Legal Knowledge and Information Systems
J. Savelka et al. (Eds.)

doi:10.3233/FAIA241243

Robots in the Middle: Evaluating LLMs in Dispute Resolution

Jinzhe TAN [a,1], Hannes WESTERMANN [b], Nikhil Reddy POTTANIGARI [c], Jaromír ŠAVELKA [d], Sébastien MEEÙS [a,e], Mia GODET [a], and Karim BENYEKHLEF [a]

[a] *Cyberjustice Laboratory, University of Montreal, Canada*
[b] *Maastricht Law and Tech Lab, Maastricht University, Netherlands*
[c] *Mila - Quebec AI Institute, University of Montreal, Canada*
[d] *School of Computer Science, Carnegie Mellon University, United States*
[e] *Faculty of Law and Criminology, Université Libre de Bruxelles, Belgium*

Abstract. Mediation is a dispute resolution method featuring a neutral third-party (mediator) who intervenes to help the individuals resolve their dispute. In this paper, we investigate to what extent large language models (LLMs) are able to act as mediators. We investigate whether LLMs are able to analyze dispute conversations, select suitable intervention types, and generate appropriate intervention messages. Using a novel, manually created dataset of 50 dispute scenarios, we conduct a blind evaluation comparing LLMs with human annotators across several key metrics. Overall, the LLMs showed strong performance, even outperforming our human annotators across key dimensions. Specifically, in 62% of the cases, the LLMs chose intervention types that were rated as better than or equivalent to those chosen by humans. Moreover, in 84% of the cases, the intervention messages generated by the LLMs were rated as better than or equal to the intervention messages written by humans. LLMs likewise performed favourably on metrics such as impartiality, understanding and contextualization. Our results demonstrate the potential of integrating AI in online dispute resolution (ODR) platforms.

Keywords.
large language models, artificial intelligence, online dispute resolution, access to justice, ai & law, chatgpt

1. Introduction

Intermediaries, such as mediators, arbitrators, or conciliators, can play an important role in facilitating dispute resolution. When a discussion turns emotionally charged, communication breaks down, or the dispute reaches a deadlock, intermediaries can step in to provide information to help calm emotions, clarify facts, identify the key issues in the dispute, and propose settlement options, thereby promoting the satisfactory progress of dispute resolution.

The involvement of such intermediaries is, however, typically limited to cases where the value of dispute exceeds the cost of the intermediary. Further, in some areas, there

[1] Corresponding Author: Jinzhe Tan, University of Montreal, jinzhe.tan@umontreal.ca.

may not be enough trained facilitators to handle all disputes [5]. Supporting mediation through technological tools is thus a promising avenue of increasing the scalability of facilitated dispute resolution, and enabling its use in new contexts.

Recent advancements in large language models (LLMs) have opened the door to the use of AI to assist intermediaries in understanding dispute scenarios, offering AI-suggested interventions, and even AI automated interventions [29]. However, the complex, interactive, and interpersonal nature of dispute resolution sets a high bar for AI to effectively perform such tasks [17]. Intermediaries need a nuanced skill set, including contextual understanding, emotion perception, and the ability to propose balanced, contextually appropriate solutions. While LLMs have demonstrated considerable capabilities in discrete tasks (such as contextual awareness [7] and language understanding [19]), their performance in more complex, integrated tasks remains under-explored.

To investigate the performance of LLMs like `GPT-4o` in dispute resolution tasks, we analyzed their abilities in selecting *intervention types* and generating *intervention messages* based on *dispute scenarios* from a novel corpus of 50 hypothetical disputes. In this study, we aim to investigate three research questions:

RQ1 To what extent can LLMs select appropriate intervention types in response to a dispute scenario?
RQ2 How do LLMs compare to humans in drafting intervention messages?
RQ3 To what degree are messages generated by LLMs safe and free of hallucinations?

2. Related Work

The use of computational methods to facilitate dispute resolution is a long-standing topic in the field of Legal Informatics, such as in the ICANS system, where the parties can choose their preferences through a mathematical mechanism and gradually reach an agreement with the assistance of the system [27]. Using a similar idea, Family_winner uses a game-theoretic approach that allows users to split up and resolve disputes using repeated offers [2,31,3]. Other approaches include indicating potential court outcome ranges to align the expectations of the parties [6,30,25,4]. These approaches have laid the groundwork for applying technology to the mediation process.

