Simulating Dispute Mediation with LLM-Based Agents for Legal Research

Junjie Chen[1,2], Haitao Li[1,2], Minghao Qin[3], Yujia Zhou[1,2], Yanxue Ren[4], Wuyue Wang[5], Yiqun Liu[1,2], Yueyue Wu[1,2*], Qingyao Ai[1,2*]

[1]Department of Computer Science and Technology, Tsinghua University
[2]Institute for Internet Judiciary, Tsinghua University
[3]China University of Political Science and Law
[4]Artificial Intelligence Laboratory, Bosera Asset Management Co., Ltd.
[5]University of Notre Dame
wuyueyue1600@gmail.com, aiqy@tsinghua.edu.cn

Abstract

Legal dispute mediation plays a crucial role in resolving civil disputes, yet its empirical study is limited by privacy constraints and complex multivariate interactions. To address this limitation, we present *AgentMediation*, the first LLM-based agent framework for simulating dispute mediation. It simulates realistic mediation processes grounded in real-world disputes and enables controlled experimentation on key variables such as disputant strategies, dispute causes, and mediator expertise. Our empirical analysis reveals patterns consistent with sociological theories, including Group Polarization and Surface-level Consensus. As a comprehensive and extensible platform, *AgentMediation* paves the way for deeper integration of social science and AI in legal research.

Code — https://github.com/cjj826/AgentMediation

Introduction

Legal dispute mediation, as a key form of alternative dispute resolution (ADR) (Mnookin 1998), plays an important role in the resolution of civil disputes worldwide. Compared to litigation, mediation is generally less time-consuming and more cost-effective, offering a promising approach to reducing court caseloads and accelerating the dispute resolution process (Sherman and Momani 2025).

Much research in law and sociology has tried to investigate factors affecting the effectiveness of dispute mediation, including the context of the dispute, the procedural design of the mediation process, the behavioral strategies of the parties to the dispute (disputants), and the domain expertise of mediators (Hsieh et al. 2022). However, systematically exploring dispute mediation remains challenging due to two primary factors: (1) privacy constraints, i.e., most mediation processes are confidential and not publicly accessible, making it difficult to collect large-scale real-world data for statistical analysis or experimental validation; and (2) multivariate interactions, i.e., dispute mediation outcomes are jointly affected by multiple variables, which are hard to isolate and quantify their individual effects in real-world settings.

Recent advances in LLM-based agent simulations (Park et al. 2023; Ashery, Aiello, and Baronchelli 2025) have shown that multi-agent systems can serve as controllable and observable social laboratories for modeling complex human interactions. More specifically, studies like AgentsCourt (He et al. 2024) demonstrate the feasibility of simulating legal scenarios with LLM-based agents. However, simulating dispute mediation with LLM-based agents remains a non-trivial task because of three reasons: (1) the scarcity of high-quality and publicly available mediation data, which limits the realistic and effective simulations; (2) the absence of a comprehensive and structured framework to model key factors in mediation and analyze their impact on outcomes; and (3) the lack of reliable and efficient quantitative evaluation methods for assessing mediation performance.

To this end, we introduce *AgentMediation* (Figure 1), the first LLM-based agent framework for simulating dispute mediation that offers a controllable, reproducible and extensible platform. Specifically, this consists of three core components: data preprocessing, mediation simulation framework, and evaluator. First, to establish a reliable and open data foundation, we processed 330 civil disputes from the *Dispute Resolution Case Database*[1], an authentic yet semi-structured dataset released by the Supreme People's Court of China after ten months of collection, with the mediation process omitted to preserve privacy. We applied automatic extraction followed by manual refinement to construct structured representations that support realistic mediation simulation and evaluation. Second, building on the above data, *AgentMediation* provides a unified and controllable framework that follows a general five-stage mediation process inspired by the Harvard handbook of dispute resolution (HDR) (Moffitt and Bordone 2012). We model three core aspects of mediation dynamics based on the HDR framework, namely the disputant behavior, dispute causes, and mediator expertise. Third, to support reliable and efficient evaluation, we design a dual-perspective assessment framework that captures mediation outcomes and solution quality. From the perspective of dispute mediation outcome, we adopt metrics such as mediation success rate, participant satisfaction, consensus level, and litigation risk. From the perspective of dispute mediation solution, we assess how well the media-

*Yueyue Wu and Qingyao Ai are the corresponding authors.

