KODIS: A Multicultural Dispute Resolution dialogue Corpus

Anonymous ACL submission

Abstract

We present KODIS, a dyadic dispute resolution corpus containing thousands of dialogues from diverse cultures worldwide. A novel contribution, this corpus adopts a dispute resolution setting in which a *buyer* accuses a *seller* of delivering an incorrect item from an online retail store; tensions further escalate as the two exchange negative reviews. Participants match online to role-play one side and argue over pertinent issues — *refund*, *removing reviews*, and *receiving an apology*. Further, we analyze the cultural differences present and how emotional differences shape outcomes. We make this corpus and data collection framework available to the community upon acceptance of this article.

1 Introduction

Conflicts ubiquitously arise between individuals, organizations, nations, and cultures. Conflicts can help individuals recognize and appreciate differences and learn essential social skills. Too often, conflicts escalate to verbal, legal, or physical violence (Brett et al., 1998; Halperin, 2008). Individual conflicts can damage relationships and incur costly legal fees. National conflicts cost the global economy USD $19 trillion in 2023 (Institute for Economics and Peace, 2024). "Culture wars" within and between nations give rise to different conceptions of reality that can perpetuate generational conflict (Marsella, 2005).

Interest grows in using natural language processing (NLP) methods to understand how conflicts arise and are resolved through conversation (Chawla et al., 2023b; Shaikh et al., 2024; Davani et al., 2023). Conflict dialogues are task-oriented non-collaborative conversations: task-oriented as parties hold specific goals (e.g., extract material concessions or influence beliefs), offering a concrete measure of task success (i.e., what goals were achieved?); non-collaborative as goals are misaligned, though not necessarily zero-sum (while parties try to maximize self-interest, they can discover mutual benefit by considering each other's interests).

This paper introduces a large (4,061 participants) and novel corpus designed to offer multicultural insights into how conflicts escalate or resolve through conversation. It is novel in that we examine *dispute resolution* rather than *deal-making*. Deal-making is a major recent focus of NLP research (Lewis et al., 2017; He et al., 2018; Cheng et al., 2019; Chawla et al., 2021; Kwon et al., 2024), though this literature has favored the term negotiation over deal-making. This creates confusion as both deal-making and dispute-resolution involve negotiation (parties converse to influence each other and extract concessions, often over multiple issues), but disputes involve unique social processes (Brett, 2007) and have received less attention within the fields of AI and NLP. By prioritizing deal-making over dispute resolution, the NLP community risks overlooking key processes that shape conflict (which this corpus seeks to address).

Deal-making is forward-looking as parties focus on opportunities for gain and try to establish a new relationship. As the relationship is not yet established, parties have greater opportunities to explore alternatives. When parties fail to reach an agreement, they can seek other partners – e.g., if unsatisfied with one car dealer, one can always negotiate with another. In contrast, disputes are backward-looking, typically involving an existing relationship that has gone badly. As parties are already linked, success depends on managing the costs of ending the relationship rather than opportunities moving forward (Brett, 2007). As a result, disputes evoke much stronger emotions, and positions are more entrenched than deal-making. This distinction is crucial as it shapes the consequence of influence attempts. For example, whereas expressions of anger promote compromise in deal-making (Van Kleef et al., 2004), they evoke escalation in disputes (Pruitt, 2007). As disputants often become entrenched in their positions, rather than seeking compromise, they seek to overpower their opponent through appeals to justice ("you violated my rights!") or by threatening harm ("I will sue you!"), leading to a spiral of further escalation, including threats of physical violence (Brett et al., 1998; Halperin, 2008; Pruitt, 2007). Thus, the costs of disputes can greatly exceed the original perceived injury, engulfing not only the disputants but other related parties and even society at large.

The corpus is also novel in measuring how culture shapes disputes. We collected a diverse sample of participants from over ninety countries, and participants were matched within and across countries. We measure cultural and individual differences that have previously been shown to shape negotiated outcomes. For culture, we focus on differences between Dignity, Face, and Honor cultures (Leung and Cohen, 2011; Yao et al., 2017). This theoretical framework distinguishes cultures by the degree to which people's social identity is independent versus interdependent and thus shapes the importance given to norms of reciprocity and honesty. Dignity cultures (typically Western society) might respond to a norm violation with a shrug or even a smile. Honor cultures (typically the Middle East or South America) might respond with hot anger (especially if the violation involves family). Face cultures (typically East Asia) might react by shutting down all emotional expressions (Aslani et al., 2016). Additionally, we measure individual differences such as risk-propensity (Meertens and Lion, 2008) and the specific goals parties bring to a dispute. Finally, we assess several theoretical mechanisms surrounding the dispute, including process variables (e.g., what tactics did parties use, what emotions were expressed, and did parties understand their partner's interests?) and outcome variables, including objective and subjective measures concerning the outcome of the dispute.

Finally, this corpus' novelty partially stems from including a mix of human-human and human-AI disputes. Though we focus on human disputes, when participants could not match with a partner promptly, they matched with a large language model (GPT-4), assuming their partner's role. Post-conflict measures include beliefs about whether their partner was human or AI and attitudes towards using AI technology for such applications. Analysis of this data could yield insight into the current limitations of GPT-4; differences between human and AI dialogs; and how these differences shape dispute processes, outcomes, and perceptions.

We envision a wide range of theoretical and applied uses for the corpus. From the social science perspective, disputes offer a unique test-bed to study critical social processes such as the social function of emotional expressions (Friedman et al., 2004), the importance of perspective-taking (Klimecki, 2019; Galinsky et al., 2011), the dynamics of escalation and deescalation (Pruitt, 2007), and the role of cultural and cultural misunderstandings (Tinsley, 2004; Gelfand et al., 2001). From a general artificial intelligence perspective, such disputes provide a vital test-bed for evaluating the social competence of AI systems (Gratch et al., 2015; Kwon et al., 2024; Yongsatianchot et al., 2024) and uncovering the potential AI's potential to propagate cultural stereotypes (Havaldar et al., 2023). From an application perspective, there is growing interest in creating artificial role-players where students can practice and receive feedback on their dispute-resolution skills (Shaikh et al., 2024; Murawski et al., 2024), in creating dialogue agents that monitor interactions between people and intervene to mitigate conflict (Cho et al., 2024), or even replace people in customer service disputes with AI (Ebers, 2022).