In recent years, with advances in natural language processing (NLP), discussions and attempts to use language models to facilitate dispute resolution have emerged [17,18]. For example, Branting et al. use the example of Utah's ODR system to analyse how language models can be used to analyse the stages of a dispute and provide facilitators with recommendations based on standard text message [5].

The evolution of LLM marks a significant shift from earlier domain-specific models. LLMs have demonstrated strong foundational capabilities [1], contrary to conventional models that are specific to particular domains. LLMs excel in tasks such as language understanding and generation [8], sentiment analysis [32], and reasoning [13]. These foundational capabilities allow LLMs to be adapted to various domains through techniques such as fine-tuning or prompt engineering. This adaptability has already led to diverse applications in the legal field, including providing legal information [26,28], serving in fiduciary roles [21], conducting empirical research [11], analyzing legal text data [23,24,12], and developing legal expert systems [14].

The main direction of our exploration in this paper is to assess how well LLMs like `GPT-4o` perform in selecting intervention types and generating intervention messages. By analyzing these tasks, we contribute to understanding how LLMs can support intermediaries in legal contexts, without focusing on immediate real-world deployment.

3. Proposed Framework

We use the LLMediator framework presented in [29] to set up a dispute scenario involving two disputing parties and a mediator. The parties can communicate with each other through text messages, and the mediator can intervene in the dispute through intervention messages in order to assist the disputants in reaching a resolution. In the LLMediator framework, this functionality can be performed by a human, by a human assisted by an LLM, and potentially in a fully automated fashion [29]. In this paper, we investigate how well LLMs can select appropriate intervention types and generate corresponding messages, compare humans.

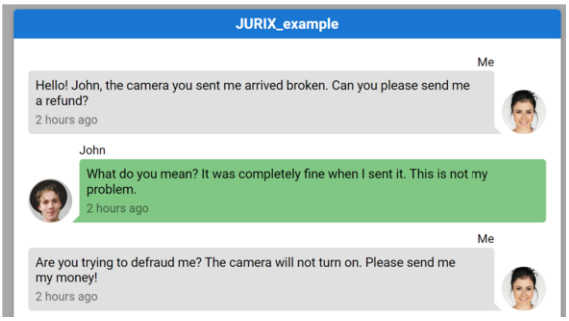

Figure 1. A screenshot from the LLMediator, showing a dispute prior to the mediator's intervention.

Figure 1 shows a screenshot of a dispute. At this point, the mediator may decide that it is time to intervene, in order to help the disputants find an amicable solution. When intervening, the mediator needs to perform two important steps:

Step 1 - Decide intervention type. Depending on the state of the discussion, the mediator may want to include different types of interventions in the messages they send. For example, they may want to calm down the discussion, encourage exchanges of information or help the parties evaluate their alternatives. We adopted a list of types of possible interventions from [15], as shown in Table 2. Deciding on the type of intervention requires an understanding of the dispute context and empathy towards the parties.

Step 2 - Draft the intervention message. The mediator has to decide which specific words to use to perform the types of intervention they have chosen. Clearly and empathically communicating these ideas is important to help the parties achieve their goals, while avoiding mistakes and ambiguities.

These steps need to be carried out implicitly by the mediator intervening in a dispute. We chose to adopt them as a framework to compare the messages created by human annotators and LLMs. By splitting the process into these two steps, we can compare the performance of the two across multiple steps, giving us a deeper understanding of their differences.

4. Experimental Design

Communications between parties during disputes often involve sensitive and private information leading to a scarcity of accessible data on dispute scenarios. Therefore, we manually constructed a dataset comprising 50 *dispute scenarios* for our experiments (Section 4.1). Afterwards, human mediators and LLMs intervened in the dispute scenarios (Sections 4.2 & 4.3), following the same instructions, and both were assessed in a blind evaluation for their performance in the same scenarios (Section 4.4).