[1]open at https://dyjfalk.court.gov.cn/site

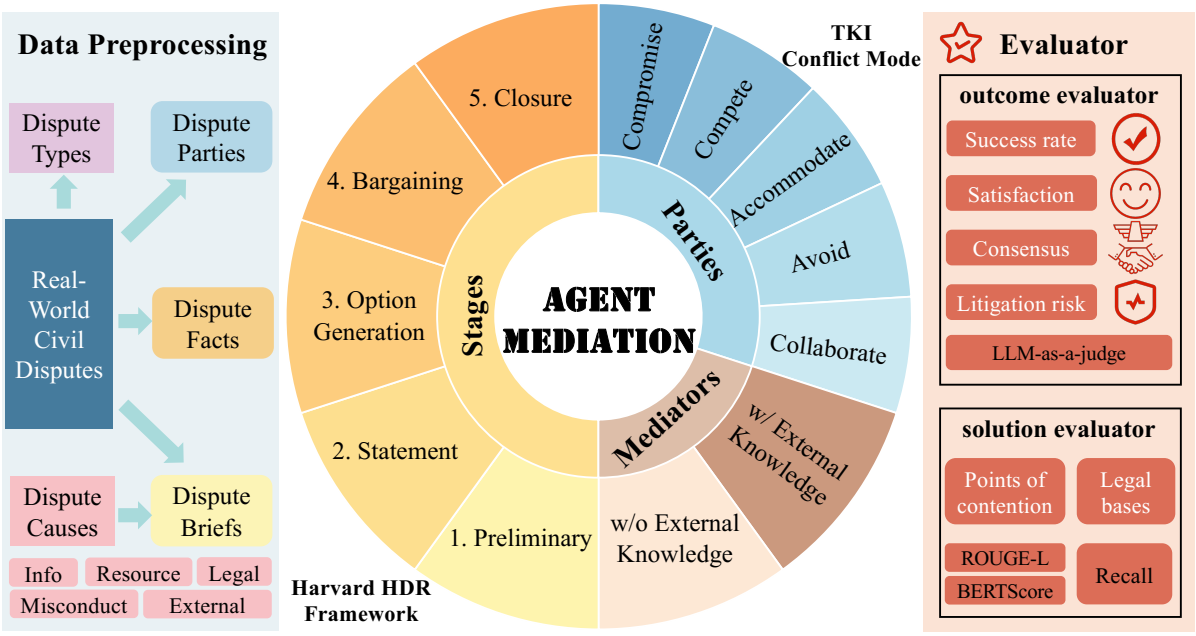

Figure 1: Overview of the *AgentMediation* framework: (1) data preprocessing from real-world civil disputes to extract structured inputs for simulation; (2) a five-stage mediation process inspired by the Harvard HDR framework; (3) configurable agent roles, including parties with TKI-based conflict modes (Thomas 2008) and mediators with or without access to external knowledge; and (4) a dual evaluation system for assessing both mediation outcomes and solution quality.

tors identify the points of contention and provide sound legal bases when necessary.

Our contributions fall into three aspects:

1. Comprehensive Framework: We build *AgentMediation*, the first framework that uses LLM-based agents to simulate full dispute mediation processes, enabling fine-grained and dynamic analysis of how various factors affect mediation outcomes. Through carefully designed human evaluations and annotations, we show the effectiveness and reliability of our system in simulating real dispute mediation processes.

2. Large-Scale and Extensible Dataset: Based on 330 real civil disputes, we create a large-scale dataset with over 14,000 simulated mediation processes, including detailed dialogues and solutions. Our framework further supports on demand generation of diverse scenarios for future research.

