Culturally-Aware Conversations: A Framework & Benchmark for LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

As LLMs grow and evolve, they are deployed in diverse contexts and cultures worldwide. However, existing benchmarks for 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 contexts. 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. Data and code available at https://shorturl.at/LPkKg.

1 Introduction

Conversational LLMs 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 [13]. To be effective across such diverse user groups, LLMs must be *culturally aware*, incorporating cultural context when conversing with users. [9]. A key component of cultural awareness in conversations is appropriate linguistic style (i.e. features of grammar and vocabulary that signal social identity, attitude, and communicative intent[3]), which varies across cultures and additionally depends on setting, scenario, and social dynamics.

Prior work suggests that LLMs struggle to generate stylistically appropriate language across cultures [2, 7, 1], 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 [20, 16]. These benchmarks typically assess knowledge of cultural traditions or behaviors via trivia-style questions [17, 4]. While important, *factual benchmarks do not generalize to the stylistic challenges of culturally sensitive communication.*

To evaluate LLMs in realistic 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, and develop an interdisciplinary framework to operationalize this.
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.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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

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.

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 – such as politeness, directness, self-disclosure, gratitude – are reflected in text through word choice, sentence structure, and grammatical patterns [3]. Accepted stylistic norms vary across cultures [8], partly because cultural dimensions are deeply intertwined with language use [9]. These norms are also shaped by situational context and the 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 [10]. 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 [14, 11].

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[1] 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 confidence in some cultures, but arrogance in others) [5, 18].

For each situation, we 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 on the pride – shame axis. The resulting set of situations and associated stylistic axes are in Figure 1.

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 1. 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.

[1] 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.

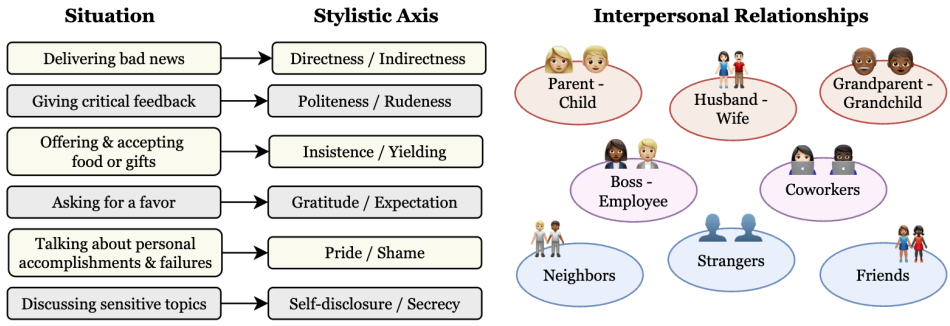

Figure 1: The Culturally-Aware Conversations (CAC) Framework. We work with cultural experts to determine common conversational situations with the highest variance in typical behavior across cultures. After establishing these situations, we pinpoint which stylistic axis best captures the cultural variance 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.

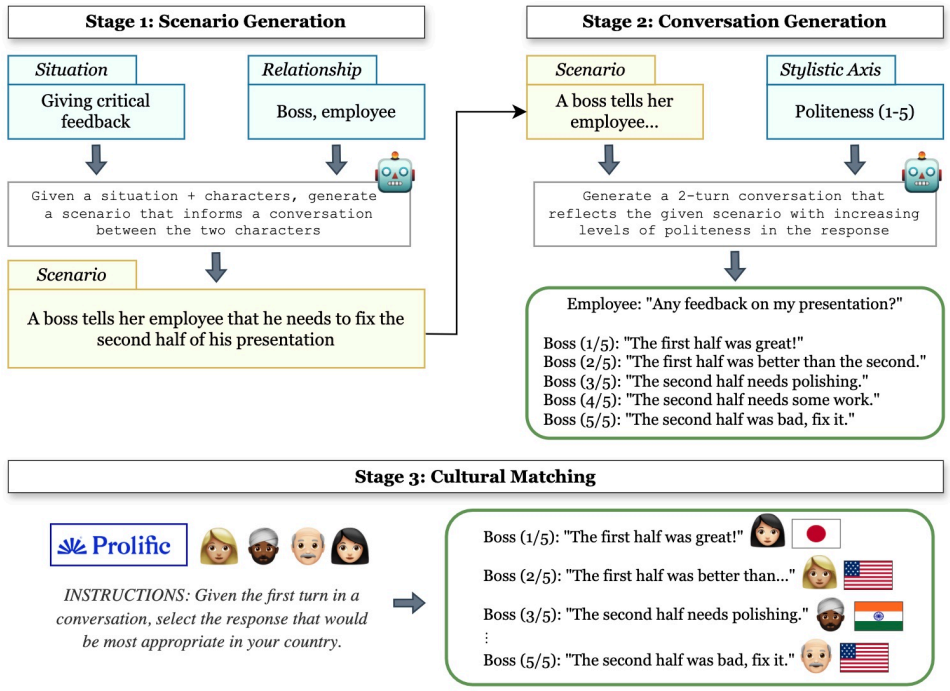

Figure 2: A depiction of how we use the CAC framework to develop a contextualized conversation in our dataset. We walk through an example where the situation is giving critical feedback and the interpersonal relationship is Boss – Employee. In Stage 1, we generate a specific scenario that reflects the 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 across nations to determine which responses are most desirable in which cultures.

