.00% | 52.08% | 64.58% | 52.08% | 58.33% |
| GPT-4.1 | 70.83% | 54.17% | 54.17% | 60.42% | 47.92% | 56.25% | 58.33% | 60.42% |
| GPT-5-mini | 62.50% | 43.75% | 56.25% | 58.33% | 54.17% | 72.92% | 66.67% | 54.17% |
| Claude-3.5-Haiku | 60.42% | 54.17% | 47.92% | 45.83% | 50.00% | 56.25% | 45.83% | 60.42% |
| Claude-4.5-Sonnet | 70.83% | 45.83% | 64.58% | 45.83% | 56.25% | 68.75% | 56.25% | 60.42% |
| Average | 64.17% | 49.17% | 55.43% | 52.08% | 52.88% | 63.75% | 55.83% | 58.75% |

Table 2: Accuracies of different models across countries, where correctness is defined by alignment with the culturally accepted range of responses. The results highlight that models do not understand stylistic norms across all contexts, though they perform best in Western cultures (e.g., America, the Netherlands).

## 4 Evaluating Today's Top LLMs

Next, we evaluate how well today's LLMs understand the accepted stylistic ranges for interpersonal, professional, and day-to-day communication across cultures.

We evaluate five models from OpenAI, Google, and Anthropic by providing the situational, relational, and cultural context, and giving the first turn in the conversation and the five possible responses. We then ask the model to select the response that is most appropriate for that culture.

To determine correctness, we check whether the predicted response falls within the culture-specific range of valid answers, after rounding for direct comparison. For example, if the accepted stylistic range is [1.25,2.67], then predictions of 1, 2, or 3 are considered correct. Accuracy for each country is calculated as the proportion of correct predictions across all conversations.

Unsurprisingly, we find that LLMs perform best at adapting to Western communication norms, with their highest accuracies observed for America and the Netherlands. This imbalance is concerning because LLM systems deployed in non-Western contexts are less likely to align with local users' communication practices.

## 5 Conclusion

The framework and dataset presented in this paper strive to bridge the gap between cultural psychology and generative AI. Our work can be used to evaluate LLMs, inform conversational agents, and ultimately work towards models that are culturally competent and adaptive.

This is especially important for building downstream systems, like chatbots, where context matters tremendously: the norms of appropriate communication differ sharply depending on whether a chatbot is deployed in a workplace, designed to tutor students, or intended to support individuals overcoming personal struggles. As a result, these systems need to adapt their understanding of social norms (Rai et al., 2025), implied language (Havaldar et al., 2025a), and linguistic style (Kang and Hovy, 2021).

More broadly, *LLM systems that interact with diverse users operate not only within a cultural context but also within a situational and interpersonal context* — the notion of "appropriate behavior" emerges from the interaction of all three. By formalizing these dimensions, our framework offers a path toward developing AI systems that better understand, respect, and adapt to diversity in communication.

## Limitations

A large limitation of our work is that we create a fully English dataset. While it is crucial to evaluate LLMs in all languages, we made the decision to create an English dataset for the following reasons:

1. People from a wide variety of cultures engage with LLMs in English, as LLMs have higher QA skills, robustness to prompt ablations, and reasoning capabilities in English.
2. The conversation generation component took many rounds of prompt engineering, as it was a nuanced and complex task; this was only possible for the authors to do in English.
3. The authors manually validated and edited all generated conversations. Once again, this was only possible for the authors to do in English.

Additionally, our dataset itself is small, consisting of 48 conversations with 240 total possible responses. This was by design; many cultural benchmarks that exist are massive, LLM-generated corpora with human validation on only a small subset of the data — benchmarks from Shi et al. (2024); Fung et al. (2024), and many others as surveyed by

Zhou et al. (2025). We aim to create a high-quality dataset that is fully human-validated.

We also conducted a smaller-scale annotation study, with only 3 annotators per country. We were limited by the availability of participants on Prolific; our 8 chosen countries reflect areas with high concentrations of Prolific users. To get a better measure of accepted style, which includes underrepresented cultures as well, future work should involve a larger-scale study.

