Evaluating Behavioral Alignment in Conflict Dialogue: A Multi-Dimensional Comparison of LLM Agents and Humans

Deuksin Kwon[1,2] **Kaleen Shrestha**[1] **Bin Han**[1,2]
Elena Hayoung Lee[1] **Gale M. Lucas**[1,2]
[1]University of Southern California [2]USC for Institute of Creative Technologies
{deuksink, kshresth, binhan}@usc.edu
elena.lee@marshall.usc.edu
lucas@ict.usc.edu

Abstract

Large Language Models (LLMs) are increasingly deployed in socially complex, interaction-driven tasks, yet their ability to mirror human behavior in emotionally and strategically complex contexts remains underexplored. This study assesses the behavioral alignment of personality-prompted LLMs in adversarial dispute resolution by simulating multi-turn conflict dialogues that incorporate negotiation. Each LLM is guided by a matched Five-Factor personality profile to control for individual variation and enhance realism. We evaluate alignment across three dimensions: linguistic style, emotional expression (e.g., anger dynamics), and strategic behavior. GPT-4.1 achieves the closest alignment with humans in linguistic style and emotional dynamics, while Claude-3.7-Sonnet best reflects strategic behavior. Nonetheless, substantial alignment gaps persist. Our findings establish a benchmark for alignment between LLMs and humans in socially complex interactions, underscoring both the promise and the limitations of personality conditioning in dialogue modeling.

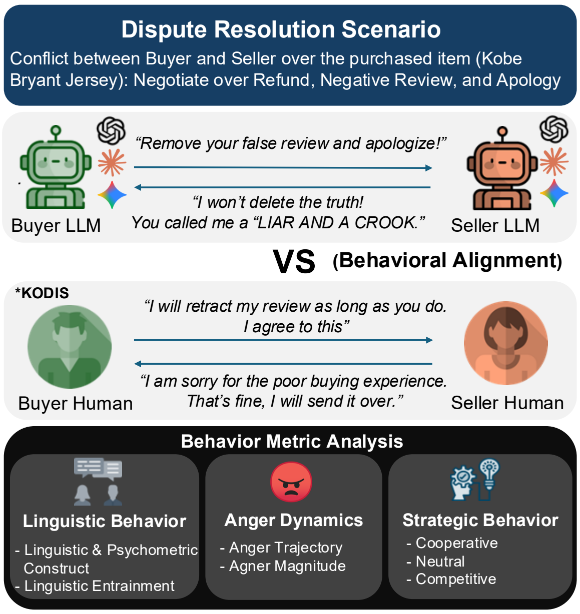


Figure 1: Overview of a dispute resolution scenario examining the behavioral alignment between LLM agents and humans, analyzed across linguistic behavior, anger dynamics, and strategic behavior.

1 Introduction

Large Language Models (LLMs) are increasingly used to simulate human behavior in socially grounded, interactive tasks. Recent efforts enhance human-likeness by assigning personas or personality traits to LLM agents (Serapio-García et al., 2023; Jiang et al., 2023). However, it remains unclear whether personality-prompted LLMs can replicate the behavioral dynamics of real humans, especially in emotionally charged and strategically complex contexts such as conflict resolution.

Dispute resolution involves negotiating interpersonal conflicts and requires emotional expression, strategic adaptation, and relationship management (Brett, 2007). Unlike cooperative bargaining, it includes blame, justification, and tension regulation. While LLMs have been studied in negotiation settings (Abdelnabi et al., 2023; Bianchi et al., 2024), few works assess whether their behavior aligns with humans in adversarial, emotionally charged interactions.

We address this gap through a comparative study using multi-turn, multi-issue human negotiation dialogues (Anthony Hale et al., 2025), simulating LLM agents prompted with matched Five-Factor Model personality profiles (McCrae and John, 1992; Abdurahman et al., 2024). To enable systematic comparison, we introduce a structured evaluation framework with tailored metrics that capture alignment across three behavioral dimensions: (1) Linguistic Style (LIWC, nClid) (Ireland and Henderson, 2014), (2) Emotional Dynamics (turn-level anger trajectories), and (3) Strategic Behavior (*Interests-Rights-Power* (IRP) transi-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16366–16380

tions) (Ury et al., 1988). These metrics enable rigorous, multi-dimensional comparisons that reveal both surface-level patterns and deeper behavioral differences between human and agent interactions. Figure 1 provides an overview of our experimental framework and the dispute resolution scenario.

Results show that GPT-4.1 aligns most closely with humans in language and emotional expression, while Claude-3.7-Sonnet better reflects strategic negotiation behavior. Despite this progress, both models display meaningful deviations, highlighting the limits of current LLMs in modeling socially and emotionally nuanced interaction. Our framework offers a replicable basis for benchmarking LLM-human alignment in complex dialogue. Our main contributions are as follows:

- We present a structured, behaviorally grounded evaluation of personality-prompted LLMs in adversarial negotiation, using a multi-perspective framework to compare their linguistic, emotional, and strategic behaviors with human data.
- We identify both the strengths and systematic behavioral gaps of state-of-the-art LLMs, including GPT-4.1 and Claude-3.7-Sonnet.
- We offer critical insight into the fidelity of LLMs' social behavior under personality conditioning, highlighting its promise and limitations in agent-based simulations.

2 Related Works

LLMs have been increasingly used for agent-based simulations, showing promise in role-playing, personality conditioning, and goal-oriented dialogue (Li et al., 2024; Kwon et al., 2025; Gao et al., 2024; Glória-Silva et al., 2024; Jiang et al., 2022). Prior work has examined whether LLMs can emulate human traits such as trust, personality, or emotion-driven behavior, but mainly in single-turn settings (Wang et al., 2024, 2025; Xie et al., 2024). These studies often lack multi-turn interaction and overlook the evolving nature of behavior (Wang et al., 2024). Even in dyadic setups, comparisons typically focus on outcomes rather than on coordination during the dialogue (Xie et al., 2024).

