

Emotionally-Aware Agents for Dispute Resolution

Sushrita Rakshit
University of Michigan
Ann Arbor, MI
sushrita@umich.edu

James Hale
University of Southern California
Los Angeles, CA
jahale@usc.edu

Kushal Chawla
Capital One
New York City, NY
kushal.chawla@capitalone.com

Jeanne M. Brett
Northwestern University
Evanston, IL
jmbrett@kellogg.northwestern.edu

Jonathan Gratch
University of Southern California
Los Angeles, CA
gratch@ict.usc.edu

Abstract—In conflict, people use emotional expressions to shape their counterparts’ thoughts, feelings, and actions. This paper explores whether automatic text emotion recognition offers insight into this influence in the context of dispute resolution. Prior work has shown the promise of such methods in negotiations; however, disputes evoke stronger emotions and different social processes. We use a large corpus of buyer-seller dispute dialogues to investigate how emotional expressions shape subjective and objective outcomes. We further demonstrate that large-language models yield considerably greater explanatory power than previous methods for emotion intensity annotation and better match the decisions of human annotators. Findings support existing theoretical models for how emotional expressions contribute to conflict escalation and resolution and suggest that agent-based systems could be useful in managing disputes by recognizing and potentially mitigating emotional escalation.

Index Terms—Dispute, Mediation, Affective Computing, Emotional Modeling, Emotion Recognition, Human-Agent-Interaction

I. INTRODUCTION

Emotional expressions serve essential social functions in human relationships. They convey one’s beliefs, desires, and intentions — shaping the beliefs, desires, and intentions of interaction partners [1], [2]. People high in emotional intelligence achieve more success in navigating emotional relationships [3], and there exists growing interest in creating AI agents that understand and enact these social functions [4], [5]. Prior work suggests that emotionally-aware agents are suitable for a growing list of applications, including teaching people to convey emotions effectively [6], improving human-agent interaction [7], detecting and moderating toxic communication [8], and serving as methodological tools for studying human emotion [9].

This paper examines the capacity of agents to understand human emotional expressions in the context of text-based dispute resolution. Disputes arise when one party in a relationship (an individual, group, or nation) levies a claim that another party refuses to accept, thus threatening the future of the relationship [10]. Though related to negotiation, which the AI community has studied extensively [11]–[16], disputes involve unique social processes [17] and have received less attention. Negotiation (or deal-making) involves coming together to

create a new relationship (i.e., focus on opportunities for gains), and when parties can’t reach a deal, they can seek other partners — e.g., if they cannot deal with one car dealer, one can always visit another. In contrast, disputes involve an existing relationship that has gone badly. As parties are already linked, success depends on managing the costs of ending the relationship rather than on opportunities moving forward. As a result, disputes evoke much stronger emotions than negotiations, particularly anger. Unlike negotiation, where expressions of anger promote compromise [18], [19], anger provokes retaliation in a dispute [20]. Disputants become entrenched in their positions and seek to overpower their opponent through appeals to justice (“You violated my rights!”) or threatening harm (“I will sue you!”), leading to a spiral of escalation, sometimes including threats of physical violence [20]–[22]. Thus, the costs of disputes can greatly exceed the original perceived injury, engulfing not only the disputants but also other related parties and even society at large. Agents who understand emotional dynamics in disputes and how to diffuse destructive behavior could have an enormous societal impact.

Our work makes several contributions to emotion-aware agents. Prior research on emotions in text focuses on a document as the unit of analysis and outputs coarse-grained “sentiment” rather than specific emotions (e.g., *anger*) [23]. In contrast, the dispute literature shows how specific emotions convey particular intentions and examines how these conveyed intentions change *within* a text. This presents a challenge as interpreting each utterance depends on the context of prior utterances [24]. To our knowledge, emotion recognition has not yet been systematically studied in dispute resolution. In negotiations, emotion recognition has shown only weak promise to predict outcomes — e.g., Chawla *et al.* [25] find that recognized *anger* predicted satisfaction in the final agreement, but only explained 5% of the variance. Though two observations suggest we can do much better in disputes. First, as emotions, like *anger*, play a much larger role in disputes than negotiations, we hypothesize that emotion recognition will yield significantly more insights into disputes relative to negotiations. Second, recent research in large language models suggests they are remarkably adept at reasoning about

human emotion, including the ability to incorporate contextual information [26]–[30]. Thus, the current paper assesses the ability of LLMs — specifically GPT4o, but we also survey others — to annotate emotion intensities in dispute dialogues; using these intensities in a linear model to predict disputants’ outcome satisfaction; and yielding insight into how emotional expressions shape these outcomes. To foreshadow our findings, we find that automatically recognized emotional expressions can, in some cases, explain over 40% of the variance in dispute outcomes (compared with about 5% in prior negotiation research). This shows the importance of emotions in disputes and implies the possibility of agent-based methods that understand and influence human disputes — e.g., mediation.

For the current study, we examine a large corpus of 2,025 text conversations where online participants act as buyers or sellers in a simulated purchase dispute. The dispute was crafted in collaboration with a dispute resolution expert (Jeanne M. Brett) to evoke strong emotions while adhering to the ethical guidelines for human experimentation. The corpus was collected using the same framework as Chawla et al.’s [31] CaSiNo — specifically, we created the dialogue collection application using Lioness Labs [32]. This allows participants to match online and engage in a dispute via text chat. Prior analysis of negotiations collected with CaSiNo found that emotional expressions, recognized by a large fine-tuned model (Google’s T5), could predict participants’ satisfaction with their negotiated agreement [25]. Thus, our paper seeks to extend these findings to disputes and contrast T5 with more recent models (out-of-the-box LLMs).

This paper reports our success in predicting the consequences of emotional expressions in disputes. First, after validating our approach against self-reports, we show that automatically recognized expressions predict subjective outcomes — i.e., each participant’s feelings about the dispute’s outcome. Secondly, we illustrate how textual emotion recognition gives insight into how disputes unfold. Our results replicate prior social science findings on the role of *anger* in escalation and lay a foundation for AI agents that could detect escalation and intervene before a dispute derails. Further, while social science findings have emphasized the role of *anger* in disputes, our findings highlight the importance of additional emotional expressions, such as *compassion*. Finally, we survey T5 against popular LLMs predicting subjective outcomes and matching human annotators.