4.1. Constructing disputes

The 50 dispute scenarios we created followed the same structure, with each scenario consisting of two dialogues between Party A and Party B, thus featuring a total of four textual messages. These scenarios were written by the authors of this paper, who have varying levels of legal expertise. In order to ensure a diverse set of disputes, we wrote disputes with varying characteristics, as described in Table 1. These dispute characteristics were intentionally chosen to cover a wide range of situations that challenge mediators in different ways. Specifically, they test the mediator's ability to recognize emotions, understand complex situations, clarify facts, and propose solutions.

After reviewing the dispute scenarios we created, we found them to be diverse in both communication style and legal areas, covering areas such as parcel delivery, land property rights disputes, noise complaints, and so on. This diversity contributes to helping us perform robust evaluations of the interventions.

Charac.	Explanation	Example
Emotional	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.
Complex	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.
Confusion	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.
Impossible	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.
Evidential	The dispute centers around conflicting evidence or claims.	One party insists that an agreement regarding a computer sale was reached, while the other disagrees.

Table 1. Description of dispute characteristics

4.2. Human interventions

After creating the dispute scenarios, interventions were manually drafted by the authors and six additional law students, who acted as annotators. During the intervention drafting phase, annotators did not annotate the dispute scenarios they had personally created, in order to avoid potential biases. For each dispute, the annotators followed the steps described above in Section 3. First, they selected one to three intervention types from

Table 2. We used the mediator's guide given on the website of Department of Justice of Canada [15] to create the prompt, which covers the disputed conversation and the 13 types of interventions. Second, they drafted intervention messages based on the chosen types. For coherence with the LLM-generated messages, the annotators maintained a one-to-one correspondence between each intervention type and the message, writing 1-2 sentences per intervention type.

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

Table 2. List of intervention types to facilitate mediation from [15]

4.3. LLM Interventions

For LLMs interventions, we used the `gpt-4o-2024-05-13` model via the `openai` Python library[2], which was the state-of-the-art model at the time of the experiment. The default parameters were used throughout the experiment. The full data, code, and prompts for reproducing this experiment are available in our GitHub repository.[3]

Step 1 - Decide intervention type. First, we asked the model to select between one and three intervention types to respond to a provided dispute. (see Table 2)

Step 2 - Generate intervention message. We then provided *human-selected* intervention types as inputs to the models and asked them to write intervention messages based on the intervention types. The LLM was also asked to maintain the correspondence between the intervention types and the text (see Section 4.2). Table 3 shows an example of chosen intervention types and resulting messages from both humans and LLMs.

We always use the intervention types chosen by the human annotator to allow for a direct comparison of the quality of the written intervention. Thus, we are able to compare LLMs to humans on two tasks: choosing the appropriate intervention types, and generating an intervention message based on chosen intervention types. Our choice of using the human messages does not imply that we considered the intervention types selected by the humans to be superior to those selected by the LLMs. In fact, during evaluation, we found that the evaluators often preferred the intervention types selected by the LLMs.

[2]Github: OpenAI Python Library. Available at: https://github.com/openai/openai-python [Accessed 2024-08-26]

[3]https://github.com/JinzheTan/Robots_in_the_middle

Dispute summary	Intervention types chosen	Interventions
Party A asks Party B to delete their picture from social media. B refuses and makes fun of A.	**Human**: 4. Promote a productive level of emotional expression, 13. Propose solutions that meet the fundamental interests of all parties. **(preferred)**	**Human:** Let's not insult each other or downplay's one feelings. B, you must delete the picture, since keeping it without A's consent is illegal.
	LLM: 2. Help the parties understand each other's views, 3. Let the parties know that their concerns are understood, 9. Encourage flexibility and creativity	**LLM:** Let's ensure the conversation is respectful. Here's a proposition: Party 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. **(preferred)**

Table 3. Example of selected intervention types and written interventions, both by human and LLM. Here, the evaluator preferred the human choice of intervention types, but the LLM-generated intervention message.

4.4. Evaluation

E1 - Evaluation of intervention types. After obtaining the intervention types chosen by humans and LLMs based on the dispute scenarios (as described in Section 4.2 & 4.3), we conducted a blind evaluation of the chosen types. We asked evaluators to compare the two intervention type choices after reading the dispute scenario and to rate which choice they found superior on a 5-point Likert scale.