3. Theory-Aligned Empirical Insights: Using the *AgentMediation*, we find results that match social science theories, including Group Polarization (Moscovici and Zavalloni 1969), Surface-level Consensus (Kelman 1958; Habermas 1985), Moral Emotions Theory (Haidt et al. 2003), Realistic Conflict Theory (Sherif et al. 1961), and the Common In-Group Identity Model (Gaertner et al. 1993). These insights demonstrate the potential of LLM-based agent simulation to capture legal-social dynamics, contributing to the development of future intelligent mediation systems.

Related Work

Research on Legal Dispute Mediation. Dispute mediation plays a crucial role in resolving civil disputes efficiently and reducing judicial burdens (Mnookin 1998; Sherman and Momani 2025). Prior studies in law, sociology, and computational methods have tried to examine how factors such as dispute context, procedural design, party behavior, and mediator expertise affect mediation outcomes (Hsieh et al. 2022). However, existing research often relies on datasets that are either private or limited in scale (Tan et al. 2024), restricting reproducibility and broader empirical analysis. In addition, these datasets are typically static and non-scalable (Chawla et al. 2021; Hale et al. 2025), making it difficult to capture and disentangle the effects of complex, interacting variables. As a result, studying the dynamic nature of mediation processes remains a significant challenge.

LLM-Based Agents. Recent work has shown that LLMs can serve as autonomous agents capable of reasoning, role-playing, and interacting within multi-agent settings. Studies such as Generative Agents (Park et al. 2023) simulates social dynamics in virtual towns, while AgentsCourt (He et al. 2024) models courtroom debates. These advances (Jin et al. 2024; Wu et al. 2023) reveal LLMs' potential to reproduce complex human behaviors in controlled, repeatable settings. However, as Table 1 shows, using LLM-based agents to simulate dispute mediation remains highly challenging, primarily due to the scarcity of real-world data, the lack of frameworks tailored to mediation, and the absence of reliable and efficient quantitative evaluation methods.

Methodology

As Figure 1 shows, *AgentMediation* is designed as an extensible testbed to study the impact of various factors on mediation outcomes. In this section, we introduce the details

Feature	GENTEEL NEGOTIATOR (Priya et al. 2025)	LLM-Stakeholders (Abdelnabi et al. 2024)	WarAgent (Hua et al. 2023)	AgentsCourt (He et al. 2024)	AgentMediation (Ours)
Negotiation Dialogue Generation	✔	✔	✔	✔	✔
Multi-Agent Simulation	✘	✔	✔	✔	✔
Real-World Civil Disputes	✘	✘	✘	✘	✔
Mediation Simulation Framework	✘	✘	✘	✘	✔
Mediation Evaluation Metrics	✘	✘	✘	✘	✔

Table 1: Comparison of existing LLM-based agent simulation works. Current approaches are limited in modeling the dynamics of dispute mediation, which our approach is designed to address.

Attribute	Value	Attribute	Value
# of D	330.00	Avg. DB Len.	160.0
# of DT	76.00	Avg. # of DF	4.72
Min # of DP	2.00	Avg. # of $DBases$	3.42
Max # of DP	6.00	Avg. $DPoints$ Len.	81.35

Table 2: Summary statistics of our dataset. "#" denotes "number"; "Len." denotes length in Chinese characters.

of *AgentMediation*, focusing on five core components: Data Preprocessing, Mediation Process, Roles, Dispute Causes Taxonomy, and Evaluator.

Data Preprocessing

As previously discussed, we use the *Dispute Resolution Case Database* as the foundation of our study. We denote this data resource as $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$, where each sample D_i represents a civil dispute. Each dispute D_i follows a fixed schema and can be formalized as:

$$D_i = (\text{Title}_i, \text{Keywords}_i, \text{Brief}_i, \text{Method}_i, \text{Bases}_i), \quad (1)$$

where Title_i is the case title, Keywords_i is a set of descriptive tags, Brief_i provides the case background, Method_i outlines the resolution approach (excluding detailed mediation dialogues), and Bases_i contains the corresponding legal bases (i.e., statutory provisions) annotated by human experts.