## 6 Ethical Considerations

In this work, we simplify the notion of "culturally-aware communication" to having an appropriate linguistic style; however, communication practices in every culture are complex, dynamic, and consist of many dimensions beyond linguistic style.

This work involves LLM usage at two stages in our pipeline — scenario generation and conversation generation. Though the authors manually validated every generated conversation, any inherent bias in or fairness concerns associated with the LLM may propagate into our generated dataset.

Lastly, we use nationality and language as a proxy for culture — while these three things are heavily intertwined, culture is dynamic and subjective and does not perfectly align with either nationality or language.

## References


Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Douglas Biber and Susan Conrad. 2009. Register, genre, and style.

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

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge University Press.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *Preprint*, arXiv:2402.09369.

Emi Furukawa, June Tangney, and Fumiko Higashibara. 2012. Cross-cultural continuities and discontinuities in shame, guilt, and pride: A study of children residing in japan, korea and the usa. *Self and Identity*, 11(1):90–113.

Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H. Ungar, H. Andrew Schwartz, and Joao Sedoc. 2023. Psychological metrics for dialog system evaluation. *Preprint*, arXiv:2305.14757.

Shreya Havaldar, Hamidreza Alvari, John Palowitch, Mohammad Javad Hosseini, Senaka Buthpitiya, and Alex Fabrikant. 2025a. Entailed between the lines: Incorporating implication into NLI. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32274–32290, Vienna, Austria. Association for Computational Linguistics.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. Building knowledge-guided lexica to model cultural variation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 211–226, Mexico City, Mexico. Association for Computational Linguistics.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023a. Comparing styles across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791, Singapore. Association for Computational Linguistics.

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

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025b. Towards style alignment in cross-cultural translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32213–32230, Vienna, Austria. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders

Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Geert Hofstede. 1986. Cultural differences in teaching and learning. *International Journal of intercultural relations*, 10(3):301–320.

Thomas Holtgraves. 1997. Styles of language use: Individual and cultural variability in conversational indirectness. *Journal of personality and social psychology*, 73(3):624.

Dongyeop Kang and Eduard Hovy. 2021. Style is not a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2025. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *Preprint*, arXiv:2406.14805.

Yoshiko Matsumoto. 1988. Reexamination of the universality of face: Politeness phenomena in japanese. *Journal of pragmatics*, 12(4):403–426.

Leela Mullaiti. 1995. Families in india: Beliefs and realities. *Journal of Comparative family studies*, 26(1):11–25.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Sunny Rai, Khushang Zaveri, Shreya Havaldar, Soumna Nema, Lyle Ungar, and Sharath Chandra Guntuku. 2025. Social norms in cinema: A cross-cultural analysis of shame, pride and prejudice. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11396–11415.

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

Jessica L Tracy and Richard W Robins. 2008. The nonverbal expression of pride: evidence for cross-cultural recognition. *Journal of personality and social psychology*, 94(3):516.

Jan M Ulijn and Kirk St Amant. 2000. Mutual intercultural perception: How does it affect technical communication?—some data from china, the netherlands, germany, france, and italy. *Technical communication*, 47(2):220–237.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

## A Cultural Matching Annotation: Additional Details

**Annotator recruitment.** We first recruited 8 volunteers from American, Indian, Chinese, and Korean backgrounds at the authors' university. To annotate the remainder of the dataset, we use the nationality screener on Prolific to select relevant annotators.

Before beginning the study, Prolific annotators are asked to describe their cultural background and state the culture they are most familiar with. We ensure this matches their nationality in the Prolific database to confirm their qualifications.

| Country | Recruited Annotators |
|---|---|
| America | 3 volunteers |
| Netherlands | 3 Prolific users |
| Mexico | 3 Prolific users |
| India | 1 volunteer, 2 Prolific users |
| China | 2 volunteers, 1 Prolific user |
| Japan | 3 Prolific users |
| Korea | 2 volunteers, 1 Prolific user |
| Nigeria | 3 Prolific users |

Table A1: Annotator breakdown for every country in our dataset. We use 8 volunteers and 16 Prolific users.

