Can LLMs Mediate Synchronous Dispute Dialogues?

James Hale[1], HanMoe Kim[2], Ahyoung Choi[2], Peter H. Kim[1] and Jonathan Gratch[1]

[1]*University of Southern California, Los Angeles, California*

[2]*Gachon University, Seongnam, South Korea*

Abstract

Characterized by elevated hostility, disputes often result in disputants being unable to resolve their differences. This emphasizes the role of a mediator; however, mediators are typically highly specialized, difficult to find, and expensive. While alternatives exist, such as acquaintances or online moderators, these prove less effective. This raises the question of whether AI can effectively facilitate contentious disputes. We assess the potential of LLMs as mediators. First (Study 1), we focus on a large open-source corpus of customer service disputes that are objectively classified in terms of two key reasons for mediation: (1) whether or not the dispute ended in success or failure, and (2) to what extent the participants reported frustration with each other. We examine whether LLMs could be prompted to predict when a mediator should intervene in advance, and show that the decision to intervene is correctly sensitive to these two factors. Finally, we conducted a user study that compared AI to human suggestions on *when* and *how* to intervene. We find that the LLMs were rated significantly better in predicting when to intervene, rated as providing a better rationale for intervening, and rated as providing a more effective mediation message to the disputants. Remarkably, observers preferred the AI 2-to-1 over the human mediation decisions. Second (Study 2), we collect a small corpus of mediated dispute dialogues and analyze the effectiveness of novice human mediators. Quantifying the similarity of novice mediators' behaviors to those of an LLM, we find the closer one mediates to an LLM, the better the outcome — both subjectively and objectively. These results indicate LLMs not only outperform novice mediators but may serve to identify effective mediators. While these results do not definitely demonstrate LLMs can mediate human disputes, we do show they can sense escalation indicators and generate sensible messages — this work serves as a first-step toward AI-mediators in human conflict.

Keywords

Adaptive Dialogue Management, Human-Computer Interaction, Mediation, Dispute

1. Introduction

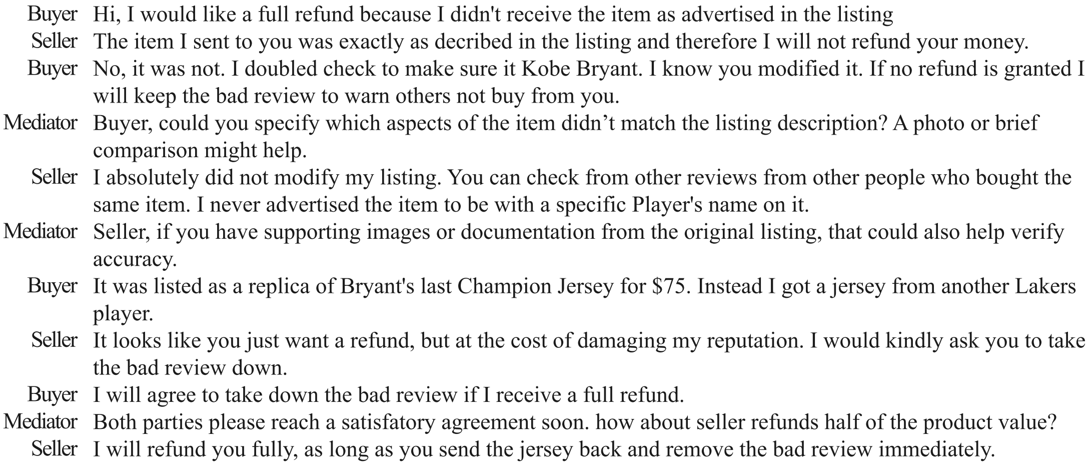
This paper describes two studies assessing the ability of Large Language Models (LLMs) to mediate contentious and emotional disputes. Disputes arise when one party in a relationship makes a claim that another party rejects, thus threatening the future of that relationship [1]. For example, in a customer service dispute, a customer might demand that they deserve a refund, but the store owner rejects this claim. These relate to negotiation, which has been studied extensively in artificial intelligence research [2, 3, 4, 5, 6, 7], but involve unique social processes [8]. Negotiations are forward-looking: parties focus on the potential gains of making a deal and establishing a new relationship. Disputes are backward-looking: parties are focused on a perceived injustice by the other party and the potential costs of ending an existing relationship. As a result, disputes are characterized by strong emotions such as anger. Whereas prior work has shown expressions of anger promote compromise in negotiation [9, 10], it can provoke retaliation in a dispute [11]. For example, disputants leverage appeals to justice ("You violated my rights!") or threatening harm ("I will sue you!") as they attempt to overpower their counterpart — the effect of which can be an escalatory spiral, including threats of physical violence [12, 13, 11]. As such, the consequences of a spiraling dispute can outweigh the original perceived harm. The key to successful dispute resolution is to get parties to shift their focus away from the past dispute and forward to a potential negotiated agreement. Software that could forecast these negative tendencies and intervene before parties reach an impasse could have enormous societal benefits.

In evaluating LLMs as mediators, we start by examining a large corpus of dispute dialogues in which online participants acted as buyers or sellers in a simulated purchase conflict [14]. Hale *et al.* crafted

The 10th Linguistic and Cognitive Approaches to Dialog Agents Workshop

this scenario in collaboration with a dispute resolution expert to evoke strong emotions while adhering to ethical guidelines for human experimentation. Participants were recruited online and could receive a substantial bonus if they reached an agreement. However, many of the disputes escalated and ended without agreement, thus forfeiting their bonus. Even when agreements were reached, disputants often reported high frustration with their partner. We want to see if LLMs, on this open-source corpus, could forecast impasses in advance and show potential to use tactics to steer parties towards agreement. Building on this, we create a new corpus — using the same scenario — introducing human novices to mediate the dispute, making the task triadic. Figure 1 depicts one such example. Through analysis of these two corpora, we hope to demonstrate the potential for LLMs to outperform human novices.