II. RELATED WORK

Disputes and negotiations are task-oriented, mixed-motive interactions. They are task-oriented as each party comes to the situation with specific goals — thus, disputes have concrete measures of success (i.e., “What goals were achieved?”). They are mixed-motive because these goals, while not necessarily zero-sum, typically misalign — i.e., parties can find better solutions if they exchange information about each other’s interests and find solutions to maximize joint gains [33], [34].

AI researchers have long worked on non-collaborative task-oriented interactions, mostly focusing on negotiation framed as

bilateral (occasionally multiparty) multi-issue bargaining [35]: parties are concerned with multiple issues that combine to form an overarching (private) objective function they seek to optimize. Parties reach agreements by exchanging offers and possibly other information. Most research examines agents that negotiate with other agents, but in recent years, there has been growth in agents that negotiate with people. Human-agent negotiation has focused on structured menu-based communication (e.g., [36]), but LLMs have sparked increased interest in text-chat negotiations [37]–[39]. Our work brings AI methods to the study of text-chat human-human disputes.

Though disputes share many elements with negotiations, they are less studied than negotiations, as they involve argumentation over facts; most negotiating agents only exchange offers and do not allow explanations or justifications. An exception is research on argument-based negotiation [40], though that line of work traditionally focuses on agent-agent interaction. Other work on disputes has focused on mediation wherein parties are prevented from interacting with each other but rather communicate to an impartial mediator that proposes fair solutions (e.g., [41]). Neither line of work considers direct interactions between people, nor do they consider the role of emotional expressions. Thus, we focus on emotional expression in human disputes as a contribution of our work.

Expressions of emotion play an important role in shaping negotiations and disputes [42], [43] though relatively little work has used automated methods to investigate emotional expressions, and most of this work has focused on negotiation. Most work uses simple dictionary-based approaches [44], though more recent work takes advantage of powerful transformer-based models [25]. A crucial consideration in emotion research is the selection of labels. Emotion recognition has largely focused on sentiment (positive vs negative) or the six “basic emotion” labels proposed by the emotion psychologist Paul Ekman [45]. Within affective science, there is a growing critique of these basic emotions for recognizing how people truly feel [46]; however, they serve as a common vocabulary for labeling the expressions of others, and much of the social science work on emotion in disputes uses these terms. Thus, this paper adopts basic emotional labels as a starting point to allow direct comparison with prior work, with minor alterations justified later on, while acknowledging that other schemes may be more appropriate, particularly in cross-cultural contexts where common emotion terms may have differing meanings [47].

While our work focuses on recognizing and analyzing emotion in conflict, other research highlights how agents could intervene or assist disputing humans. Cho et al. [48] evaluated conversational AI that intervenes in emotional online disputes — finding prompting models using insights from social science provided specific and fair feedback on toxic behavior. Other studies used AI agents to teach dispute-resolution skills. Murawski *et al.* developed an agent-based framework to teach family caregivers how to manage disputes with insurance companies and service providers through AI role-play [49]. The Rehearsal system extends this approach

Buyer: *I am seeking a refund for the order I received that was different from the item I requested.*

Seller: *Sorry, I sent you what you ordered. I will give you \$50 if you return it. I will drop my bad review if you drop yours, and I want an apology.*

Buyer: *I disagree. I ordered a Kobe Bryant jersey, and that is not what I got. I will happily return the item for a full refund. I will remove the bad review if you remove yours of me, and I will not apologize because I did nothing wrong.*

Seller: *I never offered a Kobe Bryant jersey — I never had one, and that jersey would have cost at least \$3900.*

Buyer: *I have a screenshot of the webpage on your site that I ordered the jersey from, and it clearly states that the item was a Kobe Bryant jersey for \$75.00.*

Seller: *Now you are lying. I am a dealer. I know what that jersey is worth. If you do not agree to my terms, I will update your bad review.*

Buyer: *You are wrong to accuse me of lying. I do have the screenshot and can provide you with a copy if you wish, or I can simply post it alongside my review of your business.*

Seller: *If you do that, I will report you for fraud. On the web, anyone can create untrue posts.*

Buyer: *I do not know what “fraud” is supposed to mean in this context, and apparently you do not intend to do the right thing. I will be forced to report this to the Attorney General and the Consumer Protection Bureau.*

Seller: *I Walk Away.*

Fig. 1: A KODIS dialogue snippet illustrating how an initial focus on interests escalates into appeals to moral norms and ultimately threats of retaliation, resulting in an impasse.

using large-language models (LLMs) to simulate emotion-laden interactions [50]. Rehearsal recognizes when statements likely escalate the conflict and provides personalized feedback on more cooperative dialogue moves — results suggest the system improves students’ dispute resolution abilities. The results of our work may serve to improve such systems.

III. DISPUTE RESOLUTION CORPUS

We analyze KODIS [51], a corpus of human dyadic disputes collected in English on Prolific. Potentially identifiable information from participants was removed. The University of Southern California’s Institutional Review Board approved the experimental design and classified it as minimal risk. Participants gave informed consent and could opt to leave the process at any point without penalty. The entire task, including pre- and post-questionnaires, took approximately twenty minutes. Participants were compensated a base rate of \$3.50 USD (\$10.50 per hour), plus a bonus of up to \$3 USD depending on how well they achieved their objectives.

A. Scenario

The role-playing scenario involves a dispute over purchasing a Kobe Bryant basketball jersey. Before the task, each side (buyer and seller) received unique role-playing instructions designed to provoke a dispute and argumentation over facts. The buyer read that they purchased the jersey for their sick nephew, and when the wrong item arrived, the seller claimed it was never advertised as a Kobe jersey and refused a refund. The seller read that the buyer bought a generic Jersey and now complains about the item. Lastly, each side posted negative reviews about the other, accusing them of deception. In sum, the disputants discuss whether the seller should grant a refund, each side should drop their review, and each side should

apologize. If they cannot agree on a resolution, they can choose to walk away. See Fig. 1 for an example dialogue.