Researchers must be mindful of the potential ethical pitfalls of utilizing such corpora in pursuing these uses, particularly when corporate applications race ahead of basic science. The growth in NLP for social science risks creating a GPTology that adversely impacts scientific knowledge production and understanding (Abdurahman et al., 2024; Messeri and Crockett, 2024) and researchers need to take care in drawing general conclusions from NLP analysis of this single corpus. Concerning applications, AI developed to resolve conflict might reinforce structural inequalities across cultures (Lin and Chen, 2022) or be repurposed to create conflict. As a result, we limit the access of this corpus to non-commercial uses.

The article first describes the creation and nature of the `KODIS` corpus. We then summarize a recent study using the corpus to illuminate how emotional expressions shape disputes unfold. These results replicate existing findings in the dispute literature (giving confidence in the quality and scientific potential of the corpus). We further extend these findings to show how culture shapes these processes.

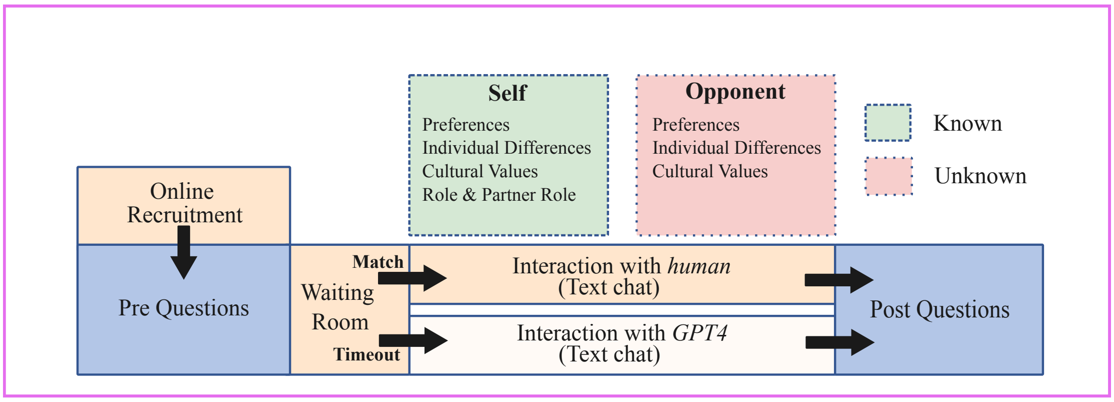


Figure 1: Study flow where participants did pre / post-dispute questionnaires and interacted with their counterpart

2 KODIS Corpus

We introduce *KObe DISpute corpus* (KODIS), a corpus of dyadic disputes. Our data collection was inspired by the CaSiNo framework of Chawla et al. (2021), which allows pairs of human participants to match online and engage in a deal-making exercise via text chat. Prior NLP-based analysis of CaSiNo dialogues found that emotional expressions predict participants' satisfaction with their negotiated agreement (Chawla et al., 2023a), that information expressed in the dialog revealed participant's private goals for the negotiation (Chawla et al., 2022), and that agents could be trained via reinforcement learning to perform effectively negotiate against people (Chawla et al., 2023c). We adapt this framework to dispute resolution and extend it to allow human-agent negotiation.

2.1 Corpus Collection Framework

Fig. 1 illustrates the KODIS data collection framework, which allows dyadic text-based interaction between two human participants or between a human and AI partner (but is easily extensible to support multi-party interactions). The framework utilizes existing software tools. Participants are recruited through an online service such as Prolific or Amazon Mechanical Turk. Participants first enter a survey implemented in Qualtrics (qualtrics.com). This survey administers the consent form, pre-task measures, and describes the dispute scenario. (Note that some information about the task is common knowledge between participants but some is private to each party.) Participants are next directed to the dispute task implemented in Lioness Labs, a software framework used for multi-participant behavioral economic experiments (Giamattei et al., 2020). Here, they first enter a virtual waiting room, which displays a timer counting down from eight minutes and a message asking them to wait while another participant joins. If another participant with an opposing role joins in the first seven minutes, the two meet in a chat interface and work on the dispute; otherwise, at one minute left, the participant moves on and converses with an AI counterpart. After completing the task, participants return to Qualtrics, where they complete post-task measures. Finally, they are routed to the online recruitment service, where they enter a completion code and receive compensation.

When performing the task, we leave the partner's nature (e.g., human or AI) ambiguous but imply to be human. At the end of the study, participants are debriefed about the partner's true nature.

2.2 Application Architecture

We use Giamattei et al. (2020)'s Lioness Lab to implement our design. Lioness allows experiment design with matching participants in dyadic or group studies online. Other research in dialogue collection used this application (Chawla et al., 2021). Figure 10, in the appendix, displays the interface participants communicated through during the dispute — the left side shows the chat box, while the right side hosts the table wherein they can finalize their agreement or walk away.

2.3 Dispute Resolution Task

Participants engage in a role-playing exercise simulating a bilateral multi-issue buyer-seller dispute. We crafted the exercise in collaboration with an expert in dispute resolution (one of the authors) and designed it to evoke strong emotions and entrenched positions while adhering to the ethical guidelines for human experimental research. The dispute centers on a buyer who purchased a basketball jersey for their sick nephew and claims they received the wrong item. The seller disputes this claim, arguing the correct item was sent and no refunds are allowed. Each side is told the other has posted negative reviews attacking their coun-

Buyer: I did not deserve a negative review. Please remove it.
Seller: I am really sorry about that, but I only let the other sellers know the truth, that there was no specific player on the jersey, and that all sales were final.
Buyer: Your site said it was a Kobe jersey, and you sent me the wrong one and lied about it. That's fraud. At least give me a refund.
Seller: It was stated on the website that there is NO specific player, and that all sales are FINAL! I also saw you had an issue with another seller, making up things like you are with me!
Buyer: I know for a fact that it said it was a Kobe jersey. Why are you lying?
Seller: Go to the website and you can see what it says. It is easy to see and easy to read. If you cannot find it, I would be happy to send you the link. It is CLEARLY stated!
Buyer: You changed the site after I purchased it. Expect more negative reviews.
Seller: You are in the wrong completely, and now you are trying to lie about it. I will contact every seller and let them know about you, as well as post on many other sites. I will give you one chance. I will offer you no refund. I will offer you no apology. I don't want an apology from you. What I do want is for you to take down the bad review of me, and I will take down the bad review of you. This is my one offer. Take it or leave it!
Buyer: Go screw yourself.
Seller: You are a liar and a thief!
Buyer: Submitted agreement: Buyer gets no refund, buyer kept their review, seller kept their review, buyer didn't apologize, and seller didn't apologize.
Seller: Reject Deal.
Seller: Submitted agreement: Buyer gets no refund, buyer retracted their review, seller retracted their review, buyer didn’t apologize, and seller didn’t apologize.
Buyer: I Walk Away.