E2 - Evaluation of intervention messages. Afterward, we assigned another evaluator to each dispute. We asked them to assess (in a blind fashion) which of the two intervention messages they preferred. The evaluators first provided their overall preference using a 5-point Likert scale, and wrote comments motivating their choice. Then, they compared the messages in terms of specific evaluation criteria, including understanding and contextualization, neutrality and impartiality, empathy awareness, and resolution quality. After completing the evaluation based on these criteria, the evaluators were asked to write additional notes highlighting any notable points.

E3 - Safety evaluation of LLM interventions. Finally, we conducted safety and quality checks of the messages generated by the LLMs. We assessed whether the model hallucinated and whether there were any safety issues with the generated messages.

5. Results

5.1. E1 - Evaluation of intervention types

While multiple intervention type choices could be suitable for each dispute scenario, some options may be more suitable depending on the context. For example, if the parties use impolite language or express strong emotions, selecting intervention type No. 4, 'Promote a productive level of emotional expression,' would be more appropriate. In situations where negotiations are at a deadlock, choosing intervention type No. 8, 'Suggest that the parties take breaks when negotiations reach an impasse,' or No. 12, 'Hold caucuses with each disputant if there is deadlock or a problem,' would be more fitting.

Table 4 shows the results of the blind evaluation. We found that evaluators generally preferred the intervention types chosen by LLMs. However, there were instances when

these choices showed strong variance. Figure 2 shows the distribution of intervention types chosen by humans, compared to those suggested by LLMs.

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

Table 4. We used a 5 point Likert scale to compare human evaluators' preferences for LLM and human-selected intervention types.

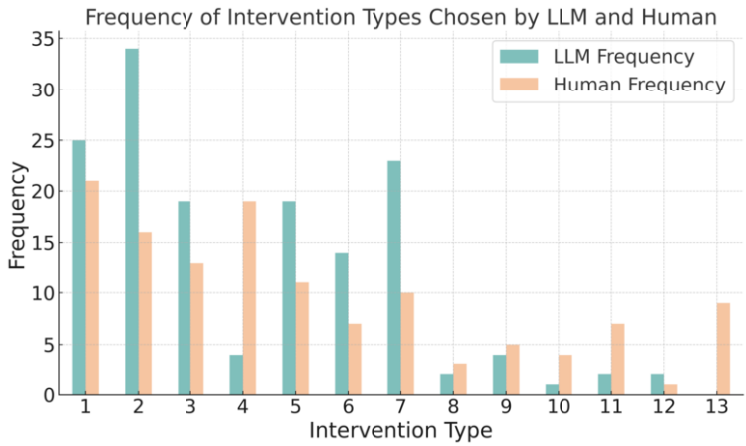

Figure 2. Frequency of intervention types chosen by LLM and human

5.2. *E2 - Evaluation of intervention messages*

Figure 3 shows the blind evaluation preferences of the evaluators on the different axes between the human and LLM-generated messages. As we can see, there was a strong preference for the messages written by the LLMs, across the all categories. In terms of overall evaluation, 84% of evaluators believed that the intervention messages generated by LLMs were either superior or equivalent to those created by human mediators, with LLMs significantly or slightly outperforming humans in 60% of cases. Further, the LLM-generated messages were scored as equal to or better than human messages in between 80% and 96% of the cases in all categories (see Figure 3).

5.3. *E3 - Safety evaluation of LLM-generated intervention*

After manually checking all the LLM-generated messages, we did not find the appearance of harmful messages and hallucinations in the scenario of this experiment.

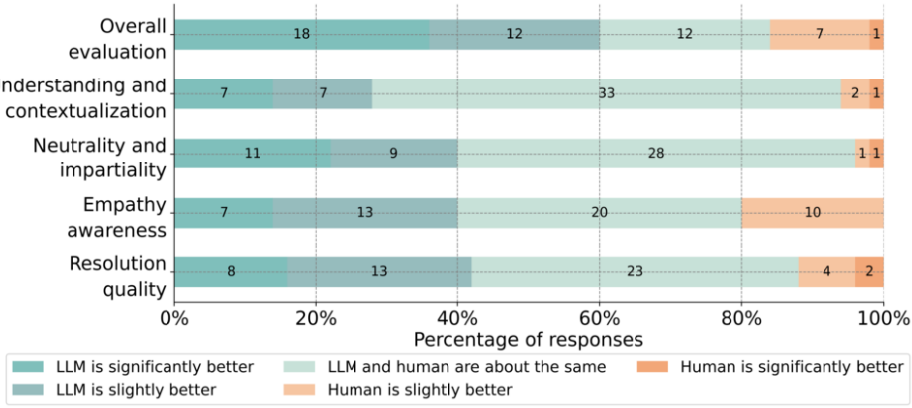

Figure 3. The bar chart shows the distribution of responses evaluating the performance of LLMs compared to humans across the five metrics.