To prepare the data for simulation, we use GPT-4o (2024-08-06) (Achiam et al. 2023) to extract key components, followed by manual verification to ensure quality. We first extract the dispute type (DT_i) from the Keywords_i. Then the Brief_i is used directly as the dispute brief (DB_i) for our mediation simulation. From the Method_i, we extract objective factual statements as dispute facts (DF_i), providing supporting evidence for the mediator. Simultaneously, we identify the dispute parties (DP_i) from the Brief_i for role-based interaction modeling. In addition, we extract the points of contention ($DPoints_i$) from Brief_i and Method_i, and normalize the Bases_i into a structured form ($DBases_i$). The overall data preprocessing can be summarized as:

$$Di \xrightarrow{\text{Preprocess}} \big(DT_i,\ DB_i,\ DF_i,\ DP_i,\ DPoints_i,\ DBases_i\big). \quad (2)$$

Table 2 presents the detailed statistics of the final preprocessed dataset. It is worth noting that *AgentMediation* also supports user-defined disputes, allowing users to conveniently simulate customized mediation scenarios based on their specific needs. A concrete example can be found in Appendix Section 1.

Mediation Process

Inspired by the Harvard HDR framework (Moffitt and Bordone 2012), *AgentMediation* adopts a structured five-stage pipeline to simulate the mediation process:

I. Preliminary. The system presents the dispute brief (DB) to help the mediators and parties build a shared understanding of the dispute.

II. Statement. All parties and the mediators introduce themselves and state their positions and claims, preparing for further discussion.

III. Option Generation. The mediators need to gather information and verify key facts. In our simulation, the preprocessed dispute facts (DF) are provided to simulate the outcome of the mediator's fact-finding process (our system also supports user-defined dispute facts). Based on DF, the mediators are required to generate a preliminary proposal that includes the points of contention ($MDPoints$), supporting legal bases ($MDBases$), and a solution.

IV. Bargaining. The system simulates multi-turn interactions where parties and mediators discuss the proposed resolution. Participants may agree, disagree, or propose alternatives, mimicking real-world negotiation dynamics.

V. Closure. The bargaining stage ends when one of the following conditions is met: an agreement is reached, a disagreement is confirmed, or the round limit is reached. The mediator is then required to output a final solution.

Roles

In our *AgentMediation* framework, we model two key roles: dispute parties (disputants) and mediators.

Parties. Parties serve as the primary roles for expressing disputes in the mediation process. Each party represents a distinct position, which may correspond to an individual, a group, or an institution. Even under the same dispute context, parties adopting different behavioral strategies can lead to significantly different mediation outcomes. To systematically investigate the impact of behavioral strategies on mediation performance, we follow the Thomas-Kilmann Conflict Mode Instrument (TKI) (Thomas 2008), which characterizes individual behaviors in dispute scenarios along two dimensions: assertiveness (the degree to which one seeks to satisfy

their own concerns, α) and cooperativeness (the degree to which one attempts to satisfy others' concerns, β). Based on different levels of α and β (High: H, Moderate: M, Low: L), five typical behavioral strategies are defined:
(1) **Competing** (H α, L β): prioritizing self-interest and often dominating others;
(2) **Collaborating** (H α, H β): seeking win-win outcomes through deep engagement;
(3) **Compromising** (M α, M β): aiming for middle-ground solutions through mutual concession;
(4) **Avoiding** (L α, L β): withdrawing from conflict;
(5) **Accommodating** (L α, H β): placing others' interests ahead of one's own.

Mediators. Mediators play a central role in guiding the mediation process, facilitating communication, and proposing resolutions. Among the various factors that influence mediation quality, domain expertise is particularly crucial, especially for accurately interpreting and citing relevant legal bases. To investigate the impact of domain expertise, we simulate mediators with varying levels of domain expertise by controlling their access to an external legal knowledge source. In the *without External Knowledge (w/o EK)* condition, mediators rely solely on the case context and their internal knowledge to cite legal bases. In the *w/ EK* condition, mediators have access to a retrieval module[2] that takes case descriptions as input and returns relevant legal bases.