Figure 2: Contentious dispute dialogue example

terpart’s reputation. Each side receives different versions of events to encourage a dispute over facts.

We frame the task as a multi-issue negotiation where participants can discuss and potentially reach a compromise. Participants are told they can discuss the refund, drop their review, request their opponent drop their review, and discuss who, if anyone, should apologize. Figure ?? illustrates a truncated version of the instructions presented to a participant slated to act as a buyer, along with issue descriptions shown to them — a seller sees a similar version, which one may find in the appendix.

Participants are incentivized to take the dispute seriously by offering a substantial bonus based on performance. Participants receive a base pay for attempting the task, but this can be nearly doubled if they achieve all their objectives in the dispute. Many such role-playing scenarios use “assigned preferences,” meaning participants are given a payoff matrix that defines their goals in the negotiation (e.g., they might receive the most bonus if they achieve a refund). To allow cultural variability, we use “elicited preferences,” meaning that participants are provided a fixed number of points they can allocate across the four issues: e.g., buyers might assign 70% of their points to receiving a refund but 30% to receiving an apology. They are compensated based on the actual resolution of the dispute or receive a small portion of this bonus if the dispute ends in an impasse. We motivate performance by paying participants bonuses based on their performance — i.e., how well they did on each issue, considering their preferences.

Participants are given a standard chat interface to engage in the dispute and a menu that reminds them of the issues under discussion and allows them to finalize a deal. After at least eight messages, and when the participant wishes, they can use this menu to send an offer by inputting the proposed level for each issue (via drop-down menu) and sending it to their counterpart — who will then accept or reject the proposal. Participants also have the option to “walk away” from the dispute after at least eight messages by clicking that button under the table — we consider these impasses. Buyers always begin the negotiation (unless one participant is an AI, then the AI always goes first). Participants alternate sending messages; the participant’s interface becomes inactive while awaiting a new message.

2.4 Measures

We measure several theoretical constructs claimed to shape negotiation processes and outcomes, focusing on those used in research on multicultural deal-making (Aslani et al., 2016). This research suggests that cultural and individual variables (e.g., Dignity, Face, and Honor) shape negotiation tactics (e.g., tendency to express emotion, willingness to exchange information). Tactics shape perspective-taking (e.g., less sharing means less understanding of your opponent’s goals), shaping the likelihood of reaching an agreement and the quality of the resulting agreement. Agreement quality can be measured in objective terms, but subjective feelings about the agreement and partner are more predictive of subsequent behavior, such as willingness to follow through on agreements and maintain a future relationship with the partner (Curhan et al., 2006; Brown and Curhan, 2012).

2.4.1 Pre-Dispute Measures

Screening: Participants are screened using techniques shown to improve data quality (Geisen, 2022). They are first asked a *commitment request* asking if they could commit to providing thoughtful responses. They next respond to a series of attention checks, including multiple choice questions (how many legs does a cat have?) and open-ended responses (describe the flavor of a tomato). Participants who fail these checks are immediately excluded without compensation.

Questionnaires: Participants complete a demographic survey including gender, education level, country where they spent the most years of their life (and the number of those years).

They next complete an 18-item Dignity, Face, Honor measure of cultural attitudes (Leung and Cohen, 2011). This measures general tendencies within a culture around the interpretation of identity and threats to identity. Dignity relates to the Western ideal that each individual has intrinsic self-worth and thus tends to be impervious to threats from others. Honor cultures tend to view self-worth as something that must be claimed and defended from external threats. Face cultures also see self-worth as conferred by others but see retaliation as further eroding self-worth. The scale asks participants to consider their culture and answer questions on how frequently they hold certain attitudes on a 7-point Likert scale. Examples include "People must always be ready to defend their honor" and "People should be very humble to maintain good relationships."

Finally, Participants complete a short measure of risk propensity (Meertens and Lion, 2008).

Preference Elicitation: After reading the scenario, participants self-report their goals for the upcoming dispute by allocating 100 points over four issues to be discussed (see Fig. **??** — (*refund*, *other drops negative review*, *you drop negative review*, and *receive apology*). Points indicate the utility of fully or partially achieving this goal and directly map onto the participant's monetary bonus. For example, assigning more points to an apology than to a refund indicates that the participant would experience greater reward from receiving an apology than a refund. Participants are also asked to write a one-sentence justification for the importance assigned to each issue. For example, one participant wrote, "He's a crook and will defraud others" as a justification for wanting to keep his bad review of the other party, suggesting they assign importance to reputational concerns.

Integrative Potential: The preferences of the Buyer and Seller determine the structure of the dispute. Depending on each party's interests, there may be an opportunity to "grow the pie" by finding mutually beneficial solutions. For example, if the Buyer only cares about a refund and the Seller only wants the Buyer to drop their negative review, both sides can maximize their bonus. However, just because a win-win solution exists doesn't mean the parties can find it. The concept of "integrative potential" measures the potential for joint gains (which may or may not be realized). We operationalize integrative potential using the preferences elicitation from each side. Given two vectors of preferences $\vec{X}_{buyer}$ and $\vec{X}_{seller}$:

$$\text{IP} = 1 - \frac{\vec{X}_{buyer} \cdot \vec{X}_{seller}}{\|\vec{X}_{buyer}\| \|\vec{X}_{seller}\|} \quad (1)$$

Notably, above, we invert the cosine similarity of the two vectors — those with dissimilar values have greater potential for joint gains.

2.4.2 Post-Dispute Measures

Perspective-Taking: Participants are asked a series of questions to assess if they accurately understand their partner's goals in the dispute. Participants are first asked about their preferences as an attention check. They rank each issue's importance (most, middle, or least). This is contrasted with preferences they provided during the preference elicitation phase to check if they recall their initial preferences. They are asked to do the same for their partner. The distance between these estimates and their partner's actual preferences can serve as a measure of perspective-taking accuracy.

Tactics: Participants are asked a 10-item questionnaire (Aslani et al., 2016) about the tactics they and their partner used during the dispute. On a five-point Likert scale, participants answer if they agree or disagree with a series of questions about the dispute process. E.g., "I expressed frustration," "The OTHER PARTY expressed frustration," and "I shared my preferences with the other party."

Subjective Value of the Outcome: Subjective perceptions of the outcome of a dispute are a better predictor of future negotiation decisions than the actual economic result (Brown and Curhan, 2012). We capture these impressions with an 8-item version of the Subjective Value Inventory (SVI) (Curhan et al., 2006). This measures four dimensions of subjective value, including feelings about the *instrumental outcome* (e.g., "Did I get a good deal?"), the *process* (e.g., "Was the process fair?"), feelings about the *self* ("Did I lose face?"), and feelings about *relationship* with the partner ("Would I work with my partner again?").