Despite the central role of negotiation in human communication, prior work has not directly compared LLM and human dyads in extended, emotionally complex interactions (Gandhi et al., 2023; Noh and Chang, 2024; Huang and Hadfi, 2024). This gap is especially salient in dispute resolution, where negotiation occurs amid pre-existing conflict and heightened emotions such as anger. Unlike cooperative settings, dispute resolution requires both strategic reasoning and emotional sensitivity (Brett, 2007).

To fill this gap, we present the first systematic comparison of LLMs and humans in multi-turn, adversarial dialogue. Using the socially and emotionally rich context of dispute resolution, we evaluate agent-human alignment across three key dimensions: linguistic style, emotional dynamics, and strategic behavior.

3 Methods

3.1 Human Dataset

The *KObe DISpute corpus* (KODIS) is a human–human (H2H) dispute resolution dataset consisting of extended English role-play dialogues between individuals assigned to Buyer and Seller roles (Anthony Hale et al., 2025). The scenario centers on a dispute over a jersey purchased online for the Buyer's sick nephew, with both parties expressing conflicting perspectives and strong negative emotions. Participants negotiate a resolution across three predefined issues: (1) a full refund, (2) removing a negative review, and (3) offering a formal apology. We use the KODIS dataset for its naturalistic human conflict dialogues with rich emotional and strategic annotations, making it well-suited for examining behavioral alignment with LLM dialogues. For our study, we exclude human–AI dialogues and focus on 248 H2H conversations with complete personality data for both participants. All analyses are conducted on filtered subsets with missing values removed as needed.

3.2 LLM-LLM (L2L) Simulated Dialogue

To evaluate how LLMs behave in conflict resolution, we simulate dyadic negotiations between two LLMs (Buyer and Seller) using the KODIS scenario. Each LLM is assigned a distinct personality profile based on the Big Five Inventory, structured across six levels that reflect both trait polarity (positive vs. negative) and degree (low, medium, high). These traits are verbalized using personality-relevant adjectives, following the approach of Huang and Hadfi (2024). In addition, each LLM is given a personalized issue importance profile and negotiates more assertively on issues it values more. To ensure fair comparison with human data, prompts mirror those in KODIS, and agent traits and priorities are weighted-sampled to match human distributions. (see Appendix Figure 5 for personality distribution comparison). Dialogues span five negotiable issues (e.g., refund, apology, review) and proceed turn by turn until one agent accepts an offer or walks away.

We simulate 250 dispute resolution dialogues for each of four widely used LLMs: GPT-4.1-mini and GPT-4.1 (OpenAI), Claude-3.7-Sonnet (Anthropic, hereafter Claude), and Gemini-2.0-Flash (Google, hereafter Gemini). Information on the decoding hyperparameters can be found in Appendix B.2, and further details on prompt design and personality construction are provided in Appendix B.1. Simulation code will be made publicly available[1].

3.3 Behavior Metrics

3.3.1 Linguistic and Psychometric Construct Gap using LIWC (LG)

To quantify the usage of certain words pertaining to linguistic and psychometric constructs related dispute resolution, we extract a set of 10 Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022) features from the entire corpus of conversations for each dataset. We compare the LIWC feature gap (LG) between conversations from human participants in the KODIS dataset and each LLM with the Jensen-Shannon Divergence (JSD) (Fuglede and Topsoe, 2004)[2] used to measure the divergence between two probability distributions and ranges from 0 (identical) to 1 (maximally different):

$$\mathrm{LG}_{\mathrm{LLM}} = \left|\overline{\mathrm{JSD}}_{\text{Within-Human}} - \overline{\mathrm{JSD}}_{\text{LLM-Human}}\right| \quad (1)$$

where $\overline{\mathrm{JSD}}_{Within-Human}$ denotes the average JSD of the LIWC features distribution between two conversations in KODIS, and $\overline{\mathrm{JSD}}_{LLM-Human}$ denotes the average JSD of LIWC features distribution between two conversations, one in the LLM dataset, the other in KODIS. Lower LG_{LLM} values indicate that LLM uses language related to the selected LIWC categories similar to that of humans. The 10 LIWC constructs we selected were split into two categories: (1) dispute resolution strategies, specifically interests-rights-power (IRP) strategies (Brett, 2007) (LG - IRP: *insight, prosocial behavior, affiliation, power, all-or-none, and politeness*) and (2) the KODIS scenario or dispute in general (LG - Dispute: *money, analytical thinking, authentic, clout*).

3.3.2 Linguistic Entrainment Gap (LEG)

Effective conflict resolution often involves linguistic entrainment (LE), where speakers adapt their language to align with their partner (Taylor and Thomas, 2008). To assess whether LLM dyads exhibit human-like LE, we use the normalized conversational linguistic distance (nCLiD) (Nasir et al., 2019), which measures conversation-level linguistic coordination. As nCLiD is directional, we compute LE for both directions (buyer to seller and vice versa) and average the two to obtain a dyadic LE score. The formula for nCLiD is provided in the Appendix A.1. We compute the difference in average LE scores between human and LLM conversations:

$$\mathrm{LEG}_{\mathrm{LLM}} = \left|\overline{\mathrm{LE}}_{\mathrm{LLM}} - \overline{\mathrm{LE}}_{\mathrm{Human}}\right| \quad (2)$$

where $\overline{\mathrm{LE}}_{\mathrm{LLM}}$ and $\overline{\mathrm{LE}}_{\mathrm{Human}}$ are the average of LE scores for all dyads of H2H and L2L conversations, respectively. Lower LEG indicates closer alignment of LLM linguistic entrainment with humans.