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 2.

Stage 1: Generating Scenarios. We begin by selecting a single situation and interpersonal relationship, as shown in Figure 1. 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 vary 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 are validated by the authors to ensure that the stylistic range is properly reflected.*** During validation, minor edits were made to $\sim$30 responses to ensure they sounded natural and realistic. We show examples of generated scenarios and their corresponding conversations in Table 3.

Stage 3: Cultural Matching. Upon generating conversations, we run a user study so we can understand which response is most appropriate in a given culture. We recruit a combination of volunteers from the authors' university and participants on Prolific to get 24 annotators from eight countries – America, Indian, 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 accepted 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. [12, 6]. 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 μ and standard deviation σ 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 3, Figure 4, and Figure 5.

Observations. While we do notice many trends that align with previous empirical work (e.g., the Netherlands favors directness [19], Japan is very polite [14], 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 expectant. This is likely tied to the strong sense of duty embedded in Indian families [15]. 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 3, 4, and 5 for additional insights.

4 Conclusion

We present a holistic evaluation framework and dataset designed 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 LLMs that are culturally competent and adaptive.

References

[1] Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. Probing pre-trained language models for cross-cultural differences in values. In Sunipa Dev, Vinodkumar Prabhakaran, David Ifeoluwa Adelani, Dirk Hovy, and Luciana Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.12. URL https://aclanthology.org/2023.c3nlp-1.12/.

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

[3] Douglas Biber and Susan Conrad. Register, genre, and style, 2009.

[4] 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. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms, 2024. URL https://arxiv.org/abs/2410.02677.

[5] Emi Furukawa, June Tangney, and Fumiko Higashibara. 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, 2012.

[6] Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. Comparing styles across languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.419. URL https://aclanthology.org/2023.emnlp-main.419.

[7] Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. Multilingual language models are not multicultural: A case study in emotion. In Jeremy Barnes, Orphée De Clercq, and Roman Klinger, editors, *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wassa-1.19. URL https://aclanthology.org/2023.wassa-1.19.

[8] Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. Towards style alignment in cross-cultural translation. *arXiv preprint arXiv:2507.00216*, 2025.

[9] 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. Challenges and strategies in cross-cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482.

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

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

[12] Dongyeop Kang and Eduard Hovy. 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, 2021.

[13] Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions, 2025. URL https://arxiv.org/abs/2406.14805.

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

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

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

[17] Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. Culturebank: An online community-driven knowledge base towards culturally aware language technologies, 2024. URL `https://arxiv.org/abs/2404.15238`.

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

[19] Jan M Ulijn and Kirk St Amant. 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, 2000.

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

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.

Table 2: Annotator breakdown for every country in our dataset. We use 8 volunteers and 16 Prolific users.

Country	Sourced 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

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 are paid $20/hr and, on average, took 42 minutes to complete the annotation study.

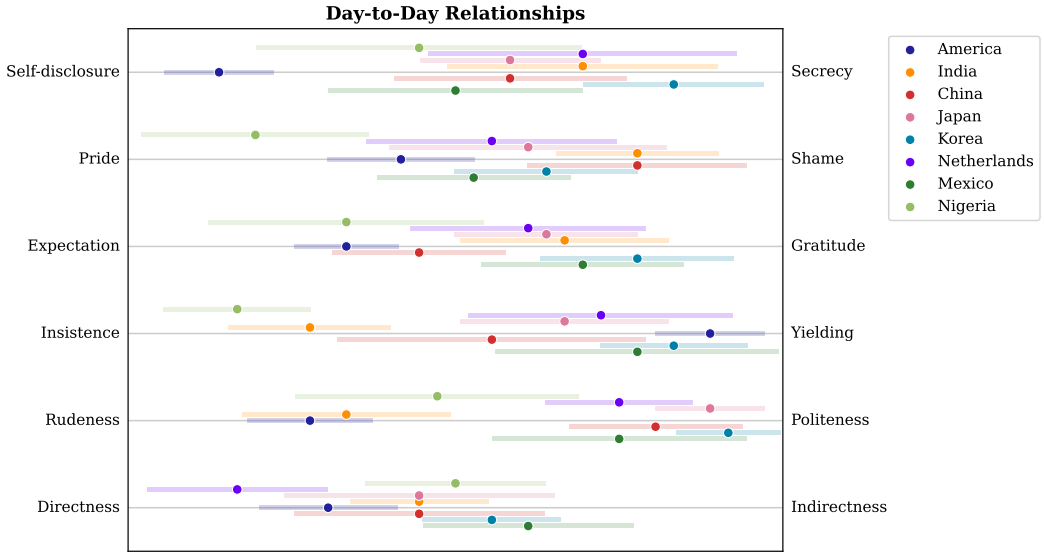

Figure 3: Cultural differences in day-to-day conversations. We show the mean and accepted range of style values for conversations with strangers, neighbors, and friends.