Objective Individual Outcome: The objective value of each side's outcome is calculated from the agreed-upon outcome using the points provided in the preference-elicitation phase. Note that each side may assign different values to the same outcome (e.g., the Buyer may assign more weight to receiving a refund, whereas the Seller might care more about their negative review). These points are also used to determine the bonus. Specifically, the objective value of the outcome is derived using the following linear additive utility function:

$$U_a = \sum_{i \in I} w_{i,a} * \ell_{i,a} \quad (2)$$

Where U_a denotes the points participant a earns; I represents the set of all issues; $w_{i,a}$ holds the value participant a gave to issue $i \in I$; and $\ell_{i,a} \in [0, 1]$ holds the level agreed upon in the dispute — typically binary, though the *refund* issue uses .5 for a partial refund. Notable, this formula only applies in disputes ending in agreement — those resulting in an impasse yield a fixed amount to each side.

Objective Joint Outcome: The sum of individual outcomes in a dispute is a measure of the collective benefit achieved by the two parties. If parties are effective negotiators, the joint outcome should be positively correlated with the integrative potential for the task, but if parties misunderstand each other (e.g., come from different cultures), they may fail to realize this potential.

2.5 Participants and Corpus Characteristics

2.5.1 Demographics

We collected responses from the crowd-sourcing platform Prolific from November 2023 to June 2024. We recruited participants from countries worldwide (see Table 1). Further, our dataset comprises of 50% Female, 49% Male, and 1% other.

2.5.2 Dignity, Face, Honor, & Risk

We analyze the correlations between the various dispositional measures we captured — specifically, dignity, face, honor, and risk. We see Dignity significantly positively correlates with Face ($r = .17$, $p < .001$), and Risk-seeking ($r = .13$, $p < .001$); Face significantly positively correlates with Honor ($r = .07$, $p < .01$) and Risk ($r = .05$, $p < .05$); and Honor significantly correlates with Risk ($r = .22$, $p < .001$). The recruited pool contains respondents of varying dignity ($M = 31.22$, $SD = 4.03$), face ($M = 25.95$, $SD = 5.15$), honor ($M = 24.33$, $SD = 7.04$), and risk-seeking ($M = 4.03$, $SD = 1.67$) propensities.

	Dyad		
Country	Within	Mixed	LLM
United States	826	116	498
United Kingdom	148	90	41
Canada	135	75	15
Mexico	96	62	95
South Africa	84	43	51
Portugal	9	39	5
Poland	6	27	10
Italy	3	13	4
Netherlands	3	22	0
France	1	20	2
Germany	1	14	4
Nigeria	1	17	9
Sweden	1	11	0
<10 appearances	1	133	15

Table 1: Counts of dyads and AI disputes collected

2.5.3 Culture

We attempt to delineate culture using the previously described Dignity, Face, Honor scale, using K-means to cluster similar countries. We only consider countries with $N \geq 10$ appearances in the dyads and calculate each country's mean dignity, face, and honor. Using the elbow method, we run K-means using five clusters — Figure 4 depicts the resultant clusters. For easy visualization, the x and y axis of this figure use PCA decomposition to show the otherwise three-dimensional data; Dignity has positive loadings with PC1 (0.63) and PC2 (0.77); Face has a negative (-0.18) loading with PC1 and a positive (0.05) one with PC2; and Honor has a positive (0.75) loading with PC1 and a negative (-0.63) one with PC2. We see the first cluster comprised of Poland, Greece, Mexico, and Turkey; the second has the Netherlands, Germany, the United Kingdom, the United States, and Canada; the third has Italy, France, and Portugal; the fourth has Nigeria, and South Africa; while the fifth has Sweden.

3 Using Emotion to Analyze Disputes

We summarize an initial use of the corpus to provide insight into how emotions and culture shape dispute processes and outcomes. This highlights an example use case and also demonstrates the corpus has "face validity" in that we replicate some expected findings in the dispute resolution literature. A paper that gives a more detailed explanation of

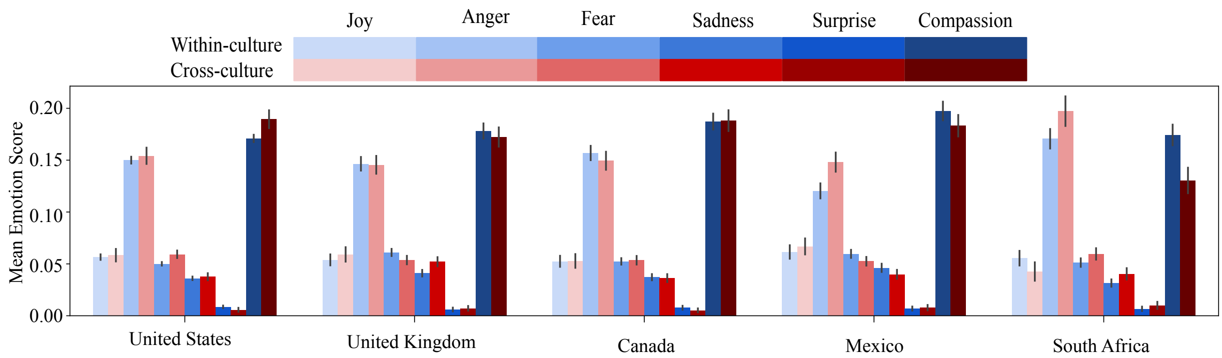


Figure 3: Average emotion scores for each of the five most common countries broken by within or cross-culture

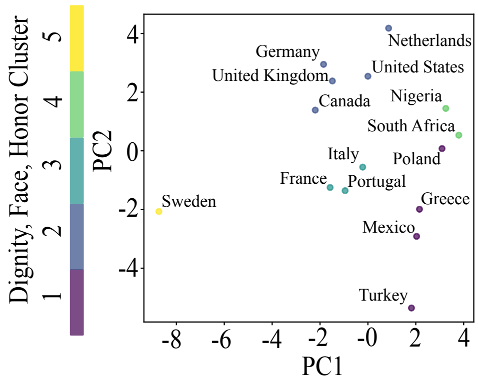


Figure 4: K-means clusters of countries ($N \geq 10$)

these findings is currently omitted for blind review but some details are included in the appendix.

Disputes often evoke strong emotions like anger and expressed anger has different social consequences in disputes than in deal-making. In deal-making, expressions of anger often signal that one side has reached their limit and the other must make concessions to reach an agreement (Van Kleef et al., 2004). Thus, expressions of anger can lead to concession-making. In contrast, anger often provokes escalation in retaliation in disputes (Pruitt, 2007). Culture also plays a role in the consequences of expressions. In dignity cultures, expressions of anger are often viewed as acceptable expressions of self-interest, whereas anger can provoke retaliation in cultures where self-worth is conferred by others (Adam et al., 2010).

