

# Robots in the Middle: Evaluating LLMs in Dispute Resolution

Jinzhe Tan, Hannes Westermann, Nikhil Reddy Pottanigari, Jaromír Šavelka, Sébastien Meeùs, Mia Godet and Karim Benyekhlef

# Contents

---

## INTRODUCTIONS

**1** BACKGROUND

**2** CHALLENGES

## PROPOSED FRAMEWORK & EXPERIMENT SETUP

**3** LLMEDIATOR

**4** EVALUATION

**5** SCENARIOS

**6** S1-TYPES

**7** S2-MESSAGES

## RESULTS

**8** KEY RESULTS

**9** E1-TYPES

**10** E2-MESSAGES

## DISCUSSION

**11** TAKEAWAYS

**12** LIMITATIONS

**13** FUTURE WORK

# Introductions

1

# Background

---

- **Disputed parties may:**
  - Be overwhelmed by emotions.
  - Struggle with complex situations.
  - Misunderstand or confuse each other.
  - Reach a deadlock.
  - Argue without clear evidence.
- **Intermediaries, such as mediators, arbitrators, or conciliators, can:**
  - Calm tensions.
  - Clarify misunderstandings.
  - Identify key issues.
  - Propose solutions.

# Challenges

---

- **Restricted Accessibility**
  - The mediation has to be worth it
- **Resource Constraints**
  - Lack of trained mediators (Branting et al., 2023)
- **Technological solutions?**
  - Game-theoretic methods (Bellucci et al., 2001)
  - Computational methods (Larson, 2010, Branting et al. 2023)
  - ...

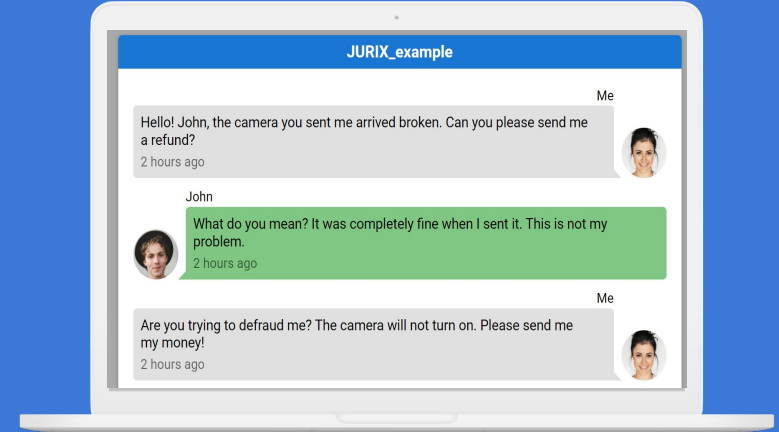
# Proposed Framework

2

# LLMediator Framework

- F1 - Reformulating inflammatory messages
- F2 - Drafting messages for the mediator
- F3 - Autonomously intervening in the negotiation?

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



# Evaluation

---

- Blind comparison with human mediators
  - Hard to maintain a structured message format
  - Hard to introduce the requirement within the prompt
- Two step evaluation approach
  - S1 - Decide intervention types
  - S2 - Draft intervention message



# Experiment Design

3

# Drafting Disputes scenarios

---

<b>Charac.</b>	<b>Explanation</b>	<b>Examples</b>
<b>Emotional</b>	The parties have strong emotional expressions in the conversation.	A person asks their neighbour to keep their dogs quiet, resulting in an escalating conversation with threats.
<b>Complex</b>	The dispute has a high degree of complexity and the facts of what happened are difficult to clarify.	A person asks an insurance company to pay for a car accident, resulting in a discussion of legal and technical nuances.
<b>Confusion</b>	The parties are confused, leading to difficulties in communication.	A customer and merchant disagree on the details of an undelivered order, leading to repeated requests for more information.
<b>Impossible</b>	The dispute features strong disagreements, resulting in a deadlock.	A customer requests a laptop to be repaired, but the manufacturer argues that the damage is caused by the user, refusing the warranty.
<b>Evidential</b>	The dispute centers around conflicting evidence or claims.	One party insists that an agreement regarding a computer sale was reached, while the other disagrees.