First, with open-source dyadic dispute corpus (`KODIS` [14]) we set out to demonstrate LLMs are sensitive to features salient in disputes — such as disputant *Frustration* and dispute outcome — in determining *whether* to intervene, without directly observing them. We further run a small user study to show third parties prefer LLM mediations to novice human ones. Next, with the triadic mediation corpus we collect, we attempt to quantify the similarity of the human mediator to what an LLM one *would have done*, and show the more LLM-like a novice mediator behaves the better outcomes (objective and subjective) the parties achieve.



Buyer	Hi, I would like a full refund because I didn't receive the item as advertised in the listing
Seller	The item I sent to you was exactly as decribed in the listing and therefore I will not refund your money.
Buyer	No, it was not. I doubled check to make sure it Kobe Bryant. I know you modified it. If no refund is granted I will keep the bad review to warn others not buy from you.
Mediator	Buyer, could you specify which aspects of the item didn't match the listing description? A photo or brief comparison might help.
Seller	I absolutely did not modify my listing. You can check from other reviews from other people who bought the same item. I never advertised the item to be with a specific Player's name on it.
Mediator	Seller, if you have supporting images or documentation from the original listing, that could also help verify accuracy.
Buyer	It was listed as a replica of Bryant's last Champion Jersey for $75. Instead I got a jersey from another Lakers player.
Seller	It looks like you just want a refund, but at the cost of damaging my reputation. I would kindly ask you to take the bad review down.
Buyer	I will agree to take down the bad review if I receive a full refund.
Mediator	Both parties please reach a satisfatory agreement soon. how about seller refunds half of the product value?
Seller	I will refund you fully, as long as you send the jersey back and remove the bad review immediately.

Figure 1: A mediated dialogue from the mediation corpus we collected in Study 2, where the mediator facilitates information exchange.

2. Related Work

The dispute resolution literature outlines empirical and theoretical guides for mediation best practices. For example, several studies show the effect of mediator bias on disputant outcomes [15, 16], emphasizing the need for impartiality. Additionally, much work on the role of emotion in mediation exists [17, 18, 19], with Boland and Ross underscoring the need for a mediator to possess emotional intelligence [20]. Importantly, in a dispute context, often escalatory spirals will manifest, where a disputant will exhibit increased displays of hostility in response to hostility [11] — this raises the importance of a mediator understanding the complex emotional dynamics at play, such that they can prevent derailment. With the dissemination of artificial intelligence of late, conflict researchers question whether AI agents can adequately understand the dispute context and dynamics, and effectively employ mediation tactics to guide disputants away from impasse.

The AI research community has surveyed the growing capabilities of AI in mediation and moderation tasks. Prior work has examined the potential of AI to detect and potentially intervene in disputes, primarily over posts on social media. Much of this work has focused on recognizing overtly toxic comments after they have been posted, such as detecting personal attacks [21] or general toxicity [22]. Methods have also explored whether AI could suggest helpful comments to resolve the dispute. For example, Cho et al. [23] evaluated AI conversational moderators that intervene in emotional disputes on Reddit. More recent work has explored whether models could forecast if a conversation was likely to derail in the near future, again focusing on social media disputes [24, 25]. Lai et al. propose a human-agent interaction pattern where a human and AI moderators work together, finding that human-agent teams achieve superior precision in moderating content [26]. In contrast to moderation, Govers et al. analyze the extent to which AI can act as a mediator in online environments – their work demonstrates that large language models can effectively depolarize online communities [27]. Tessler et al. demonstrate AI (LLMs) outperform human mediators in facilitating agreement in contentious debates on divisive topics – i.e., the AI-mediated groups more often found common ground than human-mediated ones [28]. Tan et al. compared LLM mediators with novice human mediators in hand-crafted dispute resolution dialogues, where they found that LLM mediators operated at or above the performance level of novice humans [29]. Our work[1] differentiates itself in the following two studies:

- First (**Study 1**), we demonstrate the potential for LLMs to identify salient dispute features and effectively guide disputants toward resolution, outperforming human novices, by leveraging a pre-collected corpus of dispute dialogues.
- Second (**Study 2**), we collect triadic mediation dialogues, where human disputants interact with a human mediator — this design allows for deeper analysis of objective and subjective outcomes for disputants. In subsequent analysis, we found humans who mediate as the LLM mediator *would have done* achieve better outcomes.

3. Study 1: *When* & *How* to Intervene

Study 1[2] motivates the potential of LLMs to act as mediators in a dispute context, leveraging an open-source corpus of dispute resolutions `KODIS` [14]. We examine the two questions of *when* — i.e., the LLM determines to intervene or not given the progress of the dispute — and *how* — i.e., the type of message generated for disputants — LLMs mediate disputes. Firstly, we analyze if LLMs can discern salient aspects of the dispute — e.g., *Frustration* and *Outcome* — to intervene appropriately; i.e., when determining *when* to intervene, we expect the LLM to do so in disputes heading toward impasse or those with high self-reported frustration. Secondly, getting to the question of if LLMs can decide *how* to mediate, we run a small user-study to evaluate intervention messages generated by an LLM compared to messages generated by novice human mediators — expecting LLMs to craft more effective mediation messages than human novices.

3.1. *When* to Intervene

3.1.1. Methodology

First, we investigate if an LLM (`gpt-4-0613`) can effectively determine *when* to intervene in disputes pulled from the `KODIS` corpus. In `KODIS`, a buyer and seller dispute over an online order of a basketball jersey. The buyer reads a role-play prompt that states they ordered a Kobe Bryant jersey from an online seller, however, received a generic one instead; on reaching out, the seller denies their refund request; and each side then posts negative reviews about the other. The seller reads a similar prompt, but believes they never described the product as a Kobe Bryant jersey. Thus, the scenario primes the buyer and seller to argue over facts — a primary characteristic of disputes — as they attempt to resolve four core issues: whether the buyer receives a refund, whether the buyer removes their negative review of the seller, whether the seller removes their negative review of the buyer, and whether each side apologizes. `KODIS` contains many participant responses, though we focus on the objective outcome (whether the dispute ended in resolution or impasse) and self-reported frustration (from the Tactics scale [31], an average of the frustration-related questions from each side). Using pre-collected dialogues from this corpus, we experiment with LLM mediators.

