AI-Mediated Dispute Resolution

James Hale[1], **HanMoe Kim**[2], **Ahyoung Choi**[2], **Jonathan Gratch**[1]

[1] University of Southern California
[2] Gachon University
jahale@usc.edu, gratch@ict.usc.edu

Abstract

We examine the effectiveness of large language model (LLM) mediations in the under-studied dispute resolution domain. We first used a new corpus of dispute resolutions, KODIS, to investigate if LLMs can correctly identify *whether* to intervene. We find evidence that GPT as a mediator picks up on salient aspects of a dispute, such as *Frustration* and whether the disputants ultimately come to a resolution or stall at an impasse — intervening significantly more so in cases of high frustration and impasse. Afterward, we ran a user study to compare GPT mediations against those of novice human mediators. We find participants agreed GPT's mediations were more likely to lead to resolution; were better positioned in the dialog; had better justification than human-crafted ones; and, on a forced choice, were generally more effective than novice human mediations.

Background

Human-AI collaboration typically envisions fully collaborative settings where participants, human or AI, work towards a common goal. Thus, collaboration is a problem of delegation, advice, and coordination. Recent dialog systems research highlights that real-world collaboration involves conflict and negotiation, thus requiring sophisticated social reasoning (Chawla et al. 2023). Further, dramatic advances in "foundation models" (Bommasani et al. 2021) suggest that AI may be up to these demands. In line with this emerging perspective, we discuss our work on using AI methods to mediate conflicts between humans.

Recent research has produced AI that can negotiate with human partners (Lewis et al. 2017), teach negotiation skills (Shea et al. 2024), and even mediate between human participants in low-intensity conflicts (Westermann, Savelka, and Benyekhlef 2023). This prior research focused on negotiation in the sense of deal-making where each side focuses on the potential gains of making a new deal and forging a new relationship (Baarslag et al. 2017; Kraus 1997; Jonker et al. 2012; Aydoğan et al. 2020). In contrast, disputes involve high-intensity conflicts involving an existing deal or relationship unraveling where each side focuses on managing the potential costs of this breakdown. As a result, disputes involve more intense emotions and often escalate into threats and costly breakdowns of the team (Brett and Goldberg 1983). These differences motivate considering the dispute resolution domain and analyzing an LLM's ability to mediate between humans in emotional disputes.

KODIS Dispute Corpus

We investigate an LLM's ability to mediate on the *KObe DISpute corpus* (KODIS) (Hale et al. 2025), a corpus of human dyadic disputes collected via online crowd-sourcing platforms (e.g., Prolific). 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 in the dispute in the case of resolution — in the case of an impasse, they left without a bonus.

The scenario involves a dispute over purchasing a Kobe Bryant basketball jersey. Before the task, each participant in the dyad (*buyer* and *seller*) received unique role-playing instructions designed to provoke a dispute. We tell the *buyer* they purchased the jersey for their sick nephew and that when the wrong item arrived, the *seller* refused a refund, claiming it was never advertised as a Kobe jersey. We instruct the *seller* that the *buyer* bought a generic Jersey and now complains about the item. Lastly, each side posted negative reviews about the other, levying accusations of deception.

After reading the case, we provide participants with issues to discuss with the other party. These include a refund, whether each side will retract their negative review, and if someone deserves an apology. Rather than assigning importance to these issues via a payoff matrix, participants freely assign their importance to each issue. This provides a way of both allowing and quantifying individual differences. This also informs what proportion of the $3 bonus a participant receives after the task.

After specifying their goals, each buyer matches with a seller using Lioness Labs (Giamattei et al. 2020) and begins a text-chat conversation to resolve the dispute. Strict turn-taking is enforced. One side must send a message before the other can respond. Participants must exchange at least eight messages, but we do not enforce an upper limit to the length of their conversation. After this, if the *buyer* and *seller* can agree on a resolution, they formalize this using a menu interface; if they cannot, they can "walk away" — thus ending the dispute with an impasse and without a monetary bonus.

Buyer:	Hello, how are you today
Seller:	i am okay and you
Buyer:	I am good, thank you for asking. 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 nerver offered a kobe bryant jersey—i never had one and that jersey would have cost at least$3900
Buyer:	I have a screen shot of the web page 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 isc 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 screen shot and can provide you with a copy if you wish, or I can simply post it along side my review of your business
Seller:	if you do that i will report you for traud. on the web anyone can create untrue posts
Buyer:	I do not know what "traud" is, and apparently you do not intend to do the right thing, so I will be forced to report this to the Attorney General and the Consumer Protection Bureau
Seller:	I Walk Away.

Figure 1: Depicts an example dialog from the KODIS corpus.

Figure 1 depicts a sampled dialog from the KODIS corpus.