B. Measures

Participants complete several scales assessing their experience during the dispute. This includes a 10-item scale [52] about tactics used during the dispute —e.g., “I expressed frustration,” “The OTHER PARTY expressed frustration.” Here, we only consider the 2-item *frustration* sub-scale, averaging each side of the dyad to quantify the frustration expressed in the dialogue. As subjective perceptions of dispute are a better predictor of future negotiation decisions than the objective result [53], participants completed the Subjective Value Inventory (SVI) scale [54], measuring four dimensions of subjective value. This includes feelings about the instrumental outcome, feelings about the process (*Was it fair?*), feelings about the relationship with the partner, and feelings about the self (*Did I lose face?*).

C. Corpus Overview

The 2,025 disputes¹ consist of 21,681 dialogue turns, averaging 10.7 turns per dispute. Of these, 19% failed to reach a resolution — despite a potentially sizeable monetary bonus. This contrasts with Chawla et al.’s [31] CaSiNo negotiation corpus, where only 3% of negotiations ended in an impasse. Consistent with prior research, this suggests greater impasse rates for disputes relative to negotiations. Further, examining self-reported responses, we find strong evidence that expressed emotion influenced the disputes. For example, although 19% of participants failed to reach an agreement, this climbs to 33% in the dialogues in the top quartile of self-reported frustration and drops to only 6% in the dialogues with the lowest.

¹Gender breakdown of 50% Female, 49% Male, and 1% other.

Sample Utterances	T5-Twitter	GPT4o
I have proof, and you’re not willing to look at it.	Joy	Anger
But are you sure you will refund me completely?	Joy	Fear
What if the shirt has been damaged?	Joy	Fear
If I may return the goods, will I get my money back?? Or I just lost it?	Sadness	Fear
You want to keep the item with no refund?	Anger	Surprise
Huh?	Joy	Surprise
Oh, I’m sorry to hear that, what shirt did you received?	Sadness	Compassion
Hello. I need you to make this right. This is for my nephew that’s sick.	Joy	Fear

TABLE I: Utterances with T5-Twitter and GPT Labels. We portray the top-score associated label from GPT’s intensity vector.

IV. IMPROVING EMOTION RECOGNITION

Though prior work has demonstrated the salience of emotions in shaping disputes, automatically recognizing emotions in an unfolding conversation remains a challenge — e.g., annotating the emotion of each utterance depends on the context of prior utterances [24]. Prior work on negotiation has shown limited success in predicting outcomes from recognized expression, but this work predated the advances afforded by LLMs. In this section, we outline our proposed approach to leverage LLMs as automatic emotion recognition tools; though, first, we overview our selected baseline.

A. Baseline: T5

As a baseline comparison against large language models, we adopt the best-performing approach used by Chawla et al. [25] in their negotiation research — the T5 model finetuned for emotion on the Twitter corpus. This pre-trained model classified utterances as one of *joy*, *anger*, *love*, *sadness*, *fear*, or *surprise*; the model classified each utterance of the dispute in isolation, and created an intensity vector using one-hot encoding. We adopt it for this dispute setting, as other researchers have popularized its use in conflict research — specifically negotiation [25], [55].

One concern relates to the choice of emotion labels provided by T5-Twitter. Its labels arguably improve over Ekman’s more common six basic labels, as T5-Twitter distinguishes positive emotions (*love* vs. *joy*), whereas Ekman [45] only differentiates amongst negative emotions. That said, expressions of *love* are rare in intense disputes. Instead, the conflict literature suggests that expressions of *compassion* are a powerful predictor of negotiated outcomes [56], [57]. This indicates that an alternative labeling scheme might improve results, an adaptation that general LLMs easily allow.

B. LLM-Based Emotion Recognition

We propose an alternative approach utilizing LLMs and explore several prompting strategies. In preliminary efforts, we attempted to apply annotation methods developed for deal-making dialogues in this dispute setting, though these underperformed, highlighting the need to treat disputes as a distinct genre. We limit our initial reported analysis to OpenAI’s GPT

series of models [58] (GPT4o, run on 06/28/2024) as a recent extensive evaluation found GPT-4 yielded the best performance compared with a wide range of other LLMs on making inferences from negotiation dialogues [55]. Later, we survey additional LLMs to gauge relative performance. Motivated by prior findings [59], we assign a specific role in the prompt (“You are a good emotion classification tool”), and prompt the model to output instructions in JSON output format. Further, we leverage the following prompting strategies:

1) *Incorporating Dialogue Context*: Given that emotions of each dialogue turn depend on prior turns [24], we include dialogue history when prompting the LLM to generate emotion annotations.

2) *Labeling Scheme*: To address concerns with T5-Twitter’s labels, we substituted T5-Twitter’s *love* with *compassion* and added a *neutral* label to allow annotation of emotionless utterances. We prompt the LLM to generate soft labels for each utterance — i.e., it allocates weight ($\in [0, 1]$) over *joy*, *anger*, *fear*, *surprise*, *compassion*, *sadness*, and *neutral* labels such that the resulting vector sums to one.

3) *In-context Learning*: Recent research suggests performance can be improved via “in-context learning” [60]. Thus, we include in the prompt several dialogue turns from the corpus — hand-annotated with emotion labels by the researchers.

V. ANALYZING THE DISPUTE DIALOGUES

We examine whether automatic emotion recognition approaches can offer insights into the corpus in three steps. First, we contrast how well T5-Twitter and our proposed LLM prompts recognize expressed emotion compared against disputants’ self-reports. Second, we assess how well these automatic labels can predict disputants’ subjective feelings about the outcome of the dispute. Third, we examine how emotions unfold during the dispute, and if differences in emotional expression can indicate outcomes in advance.