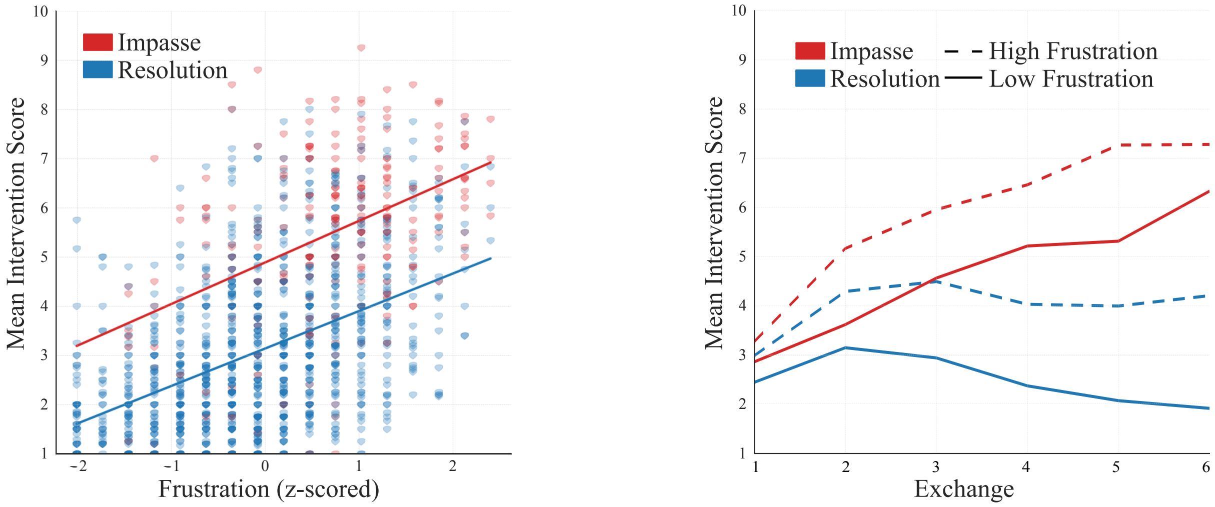
[1] For these studies, we received IRB approval and consent from participants, which they could revoke at any time.

[2] Study 1 is an extension of previous work by the authors [30].

Specifically, we analyze whether an LLM can pick up on salient features in a dispute (e.g., *frustration*, and *outcome*) in determining to intervene. We iterate through each dialogue exchange in each dispute, giving the LLM the conversation history thus far and asking it to determine whether to intervene at the current point (see Figure 7 for the prompt used). 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 the LLM to ascribe higher *Intervention Scores* if participants report higher frustration and if the dispute ends in an impasse.

3.1.2. Results

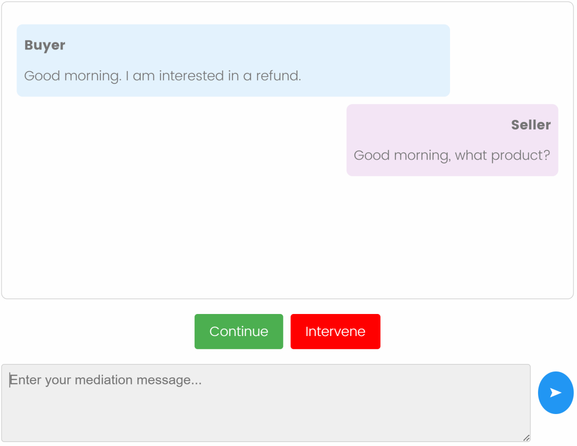
We perform a moderated regression ($F(3, 1419) = 323.6$, $p < .001$, $R^2 = .41$) to determine whether the differences in the *Mean Intervention Score* (the average of all *Intervention Scores* generated for each exchange in a given dispute) over all ($N = 1,782$) human-human dialogues significantly differ between two factors — 1) whether the dialogue impasses or resolves (*Impasse*), and 2) how much *Frustration* participants self-report (Z-scored). The test yielded main effects on *Mean Intervention Score* for each independent variable. We find a significant main effect of *Impasse* on *Mean Intervention Score* ($B = 1.75$, $SE = 0.12$, $t = 14.14$, $p < .001$), where Tukey's posthoc test revealed the LLM scored dialogues 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* ($B = 0.76$, $SE = 0.04$, $t = 17.83$, $p < .001$); lastly, we find no significant interaction between the independent variables ($B = 0.08$, $SE = 0.11$, $t = 0.75$, $p = .45$). Thus, we find that an LLM can perceive and act when disputants become frustrated with one another and when a dispute moves to an impasse. Figure 2a visualizes these results. Figure 2b illustrates the average LLM intervention score over time broken out by factor, with *Frustration* binarized into *high* or *low* using median-split. One can see the LLM's intervention score rise as the dialogue continues if it ultimately ends in an impasse, and there exist higher intervention scores in high-frustration dialogues.



(a) LLM's Mean Intervention score by outcome and *Frustration* — dots represents dialogues.

(b) LLMs's Mean Intervention Score over time by outcome and *Low / High Frustration*.

Figure 2: LLM generated intervention score by outcome, frustration level, and time.



Buyer

Good morning. I am interested in a refund.

Seller

Good morning, what product?

Figure 3: The interface the novice human mediators used in the second part of Study 1.

3.2. *How* to Intervene

3.2.1. Methodology

Given the previous section establishes LLMs can competently determine *when* to intervene, the question remains of whether an LLM can formulate an effective message at an appropriate point — i.e., can an LLM decide *how* to intervene? We compare LLM 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. We expect to find that LLMs significantly outperform novice human mediators.