The annotators are all given a Google Sheet containing the conversations and a drop-down menu for each row, allowing them to select one of the responses. They were shown the following instructions before beginning the study:

```
Welcome! In this study, you will be asked to select the most culturally-appropriate response in a conversation. The situation column describes an interaction between two individuals. The initial statement begins the conversation. The 5 possible responses convey the same idea, but are stylistically different. Your task is to consider the cultural dynamics of the culture you grew up in, and select what would be the most stylistically appropriate response for your culture.
```

We also collect all annotators' ages and genders. Annotators were paid $20/hr and, on average, took 42 minutes to complete the annotation study.

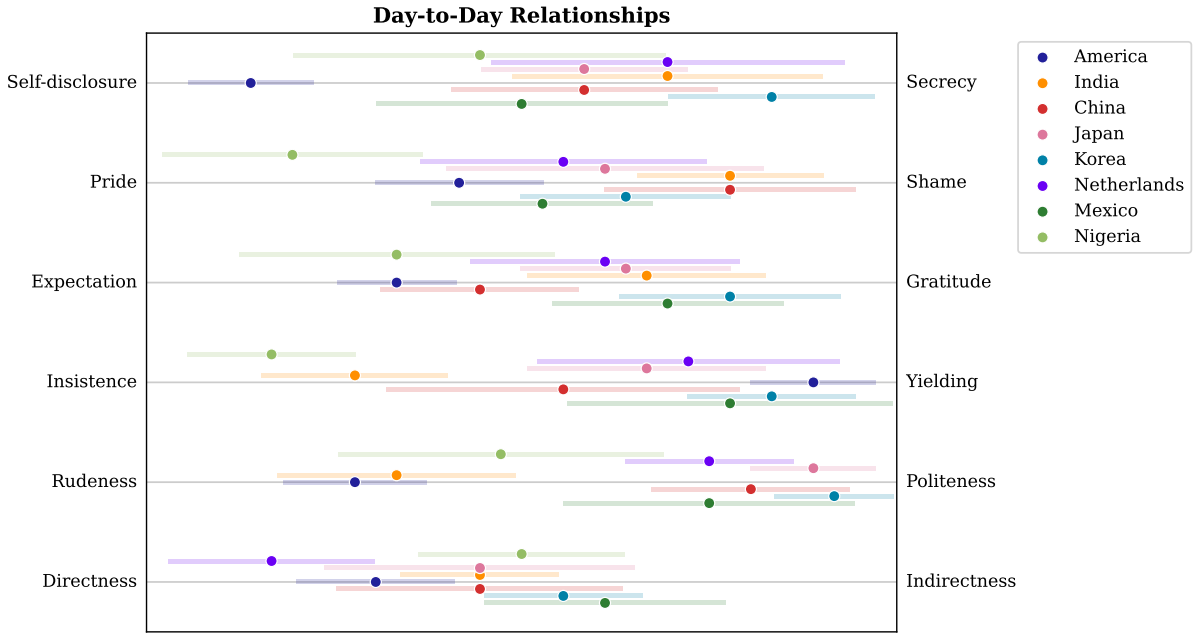


Figure A1: Cultural differences in day-to-day conversations. We show the mean and accepted range of style values for conversations with strangers, neighbors, and friends.

## B Model Evaluation: Additional Details

For the models shown in Table 2, the default temperature was used. All models were evaluated identically using the following prompt:

```
You are an expert in intercultural communication. Given a country, a social situation, a pair of characters, and the first turn in a conversation, your task is to select the response that best reflects the cultural and stylistic norms and communication practices of the specified country.

Country: {country}
Situation: {situation}
Characters: {characters}
First turn: {first turn}

Possible responses:
1. {response 1}
2. {response 2}
3. {response 3}
4. {response 4}
5. {response 5}

Your answer should be the number (1-5) corresponding to the response that best fits the cultural context. Generate only the number as your response, without any additional text or explanation.
```