6. Discussion

6.1. RQ1 - To what extent can LLMs select appropriate intervention types in response to a dispute scenario?

Table 4 shows that our human evaluators preferred the intervention types chosen by the LLMs in 22 cases, were ambivalent in 9 cases, and preferred the human-chosen types in 19 cases. Overall, this is a strong result suggesting viability of LLMs on this complex task, requiring nuanced understanding of a dispute and empathy to determine next steps. However, it's important to note that the evaluators were not expert mediators, and derermining the most appropriate intervention type may be subjective (see Section 7).

Figure 2 shows the distribution of intervention types chosen by humans and LLMs. Here, differing patterns are revealed. The top three types of interventions chosen by the LLMs are helping the parties to understand each other's views, the encouragement of exchanging information and the helping of parties to evaluate alternatives (2,1,7). The human annotators, on the other hand, preferred the encouragement of exchanging information, the promotion of a productive level of expression, and helping the parties understand each others views (1,4,2).

These preferences may reveal a different understanding of what is important in mediation. Additionally, prior work has shown that LLMs may be affected by the order of presented options [22]. This may partially explain why the LLMs seem to prefer the early items in the list, although a similar preference also seems present for the human mediators.

6.2. RQ2 - How do LLMs compare to humans in drafting intervention messages?

Our experiment shows that LLMs can perform at a level comparable to or even better than our human annotators. The LLM-generated messages were rated higher or equal to the human-written ones in 84% of the scenarios. While certain caveats apply (see Section 7), these results highlight the impressive capability of LLMs in drafting appropriate

intervention messages. Various reasons were given as to why the evaluators preferred the messages written by the LLM.

First, the evaluators often found the LLM-generated messages to be more smooth and clear than the human-written ones. The general tone used by LLMs, involving frequent messages such as "I completely understand" or "It seems like there are problems," seems to work well in a mediation context, and may have contributed to high scores.

Second, while LLMs are known to frequently "hallucinate" information [10,9], in our case the humans more often misunderstood the dispute or were confused about the party intentions or factual occurrences. This could be due to factors such as fatigue, emotional bias, or a misunderstanding of the role of the mediator. In contrast, LLMs demonstrated consistent and coherent interventions across multiple cases, with fewer instances of judgment errors.

Third, we found that human annotators would often propose very specific solutions or even indicate fault, which received a lower rating as it may not be appropriate for the role of the mediator.

Overall, while it is important to highlight the caveat of none of the annotators and evaluators having experience in mediation, it seems like the messages generated by the LLMs capture the dispute well, use an appropriate tone, are clear and do not overreach, making them compare favourably to the messages written by our human annotators.

6.3. RQ3 - To what degree are messages generated by LLMs safe and free of hallucinations?

We did not observe any unsafe messages or hallucinated information in the generated messages. While this of course does not rule out such issues, it is nonetheless a promising result for the use of LLMs in a dispute resolution context. The approach discussed in [29], where the generated messages are provided to a human mediator before being sent to the parties, could further mitigate such concerns.

6.4. On the use of gold-standard data

Using human-generated answers as Ground Truth ('gold standard') is a common practice in machine learning research, which helps us create benchmarks for evaluating the performance of algorithms or models. Here, we took a different approach by comparing human-written messages to those generated by LLMs, rather than assuming that the human data represented a reliable gold standard. With good reason - looking at the results, the LLM generated messages were consistently rated higher than the messages written by the humans.