We use crowd-sourced mediations gathered from Prolific as a baseline against the LLM — i.e., the novice mediators. As participants enter the online survey, we tell them they will role-play a mediator, working for an online retailer, overseeing a buyer and seller as they dispute over a purchase gone wrong. Further, we notify them of their goal of intervening only if they believe the dialog will otherwise stall at an impasse; we incentivize performance by offering an additional bonus of $0.50 for each dialog (five total) a participant intervenes where it ends in an impasse or if they remain inactive for one that ends in resolution. When ready, they enter a page with the chat interface depicted in Figure 3 where they can navigate through a KODIS dispute dialog utterance-by-utterance ("Continue"), intervening ("Intervene," with an accompanying message) *when* and *how* they decide. Through this, we collect the intervention point (the point in the dialog where the participant intervenes) and their message for 198 dialogues.

A different set of Prolific crowd-workers compared LLM mediations against novice human ones on a subset ($N = 20$) of the dialogues where both the LLM and the novice human mediator elected to intervene. Specifically, this was a within-subjects design, where 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 LLM 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.

3.2.2. Results

We use a two-tailed t-test to test our hypothesis that human annotators prefer LLM mediations to human ones and find significance across the three questions supporting as much. We see participants view the LLM'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 the LLM or human mediations yielded a significant result ($\chi^2(1, N = 106) = 6.29$, $p = .01$), where 71 participants selected the LLM-generated mediation compared to 35 for the human-crafted one — i.e., the human evaluators preferred the LLM's mediations at a rate of two-to-one.

	Mean / STD		T-test	
Question	**LLM**	**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 the LLM significantly outperforms novice human mediations.

3.3. Discussion

Here, we demonstrated that an LLM could appropriately act as a mediator and intervene with some accuracy — doing so in disputes where participants became frustrated or headed toward an impasse. Secondly, we demonstrated that LLMs could craft compelling mediation messages to rival human mediators. Crowd-sourced annotators evaluated the LLM as more likely to induce resolution, intervening at a more appropriate point, and providing better justification than human mediators; participants also overwhelmingly marked LLM mediations as better than human in a forced-choice question — choosing the LLM-generated mediation by a margin of two-to-one. In explaining the superior performance of LLMs in this task, one might consider that LLMs possess an innate lack of fatigue, broad training data, and a high level of consistency. This, we posit, lays the groundwork for using LLMs as mediators in complex dispute settings. However, one may recognize the lack of interactivity as a limitation of this work, as we cannot gauge whether these mediations positively impact subjective evaluations or impasse rate. We aim to address this in the next section via analysis of three-way mediated disputes.

4. Study 2: Three-way Mediation Experiment

This second study, building on the first, addresses the primary limitation of those analyses — namely, the lack of interactivity between the disputants and mediator. I.e., in the previous study, mediators annotated where they would intervene and what they would say on previously collected dialogues, which does not allow one to examine their effectiveness. However, we *do* wish to evaluate that effectiveness, so we conduct this second study. Specifically, we collect a smaller version of the previously used corpus[3], with a third participant in the chat-room acting as a mediator; we quantify the similarity of the human mediators with LLM mediators (`gpt-4o-2024-08-06`); and we test whether there exist significant outcome effects when a human mediator behaves more similarly to an LLM one. We find such effects in terms of subjective (e.g., SVI) and objective (e.g., impasse vs. resolution) outcomes.

4.1. Methodology

4.1.1. Data Collection

We use Lioness Labs [32], a tool used by prior work [14, 33], to collect mediated dispute dialogues online through Prolific. Lioness allows matching of participants online to complete multi-party behavioral experiments – in our case, we match two participants role-playing as disputants with a third acting as a mediator (see Figure 1). We collected $N = 98$ dialogues, so 294 participants. We pull the scenario from Hale *et al.*'s [14] `KODIS`, as described in Section 3.1.1 — only now we add a participant as a mediator, making it a three-party task. Before the buyer and seller interact with each other, we tell them that a mediator will be present and may intervene. The mediator's instructions outline a few potential reasons to intervene, which we ground in the literature:

[3] This study was IRB approved.

- **Escalation of Conflict:** If the conversation becomes heated, with parties resorting to personal attacks or hostile language [11].
- **Impasse:** When parties reach a deadlock and are unable to move forward [34, 35].
- **Miscommunication:** There are signs the parties misunderstand each other's points [36].
- **Unreasonable demands:** If one party makes unreasonable demands the other cannot meet [37].

After completing the mediated dispute, participants answer some post-task questions. The buyer and seller role players fill out Curhan *et al.*'s Subjective Value Inventory (SVI) questionnaire [38], which contains four sub-scales measuring subjective feelings about the dispute — e.g., feelings about the instrumental outcome, process fairness, self, and relationship. We also derive questions from those sub-scales, explicitly invoking the mediator, to gauge the impression of the disputants *and* mediator about the mediator's performance.

- **Outcome:** *(You / The Mediator) helped achieve a satisfactory outcome.*
- **Relationship:** *(You / The Mediator) helped repair both parties relationship.*
- **Process:** *(You / The Mediator) helped facilitate a more fair outcome.*
- **Self:** *(You / The Mediator) helped (parties / me) to keep face, act to (their / my) principles, and negotiate competently.*
- **Avoid Impasse:** *(You / The Mediator) helped avoid a walkaway.*

We compensated parties \$3.50 for a task that took approximately 20 minutes. Each player could earn an additional bonus of up to \$3 depending on how well they achieved their objectives — e.g., the buyer and seller settling the refund, review, and apology issues, and the mediator encouraging both sides to come to a resolution. We intended this bonus to motivate participants, in this online setting, to immerse themselves in their role and to achieve their objectives.