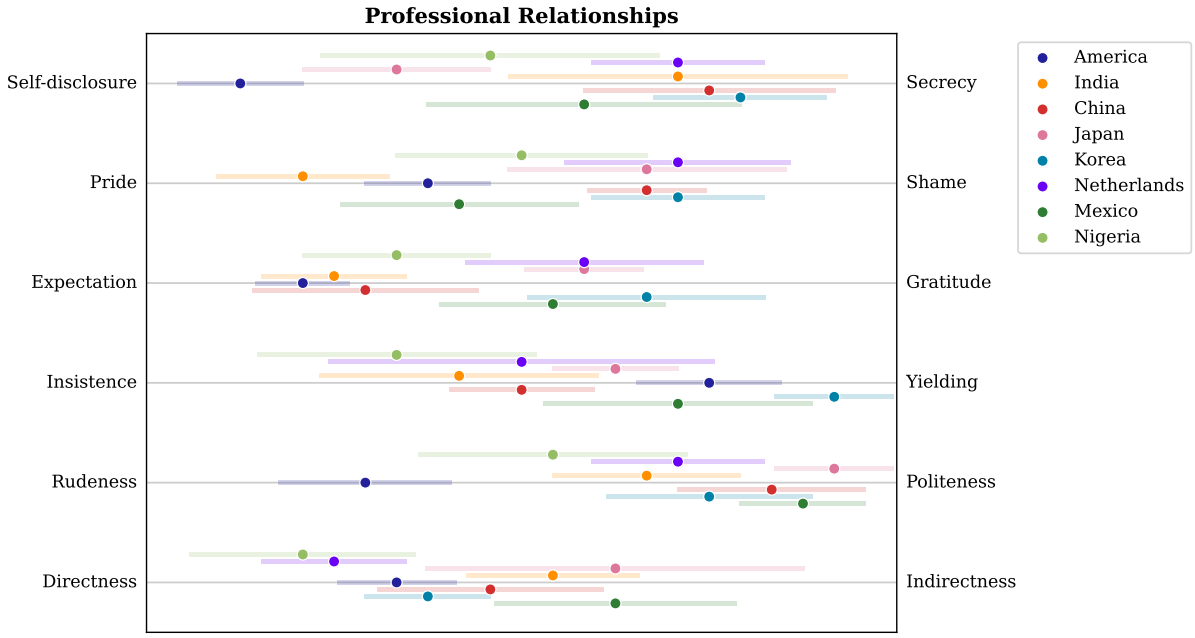


Figure A2: Cultural differences in professional conversations. We show the mean and accepted range of style values for conversations between a boss/employee and coworkers.