We use the KODIS to address several theoretical claims about the role of emotion in disputes: do expressions of anger provoke escalation and impasses in disputes (as previously claimed), can negotiation satisfaction be predicted by emotional expression alone, and how does culture shape these findings?

3.1 Approach

Prior research on the CaSiNo corpus used NLP techniques to examine the impact of emotion in deal-making, and we closely follow that approach (Chawla et al., 2023a), allowing qualitative comparisons with a similar deal-making corpus. Following research demonstrating that GPT-4 currently yields the most accurate inferences on negotiation dialoges (Kwon et al., 2024), we limit our reported analysis to GPT-4o.

GPT4o (run on 06/28/2024) (Achiam et al., 2023) is prompted to annotate each dialogue turn. Following detailed experiments on different prompting methods and emotion labels, GPT is prompted to annotate the intensity ([0..1]) of six emotions expressed in the turn (anger, compassion, sadness, joy, fear, and surprise). Anger and compassion, in particular, were chosen based on their role in prior negotiation research (Van Kleef et al., 2004; Allred et al., 1997; Zhang et al., 2014). Each utterance is presented along with the preceding dialog context, and a small amount of in-context learning guides responses. It also helps GPT report results in a machine-readable JSON format. See the appendix for justification of these choices and evidence that GPT substantially improves predictive accuracy compared to earlier methods (specifically, emotion fine-tuned Google's T5). We only examine human-human dyads; the final agreement (or non-agreement) is excluded from the dialogues.

3.2 Escalation

The negotiation literature suggests that disputes are more likely to end with non-agreement than deal-making due to increased anger and entrenched positions. By comparing KODIS disputes with the CaSiNo deal-making dialogues, we replicate this finding: 19% of the KODIS dialogues ended with an impasse (one side walked away) whereas only

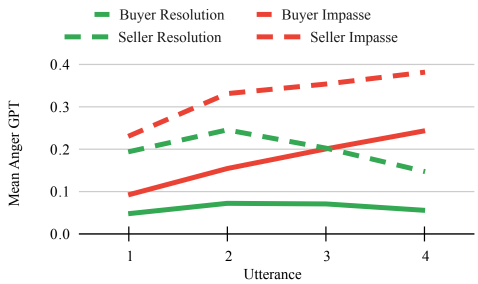


Figure 5: Anger by role, outcome and dialog turn

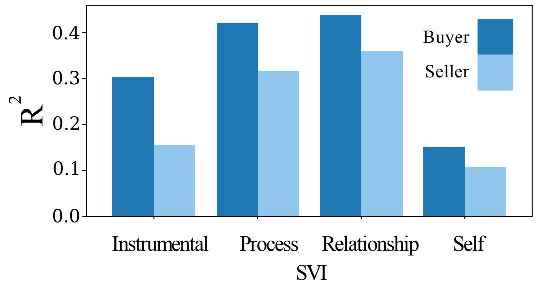


Figure 6: R^2 using GPT emotion to predict outcome

	R^2	
Country	**Within-culture**	**Cross-culture**
US	0.236	0.227
UK	0.188	0.215
Canada	0.211	0.214
Mexico	0.137	0.170
SouthAfrica	0.047	0.157

Table 2: Testing US regression on mixed and pure dyads

3.5% of the CaSiNo dialogues ended with an impasse. This is remarkable as participants forfeit a cash bonus if they fail to achieve an agreement, even though this was merely a simulated dispute.

To examine escalatory dynamics, we divide dialogues based on whether they ended in an agreement or impasse. We then examine expressed emotion by role over time (see Fig. 5). Dialogues that ended in an impasse are in red; those that ended in an agreement are in green. The results show evidence of escalation. Buyers always enter the dialog with greater anger. Dialogues that end with an impasse show evidence of escalation: sellers reciprocate anger, leading to even greater anger by the buyer. In contrast, sellers avoid reciprocating anger for dialogues ending in agreement, and buyers subsequently express less anger. This supports prior research on conflict spirals (Pruitt, 2007).

3.3 Predicting Subjective Outcome

We next examine if emotions expressed during the dialog can predict participants' subjective feelings about the result. We additionally examine how culture shapes these predictions, which could lead to systematic bias, by training a model on Western culture and testing its predictive accuracy on other cultures. Given imbalances in our dataset, we use the nation as a proxy for culture and focus our analysis on the most prevalent countries in our sample (US, UK, Canada, Mexico, and South Africa). Specifically, we train a model on a subset of participants from the dominant nation in our sample (US) and test its performance on unseen participants across the five national groups. Figure 3 illustrates the differences in occurrences of emotions for the five countries we consider when disputing against the same or a different country.

We used multiple linear regression on a random subset of 406 dialogs to predict the four facets of subjective value from expressed emotion. These include feelings about the outcome, self (did I lose face?), process (was it fair?), and relationship with the partner. We construct one model over the entire dataset to assess the relationship between emotion and subjective outcomes. Overall, expressions are surprisingly accurate at predicting subjective feelings. Examining R^2 of the regression models, emotional expressions predict almost 50% of the variance in feelings about the process and the partner. This is particularly remarkable as these models ignore the content of the dialog. Regarding the coefficients, all emotions play a significant role in the model prediction (see Figure 6).

To examine cultural differences, a model was trained on a random 50/50 split of US v. US dialogues and tested on pure and mixed-nation dialogues from the other countries – we ran this 1,000 times on different splits of the US-US dyads and Table 2 displays resultant differences. We see the US within-culture regression explain the variance of the subjective outcome better for the countries similar and worse for dissimilar ones — as uncovered during the cultural K-means.

4 Conclusion & Future Work

We collected a corpus of human disputes and showed some promise using NLP methods to illuminate processes. Current analyses predict cultural tendencies from the dialogues and develop models to assist human disputants (e.g., can an algorithm recognize an escalatory spiral and suggest an intervention to help parties reach an agreement?). The corpus continues to expand, focusing on additional languages and refining the theoretical measures.

5 Limitations

We design the corpus to address existing limitations in the literature on NLP and negotiation by emphasizing the distinction between dispute resolution and deal-making and providing a substantial corpus of disputes. However, it is important to emphasize that the data is drawn from a single artificial scenario. The dispute focused on a standard consumer economic dispute, and participants were asked to role-play. Thus, care must be taken in generalizing these findings to real-world interactions.