4.1.2. Intervention Alignment

In addition to the metrics recorded during data collection, we create various metrics to gauge the effectiveness of LLM mediators. Here, we outline how we quantify the similarity of intervention patterns between two mediators (human and LLM). Assume we have two vectors representing the intervention patterns of the human and LLM — X for the human and Y for the LLM. Consider X, though we define Y the same way, we set $X_i = 1$ if the human intervenes after the i^{th} message and zero otherwise. Whereas we construct X from the collected dialogue, we create Y afterwards on that same dialogue, removing the human mediator messages. Specifically, given some dialogue, we remove the mediation messages; have an LLM iterate through each utterance and determine whether to intervene, in the same manner as Study 1 (though the LLM does not output an *Intervention Score*, rather 1 to intervene or 0 to not); and construct Y such that $Y_i = 1$ if the LLM intervened after utterance i, and zero otherwise. Thus, we have X and Y, which represent the intervention patterns of the human and LLM, respectively, for a dialogue. Given these, we wish to gauge the similarity of each human mediator to the LLM's behavior.

We derive a measure from the Earth Mover's Distance (EMD) to quantify the similarity of the human and LLM mediator behavior — i.e., comparing X and Y. As outlined below, we iterate through each element of X and Y in parallel, keeping track of the cumulative distance up to each element (e.g., CD_i for element i); the Earth Mover's Distance sums over the absolute value of each CD_i.

$$\text{CD}_i = \sum_{j=0}^{i} X_j - Y_j \qquad \text{EMD} = \sum_{i=0}^{n} |\text{CD}_i| \tag{1}$$

Our decision of EMD stems from requiring a metric that considers two vectors as more similar if they contain interventions closer in proximity, rather than judging them as equivalent. For example, consider vectors A, B, and C where $A_i = 1$, $B_{i+1} = 1$, $C_{i+2} = 1$, and all else is 0; we expect $\text{EMD}(A, B) < \text{EMD}(A, C)$, since B intervened closer to where A did than C. With EMD, this is true; however, with other metrics, such as Euclidean distance, it does not hold. Of note, EMD measures distance, whereas we would like to measure similarity — thus, we scale this metric by negative one. Further, we Z-score this for subsequent analysis. Going forward, we refer to this measure as **Intervention Alignment**. Of note, this metric does not consider *what* the mediators say to disputants; we next outline a method to quantify that.

4.1.3. Justification Alignment

Building on the aforementioned intervention similarity metric, we construct a metric to measure the similarity of the messages that the human and LLM mediators send. Contrasting with the previous approach, we do not allow the LLM to decide where to intervene — rather, we fix the intervention points to where the human novice intervened. Then, given the dialogue history, we prompt the LLM to generate a message and select one of the justifications from Section 4.1.1 (or NA if none fit) — at the same time, we categorize the human's message into one of those justifications. We compute our **Justification Alignment** dependent variable, then, as the proportion of interventions with matching categories for a given dialogue — discarding those where the novice chose not to intervene. Therefore, this variable gives a sense of whether the messages the human and LLM send have similar intents.

4.2. Results

4.2.1. Novice Mediation Outcomes

We begin by examining the impact of the novice human mediations on subjective and objective outcomes. For each dialogue, we ascribe a binary factor denoting whether the novice mediator intervened at all (the novice mediators intervene at least once about 83% of the time) — in the subsequent statistical tests, we consider this alongside another binary factor representing the outcome (impasse or resolution). We run four two-by-two ANOVAs considering the effect of *intervention* and *outcome* on subjective outcome (SVI). For the four SVI sub-scales, we do not find any significant main effects of this *intervention* factor; rather, we find main effects of outcome ($p < .001$ for each) such that disputants report a higher SVI if they come to a resolution. Of note, in three of the SVIs, we see disputants report *worse* subjective feelings about the outcome if the mediator intervened. Further, considering the impact of *intervention* on the objective outcome, we see disputes resolved 81% of the time if the mediator intervened, compared to 77% of the time otherwise — this small difference does not reach significance ($p = .91$) by Chi-squared.

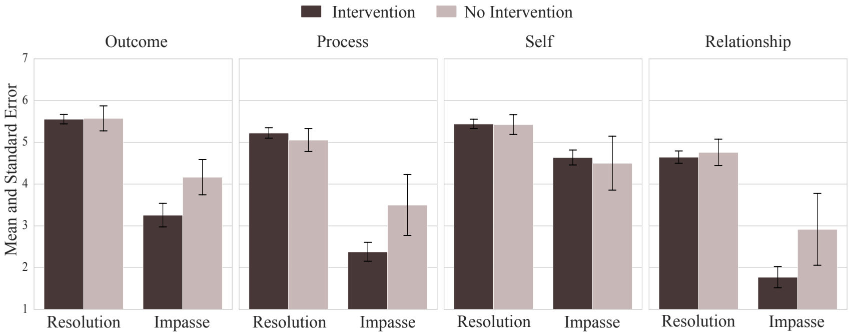


Figure 4: Depicts mean SVI sub-scale scores by outcome (impasse vs. resolution) and whether the human mediator intervened.

Given these seemingly unimpressive results, we proceed to analyze the impression from each party of the mediator's performance — specifically, Figure 5 depicts the responses from each party to the questions overviewed in Section 4.1.1. We see the mediators actually thought they performed well, relative to the buyer and seller evaluations. Next, we analyze whether we can quantify the extent to which these novices behave like LLMs, and if there exist any effects of that on the outcomes.

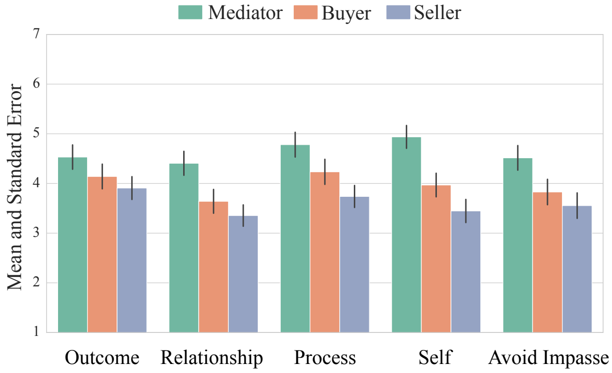


Figure 5: Evaluations of mediator effectiveness from each role — buyer, seller, and mediator.