A. Emotion Recognition Performance Validity

We first analyze how well GPT and T5’s labels — see TABLE I for comparative annotations — match disputants’ actual emotions ($N = 1,938$). Without the self-reports of each emotion, we assess how well T5 and GPT4o’s emotion

intensities², averaged over the entire dialogue, correlate with self-reported *frustration*. GPT4o outperforms T5-Twitter at capturing the magnitude and direction of the correlation coefficients — likely because T5-Twitter fails to annotate instances of *fear* and *sadness* accurately, giving more consideration to *joy* and *anger*. Examining the Pearson correlation coefficients in Table II, we see that *fear*, as annotated by GPT4o, yields a coefficient of 0.360; while T5-Twitter’s, however, stands at a low 0.026. Notably, we see a much stronger correlation between GPT’s *anger* annotations and *frustration* compared to T5. As T5-Twitter cannot capture a variety of labels — such as *fear* and *sadness* — GPT4o better captures nuanced emotional trends that match self-reports.

Emotion	T5-Twitter w/ Frustration	GPT4o w/ Frustration
Anger	0.294	0.544
Fear	0.026	0.360
Sadness	0.053	0.162
Surprise	0.013	-0.010
Compassion	-	-0.181
Love	-0.024	-
Joy	-0.326	-0.357

TABLE II: Pearson correlation coefficients of emotion labels from T5-Twitter and GPT4o with self-reported *frustration*.

To explain this, we explore differences in mean emotion broken out by role between GPT and T5-Twitter. Fig. 2 shows that T5 skews towards annotating utterances as *joy* or *anger*, while GPT assigns more diverse labels and utilizes *neutral* as a dampener. GPT suggests dialogues contain far more *anger* than *joy*, especially for *buyers*. Interestingly, GPT also recognizes more *compassion* relative to T5’s *love*.

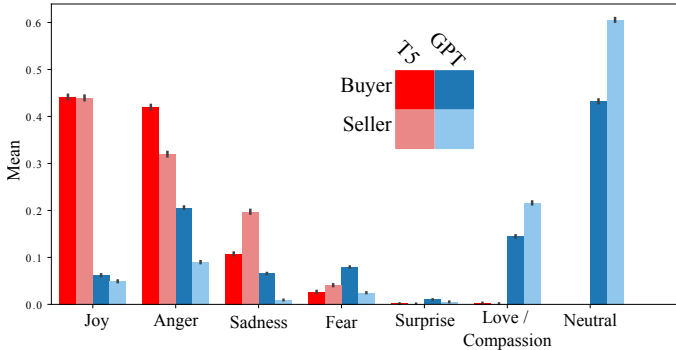


Fig. 2: Comparison of mean intensity of recognized emotions by GPT4o and T5-Twitter separated by role.

B. Predicting Subjective Outcomes

We next predict each side’s subjective feelings about the dispute’s outcome — as measured by the SVI scale (Instrumental Outcome, Process Fairness, Relationship, and Self) [54] — from the automatically recognized emotions during the dispute. To justify our prompting approach, we first contrast T5-Twitter and the various prompting techniques — ablating

²As T5 selects one emotion, we convert that to one-hot encoding.

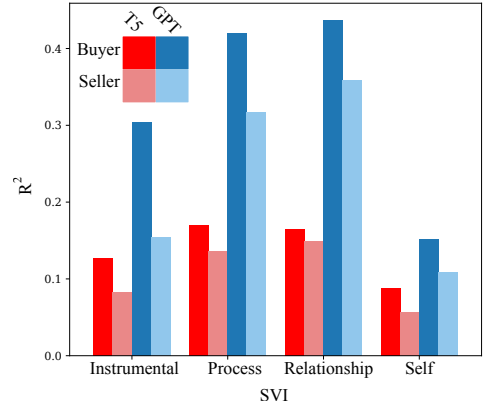


Fig. 3: R^2 values for various regressions measuring the efficacy of the emotion labels to predict SVI scores.

or modifying various aspects of the prompt — to see which method’s annotations better explain the subjective outcomes. We perform these tests on a 20% subset ($N = 406$) of the corpus. We use multiple linear regression (MLR) to predict subjective feelings for each role, regressing each of the four SVI sub-scales on the average emotion intensities from a given dialogue. The coefficient of determination (R^2) measures model fit — i.e., how much of the variance the emotion labels explain of the outcome.

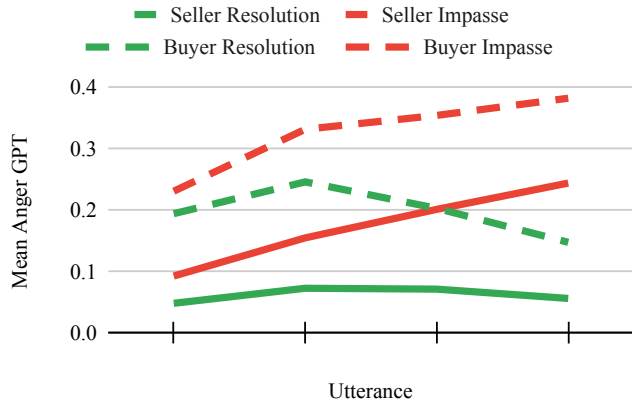
Model	IC Learning	w/ Compassion	Dialogue History	w/ Neutral	R^2
T5	×	×	×	×	0.112
GPT4o	✓	×	×	×	0.250
GPT4o	✓	✓	×	✓	0.258
GPT4o	✓	✓	✓	✓	0.267

TABLE III: Mean R^2 for predicting SVI from emotion labels for different schemes.