This analysis of cultural differences contains confounds that must be unpacked. We predict dispute outcomes from expressed emotion and show that performance degrades when models trained on US participants are applied to non-US participants. Yet it is unclear if this is due to bias in emotion recognition (e.g., does GPT-4o over-estimate anger in South African dialogues) or if emotion functions differently in different cultures. Evidence from other research suggests both factors are probably in play. Thus, the analysis should be augmented by human annotations from those target cultures.

Though we collect information on demographics, personality, and culture, all information comes from self-reports, and we cannot verify. Participants act as part of a paid service, which may shape their responses in ways that do not match real-world interactions. Participants also retained anonymity during the interaction, which can shape their responses. In real interactions, parties have an existing relationship, and there can be real-world consequences. These factors can strongly shape the expression and function of emotions.

We use commercial pre-trained models to recognize emotional expressions in our dialogues. Still, we do not have independent human annotations of what emotions will likely be perceived in the text. So, while we provide some evidence of external validity (expressed emotion impacts outcomes in theoretically predicted ways), subsequent research must verify these machine-generated emotions correspond to the intended mechanism. This is particularly fraught in a cross-cultural setting as existing work shows that large pre-trained models introduce bias in interpretations (Havaldar et al., 2023).

6 Ethical Considerations

Data Collection Our study was approved by our Institutional Review Board (IRB) and classified as exempt (minimal risk). Each participant received an Informed Consent document at the beginning of the study, which covered the purpose of the study, warned about potential discomfort, and noted the collection of data and its later use. Further, the participants were informed that they could withdraw at any time. Participants knew that withdrawing meant they would lose the opportunity to receive a performance-based bonus. This can be seen as a form of coercion, though it is viewed as in line with situations people would experience in the real world and is in line with experimental norms in behavior economics, which seek to ensure that experimental protocols match real-world decision-making (Rousu et al., 2015). The compensation was set to provide a fair wage and to conform to the guidelines of the online collection service. No personally identifiable information was collected during the collection. Potentially identifiable information such as IP addresses, worker IDs, and location information is removed before releasing the data. Any mention of demographics or personality of participants is based on self-identified information in our pre-survey and standard procedures of collecting personality metrics.

Potential Risks Our work supports using NLP methods to provide insight into psychological processes. However, there are reasonable concerns that NLP can undermine the diversity of scientific research (by over-reliance on a small number of tools), create the illusion of objectivity, and reinforce cultural stereotypes (Messeri and Crockett, 2024). This is particularly the case for research on emotion. Recent findings in affective science emphasize that emotions are perhaps best seen as cultural constructs labeled and interpreted differently across cultures. Yet many labeling schemes used in emotion databases rely on Western representational taxonomies. This is true of the labels we adopted in the evaluation experiment. This can serve to reinforce Western biases on the interpretation of the data. As noted by Dehghani et al., using LLMs "as an off-the-shelf 'one-size-fits-all' method in psychological text analysis—can lead to a proliferation of low-quality research, especially if the convenience of using LLMs such as ChatGPT leads researchers to rely too heavily on them." Augmenting our findings with diverse models and human judgments remains imperative.

Acknowledgments

OMITTED FOR ANONYMITY

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3(7):pgae245.

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*.

Hajo Adam, Aiwa Shirako, and William W Maddux. 2010. Cultural variance in the interpersonal effects of anger in negotiations. *Psychological Science*, 21(6):882–889.

Keith G. Allred, John S. Mallozzi, Fusako Matsui, and Christopher P. Raia. 1997. The influence of anger and compassion on negotiation performance. *Organizational Behavior and Human Decision Processes*, 70(3):175–187.

Soroush Aslani, Jimena Ramirez-Marin, Jeanne Brett, Jingjing Yao, Zhaleh Semnani-Azad, Zhi-Xue Zhang, Catherine Tinsley, Laurie Weingart, and Wendi Adair. 2016. Dignity, face, and honor cultures: A study of negotiation strategy and outcomes in three cultures. *Journal of Organizational Behavior*, 37(8):1178–1201.

Jeanne M Brett. 2007. *Negotiating globally: How to negotiate deals, resolve disputes, and make decisions across cultural boundaries*. John Wiley & Sons.

Jeanne M Brett, Debra L Shapiro, and Anne L Lytle. 1998. Breaking the bonds of reciprocity in negotiations. *Academy of Management Journal*, 41(4):410–424.

Ashley D Brown and Jared R Curhan. 2012. The utility of relationships in negotiation. *The Oxford handbook of economic conflict resolution*, pages 137–154.

Kushal Chawla, Rene Clever, Jaysa Ramirez, Gale M Lucas, and Jonathan Gratch. 2023a. Towards emotion-aware agents for improved user satisfaction and partner perception in negotiation dialogues. *IEEE Transactions on Affective Computing*.

Kushal Chawla, Gale M Lucas, Jonathan May, and Jonathan Gratch. 2022. Opponent modeling in negotiation dialogues by related data adaptation. *arXiv preprint arXiv:2205.00344*.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185.

Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023b. Social influence dialogue systems: A survey of datasets and models for social influence tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 750–766.

Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. 2023c. Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13078–13092, Singapore. Association for Computational Linguistics.

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3325–3335, Minneapolis, Minnesota. Association for Computational Linguistics.

Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, and Jonathan May. 2024. Can language model moderators improve the health of online discourse? 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 7478–7496, Mexico City, Mexico. Association for Computational Linguistics.

Jared R Curhan, Hillary Anger Elfenbein, and Heng Xu. 2006. What do people value when they negotiate? mapping the domain of subjective value in negotiation. *Journal of personality and social psychology*, 91(3):493.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Martin Ebers. 2022. Automating due process-the promise and challenges of ai-based techniques in consumer online dispute resolution. In *Frontiers in Civil Justice*, pages 142–168. Edward Elgar Publishing.

Ray Friedman, Cameron Anderson, Jeanne Brett, Mara Olekalns, Nathan Goates, and Cara Cherry Lisco. 2004. The positive and negative effects of anger on dispute resolution: evidence from electronically mediated disputes. *Journal of Applied Psychology*, 89(2):369.

Adam D Galinsky, Debra Gilin, and William W Maddux. 2011. *Using both your head and your heart: The role of perspective taking and empathy in resolving* *social conflict*, volume 13. Psychology Press New York.

Emily Geisen. 2022. Improve data quality by using a commitment request instead of attention checks. Accessed on Oct 13, 2024.

Michele J Gelfand, Lisa H Nishii, Karen M Holcombe, Naomi Dyer, Ken-Ichi Ohbuchi, and Mitsuteru Fukuno. 2001. Cultural influences on cognitive representations of conflict: Interpretations of conflict episodes in the united states and japan. *Journal of applied psychology*, 86(6):1059.