However, the results also reveal the general difficulty of evaluating models that can perform complex, nuanced tasks without giving obviously wrong answers. None of the messages written by the LLM contained any hallucinations or other obvious defects, which makes the overall assessment difficult and subjective. Perhaps, as discussed in [20], it is more useful to see the annotations as surveys of individual views, rather than a single "truth", when it comes to bespoke and nuanced legal tasks. Regardless, the science of evaluating large language models on legal tasks is in its infancy, and we hope that this paper can contribute some insights to this important issue.

7. Limitations

In this work, we used a structured evaluation method to compare the performance of LLMs to humans. While the results are promising, there are some important caveats. *First*, the process of selecting an intervention type and then being bound to it may not correspond to how mediators approach drafting messages in reality. Likewise, the drafting of messages in blocks organized by intervention types may also impose artificial limitations on the types of interventions that can be written.

Second, our disputes and messages were drafted and evaluated by people without specific training in mediation, and none of whom are native English speakers. This may give an advantage to the LLMs. While it seems like the ability of the LLM to select intervention types and write messages is favourable to that of average people, this paper cannot tell us about how trained mediators would approach these issues.

Third, as touched upon in Section 6.4, it may not be possible to assess which intervention type or message is "better" without observing real-world outcomes, leading to a subjective assessment. For example, it is possible that grammar mistakes and our expectations of the tone of the mediation message played an exaggerated role in our comparison of the messages, which may not make a big difference in a real context.

Fourth, our experimental design assumes that there are always 4 messages, and that the mediator should intervene next. It does not include the messages after the intervention, or the important choice on when to intervene (c.f. [5]),

While these choices were made to enable the assessment of LLMs in mediation, they also somewhat limit the general applicability of the results. Future work should focus on evaluating such tools in real-world contexts, and involve expert mediators, in order to achieve a higher "construct validity," i.e., be more closely aligned with real-world outcomes (c.f. [16]).

8. Conclusion & Future Work

In this study, we demonstrated that large language models possess significant potential in mediating disputes, performing on par or even surpassing our human annotators in selecting appropriate intervention types and crafting effective intervention messages. These findings suggest that LLMs could play a pivotal role in enhancing online dispute resolution platforms by providing scalable and cost-effective mediation services.

Our research contributes to the growing body of knowledge on AI applications in law and dispute resolution, highlighting the capabilities of LLMs in understanding complex human interactions and responding with empathy and neutrality. This advancement could significantly improve access to justice, particularly in cases where traditional mediation is inaccessible due to cost or availability constraints.

Future work should incorporate multi-modal data to better simulate real-world mediation scenarios, and conduct pilot studies within actual ODR systems to assess practical effectiveness. By continuing to refine these technologies, we move closer to a future where AI not only supports but enhances the human capacity for dispute resolution, contributing to a more accessible and efficient justice system.

Acknowledgements We acknowledge the generous support from the Cyberjustice Laboratory, LexUM Chair, and Autonomy through Cyberjustice Technologies project.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[2] Bellucci, E., Zeleznikow, J.: Representations of decision-making support in negotiation. Journal of decision systems **10**(3-4), 449–479 (2001)

[3] Bellucci, E., Zeleznikow, J.: Developing negotiation decision support systems that support mediators: a case study of the family_winner system. Artificial Intelligence and Law **13**, 233–271 (2005)

[4] Benyekhlef, K., Zhu, J.: Intelligence artificielle et justice: justice prédictive, conflits de basse intensité et données massives. Intelligence **30**(3) (2018)

[5] Branting, K., McLeod, S., Howell, S., Weiss, B., Profitt, B., Tanner, J., Gross, I., Shin, D.: A computational model of facilitation in online dispute resolution. Artificial Intelligence and Law **31**(3), 465–490 (2023)

[6] Carneiro, D., Novais, P., Andrade, F., Zeleznikow, J., Neves, J.: Online dispute resolution: an artificial intelligence perspective. Artificial Intelligence Review **41**, 211–240 (2014)

[7] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology **15**(3), 1–45 (2024)

[8] Chen, X., Ye, J., Zu, C., Xu, N., Zheng, R., Peng, M., Zhou, J., Gui, T., Zhang, Q., Huang, X.: How robust is GPT-3.5 to predecessors? a comprehensive study on language understanding tasks. arXiv preprint arXiv:2303.00293 (2023)