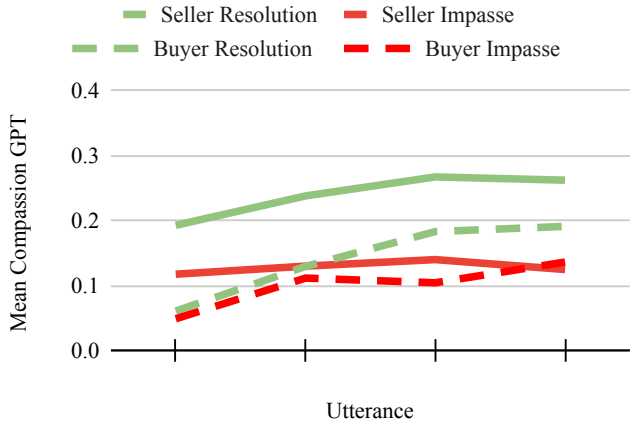
We compare T5 with different combinations of prompt enhancements for GPT4o, considering the average model fit (R^2) across all four SVI sub-scales — see TABLE III. This shows that simply replacing T5 with GPT4o yields the largest improvement. Additional modest improvements come from adopting each of our proposed prompting methods from Section IV-B. Fig. 3 compares this highest-performing GPT configuration with T5 when breaking out the SVI sub-scales, and separating predictions by role in the dispute. This shows *buyers* are more straightforward to predict: GPT explains almost half the variance in outcome about *buyers’* feelings about the *process* and *relationship*. Now that we have justified our approach, we can begin to analyze the emotional dynamics in KODIS.

C. Exploring Escalatory Spirals

We find that outcomes can be predicted quite well solely from emotions expressed throughout the dispute — in some



(a) Anger over time.



(b) Compassion over time.

Fig. 4: Average expressed *anger* and compassion across dialogue turns broken out by role and objective outcome.

cases explaining almost half of the variance. Still, if we want agents to intervene and steer the conversation away from an impasse, they must understand how expressions unfold over time. Prior research on disputes suggests that disputants often reach an impasse because anger provokes anger, leading to escalatory spirals and hardened positions [20]. Thus, we examine whether automatic emotion recognition reveals this pattern.

We examine how disputants express emotion turn-by-turn — as annotated by GPT4o — and how each side responds to the other. We do this separately for disputes that end in resolution or impasse. We focus on *anger*, given its recognized role in disputes, but also compassion, as some research highlights its role in shaping outcomes [56]. Fig. 4 illustrates the results. At the beginning of the dispute, buyers express more anger than sellers, regardless of whether the conversation ends in resolution or impasse. This may reflect that participants internalize typical customer service conversational norms where sellers are expected to express positive emotions in conflict situations [61], [62]. Emotion trajectories begin to diverge almost immediately after the first round. For disputes that end in an impasse (the red lines), sellers reciprocate the buyer’s anger, leading to an escalatory spiral. For disputes ending in resolution (the green lines), sellers apparently “stick to the script” and resist reciprocating the buyer’s anger. This seems to defuse the buyers’ initial anger, and expressions of anger become less frequent as the conversation continues.

Interestingly, the pattern of compassion also highlights the differences between success and failure. For disputes ending in an impasse, sellers begin with little compassion. In contrast, sellers who reach a resolution begin the dispute by expressing more *compassion* for the situation, and buyers reciprocate those expressions. Thus, our automatic analysis highlights a pathway to dispute resolution that has been under-emphasized in the dispute literature. Together, these patterns suggest that early differences in expressed emotion can build and shape the

outcome. This gives hope that algorithms can be developed to recognize these patterns early and intervene.

VI. BENCHMARKING

Our analysis focused on GPT4o due to its growing usage in research. We extend our analysis from GPT4o to other current generative language models to compare against human annotators. Here, we consider a subset of $N = 400$ dialogues.

A. Additional LLMs

We analyze the following additional LLMs:

- **Deepseek V3:** DeepSeek released a 671B parameter model in January 2025, which outperforms GPT4o at some tasks [63].
- **Llama3:** Meta released this open-source 405B parameter model in November 2024 [64].
- **GPT4o-mini:** OpenAI released this smaller model in July 2024 — it performs competitively at many tasks compared to larger, more expensive models [65].

B. Ground Truth

Previously, we looked at correlations between emotion predictions and self-reported frustration. We expand on this by collecting third-party human emotion intensity annotations ($N = 336$) from Prolific (US annotators) on 100 randomly sampled utterances to evaluate T5 and the aforementioned LLMs — each sampled utterance received at least two annotations from different annotators. The annotators’ task resembled that of the models in that they read a highlighted utterance, read the dialogue history, and allocated weight over the considered emotions (*anger*, *joy*, *sadness*, *fear*, *compassion*, and *neutral*) such that they summed to one. A correlational analysis allows us to gauge how well each model matches the average human. Fig. 5 illustrates these correlations. Generally, T5 (we map its *love* predictions to *compassion*) performs relatively poorly against the LLMs.

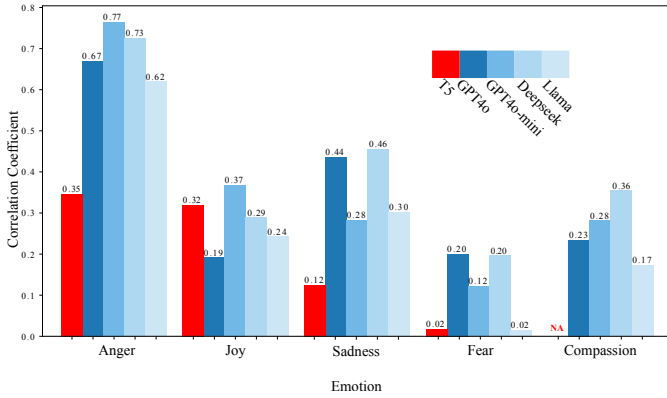


Fig. 5: Emotion intensity correlations with human annotations — LLMs better match human annotations.

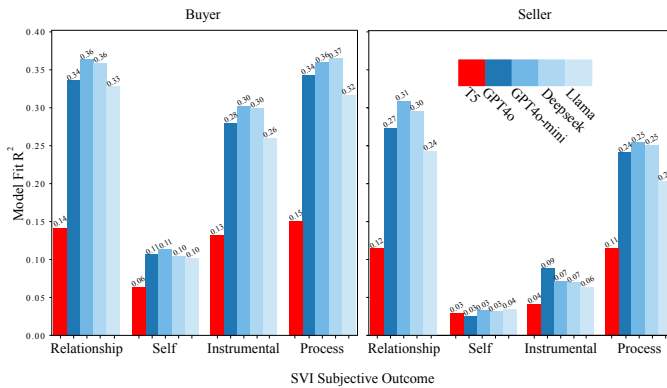


Fig. 6: Fit of emotion labels predicting subjective outcome — LLMs explain more variance in subjective outcomes.