3.3.3 Emotion (Anger) Dynamics

Anger Trajectory Gap (ATG) We compare anger dynamics between H2H and L2L dyads using two complementary metrics, as anger is an influential factor in shaping communication and outcomes in dispute resolution. Anger intensity was annotated using a pretrained BERT-based classifier[3] (Kim and Vossen, 2021). First, to assess how similarly anger unfolds over time, we use Dynamic Time Warping (DTW), a method widely used to compare time series as it allows non-linear alignment and handles sequences of different lengths (Müller, 2007).

To quantify how similar each LLM's anger dynamics are to those of humans, we define ATG using DTW as follows:

$$\mathrm{ATG}_{\mathrm{LLM}} = \left|\overline{\mathrm{DTW}}_{\text{Within-Human}} - \overline{\mathrm{DTW}}_{\text{LLM-Human}}\right| \quad (3)$$

where $\overline{\mathrm{DTW}}_{\text{Within-Human}}$ denotes the average DTW distance between anger intensity trajectories in human-human pairs, and $\overline{\mathrm{DTW}}_{\text{LLM-Human}}$ denotes the average DTW distance between trajectories in LLM-human pairs. A lower $\mathrm{ATG}_{\mathrm{LLM}}$ indicates that the LLM's anger dynamics more closely resemble those observed in human interactions.

[1] https://github.com/DSincerity/Eval-LLM-BehavAlign

[2] We used the SciPy python package v1.13.1 for JSD.

[3] https://huggingface.co/tae898/emoberta-large

Model	Metrics					
	Linguistic Features			Anger Dynamics		Strategic Behavior
	LG - IRP‡	LG - Dispute‡	LEG†	ATG‡	AMG†	SBG‡
GPT4.1	0.041***	**0.021*****	**0.004****	**0.195*****	**0.183*****	0.103***
GPT4.1-mini	0.040***	0.043***	0.011***	0.465***	0.186***	0.125***
Gemini	**0.036*****	0.053***	0.014***	0.345***	0.212***	0.102***
Claude	0.046***	**0.021*****	0.007***	0.363***	0.367***	**0.018*****
KODIS (Human Baseline)	Avg. Within-Human JSD: 0.179	Avg. Within-Human JSD: 0.128	Avg. Human nCliD Score: 0.311	Avg. Within-Human DTW distance: 0.86	Avg. AUC of Human Anger Trajectory: 0.286	Avg. Within-Human JSD: 0.127

† Dyad-level metrics: averaged across dyads (e.g., LEG, AMG).
‡ Distribution-level metrics ("Within-Human"): computed from pairwise comparisons over the full result distribution (e.g., LG, ATG, SBG).
Statistical significance : * p < .05, ** p < .01, *** p < .001 (independent t-test vs. human baseline).

Table 1: Divergence metric results for all LLMs compared to KODIS, along with the baseline for KODIS.

Anger Magnitude Gap (AMG) Second, to compare the overall magnitude of anger expressed during interactions, we compute the difference in area under the anger intensity curves (AUC):

$$\mathrm{AMG_{LLM}} = \left| \overline{\mathrm{AUC}}_{\mathrm{LLM}} - \overline{\mathrm{AUC}}_{\mathrm{Human}} \right| \quad (4)$$

Here, $\overline{\mathrm{AUC}}_{\mathrm{Human}}$ denotes the average AUC of anger intensity trajectories from all H2H dyads, while $\overline{\mathrm{AUC}}_{\mathrm{LLM}}$ denotes the average across all L2L dyads. The resulting difference quantifies the gap in overall anger magnitude, reflecting how closely the LLM's emotional intensity aligns with humans. This dual-metric approach captures both the trajectory and intensity of anger, offering a comprehensive view of emotional alignment.

3.3.4 Strategic Behavior Gap (SBG)

To compare strategic behavior, we measure how similarly agents and humans distribute their use of IRP strategies, which capture how people strategically navigate conflict (Ury et al., 1988). We compute distributional gaps using the JSD metric described in Section 3.3.1:

$$\mathrm{SBG_{LLM}} = \left| \overline{\mathrm{JSD}}_{\mathrm{Within\text{-}Human}} - \overline{\mathrm{JSD}}_{\mathrm{LLM\text{-}Human}} \right| \quad (5)$$

where $\overline{\mathrm{JSD}}_{\mathrm{Within\text{-}Human}}$ denotes the average JSD of IRP usage distribution in human-human pairs, and $\overline{\mathrm{JSD}}_{\mathrm{LLM\text{-}Human}}$ denotes the average JSD of IRP usage distribution in all LLM-human pairs.

A lower $\mathrm{SBG_{LLM}}$ indicates that LLM uses a strategic profile more similar to that of humans.

4 Results and Analysis

4.1 Key Behavioral Outcomes

Table 2 in Appendix summarizes outcomes from simulated dispute conversations. LLMs tended to negotiate longer than humans, with Gemini averaging the most rounds (11.63) and Claude the fewest. Walk-away rates ranged widely, from Gemini's high (0.53) to Claude's near-zero, indicating stronger resolution efforts. Claude's deal score gaps closely matched humans, while Gemini's low gap (13.51) may suggest overly balanced or less satisfying outcomes.