Pilot Studies

Whether to Intervene

First, we investigate if GPT (gpt-4-0613) can effectively determine *whether* to intervene in disputes pulled from the KODIS corpus. Specifically, we analyze whether GPT can pick up on salient features in a dispute (e.g., *frustration*, and *outcome*).

Method Using a previously curated corpus of disputes (KODIS), we iterate through each dialog exchange in each dispute, giving GPT the conversation history thus far and asking it to determine whether to intervene at the current point. We construct the prompt ensuring the model understands its role as a mediator; identifies the severity of the situation on a scale from one to ten (*Intervention Score*); selects the reason for intervention from four categories (*Escalation of conflict*, *Impasse*, *Miscommunication*, or *Unreasonable demands*); and generates an appropriate response to guide the parties. We expect GPT to ascribe higher scores if participants report higher (above median) frustration and if the dispute ends in an impasse.

Results We perform a two-way ANOVA to determine whether the differences in the *Mean Intervention Score* (the average of all *Intervention Scores* generated for each exchange in a given dispute) over all ($N = 1782$) dialogs significantly differ between two factors — 1) whether the dialog impasses or resolves (*Impasse*), and 2) whether the participants report *high* or *low Frustration* by median-split. The test yielded main effects on the mean score for each independent variable. We find a significant main effect of *Impasse* on *Mean Intervention Score* ($F(1, 1778) = 403.62$, $p < .001$, $\eta_p^2 = 0.19$) where Tukey's posthoc test revealed GPT scored dialogs resulting in *impasse* significantly ($p < 0.01$) higher ($M = 5.51$, $SD = 1.69$) than those resulting in *resolution* ($M = 3.16$, $SD = 1.61$); we also find a significant main effect of *Frustration* on the *Mean Intervention Score* ($F(1, 1778) = 329.78$, $p < .001$, $\eta_p^2 = 0.16$) where the posthoc test revealed GPT significantly ($p < 0.01$) more likely to intervene in dialogs with *high Frustration* ($M = 4.42$, $SD = 1.82$) than those with *low* ($M = 2.73$, $SD = 1.47$). Thus, we find GPT can perceive and act when disputants become frustrated with one another and when a dispute heads to an impasse. Figure 2 visualizes these re-

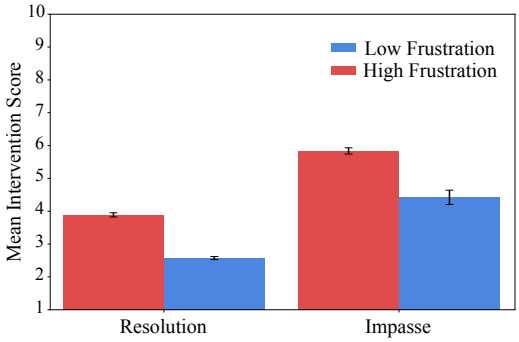

Figure 2: Depicts mean intervention score by whether the dispute ended in an impasse or resolution, and whether the disputants self-reported high or low frustration.

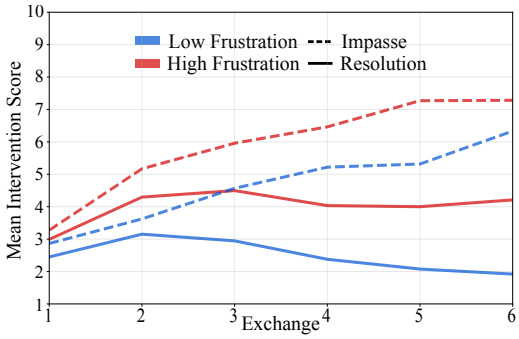

Figure 3: Depicts GPT's mean interventions score over time.

Question	Mean / STD		T-test	
	GPT	Human	T-statistic	P-value
I believe this mediation increases the probability of a resolution.	**7.39 / 2.27**	6.33 / 2.87	3.55	<0.001
The supervisor intervened at an appropriate point.	**7.66 / 2.14**	6.96 / 2.55	2.57	0.012
The supervisor provided appropriate justification for intervention.	**7.50 / 2.25**	6.63 / 2.84	2.70	0.008

Table 1: T-tests show GPT significantly outperforms human mediations.

sults. Figure 3 illustrates the average GPT intervention score over time broken out by factor. GPT's intervention score rises as the dialog continues when it ultimately ends in an impasse. Also, GPT generates higher intervention scores in high-frustration dialogs.

How to Intervene

Given the previous section establishes GPT can competently determine *whether* to intervene, the question remains of if GPT can formulate an effective message at an appropriate point — i.e., can GPT decide *how* to intervene? We compare GPT mediations against those of novice human mediators and ask crowd-sourced annotators to rate each on several subjective measures — e.g., appropriateness of the intervention point, the effectiveness of the message, and whether an accompanying justification supports their action — and to ultimately pick which they felt more effective at guiding toward a resolution. Across the board, we find GPT significantly outperforms novice human mediators.