C. Predicting Subjective Outcome

As we did in Section V-B, we probe these models’ abilities to generate emotion labels useful in downstream tasks, such as predicting subjective outcomes. Fig. 6 demonstrates these models’ differing abilities to generate emotion labels useful in this task. Again, we see T5 performs notably worse than the four LLMs at creating useful annotations — further, the LLM annotations hold quite a bit of power, in some cases explaining about 35% of the variance from emotion alone.

VII. DISCUSSION

This study analyzed emotional expressions in a large corpus of buyer-seller disputes. We find LLMs are quite effective at recognizing emotional expressions in dispute dialogues, and far more effective than T5 models used in recent studies. Remarkably, we find that quite a bit of the variance in the subjective outcome of a dispute can be explained from emotional expressions alone, ignoring the actual content of the dialogue. For example, simple regression models trained on automatic emotion labels explain, in some cases, thirty to forty percent of the variance in how participants viewed the process and their relationship with their partner.

Analysis also reveals how emotions shape disputes over time. Fig. 4a illustrates how disputes can escalate into failure if

one partner reciprocates anger with anger — i.e., an escalatory spiral. Buyers begin the dispute, expressing far more anger than their partners. Both sides become increasingly angry when sellers respond angrily, and the dispute ends in an impasse. When sellers maintain their calm, the anger of buyers dissipates, and disputants tend to resolve the dispute. This reinforces findings on escalatory spirals in the social science literature [20]. Interestingly, we also find evidence that spirals of compassion might reverse these effects (see Fig. 4b). Together, our findings suggest agents could intervene early to encourage participants to avoid costly escalation (see [49], [66]).

Our analysis provides prompt engineering insight that might improve text emotion recognition. As highlighted in Table III, simply moving from T5 to GPT substantially increased predictive power. However, the results draw attention to other factors that improved the results. For example, including a few labeled examples (in-context learning) aided prediction. Including prior dialogue history also leads to greater performance. Finally, the original T5 model did not allow utterances to be labeled as *neutral*, and it used *love* rather than *compassion* — changing these labels also enhanced performance.

Lastly, we benchmarked several LLMs against human annotations and assessed how well their predictions support modeling subjective outcomes. The four LLMs generally outperform T5 and exhibit varying relative performance across different tasks — e.g., correlating with human ground truth and explaining subjective outcomes.

VIII. CONCLUSION & FUTURE WORK

There are clear individual and societal benefits to agents understanding and helping resolve emotional disputes. Unresolved disputes can promote social divisions, erode trust in groups and institutions, and cause mental and emotional distress to the parties involved. Future work may wish to analyze how gender and culture interact to shape automatically recognized emotion. However, in our results, we do not find much evidence of interactions between those demographic factors, as one might expect if there existed a systemic issue of misclassification. Nonetheless, this remains a salient question going forward, especially as we work toward our ultimate goal of AI managing human disputes.

If AI-driven systems could recognize the emotional dynamics of how disputes arise and escalate, they have the potential to manage these negative costs. Ultimately, we will leverage the findings here to create agents that can effectively mediate human disputes. We previously published preliminary findings on the abilities of LLMs in mediation [67] — finding the model picked up on salient dialogue features to intervene more so in disputes resulting in impasse, and in those with high frustration. Further, we find crowd-source workers rated AI-generated interventions higher compared to novices. Additionally, we will move beyond *just* emotion and explore methods of annotating the semantic content of conflict dialogues, following pertinent research [55], [68].

IX. ETHICAL IMPACT STATEMENT

Several ethical concerns must be considered as this work proceeds. Prior work has suggested that LLMs internalize societal and cultural stereotypes, and any attempts to deploy such technology must carefully consider the risks of misrepresenting or misinterpreting perceived emotion. Quite possibly, certain groups might be labeled as “angry” and treated unfairly by automated methods. This is particularly a concern in cross-cultural contexts where emotions might hold different connotations and serve different functions [47]. Finally, though we focused on dispute *resolution*, our findings could inform techniques that intentionally escalate conflicts. This emphasizes the importance of increased research in conflict processes and techniques, how they might be deployed, and their consequence for society.

Secondly, we acknowledge the limited purview of our work and the uncertainty of whether these results could extend outside the realm of dispute resolution. However, we wish to reiterate that *this* kind of generalizability lies outside the scope of this work; rather, we intended to demonstrate the feasibility of using emotion annotations from LLMs in the understudied domain of dispute resolution. Here, emotion comes to the forefront compared to the more commonly adopted — at least in computer and affective science — task of deal-making. We find, in some cases, linear models can explain almost 40% of the variance in subjective outcome considering just these emotion labels, supporting the importance of affect in this domain.

This work was approved by the University of Southern California’s institutional review board (UP-24-01183-AM001).

ACKNOWLEDGMENT

This work is supported by the U.S. Government including the Air Force Office of Scientific Research (grant FA9550-23-1-0320), the Army Research Office under agreement W911NF-25-2-0040, and the National Science Foundation (grant 2150187). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