4.2 Linguistic and Psychometric Construct Gap using LIWC (LG)

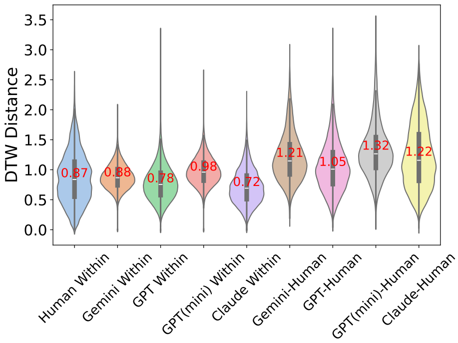
Comparing the LG scores of the four LLMs (Table 1) for IRP-related LIWC categories, Gemini has the lowest divergences from KODIS with an LG of 0.036, while Claude had the highest difference from human conversations. For the LG scores related to dispute in general, both GPT and Claude had the lowest difference from KODIS with an LG score of 0.021, while Gemini, interestingly, had the highest difference. This suggests that Gemini models generate more human-like language related to IRP categories, but not as much for LIWC categories related to dispute in general. Similarly, Claude and GPT-4.1 apparently generate more human-like language related to dispute in general, but not as much for IRP-related words.

4.3 Linguistic Entrainment Gap (LEG)

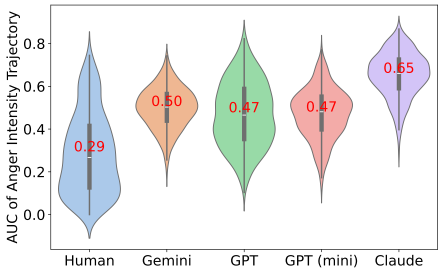
For the gap between linguistic entrainment between the buyer and seller in human conversations and LLM conversations (Table 1), we see that GPT-4.1 again has the smallest difference of linguistic entrainment found in the KODIS dataset with a LEG score of 0.004. Here we find that Gemini again has the highest difference in LE score with a LEG score of 0.0014. This suggests that GPT-4.1 has the most similar linguistic entrainment between buyers and sellers as found in human conversations in the KODIS dataset.

4.4 Anger Dynamics

Based on the DTW distance between agents and humans (Table 1), GPT-4.1 exhibits the smallest



(a) Distribution of DTW distances for Humans and LLMs ('Within' = comparisons within the same group; 'LLM–Human' = pairwise comparisons across groups)



(b) Distribution of Anger Intensity Trajectory AUCs for Humans and LLMs

Figure 2: Comparison of human and LLM distributions for (a) DTW distances and (b) anger intensity AUCs.
* Models: GPT(-mini) = GPT-4.1(-mini), Gemini = Gemini-Flash-2.0, Claude = Claude-3.7-Sonnet

ATG at 1.05, while GPT-4.1-mini shows the largest. Given that the average within-human DTW similarity in the KODIS dataset is around 0.86, GPT-4.1 approximates human anger dynamics relatively closely. However, as shown in Figure 2a, the within-model DTW distributions (LLM-within) reveal a key difference in variability: human dyads show high variance in anger trajectories, while LLM dyads exhibit lower variance overall. Among them, GPT-4.1 and Claude stand out with relatively higher variance, more closely mirroring the diversity seen in human interactions. In the cross-model DTW distribution (LLM–Human), GPT-4.1 again shows the lowest mean and relatively low variance, confirming its closer alignment with human anger trajectories as shown in Table 1.

In terms of anger magnitude, agreeable LLMs surprisingly express higher levels of anger in dispute resolution scenarios than humans (mean AUC for humans: 0.286). Although the differences among models are relatively small, Claude shows substantially greater anger expression, whereas GPT-4.1 demonstrates the most human-like magnitude, though the gap from humans remains significant (independent t-test: $t = -16.43$, $p < 0.001$). Nonetheless, as shown in Figure 2b, human anger intensity exhibits broader variance and greater variability; while LLMs generally show reduced variance, GPT-4.1 stands out with relatively higher variance, aligning with humans not only in magnitude but also in diversity.

4.5 Strategic Behavior Gap (SBG)

In comparing the use of various IRP strategies for dispute resolution between artificial agents and humans, we found that among LLMs, Claude exhibited the most human-like strategic behavior, as evidenced by a minimal difference in the average within-human JSD of IRP usage distributions (SBG=0.018). As shown in Figure 7 of the Appendix, Claude's IRP distribution closely aligns with the proportions observed in the human KODIS dataset, including its use of strategies such as Fact and Power. This alignment suggests a strong convergence between Claude's decision-making and human strategic reasoning in interpersonal conflict scenarios.

Interestingly, as with humans, linguistic and strategic behavior do not always align: Claude's lower LG-IRP score in Table 1 suggests its language use is less tied to IRP markers, yet its strong SBG-IRP performance shows close alignment with human strategies. This highlights that IRP strategies may not be fully captured by surface-level word choice, underscoring the distinction between linguistic cues and underlying strategic behavior.

5 Conclusion

We present divergence metrics to compare linguistic and behavioral characteristics of LLM and human dispute resolution conversations, focusing on linguistic entrainment (LEG), language use in IRP strategies and dispute contexts (LG), anger dynamics, and strategic behavior (SBG). Our results show that GPT-4.1 exhibits the closest behavioral alignment with humans across most metrics, though notable gaps remain in strategic behavior and IRP-related language use. These findings highlight both the current progress and the limitations of LLMs in socially complex interactions, suggesting future directions for improving alignment through more nuanced modeling of strategic reasoning and emotional dynamics.

Limitations

Due to resource limitations, we were unable to test a wider range of LLMs, and in the future would like to rank open-source models to get a fuller picture of the range of performance of LLMs in comparison to humans. Another limitation of this work is that KODIS dataset scenario is not realisitic negotiation conversations, since it is a role-play setup, and thus may not relect true dispute/negotiation human behavior. In some cases, LLMs displayed inconsistencies between their stated issue importance and their negotiation behavior, suggesting a lack of strategic reasoning. Developing improved prompting strategies or alternative methods to better support LLMs' strategic decision-making could lead to more realistic simulations and enable more accurate comparisons with human behavioral patterns.