Marcus Giamattei, Kyanoush Seyed Yahosseini, Simon Gächter, and Lucas Molleman. 2020. Lioness lab: a free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6(1):95–111.

Jonathan Gratch, David DeVault, Gale M Lucas, and Stacy Marsella. 2015. Negotiation as a challenge problem for virtual humans. In *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings 15*, pages 201–215. Springer.

Eran Halperin. 2008. Group-based hatred in intractable conflict in israel. *Journal of Conflict resolution*, 52(5):713–736.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. *arXiv preprint arXiv:2307.01370.*

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Institute for Economics and Peace. 2024. Global peace index briefing 2024. Accessed on Oct 13, 2024.

Olga M Klimecki. 2019. The role of empathy and compassion in conflict resolution. *Emotion Review*, 11(4):310–325.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale M. Lucas, and Jonathan Gratch. 2024. Are llms effective negotiators? systematic evaluation of the multifaceted capabilities of llms in negotiation dialogues. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Miami, FL. Association for Computational Linguistics.

Angela K-Y Leung and Dov Cohen. 2011. Within-and between-culture variation: individual differences and the cultural logics of honor, face, and dignity cultures. *Journal of personality and social psychology*, 100(3):507.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Ting-An Lin and Po-Hsuan Cameron Chen. 2022. Artificial intelligence in a structurally unjust society. *Feminist Philosophy Quarterly*, 8(3/4).

Anthony J. Marsella. 2005. Culture and conflict: Understanding, negotiating, and reconciling conflicting constructions of reality. *International Journal of Intercultural Relations*, 29(6):651–673. Special Issue: Conflict, negotiation, and mediation across cultures: highlights from the fourth biennial conference of the International Academy for Intercultural Research.

Ree M. Meertens and René Lion. 2008. Measuring an individual's tendency to take risks: The risk propensity scale1. *Journal of Applied Social Psychology*, 38(6):1506–1520.

Lisa Messeri and MJ Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

Alaine Murawski, Vanessa Ramirez-Zohfeld, Johnathan Mell, Marianne Tschoe, Allison Schierer, Charles Olvera, Jeanne Brett, Jonathan Gratch, and Lee A Lindquist. 2024. Negotiage: Development and pilot testing of an artificial intelligence-based family caregiver negotiation program. *Journal of the American Geriatrics Society*, 72(4):1112–1121.

Dean G Pruitt. 2007. Conflict escalation in organizations. In *The psychology of conflict and conflict management in organizations*, pages 261–282. Psychology Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Matthew C. Rousu, Gregory Colson, Jay R. Corrigan, Carola Grebitus, and Maria L. Loureiro. 2015. Deception in experiments: Towards guidelines on use in applied economics research. *Applied Economic Perspectives and Policy*, 37(3):524–536.

Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Catherine H Tinsley. 2004. Culture and conflict: Enlarging our dispute resolution framework. *The handbook of negotiation and culture*, pages 193–210.

Gerben A Van Kleef, Carsten KW De Dreu, and Antony SR Manstead. 2004. The interpersonal effects of anger and happiness in negotiations. *Journal of personality and social psychology*, 86(1):57.

Jingjing Yao, Jimena Ramirez-Marin, Jeanne Brett, Soroush Aslani, and Zhaleh Semnani-Azad. 2017. A measurement model for dignity, face, and honor cultural norms. *Management and Organization Review*, 13(4):713–738.

Nutchanon Yongsatianchot, Tobias Thejll-Madsen, and Stacy Marsella. 2024. Exploring theory of mind in large language models through multimodal negotiation. In *Proceedings of 24th ACM International Conference on Intelligent Virtual Agents*, Glascow Scottland. Association for Computational Linguistics.

Qin Zhang, Stella Ting-Toomey, and John G. Oetzel. 2014. Linking Emotion to the Conflict Face-Negotiation Theory: A U.S.-China Investigation of the Mediating Effects of Anger, Compassion, and Guilt in Interpersonal Conflict. *Human Communication Research*, 40(3):373–395.

A GPT4 Emotion Labelling

A.1 Construct Validity

We check whether our GPT emotion labels, compared with T5 (Raffel et al., 2020), accurately capture expressed emotion. We do not have ground truth, however we do have self-reported frustration. Thus, we assess how well each predicted emotion correlates with this self-report. We see GPT-4o outperform T5 in the magnitude and direction of correlation, as seen in Table 3.

Emotion	**T5-Twitter Frustration**	**GPT-4o Frustration**
Anger	0.509	0.553
Fear	-0.010	0.371
Sadness	0.120	0.178
Surprise	-0.001	-0.021
Compassion	-	-0.201
Love	-0.390	-
Joy	-0.435	-0.349

Table 3: Correlation coefficients of self-reported frustration with emotion scores between T5-Twitter and GPT-4o

A.2 Predicting Subjective Outcomes

For a random sample of 406 dialogues, we regressed the SVI scales on the emotion scores. We analyzed how well the model fits the data (R^2) with GPT4 scores compared to T5. Fig. 7 depicts the varying R^2 values across the different configurations. The biggest leap in fit (R^2)comes from using GPT-4o rather than T5, as GPT explains almost half the variance in several measures.

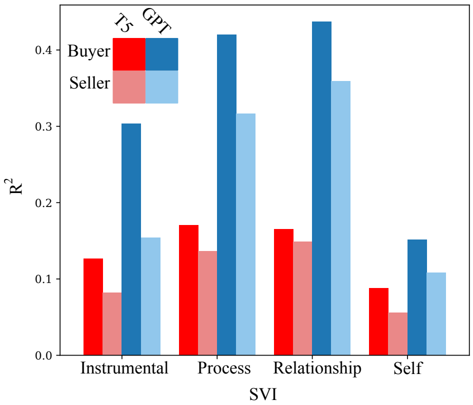


Figure 7: R^2 predicting the four subjective outcomes using GPT or T5 emotion labels

.

A.3 GPT Prompt

Figure 8 outlines the prompt used for GPT4o in the emotion labeling task.

B Task Background

This section of the appendix provides further details about the instructions for the participants, as well as the interface. Figure 9 depicts what the crowd worker would have seen when recruited to the task. Figure 11 shows instructions the participants read before the task. Figure 12 shows the interfaces through which participants would enter pre-task measures, such as their preferences or aspirations.

LLM Emotion Classifier Prompt

You are a good emotion classification tool. Your task is to classify the emotion of the last speaker based on the contextual dialogue.