- [1] B. Parkinson, A. H. Fischer, and A. S. Manstead, *Emotion in social relations: Cultural, group, and interpersonal processes*. Psychology press, 2005.
- [2] A. Scarantino, “How to do things with emotional expressions: The theory of affective pragmatics,” *Psychological Inquiry*, vol. 28, no. 2-3, pp. 165–185, 2017.
- [3] P. J. Jordan and A. C. Troth, “Managing emotions during team problem solving: Emotional intelligence and conflict resolution,” in *Emotion and performance*. CRC Press, 2021, pp. 195–218.
- [4] C. M. de Melo, J. Gratch, S. Marsella, and C. Pelachaud, “Social functions of machine emotional expressions,” *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1382–1397, 2023.
- [5] M. F. Jung, “Affective grounding in human-robot interaction,” in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 263–273.
- [6] A. Foster, N. Chaudhary, T. Kim, J. L. Waller, J. Wong, M. Borish, A. Cordar, B. Lok, and P. F. Buckley, “Using virtual patients to teach empathy: a randomized controlled study to enhance medical students’ empathic communication,” *Simulation in Healthcare*, vol. 11, no. 3, pp. 181–189, 2016.
- [7] R. Stock-Homburg, “Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research,” *International Journal of Social Robotics*, vol. 14, no. 2, pp. 389–411, 2022.
- [8] A. G. d’Sa, I. Illina, and D. Fohr, “Bert and fasttext embeddings for automatic detection of toxic speech,” in *2020 International Multi-Conference on: “Organization of Knowledge and Advanced Technologies”(OCTA)*. IEEE, 2020, pp. 1–5.
- [9] B. Kennedy, P. Golazizian, J. Trager, M. Atari, J. Hoover, A. Mostafazadeh Davani, and M. Dehghani, “The (moral) language of hate,” *PNAS nexus*, vol. 2, no. 7, p. pgad210, 2023.
- [10] W. L. Felstiner, R. L. Abel, and A. Sarat, “The emergence and transformation of disputes: Naming, blaming, claiming. . . .” in *Theoretical and Empirical Studies of Rights*. Routledge, 2017, pp. 255–306.
- [11] T. Baarslag, M. Kaisers, E. Gerding, C. M. Jonker, and J. Gratch, “When will negotiation agents be able to represent us? the challenges and opportunities for autonomous negotiators.” International Joint Conferences on Artificial Intelligence, 2017.
- [12] P. Faratin, C. Sierra, and N. R. Jennings, “Negotiation decision functions for autonomous agents,” *Robotics and Autonomous Systems*, vol. 24, no. 3-4, pp. 159–182, 1998.
- [13] S. Kraus, “Negotiation and cooperation in multi-agent environments,” *Artificial intelligence*, vol. 94, no. 1-2, pp. 79–97, 1997.
- [14] C. M. Jonker, K. V. Hindriks, P. Wiggers, and J. Broekens, “Negotiating agents,” *AI Magazine*, vol. 33, no. 3, pp. 79–79, 2012.
- [15] R. Aydoğan, T. Baarslag, K. Fujita, J. Mell, J. Gratch, D. De Jonge, Y. Mohammad, S. Nakadai, S. Morinaga, H. Osawa *et al.*, “Challenges and main results of the automated negotiating agents competition (anac) 2019,” in *Multi-Agent Systems and Agreement Technologies: 17th European Conference, EUMAS 2020, and 7th International Conference, AT 2020, Thessaloniki, Greece, September 14-15, 2020, Revised Selected Papers 17*. Springer, 2020, pp. 366–381.
- [16] J. Gratch, D. DeVault, G. M. Lucas, and S. Marsella, “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*. Springer, 2015, pp. 201–215.
- [17] J. M. Brett, “Negotiating globally: How to negotiate deals, resolve disputes, and make decisions across cultural boundaries,” 2014.
- [18] G. A. Van Kleef, C. K. De Dreu, and A. S. Manstead, “The interpersonal effects of anger and happiness in negotiations,” *Journal of personality and social psychology*, vol. 86, no. 1, p. 57, 2004.
- [19] C. M. de Melo, P. Carnevale, and J. Gratch, “The effect of expression of anger and happiness in computer agents on negotiations with humans,” in *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, 2011, pp. 937–944.
- [20] D. G. Pruitt, “Conflict escalation in organizations,” in *The psychology of conflict and conflict management in organizations*. Psychology Press, 2007, pp. 261–282.
- [21] J. M. Brett, D. L. Shapiro, and A. L. Lytle, “Breaking the bonds of reciprocity in negotiations,” *Academy of Management Journal*, vol. 41, no. 4, pp. 410–424, 1998.
- [22] E. Halperin, “Group-based hatred in intractable conflict in israel,” *Journal of Conflict resolution*, vol. 52, no. 5, pp. 713–736, 2008.
- [23] M. Wankhade, A. C. S. Rao, and C. Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.
- [24] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE access*, vol. 7, pp. 100943–100953, 2019.
- [25] K. Chawla, R. Clever, J. Ramirez, G. M. Lucas, and J. Gratch, “Towards emotion-aware agents for improved user satisfaction and partner perception in negotiation dialogues,” *IEEE Transactions on Affective Computing*, 2023.
- [26] A. N. Tak and J. Gratch, “Is gpt a computational model of emotion?” in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2023, pp. 1–8.
- [27] N. Yongsatianchot, P. G. Torshizi, and S. Marsella, “Investigating large language models’ perception of emotion using appraisal theory,” in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2023, pp. 1–8.