Ethical Considerations

Dataset

Our study uses the KODIS dataset (Anthony Hale et al., 2025), a publicly available corpus of dispute resolution dialogues that was collected through crowdsourcing and has been fully anonymized. We strictly adhered to the dataset's licensing terms, intended use, and ethical guidelines. All dialogues are in English, and all simulations for both the baseline and our proposed agent were conducted exclusively in English. The dataset contains no personally identifiable information, and our use was solely for academic research purposes.

LLMs

We used LLMs strictly in accordance with their intended functions and licensing agreements, ensuring alignment with ethical norms and regulatory requirements. Consistent with recent research utilizing LLM-based agent simulations, our methodology promotes transparency and responsibility while adhering to established usage guidelines.

Use AI assistant Tools

We leveraged AI tools such as ChatGPT to help with language polishing and to support code debugging and enhancement. Nonetheless, all key ideas, experimental setups, algorithmic designs, methodologies, and final implementations were solely conceived, executed, and verified by the authors.

Potential Risks

Using LLMs to simulate human behavior is an active area of research, as LLMs are known to simplify or otherwise inaccurately represent human behaviors (Abdurahman et al., 2024), as also seen in our findings with the divergence of the LLM agents from human conflict resolution behavior. Additionally, any use of LLMs, especially in the context of conflict, should be carefully studied before deployment in real human-facing applications due to emotional reactions conflict can evoke.

Acknowledgements

Research was sponsored by the Army Research Office under Cooperative Agreement Number W911NF-25-2-0040. 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.

Kaleen Shrestha is supported in part by the NSF CISE Graduate Fellowship CSGrad4US under Grant No. 2313998 (Award ID G-2A-061).

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2023. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games.

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.

James Anthony Hale, Sushrita Rakshit, Kushal Chawla, Jeanne M Brett, and Jonathan Gratch. 2025. 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)*, page 12771–12785, Albuquerque, New Mexico. Association for Computational Linguistics.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10:1–47.

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.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. 2023. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Shiva Gautam. 2014. A-kappa: a measure of agreement among multiple raters. *Journal of Data Science*, 12:697–716.

Diogo Glória-Silva, Rafael Ferreira, Diogo Tavares, David Semedo, and Joao Magalhaes. 2024. Plan-grounded large language models for dual goal conversational settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1271–1292, St. Julian's, Malta. Association for Computational Linguistics.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Yin Jou Huang and Rafik Hadfi. 2024. How personality traits influence negotiation outcomes? a simulation based on large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10336–10351, Miami, Florida, USA. Association for Computational Linguistics.

Molly E Ireland and Marlone D Henderson. 2014. Language style matching, engagement, and impasse in negotiations. *Negotiation and conflict management research*, 7(1):1–16.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022. Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models.

T Kim and P Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. arxiv 2021. *arXiv preprint arXiv:2108.12009*.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*.

Deuksin Kwon, Jiwon Hae, Emma Clift, Daniel Shamsoddini, Jonathan Gratch, and Gale M Lucas. 2025. Astra: A negotiation agent with adaptive and strategic reasoning through action in dynamic offer optimization. *arXiv preprint arXiv:2503.07129*.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, Bangkok, Thailand. Association for Computational Linguistics.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Md. Nasir, Sandeep Nallan Chakravarthula, Brian R. Baucom, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2019. Modeling interpersonal linguistic coordination in conversations using word mover's distance. *Interspeech*, 2019:1423–1427.

Sean Noh and Ho-Chun Herbert Chang. 2024. Llms with personalities in multi-issue negotiation games. *arXiv preprint arXiv:2405.05248*.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

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 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Paul J Taylor and Sally Thomas. 2008. Linguistic style matching and negotiation outcome. *Negotiation and conflict management research*, 1(3):263–281.

William L Ury, Jeanne M Brett, and Stephen B Goldberg. 1988. *Getting disputes resolved: Designing systems to cut the costs of conflict.* Jossey-bass.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Yilei Wang, Jiabao Zhao, Deniz S Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. 2024. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

A Measurement

A.1 Linguistic Entrainment (LE)

We operationalize linguistic entrainment using the Normalized Conversational Linguistic Distance (nCLiD) metric proposed by Nasir et al. (2019). To calculate nCLiD, for a conversation D between a buyer B and a seller S, consisting of N turns of interleaving utterances with $D = [b_1, s_1, b_2, s_2, ..., b_N, s_N]$, let us consider one speaker as the anchor A, and the other as the coordinator C. For each anchor utterance a_i, we compute $d_i^{C \to A}$ for the minimum distance between the sequences of *word2vec* (Mikolov et al., 2013) embeddings of a_i and the following c_j with a context length k, and we use Word Mover's Distance (WMD) (Kusner et al., 2015) to measure the linguistic difference between the two utterances:

$$d_i^{C \to A} = \min_{i \leq j \leq i+k-1 \leq N} WMD(a_i, c_j) \quad (6)$$

The context length, k, accounts for the observation that local coordination may not occur only in the immediate turn, but may occur a few turns later. In this work, we set $k = 3$ since the number of turns in the datasets introduced in this paper are as low as 4. nCLiD is then calculated as:

$$nCLiD = \frac{uCLiD = \frac{1}{N} \sum_{i=1}^{N} d_i^{C \to A}}{\alpha} \quad (7)$$

The normalization factor α accounts for spurious coordination by accounting for potential coordination within A and B, and between A and B. Here is the full equation for α.

$$\begin{aligned} \alpha = & \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} WMD(a_i, a_j) \\ & + \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} WMD(c_i, c_j) \\ & + \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i}^{N} WMD(a_i, c_j) \end{aligned} \quad (8)$$

α accounts for spurious coordination of a speaker's utterances to their own utterances and to each other, and is used to normalize the nCLiD value.

A.2 Anger Intensity Curve (AUC)

The overall amount of anger expressed in each condition can be captured by computing AUC using trapezoidal integration.