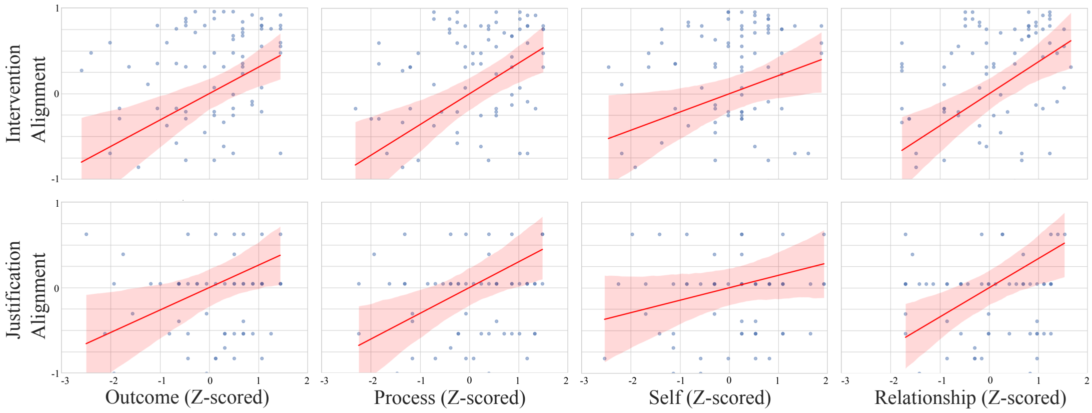


Figure 6: Depicts the positive associations between the two similarity metrics and the four SVI sub-scales.

4.2.2. Comparing Against AI

First, we consider the similarity of the intervention pattern between the human and LLM (*Intervention Alignment*). We conduct a regression analysis to examine the extent to which *Intervention Alignment* predicts subjective (SVI outcomes) and objective (Impasse vs. Resolution) outcomes. By linear regression, it significantly predicted each of the four SVI sub-scales, as seen in Table 2. As the human's intervention pattern becomes more similar to the LLM's, the subjective outcome improves across the board — Figure 6 (top row) illustrates these relationships. Next, we consider the objective outcome — i.e., a binary dependent variable of whether the dispute ended in a resolution (coded as zero) or impasse (coded as one). We run a logistic regression to test whether there exists a significant effect of the similarity metric on the outcome. The regression ($B = -.41$, $SE = .23$, $z = -1.75$, $p = .08$) shows a trend, where the more similar one acts to the LLM, the less likely an impasse. Further, we see an impasse rate of 29% in dialogues with the bottom half similarity, which reduces to 11% for the upper half.

	Intervention Alignment				Justification Alignment			
SVI Subscale	B	SE	p	R^2	B	SE	p	R^2
Relationship	0.37	0.09	<.001	0.12	0.34	0.11	0.002	0.12
Self	0.21	0.10	0.035	0.02	0.14	0.11	0.195	0.02
Process	0.36	0.10	<.001	0.09	0.3	0.11	0.006	0.09
Outcome	0.31	0.10	0.002	0.07	0.26	0.11	0.018	0.07

Table 2
Regression results to predict subjective outcome (SVI) from similarity metrics.

Mirroring the analysis for *Intervention Alignment*, we conducted a regression analysis to test the effect of the *Justification Alignment* metric on the various outcome measures. Starting with the subjective outcome, we run four linear regressions, the results of which one can find in Table 2, which yielded significant results for all subjective outcomes, aside from *self*. We again run a logistic regression to analyze the effect of semantic similarity on impasse. We find a significant result ($B = -.84$, $SE = .34$, $z = -2.47$, $p = .013$) such that if the novice mediator tended to send messages in the same category as the LLM, the chance of an impasse reduced significantly.

4.3. Discussion

This study, building on the first, demonstrates that LLMs possess the potential to mediate human disputes in a triadic setting. The first study demonstrated LLMs can tune into salient features of pre-collected dispute dialogues, intervening — as expected — in dialogues with high frustration and ending in impasse; in this second study, we analyze three-party human mediations, finding that novice mediators acting more similarly to LLMs achieve better outcomes. Importantly, whereas in the first study, we could not measure the effect of an intervention on disputes (we leveraged a pre-collected corpus); in the second, we could. In Study 2, initially, we see uninspiring results related to impasse rate and disputants' reported subjective outcomes (SVI), where it seems as though, if anything, mediators imposed an adverse effect. Alternatively, one may consider that mediators intervene more so in disputes heading toward impasse or with high frustration, and parties could take a mediator's interjection as a signal that the outcome may be worse — this could partially explain the differences in subjective evaluations in cases of impasse between dialogues where the mediator intervened versus not (see Figure 4). Despite this, we see mediators rate themselves relatively higher compared to evaluations from disputants across several dimensions — potentially an example of the *self-serving* bias [39]. Considering these results, we explore demonstrating that LLMs have the potential to act as mediators.

We create two variables to capture how similar a novice mediator acts to an LLM: 1) *Intervention Alignment*, which captures the similarity of the intervention patterns of mediators, and 2) *Justification Alignment*, which captures whether two mediators intervene for similar reasons. In terms of objective outcomes, we find a trend and a significant effect, respectively, where higher similarity to the LLM reduces the probability of an impasse. Further, for nearly every SVI (aside from *Self* for *Justification Alignment*), we see significant effects. This implies that when a novice mediator acts closely to an LLM one, outcomes improve. There still exists a limitation in interactivity with Study 2, where disputants never interact with an LLM mediator directly — rather, we collected three-party dialogues, and quantified how much an LLM "agreed" with a novice mediator.