# 50 Disputes scenarios

---

Sender	Messages
A	Hi, I received the package I ordered from your store, but the item is damaged. I'd like to request a refund or a replacement.
B	We're sorry to hear that. However, our policy states that damage incurred during shipping is the responsibility of the courier service, not ours.
A	But as a customer, my transaction is with your store, not the courier. It's so ironic that you have this kind of service attitude, it's unacceptable. I am not only going to file a complaint with the Consumer Protection Service, but I am also going to write a bad review so that everyone will know what kind of business you are!
B	That's unfortunate. We understand your frustration, but we can only offer a discount on your next purchase. We can't control what happens during shipping.

# S1 - Intervention types

---

No.	Intervention Types
1	Encourage exchanges of information
2	Help the parties understand each other's views
3	Let the parties know that their concerns are understood
4	Promote a productive level of emotional expression
5	Lay out the differences in perceptions and interests
6	Identify and narrow issues
7	Help parties realistically evaluate alternatives to settlement
8	Suggest that the parties take breaks when negotiations reach an impasse
9	Encourage flexibility and creativity
10	Shift the focus from past to future
11	Shift the focus from one of blame to a creative exchange between the parties
12	Hold caucuses with each disputant if there is deadlock or a problem
13	Propose solutions that meet the fundamental interests of all parties

# S2 - Intervention message

---

<b>Dispute summary</b>	<b>Intervention types chosen</b>	<b>Interventions messages</b>
<p>Party A asks Party B to delete their picture from social media. B refuses and makes fun of A.</p>	<p><b>Human:</b> 4. Promote a productive level of emotional expression, 13. Propose solutions that meet the fundamental interests of all parties. <b>(preferred)</b></p>	<p><b>Human:</b> Let's not insult each other or downplay one's feelings. B, you must delete the picture, since keeping it without A's consent is illegal.</p>
	<p><b>LLM:</b> 2. Help the parties understand each other's views, 3. Let the parties know that their concerns are understood, 9. Encourage flexibility and creativity</p>	<p><b>LLM:</b> Let's ensure the conversation is respectful. Here's a proposition: Part B, how about posting an image that both of you find humorous and enjoyable instead? This way, Party A won't feel embarrassed and both of you can have fun. <b>(preferred)</b></p>

# Results

# Key results

---

- **E1 - Intervention type**
  - 62% of LLMs chosen types rated equal to or better than human messages.
- **E2 - Intervention message**
  - 84% of LLM-generated messages rated equal to or better than human messages.