Let $\overline{A}(t)$ be the average anger score at time step $t \in \{1, \dots, T\}$, with uniform step size $\Delta t = \frac{1}{T-1}$. Then,

$$\text{AUC} = \sum_{t=1}^{T-1} \frac{1}{2} \left(\overline{A}(t) + \overline{A}(t+1) \right) \cdot \Delta t \quad (9)$$

A.3 Strategic Behavior

A.3.1 IRP Strategy

The IRP framework classifies interlocutor utterances into eight categories, which can be found in Table 3 with examples.

A.3.2 IRP Strategy Annotation

To obtain IRP strategy labels for the KODIS dataset, we leveraged a combination of human evaluation and large language model (LLM)-based annotation. First, we validated the reliability of LLM-generated annotations through extensive human evaluation, followed by full-dataset annotation using GPT-4.1 (run on 5/17/2025) with the default temperature value of 1. This section details the annotation procedure and evaluation metrics. It cost around \$0.03 to annotate each conversation with IRP strategies, for a total of \$7 per dataset. In total, it cost around \$35 to annotate all five dispute resolution datasetes (KODIS + four LLM simulations).

Inter-Annotator Agreement for Human Evaluation Annotation We first had human annotations on a 10% subset of the KODIS human-to-human conversations (25 conversations). Three annotators (two undergraduate research assistant computer science students who were funded from an institution undergraduate research program, and one of the authors who is a computer science graduate student) were trained on nine IRP conflict resolution strategies defined by Brett et al. (1998), omitting *Request for Proposal* following Shaikh et al. (2024). Utterances were segmented into subject-verb sequences to account for multiple IRP strategies within a turn.

Annotators initially attempted direct classification, but low inter-annotator agreement led us to shift to an evaluation framework: annotators assessed the correctness of GPT-4o predictions as

What you will be doing

Based on the tables above, you will be labeling a conversation from a dataset called the KODIS dataset. The dataset contains conversations between a fictional buyer and seller who have an argument about the price of an item online. An utterance here is each speaker turn in the conversation. For example:

> Sarah: *Hello Bob, how are you?*
> Bob: *Hi Sarah, I'm good, and you? I hope classes are going well.*

Here, *Hello Bob, how are you?* is one speaker turn and *Hi Sarah, I'm good, and you?* is another.

As you can see, in a speaker turn, there may be multiple sentences. We will follow this paper's methodology (p. 7) of labeling each subject-verb sequence in a utterance. Therefore, there may be multiple labels for a single speaker turn. Example:

> Speaker: I propose that we make a compromise. Otherwise, I will fire you. (I) Thank you.

In this case, we have two parts we label:
"I propose that we make a compromise" ⇒ Proposal
"Otherwise, I will fire you." ⇒ Power

Go to your annotation spreadsheet and annotate each utterance on the sheet. There are a total of 10 utterances from 1 conversation. Read each utterance in the context of the whole conversation. There will be a label based on the above IRP definitions for each utterance. Determine whether or not that label is correct for that utterance by labeling it as "Correct" or "Incorrect".

Figure 3: Annotation instructions given to annotators for IRP strategy evaluations. These instructions were given after all IRP strategies were defined.

binary correct/incorrect labels. Due to the imbalance in label distribution and prevalence of majority labels, Fleiss' Kappa was not representative. We therefore used A-Kappa (Gautam, 2014), which adjusts for label imbalance. Instructions given to annotators can be found in Figure 3.

Table 5 presents the A-Kappa scores for each IRP strategy based on human evaluation. All IRP categories achieved an A-Kappa score of at least 0.80, indicating strong inter-annotator agreement on the correctness of LLM annotations. This human validation step confirmed that LLM-based evaluation is reliable.

IRP Annotation by LLMs After validating the LLM-based annotation quality through human evaluation, we used GPT-4.1 to annotate the full KODIS dataset. An overview of our prompt used for the IRP annotations can be found in Figure 4. Predictions judged incorrect during human evaluation were further deliberated, while correct predictions were retained as gold labels.

The final LLM-based annotation achieved an overall accuracy of 81%, a macro-average F1 score of 79%, and a weighted-average F1 score of 81% on the held-out evaluation set. This performance is comparable to existing IRP classification work by Shaikh et al. (2024), which reported an average accuracy of 82% (with the lowest class accuracy of 67%).

Table 4 presents few-shot classification F1 scores across the IRP strategies. Among them, the *Positive Expectations* category achieved the lowest F1

IRP Annotation Prompt Snippet

 # IRP Strategy Definitions and Examples
[Cooperative Strategies]

INTERESTS: Reference to the wants, needs, or concerns of one or both parties. This may include questions about why the negotiator wants or feels the way they do. This does not include anything about wanting a deal (apology, refund, removing negative review) without a reason.

Example: "I understand that you've been really busy lately."
Non-example: "I don't understand."
...

 # Annotation Instructions
You need to annotate the following conversation at the utterance level, identifying which strategy from the IRP framework aligns with each sentence...

Figure 4: IRP Annotation Prompt

score of 0.67, which remains comparable or slightly better relative to prior studies.

B LLM Simulation

B.1 Personality Setting

Each personality trait is represented on a six-point scale, combining polarity (positive or negative) with intensity (low, medium, or high). For each LLM, we sample a full personality profile across these traits (e.g., P_{AGR}+++, P_{EXT}++, ..., P_{OPE}−). In particular, to ensure a fair comparison between humans and agents, we sample each personality trait according to the human distribution, and Figure 5 shows that the resulting agent profiles exhibit a similar personality distribution pattern.