- [28] H. Zhan, D. C. Ong, and J. J. Li, "Evaluating subjective cognitive appraisals of emotions from large language models," *arXiv preprint arXiv:2310.14389*, 2023.
- [29] J. Broekens, B. Hilpert, S. Verberne, K. Baraka, P. Gebhard, and A. Plaat, "Fine-grained affective processing capabilities emerging from large language models," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.
- [30] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu, "Emotional intelligence of large language models," *Journal of Pacific Rim Psychology*, vol. 17, p. 18344909231213958, 2023.
- [31] K. Chawla, J. Ramirez, R. Clever, G. Lucas, J. May, and J. Gratch, "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*, 2021, pp. 3167–3185.
- [32] M. Giamattei, K. S. Yahosseini, S. Gächter, and L. Molleman, "Lioness lab: a free web-based platform for conducting interactive experiments online," *Journal of the Economic Science Association*, vol. 6, no. 1, pp. 95–111, 2020.
- [33] L. L. Thompson, "Information exchange in negotiation," *Journal of experimental social psychology*, vol. 27, no. 2, pp. 161–179, 1991.
- [34] T. Baarslag, M. J. Hendriks, K. V. Hindriks, and C. M. Jonker, "Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques," *Autonomous Agents and Multi-Agent Systems*, vol. 30, pp. 849–898, 2016.
- [35] S. S. Fatima, M. Wooldridge, and N. R. Jennings, "An agenda-based framework for multi-issue negotiation," *Artificial Intelligence*, vol. 152, no. 1, pp. 1–45, 2004.
- [36] J. Mell and J. Gratch, "Iago: interactive arbitration guide online," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 2016, pp. 1510–1512.
- [37] M. Lewis, D. Yarats, Y. Dauphin, D. Parikh, and D. Batra, "Deal or no deal? end-to-end learning of negotiation dialogues," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2443–2453.
- [38] H. He, D. Chen, A. Balakrishnan, and P. Liang, "Decoupling strategy and generation in negotiation dialogues," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2333–2343. [Online]. Available: <https://aclanthology.org/D18-1256>
- [39] K. Chawla, I. Wu, Y. Rong, G. Lucas, and J. Gratch, "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*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 078–13 092. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.808>
- [40] S. Parsons, C. Sierra, and N. Jennings, "Agents that reason and negotiate by arguing," 1998.
- [41] J. Goldman and A. D. Procaccia, "Spliddit: Unleashing fair division algorithms," *ACM SIGecom Exchanges*, vol. 13, no. 2, pp. 41–46, 2015.
- [42] M. W. Morris and D. Keltner, "How emotions work: The social functions of emotional expression in negotiations," *Research in organizational behavior*, vol. 22, pp. 1–50, 2000.
- [43] K. Douglas and C. Coburn, "Attitude and response to emotion in dispute resolution: The experience of mediators," *Flinders LJ*, vol. 16, p. 111, 2014.
- [44] C. Laubert and J. Parlamis, "Are you angry (happy, sad) or aren't you? emotion detection difficulty in email negotiation," *Group Decision and Negotiation*, vol. 28, pp. 377–413, 2019.
- [45] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
- [46] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.
- [47] S. Havaladar, S. Rai, B. Singhal, L. Liu, S. C. Guntuku, and L. Ungar, "Multilingual language models are not multicultural: A case study in emotion," *arXiv preprint arXiv:2307.01370*, 2023.
- [48] H. Cho, S. Liu, T. Shi, D. Jain, B. Rizk, Y. Huang, Z. Lu, N. Wen, J. Gratch, E. Ferrara, and J. May, "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)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 7478–7496. [Online]. Available: <https://aclanthology.org/2024.naacl-long.415>
- [49] A. Murawski, V. Ramirez-Zohfeld, J. Mell, M. Tschoe, A. Schierer, C. Olvera, J. Brett, J. Gratch, and L. A. Lindquist, "Negotiate: Development and pilot testing of an artificial intelligence-based family caregiver negotiation program," *Journal of the American Geriatrics Society*, vol. 72, no. 4, pp. 1112–1121, 2024.
- [50] O. Shaikh, V. E. Chai, M. Gelfand, D. Yang, and M. S. Bernstein, "Rehearsal: Simulating conflict to teach conflict resolution," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–20.
- [51] J. A. Hale, S. Rakshit, K. Chawla, J. M. Brett, and J. Gratch, "Kodis: A multicultural dispute resolution dialogue corpus," 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)*, 2025, pp. 12 771–12 785.
- [52] S. Aslani, J. Ramirez-Marin, J. Brett, J. Yao, Z. Semnani-Azad, Z.-X. Zhang, C. Tinsley, L. Weingart, and W. Adair, "Dignity, face, and honor cultures: A study of negotiation strategy and outcomes in three cultures," *Journal of Organizational Behavior*, vol. 37, no. 8, pp. 1178–1201, 2016.
- [53] A. D. Brown and J. R. Curhan, "The utility of relationships in negotiation," *The Oxford handbook of economic conflict resolution*, pp. 137–154, 2012.
- [54] J. R. Curhan, H. A. Elfenbein, and H. Xu, "What do people value when they negotiate? mapping the domain of subjective value in negotiation," *Journal of personality and social psychology*, vol. 91, no. 3, p. 493, 2006.
- [55] D. Kwon, E. Weiss, T. Kulshrestha, K. Chawla, G. M. Lucas, and J. Gratch, "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, Nov. 2024.
- [56] K. G. Allred, J. S. Mallozzi, F. Matsui, and C. P. Raia, "The influence of anger and compassion on negotiation performance," *Organizational Behavior and Human Decision Processes*, vol. 70, no. 3, pp. 175–187, 1997. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0749597897927059>
- [57] Q. Zhang, S. Ting-Toomey, and J. G. Oetzel, "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*, vol. 40, no. 3, pp. 373–395, 07 2014. [Online]. Available: <https://doi.org/10.1111/hcre.12029>
- [58] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [59] M. Zheng, J. Pei, and D. Jurgens, "Is" a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts," *arXiv preprint arXiv:2311.10054*, 2023.
- [60] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.
- [61] A. R. Hochschild, *The managed heart: Commercialization of human feeling*. Univ of California Press, 2003.
- [62] C. Tang, C. R. Seal, S. E. Naumann, and K. Miguel, "Emotional labor: The role of employee acting strategies on customer emotional experience and subsequent buying decisions," *International Review of Management and Marketing*, vol. 3, no. 2, pp. 50–57, 2013.
- [63] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.
- [64] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
- [65] OpenAI, "Gpt-4o mini: Advancing cost-efficient intelligence," 2024. [Online]. Available: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>
- [66] O. Shaikh, V. E. Chai, M. Gelfand, D. Yang, and M. S. Bernstein, "Rehearsal: Simulating conflict to teach conflict resolution," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA:

Association for Computing Machinery, 2024. [Online]. Available: <https://doi.org/10.1145/3613904.3642159>

- [67] J. Hale, H. Kim, A. Choi, and J. Gratch, "Ai-mediated dispute resolution," in *Proceedings of the AAAI Symposium Series*, vol. 5, no. 1, 2025, pp. 67–70.
- [68] R. Friedman, J. Cho, J. Brett, X. Zhan, N. Han, S. Kannan, Y. Ma, J. Spencer-Smith, E. Jäckel, A. Zerres *et al.*, "An application of large language models to coding negotiation transcripts," *arXiv preprint arXiv:2407.21037*, 2024.