Your output should be a JSON object with an ‘emotion’ field, categorizing the dialogue with a score for each: joy, anger, fear, sadness, surprise, compassion, or neutral. These scores should sum to one. If an utterance is neutral, then neutral must be one with everything other label set to zero.

Here are a few examples of proper annotations:

```
[label=]{"statement": "Hi ! I 'd like to return my jersey.", "emotion": {"joy": "0", "anger": "0", " fear ": "0", "sadness": "0", "surprise": "0", "compassion": "0", "neutral": "1"}}, {"statement": "Please understand this was for my dear nephew he loves Kobe. I understand we had a misunderstanding, last thing I want is to hurt your business. Let's resolve this together", "emotion": {"joy": "0", "anger": "0", "fear": "0.4", "sadness": "0", "surprise": "0", "compassion": "0.6", "neutral": "0"}}, {"statement": "Thank you!", "emotion": {"joy": "1", "anger": "0", "fear": "0", "sadness": "0", "surprise": "0", "compassion ": "0", " neutral ": "0"}}, {" statement ": "I will report you to authorities for doing this .", " emotion ": {"joy": "0", "anger": "1", "fear ": "0", "sadness ": "0", "surprise": "0", "compassion ": "0", "neutral": "0"}}
```

Figure 8: Prompt used for the LLM emotion annotation task

Figure 9: Prolific recruitment for participants

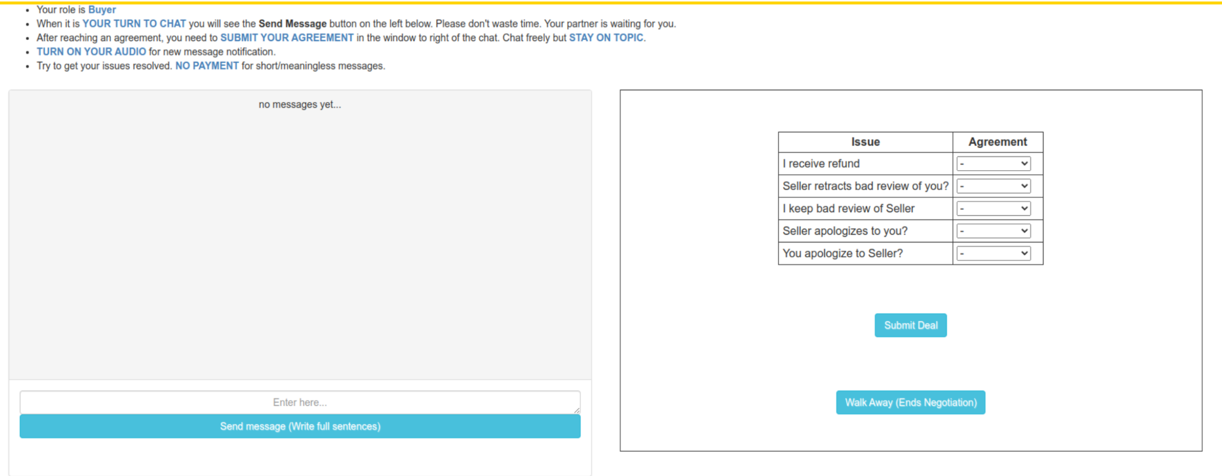
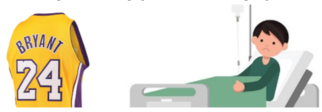


Figure 10: Interface used in the data collection from Lioness Labs

Role Play Instructions: You will play the role of a mistreated buyer in a purchase dispute with another player online. Imagine you are in the following situation:

Kobe Bryant Jersey (You are the Buyer)



Your terminally ill nephew is a huge Kobe Bryant fan so you purchased him a replica of Bryant's last Basketball Championship jersey for $75. The website clearly indicated the purchase was for Bryant's jersey. Other sites were available but this was cheaper and offered quick delivery. When the jersey arrived, it was for a different player you never heard of. You request the correct jersey be sent.

The Seller responds: "The website clearly indicated this was for a Los Angeles Lakers jersey, not for a specific player. All sales are final." You see they now removed mention of Kobe Bryant from their website but you know they are lying. To protect other customers, you post a negative review warning about the Seller's deceptive behavior.

The Seller posted a negative review about you, calling you a "SMARTASS, SLANDERER and a FRAUD." You have dozens of transactions on this site and have a near-spotless reputation. Now you worry others won't sell to you.

(a) Buyer

Role Play Instructions: You will play the role of a mistreated store owner in a purchase dispute with another player online. Imagine you are in the following situation:

LA Lakers Jersey – Seller

You have a successful online sports business and enjoy a near-spotless reputation. You are proud to offer a wide range of official sports merchandise at a fair price. You sell a customer a replica Los Angeles Lakers Basketball Championship jersey for $75. You promptly mail the item.

Two weeks later the customer demands a refund. They claim they purchased a Kobe Bryant jersey but your website clearly states it was not for a specific player. You explained that all sales were final. You saw a review that a previous seller had an issue with this customer and suspect they are trying to extort a lower price.

The Buyer has the audacity to post negative review calling you a "LIAR AND A CROOK." To protect other Sellers from this headache, you post a negative review for this Buyer warning other Sellers of your unpleasant experience.

(b) Seller

Figure 11: Participant instructions for each role-play position

Rank the importance of these issues
People usually can't get all that they want in a dispute. Of the issues discussed above, rank **how important each issue is to <u>you</u>** by assigning points. You have 100 points to distribute across the four issues. More points mean the issue is more important. You must assign all 100 points. (You will be asked about these later)

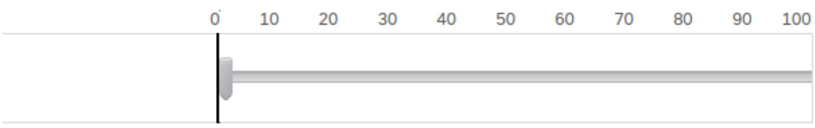
Receive a refund	0
Keep your negative review of Seller	0
Seller removes negative review of you	0
Receive formal apology	0
Total	0

(a) Mechanism for participants to input their preferences

Get ready to negotiate!
On the next page, you will be matched with another player. This may take a few minutes. A sound will play when you are matched. Please ensure your volume is up.

EARN A BIG BONUS: You gain up to ${e://Field/BonusMax} **bonus** based on how well you negotiate. You start with no bonus but gain for each issue you fully achieve. You gain more bonus for achieving your more important issues and all of your bonus if you achieve all of your issues.

With this in mind, what percent of your bonus do you expect to gain?



(b) Slider to measure aspiration of participant

Figure 12: Sliders for participants to input pre-dispute responses