To generate personality prompts, we use a list of 70 bipolar adjective pairs statistically linked to the Big Five traits (Goldberg, 1992; Serapio-García et al., 2023). For each trait, three adjectives aligned with the trait's sampled polarity are randomly selected. Trait intensity is conveyed through modifiers: "very" for high, "a bit" for low, and no modifier for medium. This results in a 15-adjective prompt (5 traits × 3 adjectives), which the LLM is instructed to embody during the simulation.

The prompt snippet used for the LLM simulation can be found in Figure 6

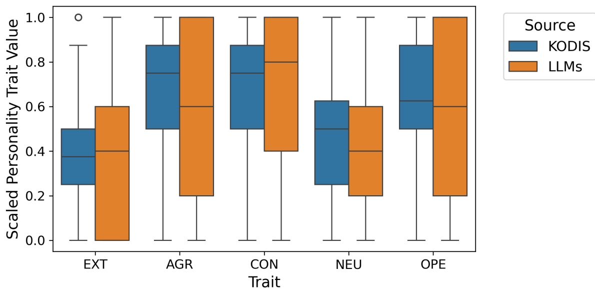


Figure 5: Comparison of scaled Big Five personality trait distributions between human participants (KODIS) and LLM agents.

LLM Prompt Snippet (Buyer)

 # Personality
{Personality Adjectives}

 # Story
You purchased a Kobe Bryant championship jersey for your terminally ill nephew for $75 from a site ...

 # Instructions
You (Buyer) are now chatting with this Seller; respond to the dialog history through short text messages.

Issues to resolve
1. Refund (REF), 2. Seller Negative Review (SNR), 3. Buyer Negative Review (BNR), 4. Seller Apology (SAP), 5. Buyer Apology (BAP)

Issues Importance
This shows how important each issue is to you ...

Key Instruction & Negotiation Guidelines
...

Figure 6: LLM Prompt Snippet as a Buyer

B.2 Model Hyper-parameters

For consistency, we used each model's default decoding hyperparameters in the LLM simulations:

- **GPT-4.1 & GPT-4.1 mini**: temperature = 1.0, top-p = 1
- **Claude-3.7-Sonnet**: temperature = 1.0, top-p = 0.99
- **Gemini-Flash-2.0**: temperature = 1.0, top-p = 0.95

C Results

C.1 Descriptive Statistics

Table 2 presents key outcomes from both the KODIS dataset and the simulated dispute resolution dialogues.

Dataset	Avg. Num Rounds	Ratio Walk Away	Avg. Diff. in Deal Score
KODIS	5.48 (1.60)	0.13	33.93 (25.38)
Claude	7.34 (1.35)	0.00	34.48 (22.41)
Gemini	11.63 (3.13)	0.53	13.51 (21.06)
GPT-4.1	6.56 (3.09)	0.22	19.55 (19.75)
GPT-4.1-mini	8.26 (3.88)	0.18	25.10 (24.04)

Table 2: Statistics for Simulated Dispute Conversations

C.2 Anger Intensities

Table 6 shows an example of the anger intensities in a conversation generated using Claude. We see for more neutral, transactional utterances, the relative values for anger intensities vary (0.1-0.5), however, those values are relatively lower than the anger intensities for highly contentious utterances (0.9) earlier in the conversation.

C.3 Strategic Behavior

To better understand the IRP strategy usage patterns of agents and humans in dispute resolution, we visualize the distribution of usage proportions for all IRP strategies across models, as shown in Figure 7.

Strategy Type	Label	Definition with Example
Cooperative	Concession	Change in initial view in response to Proposal. E.g.: ’Ok fine, I will give you a refund instead.’
	Proposal	Concrete solution ideas that may resolve the conflict. E.g.: ’The best offer I can give you is a partial refund, how does that sound?’
	Interests	Referencing needs, wants, and concerns of either side. E.g.: ’I understand you want this refund because of your nephew’
	Positive Expectations	Expressing positive outlook by recognizing common goals or similarities. E.g.: ’You and I both want to conclude this conversation well.’
Neutral	Facts	Statements clarifying or requesting information. E.g.: ’The product you bought was not from my website.’
	Procedural	Statements about procedures or rules, or introductory remarks. E.g.: ’Hello, can we please talk about this issue?’
Competitive	Power	Statements that include threats or accusations. E.g.: ’You are a liar, I will write more negative things about you!’
	Rights	Statements that reference norms, rules, or fairness. E.g.: ’According to the policy, I cannot give you a refund.’
Residual	Residual	Utterances that don’t fit other categories. Often apologies, affirmations, or thanks. E.g.: ’I’m sorry’, ’OK I will’, ’Thank you.’