[9] Dahl, M., Magesh, V., Suzgun, M., Ho, D.E.: Large legal fictions: Profiling legal hallucinations in large language models. Journal of Legal Analysis **16**(1), 64–93 (2024)

[10] Das, S., Saha, S., Srihari, R.K.: Diving deep into modes of fact hallucinations in dialogue systems. arXiv preprint arXiv:2301.04449 (2023)

[11] Drápal, J., Westermann, H., Savelka, J.: Using large language models to support thematic analysis in empirical legal studies. arXiv preprint arXiv:2310.18729 (2023)

[12] Gray, M., Savelka, J., Oliver, W., Ashley, K.: Using LLMs to discover legal factors. arXiv preprint arXiv:2410.07504 (2024)

[13] Huang, J., Chang, K.C.C.: Towards reasoning in large language models: A survey. arXiv preprint arXiv:2212.10403 (2022)

[14] Janatian, S., Westermann, H., Tan, J., Savelka, J., Benyekhlef, K.: From text to structure: Using large language models to support the development of legal expert systems. In: Legal Knowledge and Information Systems, pp. 167–176. IOS Press (2023)

[15] Department of Justice, C.: Dispute Resolution Reference Guide: Practice Module 2. Dispute Resolution Series, Department of Justice, Canada (Jan 2007), https://www.justice.gc.ca/eng/rp-pr/csj-sjc/dprs-sprd/res/drrg-mrrc/04.html, last Modified: August 25, 2022

[16] Kapoor, S., Henderson, P., Narayanan, A.: Promises and pitfalls of artificial intelligence for legal applications. arXiv preprint arXiv:2402.01656 (2024)

[17] Larson, D.A.: Artificial intelligence: Robots, avatars, and the demise of the human mediator. Ohio St. J. on Disp. Resol. **25**, 105 (2010)

[18] Larson, D.A.: Brother, can you spare a dime-technology can reduce dispute resolution costs when times are tough and improve outcomes. Nev. LJ **11**, 523 (2010)

[19] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)

[20] Ma, M., Waldon, B., Nyarko, J.: Conceptual questions in developing expert-annotated data. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. pp. 427–431 (2023)

[21] Nay, J.J.: Large language models as fiduciaries: a case study toward robustly communicating with artificial intelligence through legal standards. arXiv preprint arXiv:2301.10095 (2023)

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

[23] Savelka, J., Ashley, K.D., Gray, M.A., Westermann, H., Xu, H.: Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? ASAIL'23: 6th Workshop on Automated Semantic Analysis of Information in Legal Text (2023)

[24] Savelka, J., Ashley, K.D., Gray, M.A., Westermann, H., Xu, H.: Explaining legal concepts with augmented large language models (GPT-4). arXiv preprint arXiv:2306.09525 (2023)

[25] Susskind, R.: Online courts and the future of justice. Oxford University Press (2019)
[26] Tan, J., Westermann, H., Benyekhlef, K.: ChatGPT as an artificial lawyer? Artificial Intelligence for Access to Justice (AI4AJ 2023) (2023)
[27] Thiessen, E.M.: ICANS: An interactive computer-assisted multiparty negotiation support system. Cornell University (1993)
[28] Westermann, H., Meeùs, S., Godet, M., Troussel, A.C., Tan, J., Savelka, J., Benyekhlef, K.: Bridging the gap: Mapping layperson narratives to legal issues with language models. In: ASAIL@ ICAIL. pp. 37–48 (2023)
[29] Westermann, H., Savelka, J., Benyekhlef, K.: LLMediator: GPT-4 assisted online dispute resolution. Artificial Intelligence for Access to Justice (AI4AJ 2023) (2023)
[30] Zeleznikow, J.: Can artificial intelligence and online dispute resolution enhance efficiency and effectiveness in courts. In: IJCA. vol. 8, p. 30. HeinOnline (2016)
[31] Zeleznikow, J., Bellucci, E.: Family_winner: integrating game theory and heuristics to provide negotiation support. In: Proceedings of sixteenth international conference on legal knowledge based system. pp. 21–30. IOS Publications Amsterdam (2003)
[32] Zhang, W., Deng, Y., Liu, B., Pan, S.J., Bing, L.: Sentiment analysis in the era of large language models: A reality check. arXiv preprint arXiv:2305.15005 (2023)