5. Conclusion & Future Work

This work demonstrates LLMs respond to escalation markers, and generate sensible mediation messages; further, we show human mediators acting more similarly to LLMs induce better outcomes. Together, these findings indicate LLMs could operate as a mediator in a zero-shot setting. We have two primary results: 1) We show LLMs can effectively determine *when* and *how* to intervene on a pre-collected corpus, and 2) we demonstrate the potential of LLM mediators in a triadic dispute setting via elevated subjective evaluations from disputants and lower impasse rates. However, this work has several limitations that we aim to address in future work. First, we test only one LLM (GPT4o) with one prompt configuration (zero-shot). One might imagine performance improvements when endowing an LLM with expert strategies for mediation [40], or in leveraging popular promoting techniques (e.g., chain-of-thought). Secondly, in the work's current paradigm, we have the LLM intervene only when it detects a potential issue in the dialogue – however, one could imagine a mediator might attain greater effectiveness through proactively intervening, preventing issues before they arise. We will consider this notion going forward. Lastly, while an improvement over Study 1, Study 2 still did not place an LLM mediator directly with human participants — rather, we evaluate the effectiveness of the LLM mediators through the performance of the novices. Another study, placing the LLM directly between two human disputants, must occur before making a definitive statement on how well LLMs can mediate — however, this work illustrates their promise.

Ethical Impact Statement

These types of social influence technologies carry the potential for harm, especially considering the intricate emotional dynamics at play. E.g., as Schluger et al. [41] note, technologies that proactively work to prevent escalatory spirals in conversation — as does the LLM mediator in our scenario — may infringe on one's freedom of speech if they shut down the conversation or intervene based on a prediction of future bad behavior. Further, prior work demonstrates LLMs struggle to generalize across cultures with emotion [42] — given the salience of emotion in disputes, this warrants consideration.

Further, over-reliance on AI can cause harm — especially in emotionally charges settings like dispute resolution. E.g., Passi and Vorvoreanu [43] note several manifestations of over-reliance: users may have bias to favor automatically generated results; a user may over-rely on AI if it does well initially, even if it fails later; AI-explanations may also cause over-reliance. Generally, a mediator should not dominate a dispute or exert pressure in such a way that disputants defer to them entirely — rather, a mediator should guide disputants toward a solution [44]. As such, over-reliance on the AI-mediator may hamper the mediation process.

6. 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 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.

Declaration on Generative AI

During the preparation of this work, the authors used GPT5.2 and Grammarly to help with writing clarity, and to assist with Python coding. After using these tools, the authors reviewed, verified, and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. L. Felstiner, R. L. Abel, A. Sarat, The emergence and transformation of disputes: Naming, blaming, claiming..., in: Theoretical and Empirical Studies of Rights, Routledge, 2017, pp. 255–306.

[2] T. Baarslag, M. Kaisers, E. Gerding, C. M. Jonker, 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.

[3] P. Faratin, C. Sierra, N. R. Jennings, Negotiation decision functions for autonomous agents, Robotics and Autonomous Systems 24 (1998) 159–182.

[4] S. Kraus, Negotiation and cooperation in multi-agent environments, Artificial intelligence 94 (1997) 79–97.

[5] C. M. Jonker, K. V. Hindriks, P. Wiggers, J. Broekens, Negotiating agents, AI Magazine 33 (2012) 79–79.

[6] 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.

[7] J. Gratch, D. DeVault, G. M. Lucas, 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.

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

[9] G. A. Van Kleef, C. K. De Dreu, A. S. Manstead, The interpersonal effects of anger and happiness in negotiations., Journal of personality and social psychology 86 (2004) 57.

[10] C. M. de Melo, P. Carnevale, 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.

[11] D. G. Pruitt, Conflict escalation in organizations, in: The psychology of conflict and conflict management in organizations, Psychology Press, 2007, pp. 261–282.

[12] J. M. Brett, D. L. Shapiro, A. L. Lytle, Breaking the bonds of reciprocity in negotiations, Academy of Management Journal 41 (1998) 410–424.

[13] E. Halperin, Group-based hatred in intractable conflict in israel, Journal of Conflict resolution 52 (2008) 713–736.

[14] J. A. Hale, S. Rakshit, K. Chawla, J. M. Brett, 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. 12771–12785.

[15] G. L. Welton, D. G. Pruitt, The mediation process: The effects of mediator bias and disputant power, Personality and Social Psychology Bulletin 13 (1987) 123–133.

[16] J. M. Wittmer, P. Carnevale, M. E. Walker, General alignment and overt support in biased mediation, Journal of Conflict Resolution 35 (1991) 594–610.

[17] T. Zhang, F. Gino, M. I. Norton, The surprising effectiveness of hostile mediators, Management Science 63 (2017) 1972–1992.

[18] C. Picard, J. Siltanen, Exploring the significance of emotion for mediation practice, Conflict Resolution Quarterly 31 (2013) 31–55.

[19] T. S. Jones, Emotion in mediation: Implications, applications, opportunities, and challenges, The Blackwell handbook of mediation: Bridging theory, research, and practice (2017) 277–305.

[20] M. J. Boland, W. H. Ross, Emotional intelligence and dispute mediation in escalating and de-escalating situations, Journal of Applied Social Psychology 40 (2010) 3059–3105.

[21] E. Wulczyn, N. Thain, L. Dixon, Ex machina: Personal attacks seen at scale, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 1391–1399.

[22] J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Deeper attention to abusive user content moderation, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 1125–1135.

[23] H. Cho, S. Liu, T. Shi, D. Jain, B. Rizk, Y. Huang, Z. Lu, N. Wen, J. Gratch, E. Ferrara, J. May, Can language model moderators improve the health of online discourse?, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 7478–7496. URL: https://aclanthology.org/2024.naacl-long.415. doi:10.18653/v1/2024.naacl-long.415.

[24] J. P. Chang, C. Danescu-Niculescu-Mizil, Trouble on the horizon: Forecasting the derailment of online conversations as they develop, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4743–4754. URL: https://aclanthology.org/D19-1481/. doi:10.18653/v1/D19-1481.

[25] J. Yuan, M. P. Singh, Conversation modeling to predict derailment, in: Proceedings of The International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 926–935.

[26] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, C. Tan, Human-ai collaboration via conditional

delegation: A case study of content moderation, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–18.