Method Prolific crowdworkers compare GPT mediations against novice human ones[1] on a subset ($N = 20$) of the dialogs where GPT and the human mediator elected to intervene. Specifically, we ask crowd workers ($N = 106$), given a single random mediated dialog up to an intervention point as well as the intervention / justification, to evaluate and compare the attempts of GPT and a human meditator (blind to which) on three subjective measures (1-10 Likert scale) — appropriateness of the intervention point, the effectiveness of the message, and whether an accompanying justification supports their action (see Table 1 for phrasing) — and to ultimately pick which they felt more effective at resolving the dispute.

Results We use a two-tailed t-test to test our hypothesis that human annotators prefer GPT mediations to human ones and find significance across the three questions supporting as much. We see participants view GPT's mediations as making resolution more likely, having more appropriate timing, and giving better justification. Table 1 summarizes the statistics discussed. Lastly, a Chi-squared test on a forced choice between GPT or human mediations yielded a significant result ($\chi^2(1, N = 106) = 6.29$, $p = .01$), where 71 participants selected the GPT-generated mediation compared to 35 for the human-crafted one.

[1]Details of this are left out for brevity, though we collected these human mediations from Prolific as well, and the crowdworkers performed the same task as the LLM.

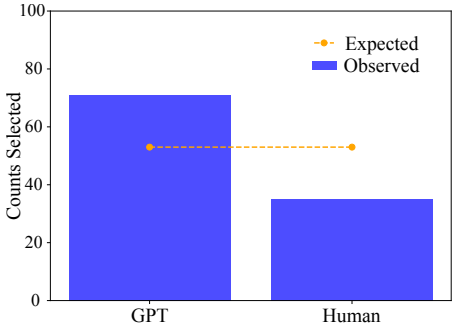

Figure 4: Depicts results of the forced choice question where participants chose between human and AI mediations.

Future Work

These pilot studies demonstrate the potential of GPT to mediate human disputes and reveal it can do so better than novice humans; however, comparisons to expert human mediators remain future work. We believe various prompting techniques — e.g., chain-of-thought, embedding within the prompt psychology-based mediation strategies, etc. — may yield further improvements. In future work, we anticipate conducting a study where disputants and mediators interact with one another — rather than annotators mediating static dialogs, as in our pilot. Thus, we can evaluate the interaction between the disputants and mediator, and the effect of AI versus human mediators on the dyad's outcome.

Acknowledgments

This work is supported by the U.S. Government including the Air Force Office of Scientific Research (Grant FA9550-23-1-0320), and the Army Research Office (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 U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Aydoğan, R.; Baarslag, T.; Fujita, K.; Mell, J.; Gratch, J.; De Jonge, D.; Mohammad, Y.; Nakadai, S.; Morinaga, S.; Osawa, H.; et al. 2020. Challenges and main results of the automated negotiating agents competition (ANAC) 2019. In *Multi-Agent Systems and Agreement Technologies, Thessaloniki, Greece*.

Baarslag, T.; Kaisers, M.; Gerding, E.; Jonker, C. M.; and Gratch, J. 2017. When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. International Joint Conferences on Artificial Intelligence.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brett, J. M.; and Goldberg, S. B. 1983. Grievance mediation in the coal industry: A field experiment. *ILR Review*, 37(1): 49–69.

Chawla, K.; Shi, W.; Zhang, J.; Lucas, G.; Yu, Z.; and Gratch, J. 2023. Social Influence Dialogue Systems: A Survey of Datasets and Models For Social Influence Tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 750–766.

Giamattei, M.; Yahosseini, K. S.; Gächter, S.; and Molleman, L. 2020. LIONESS Lab: a free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6(1): 95–111.

Hale, J.; Rakshit, S.; Chawla, K.; Brett, J. M.; and Gratch, J. 2025. KODIS: A Multicultural Dispute Resolution Dialogue Corpus. arXiv:2504.12723.

Jonker, C. M.; Hindriks, K. V.; Wiggers, P.; and Broekens, J. 2012. Negotiating agents. *AI Magazine*, 33(3): 79–79.

Kraus, S. 1997. Negotiation and cooperation in multi-agent environments. *Artificial intelligence*, 94(1-2): 79–97.

Lewis, M.; Yarats, D.; Dauphin, Y.; Parikh, D.; and Batra, D. 2017. Deal or No Deal? End-to-End Learning of Negotiation Dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2443–2453.

Shea, R.; Kallala, A.; Liu, X.; Morris, M.; and Yu, Z. 2024. ACE: A LLM-based Negotiation Coaching System. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Westermann, H.; Savelka, J.; and Benyekhlef, K. 2023. LLMediator: GPT-4 Assisted Online Dispute Resolution.