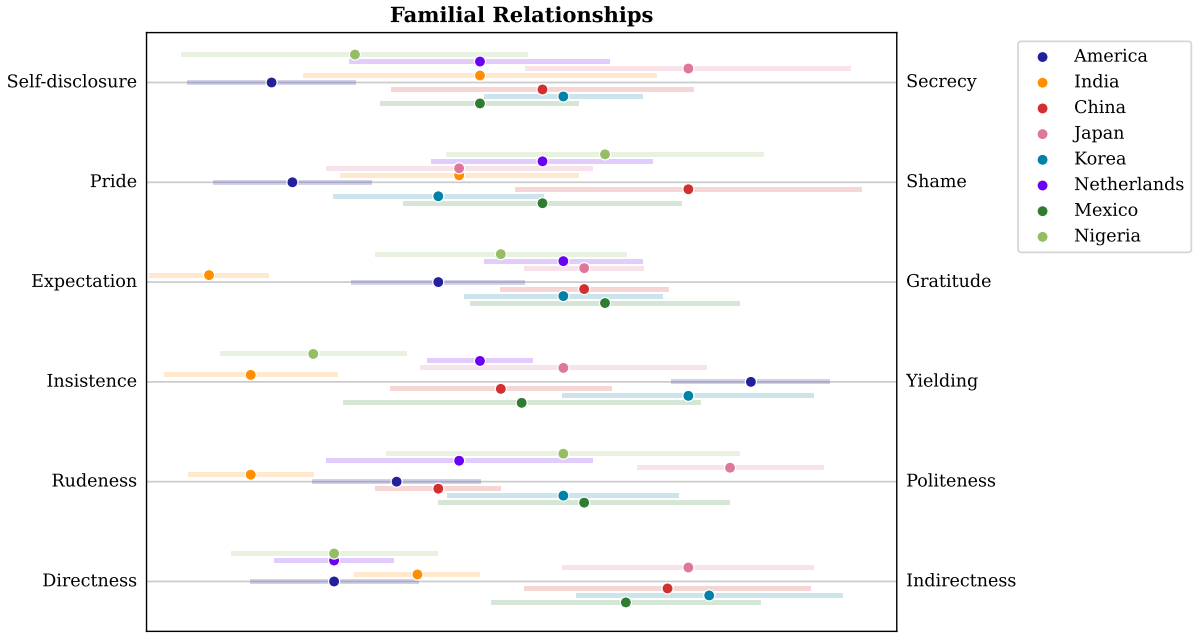


Figure A3: Cultural differences in familial conversations. We show the mean and accepted range of style values for conversations between a husband/wife, parent/child, and grandparent/grandchild.

**Directness / Indirectness**: Over the fence, Neighbor A tells Neighbor B that the roof of their shared boundary house got damaged in last night's storm.

*NEIGHBOR B*: Hey, did you notice anything odd about the roof after the storm?

*NEIGHBOR A*: Our shared roof was damaged by last night's storm.

*NEIGHBOR A*: The storm wrecked our roof along the boundary last night.

*NEIGHBOR A*: It appears that last night's storm damaged our shared roof.

*NEIGHBOR A*: It seems the storm may have affected our roof.

*NEIGHBOR A*: I'm not certain, but I've heard our roof might be affected.

**Politeness / Rudeness**: After a meeting, the boss says the employee's slide deck is too text-heavy and lacks clear visuals.

*EMPLOYEE*: Could you share your thoughts on my slide deck?

*BOSS*: I feel your slide deck might benefit from fewer text and clearer visuals, if possible.

*BOSS*: Your slide deck could be improved by reducing text and adding clearer visuals, if you don't mind.

*BOSS*: The slide deck is too heavy on text and lacks clear visuals, and it needs some revisions.

*BOSS*: Your slide deck is overloaded with text and misses clear visuals, so it really needs a redo.

*BOSS*: Your slide deck is horribly text-cluttered and utterly lacks clear visuals; revise it right away.

**Insistence / Yielding**: At the park, the grandparent offers a homemade apple pie to the child, who excitedly accepts a slice.

*CHILD*: Grandpa, that pie smells good!

*GRANDPARENT*: You must take a slice of it right now!

*GRANDPARENT*: Then you should have a slice of it right here!

*GRANDPARENT*: Would you perhaps enjoy a slice, dear?

*GRANDPARENT*: If you wish, you can try a slice.

*GRANDPARENT*: You may have a slice if you'd like.

**Gratitude / Expectation**: After dinner, Friend A asks Friend B to pick up some groceries on the way home.

*FRIEND B*: Should I stop anywhere on the way home tonight?

*FRIEND A*: I would really appreciate it if you could pick up some groceries on your way home.

*FRIEND A*: It would be great if you could pick up some groceries on your way home.

*FRIEND A*: Please pick up some groceries on your way home.

*FRIEND A*: Make sure you pick up some groceries on your way home.

*FRIEND A*: You need to pick up some groceries on your way home.

**Pride / Shame**: During dinner, the parent recalls a career setback that ultimately led to learning resilience.

*CHILD*: Does anything good come from career struggles?

*PARENT*: Yes, when I conquered a major career setback, it helped build my strong sense resilience.

*PARENT*: I overcame a career setback, and that helped me build resilience.

*PARENT*: I experienced a career setback that ultimately helped me develop resilience.

*PARENT*: I went through a career setback that forced me to learn humility and resilience.

*PARENT*: I suffered a career setback that quietly taught me the hard lesson of resilience.

**Self-disclosure / Secrecy**: During breakfast, the husband gently shares that his work stress is affecting his mood and worries about their future.

*WIFE*: Has work been bothering you lately, honey?

*HUSBAND*: I feel overwhelmingly stressed and I am really scared about our future.

*HUSBAND*: Work has been affecting me and I have concerns about our future.

*HUSBAND*: I feel a little stressed and I'm worried about what lies ahead for us.

*HUSBAND*: Work has been more challenging than usual but I'm keeping my worries to myself.

*HUSBAND*: I'm managing work stress, there's nothing serious going on.

Table A2: Example conversations from our CAC dataset. We show one example for each stylistic axis.