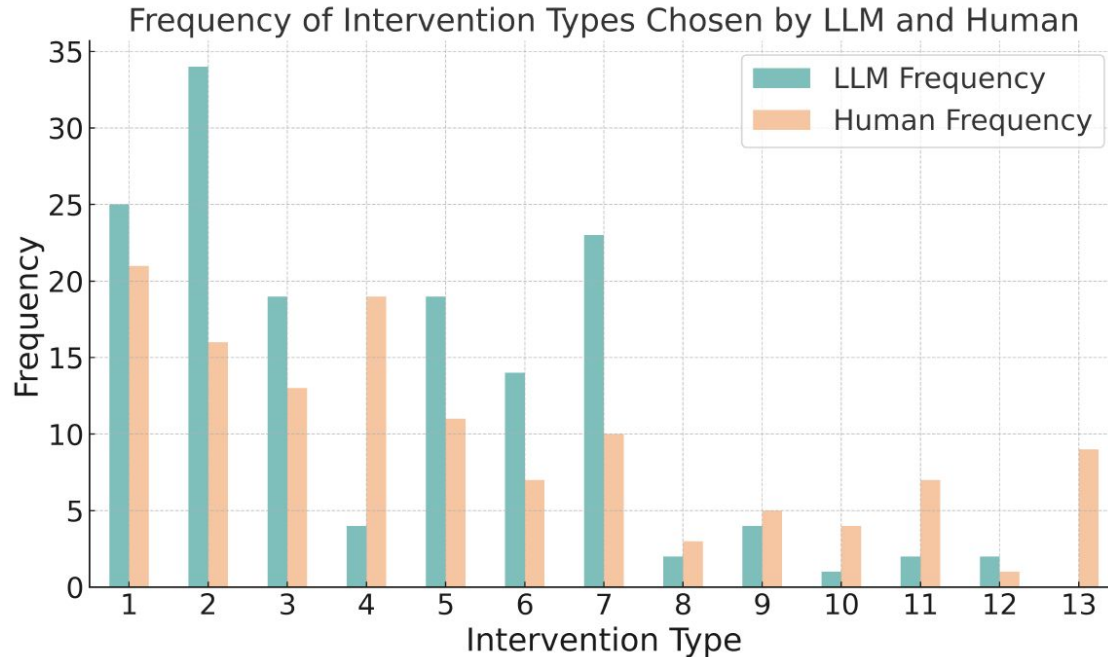
# E1 - Intervention types

---

Description	Number of responses
● LLM is significantly better than Human	11
● LLM is slightly better than Human	11
● LLM and human are about the same	9
● Human is slightly better than LLM	14
● Human is significantly better than LLM	5

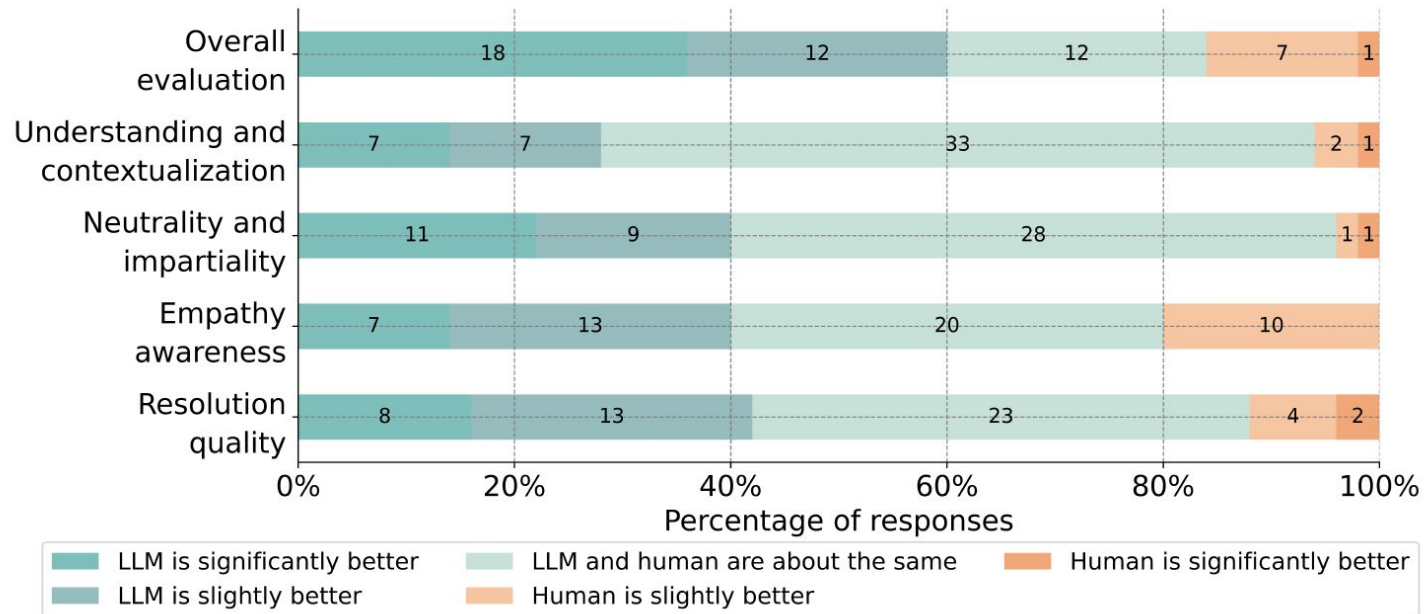


# E1 - Intervention types







Pezeshkpour, P., Hruschka, E.: Large language models sensitivity to the order of options in multiple choice questions (2023), <https://arxiv.org/abs/2308.11483>

# E2 - Intervention messages






# Discussion

# Takeaways

-  **1** LLMs provide clarity, consistent tone, provide more acceptable solutions.
-  **2** Humans more often misunderstood the dispute.
-  **3** LLMs have better performance in choosing intervention types and drafting messages.
-  **4** Assumption of humans as gold standard is facing challenges.

# Limitations

-  **1** Experimental setup differs from real-world mediation contexts.
-  **2** Annotators and evaluators have legal background but are not expert in mediation.
-  **3** Hard to tell which intervention is “better” objectively.

# Q&A

Any questions?