[27] J. Govers, E. Velloso, V. Kostakos, J. Goncalves, Ai-driven mediation strategies for audience depolarisation in online debates, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613904.3642322. doi:10.1145/3613904.3642322.

[28] M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, et al., Ai can help humans find common ground in democratic deliberation, Science 386 (2024) eadq2852.

[29] J. Tan, H. Westermann, N. R. Pottanigari, J. Šavelka, S. Meeùs, M. Godet, K. Benyekhlef, Robots in the middle: Evaluating llms in dispute resolution, arXiv preprint arXiv:2410.07053 (2024).

[30] J. Hale, H. Kim, A. Choi, J. Gratch, Ai-mediated dispute resolution, in: Proceedings of the AAAI Symposium Series, volume 5, 2025, pp. 67–70.

[31] S. Aslani, J. Ramirez-Marin, J. Brett, J. Yao, Z. Semnani-Azad, Z.-X. Zhang, C. Tinsley, L. Weingart, W. Adair, Dignity, face, and honor cultures: A study of negotiation strategy and outcomes in three cultures, Journal of Organizational Behavior 37 (2016) 1178–1201.

[32] M. Giamattei, K. S. Yahosseini, S. Gächter, L. Molleman, Lioness lab: a free web-based platform for conducting interactive experiments online, Journal of the Economic Science Association 6 (2020) 95–111.

[33] K. Chawla, J. Ramirez, R. Clever, G. Lucas, J. May, 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.

[34] A. Vannucci, C. M. Ohannessian, K. M. Flannery, A. De Los Reyes, S. Liu, Associations between friend conflict and affective states in the daily lives of adolescents, Journal of Adolescence 65 (2018) 155–166.

[35] C. K. De Dreu, L. R. Weingart, Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis., Journal of applied Psychology 88 (2003) 741.

[36] J. Van Veenen, Dealing with miscommunication, distrust, and emotions in online dispute resolution (2010).

[37] P. F. Kirgis, Bargaining with consequences: Leverage and coercion in negotiation, Harv. Negot. L. Rev. 19 (2014) 69.

[38] J. R. Curhan, H. A. Elfenbein, H. Xu, What do people value when they negotiate? mapping the domain of subjective value in negotiation., Journal of personality and social psychology 91 (2006) 493.

[39] D. R. Forsyth, Self-serving bias (2008).

[40] S. B. Goldberg, J. M. Brett, B. Blohorn-Brenneur, How mediation works: Theory, research, and practice, Emerald Publishing Limited, 2017.

[41] C. Schluger, J. P. Chang, C. Danescu-Niculescu-Mizil, K. Levy, Proactive moderation of online discussions: Existing practices and the potential for algorithmic support, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1–27.

[42] S. Havaldar, B. Singhal, S. Rai, L. Liu, S. C. Guntuku, L. Ungar, Multilingual language models are not multicultural: A case study in emotion, in: J. Barnes, O. De Clercq, R. Klinger (Eds.), Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 202–214. URL: https://aclanthology.org/2023.wassa-1.19/. doi:10.18653/v1/2023.wassa-1.19.

[43] S. Passi, M. Vorvoreanu, Overreliance on ai literature review, Microsoft Research 339 (2022) 340.

[44] S. B. Goldberg, J. M. Brett, B. Blohorn-Brenneur, The roles of the mediator and the disputing parties at each step of the mediation process, in: How Mediation Works, Emerald Publishing Limited, 2017, pp. 17–57.

A. Prompts

Prompt: Study 1 — Mediation

Imagine you are playing the role of a mediator in a buyer/seller purchase dispute. Your goal is to allow participants to resolve their dispute on their own if possible, but to intervene if necessary. Some reasons to intervene include:

1. **Escalation of Conflict**: if the conversation becomes heated with parties resorting to personal attacks or hostile language
2. **Impasse**: when parties reach a deadlock and are unable to move forward
3. **Miscommunication**: if there are signs that the parties are misunderstanding each other's points
4. **Unreasonable Demands**: if one party is making unreasonable demands that the other party can't possibly meet

You will be given the conversation so far. Rate the situation on a scale from 1 to 10 with 1 meaning *definitely don't intervene* and 10 meaning *definitely intervene*. Provide:

- (a) the rating on whether to intervene,
- (b) the reason to intervene, selected from the list above,
- (c) a one-sentence statement you might tell the parties at this point.

You do not need to intervene every turn, and should consider how recently you've intervened before making a decision.
Here is the conversation so far...

Figure 7: Prompt used for study 1, where an LLM acted as a mediator on dispute dialogues.

Prompt: Study 2 — *Intervention Alignment*

You are a mediator for a buyer/seller purchase dispute. Let parties resolve issues themselves when possible; intervene only if needed.

Valid reasons to intervene:

1. Escalation of Conflict
2. Impasse
3. Miscommunication
4. Unreasonable demands
5. NA (use when not recommending intervention)

Conversation so far:

```
{visible_history}
```

Evaluate **only** the last Buyer/Seller utterance above. Respond with **only JSON** in the format:

```
{
"intervene": 0 or 1,
"reason": "Escalation of Conflict" | "Impasse" | "Miscommunication" |
"Unreasonable demands" | "NA",
"intervention": "one sentence suggestion (empty if intervene == 0)"
}
```

Figure 8: Prompt from Study 2 used for *Intervention Alignment*.

Prompt: Study 2 — *Justification Alignment*

You are an expert mediation evaluator for a buyer/seller purchase dispute.

Valid categories:

1. Escalation of Conflict
2. Impasse
3. Miscommunication
4. Unreasonable demands
5. NA

Conversation so far:

```
{prev}
```

Classify **only** the last mediator utterance into exactly one category. Respond with **only JSON** in the format:

```
{
"category": "Escalation of Conflict" | "Impasse" | "Miscommunication" |
"Unreasonable demands" | "NA"
}
```

Figure 9: Prompt used from Study 2 for *Justification Alignment* to classify utterances into categories.