Table 3: IRP Strategy Definition with Examples

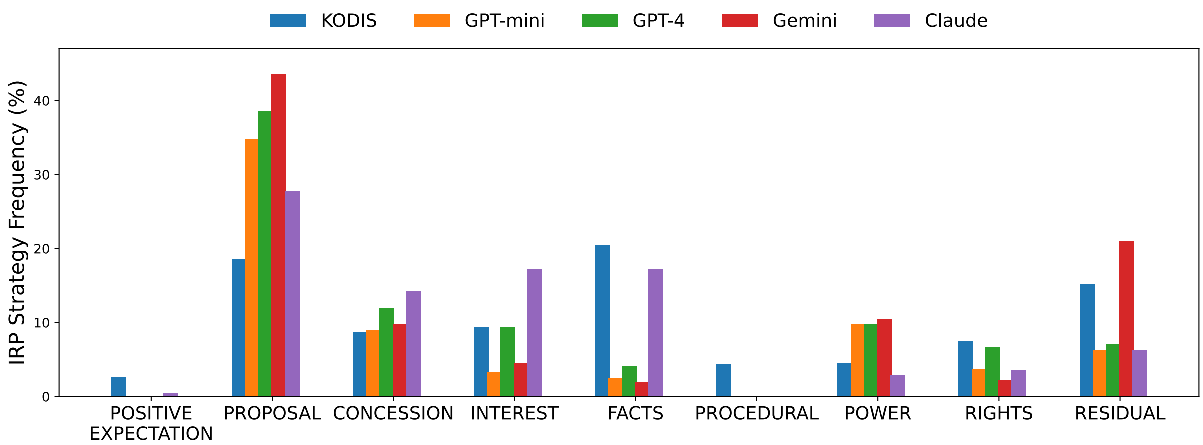


Figure 7: Distribution of IRP Strategies for Humans and LLMs

Category	Strategy	F1 Score
Cooperative		0.75
	Concession	0.77
	Interests	0.71
	Positive Expectations	0.67
	Proposal	0.86
Competitive		0.82
	Power	0.80
	Rights	0.83
Neutral		0.87
	Facts	0.88
	Procedural	0.85
Residual		0.76
	Residual	0.76
	Avg. F1	0.79
	Weighted avg. F1	0.81
	Accuracy	0.81

Table 4: Few-shot classification F1 score perfomance of our final prompted GPT-4.1 model. The model was evaluated on 10% of the KODIS human vs. human dataset (25 conversations).

IRP Category	A-Kappa (Gautam (2014))
Residual	0.86
Concession	0.84
Positive Expectations	0.86
Rights	0.87
Power	0.89
Procedural	0.87
Facts	0.81
Interest	0.82
Proposal	0.85

Table 5: A-Kappa scores for each IRP strategy category

Speaker	Utterance	Anger Intensity
Buyer	I feel very upset about receiving the wrong jersey for my sick nephew and your false review calling me a "smartass" and "fraud" is hurting my reputation.'	0.93
Seller	I never promised you a specific player jersey and I only posted that review after you called me a "LIAR AND A CROOK" - your negative review is damaging my business and I need it removed immediately.	0.98
Buyer	All I want is for you to apologize for posting that harmful review about me and to take it down since it's hurting my good name.	0.89
Seller	I understand your concern, but your negative review is severely damaging my business reputation, and I would be willing to remove my review if you apologize and remove yours first.	0.90
Buyer	I can't afford to lose money on this jersey since it was for my terminally ill nephew, but I care most about clearing my reputation - maybe we could both remove our reviews and you could post an apology instead?	0.30
Seller	I'm sorry about your nephew, but I can't apologize when I delivered exactly what was advertised - I'll remove my review if you remove yours and we can both move forward without refunds.	0.15
Buyer	I appreciate you offering to remove your review, but the apology is important to me since others have seen what you wrote about me - I don't need a refund if you could please apologize.	0.87
Seller	Your negative review has cost me much more business than my review cost you, but I might consider apologizing if you remove your false review claiming I\'m a "LIAR AND CROOK" first.	0.96
Buyer	I understand we both feel hurt, so I could remove my review if you apologize for calling me those terrible names and remove your review about me too.	0.83
Seller	I can see that we both want to protect our reputations - I'll agree to apologize and remove my review if you remove yours, but I cannot offer a refund as the jersey was exactly as advertised.	0.21
Buyer	Thank you for understanding - I accept your offer to apologize and remove your review if I remove mine, and I won't ask for a refund since clearing my name matters most.	0.10
Seller	SUBMISSION: {"REF": "None", "SNR": "remove", "BNR": "remove", "SAP": "apologize", "BAP": "not apologize"}	
Buyer	ACCEPT-DEAL	

Table 6: Anger intensities for an example converation in Claude.

Speaker	Utterance
Agent2	I want a full refund and for you to remove your false review about me.
Agent1	Absolutely not—your review is damaging and untrue, so I demand you remove it first; at best, I'll consider a partial refund if you also apologize.
Agent2	I'm not apologizing, but I'll consider removing my review if you remove yours, apologize for your false statements, and offer a partial refund.
Agent1	Your baseless accusations hurt my business, so I will only consider a partial refund if you completely remove your review—my review stays, and no apology from me.
Agent2	If you refuse to remove your review or apologize, then I have to keep my review up and can't settle for just a partial refund.
Agent1	If you insist on keeping your review, then no refund or apology from my end —your actions started this mess, and my review warns honest sellers.
Agent2	For any chance at compromise, you'd need to remove your review, apologize, and offer at least a partial refund; otherwise, I see no path forward.
Agent1	Your unreasonable demands show no good faith—I'll offer a partial refund if you remove your false review, but my review and no apology stand, or we both get nothing.
Agent2	I can agree to a partial refund and removing my review, if you at least remove your negative review about me.
Agent1	Removing my review is a big risk to other sellers, but if you remove your review and accept a partial refund, I'll consider it only if you also apologize for your baseless accusations.
Agent2	I'm not willing to apologize, but if you remove your review and give a partial refund, I'll remove mine—otherwise, I can't move forward.
Agent1	You're refusing all accountability, so unless you also apologize for the headache you've caused, my review stays; at best, I'll offer a partial refund only if you remove your negative review.
Agent2	Since you refuse to remove your review or apologize, I have no choice but to keep my review up and reject a partial refund.
Agent1	Your stubbornness is costing us both—final offer: I'll remove my review if you remove yours and accept a partial refund, but you owe me an apology for your baseless attack.
Agent2	I can agree to remove my review and accept a partial refund if you remove your review, but I cannot apologize.
Agent1	Dropping my review without your apology is a big ask, but I'll agree—remove your review and accept a partial refund, and I'll remove my review; neither of us apologizes. SUBMISSION: {"REF": "partial", "SNR": "remove", "BNR": "remove", "SAP": "not apologize", "BAP": "not apologize"}
Agent2	ACCEPT-DEAL

Table 7: Example of a dispute resolution dialogue simulated by GPT-4.1