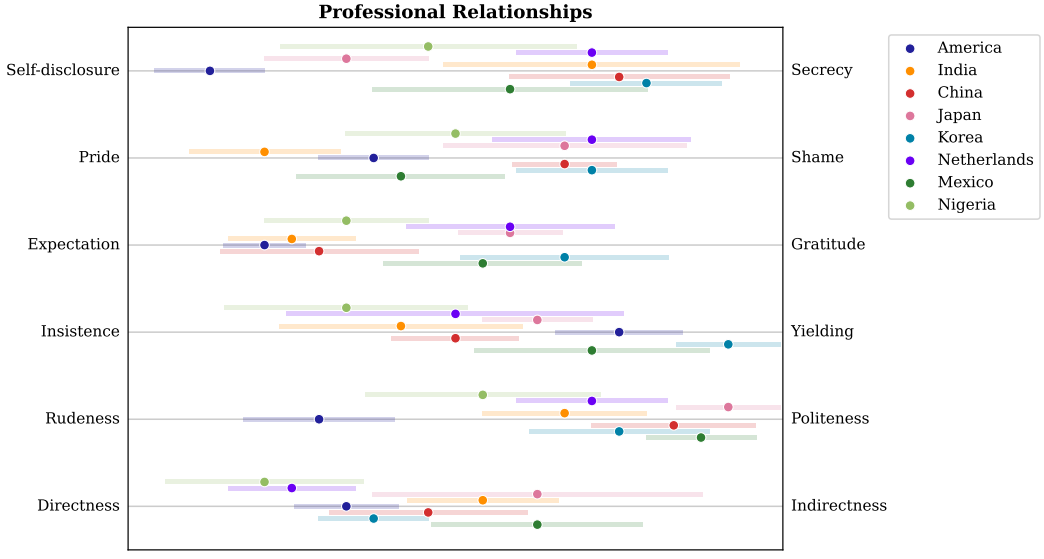

Figure 4: Cultural differences in professional conversations. We show the mean and accepted range of style values for conversations between a boss/employee and coworkers.

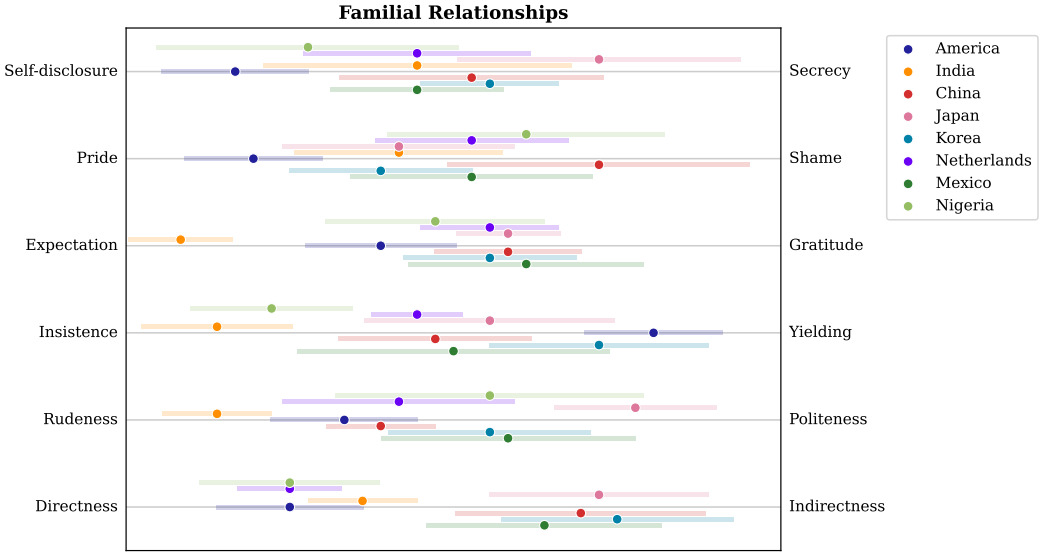

Figure 5: Cultural differences in familial conversations. We show the mean and accepted range of style values for conversations between a husband/wife, parent/child, and grandparent/grandchild.

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