Dispute Causes Taxonomy

To address the limitations of traditional dispute type (DT) labels (e.g., housing contract), which typically reflect only surface-level dispute categories and fail to support the fine-grained causal reasoning required in mediation processes, we construct a dispute causes taxonomy. Specifically, we first design an initial framework based on existing research (Vilendrer Law 2023). Building on this framework, we leverage GPT-4o to classify the underlying causes of 1,080 real-world civil court cases[3] based on their factual descriptions. When the existing taxonomy fails to accommodate certain cases, we dynamically extend the category set. The resulting taxonomy is then reviewed and refined by legal experts to ensure its completeness and interpretability. The final taxonomy consists of five top-level categories and twenty-nine subcategories. The top-level categories include: Information Conflict, Resource Conflict, Behavioral Misconduct, Legal Issues, and External Factors (e.g., natural disasters) (see Appendix Section 2 for details).

Importantly, this is not a single-label classification task but rather a multi-causal attribution process: a single case often involves multiple interrelated causes. Based on this taxonomy, we further investigate how different dispute causes impact the effectiveness of mediation.

Evaluator

Our evaluation module includes two components: an outcome evaluator for assessing mediation results and a solution evaluator for measuring solution quality.

[2]we use the API: https://legalone.com.cn/

[3]from https://wenshu.court.gov.cn/

Outcome Evaluator. To assess mediation outcomes, we adopt four complementary evaluation metrics: Success Rate, Satisfaction, Consensus, and Litigation Risk, each reflecting a distinct evaluative dimension. These metrics fall into two categories based on their perspective: Success Rate and Satisfaction are *subjective metrics*, while Consensus and Litigation Risk are *objective metrics*. (1) **Success Rate** (SR) captures explicit acceptance: at the end of each mediation, all disputants are required to indicate their stance on the final solution by selecting from {`Accept`, `Uncertain`, `Reject`}. A mediation is considered successful only if all parties choose `Accept`. Let N_{succ} and N_{tot} denote the number of successful and total mediation cases, respectively. We define the success rate as:

$$SR = \frac{N_{\text{succ}}}{N_{\text{tot}}} \times 100\%. \tag{3}$$

(2) **Satisfaction** (Sat) reflects each party's subjective evaluation of the outcome. After the mediation concludes, each disputant rates their level of satisfaction using a five-point Likert scale (Jebb, Ng, and Tay 2021), with labels {`Very Low`, `Low`, `Medium`, `High`, `Very High`} mapped to scores {0, 1, 2, 3, 4}. The final satisfaction score is computed as a normalized weighted average:

$$Sat = \frac{1}{4} \times \left(\frac{\sum_{i=1}^{5} c_i \cdot (i-1)}{\sum_{i=1}^{5} c_i} \right) \times 100, \tag{4}$$

where c_i denotes the number of responses at level i. Unlike Success Rate, Satisfaction captures a more nuanced and deeper evaluation, as even accepted cases may score low in satisfaction if parties feel emotionally unfulfilled. (3) **Consensus (Con) and Litigation Risk (LR)** are evaluated from a third-party perspective, based on the full trajectory of the mediation. Consensus measures the extent to which the disputants converge on key issues, while Litigation Risk estimates the likelihood that the dispute may escalate into formal legal proceedings. Both Con and LR are rated on the same five-point Likert scale and computed using the same normalization formula as Sat. Our evaluation follows the LLM-as-a-judge paradigm (Li et al. 2024), a widely used approach for assessing generative model outputs. To further enhance the quality and transparency of evaluation, we follow the research (Liu et al. 2023; Wang et al. 2024) by incorporating chain-of-thought reasoning into the prompts, encouraging the LLM to reason step-by-step before generating a final judgment. Prompts are in Appendix Section 9.

Solution Evaluator. This component focuses on assessing the reasonableness and legal soundness of the mediator's proposed solutions. We evaluate two core elements: (1) the points of contention, i.e., the alignment between the human-annotated contention points ($DPoints$) and the mediator-generated points ($MDPoints$); (2) the legal bases, i.e., the set of legal bases provided by human experts ($DBases$) versus those proposed by the mediator ($MDBases$). For the comparison of contention points, we apply ROUGE-L and BERTScore (Zhang et al. 2019). For the evaluation of legal bases coverage, we compute Recall.

Different Backbone LLM	Human1		Human2		Human3		Max of Three	
	BERTScore (F1)	*LLMScore*	*BERTScore (F1)*	*LLMScore*	*BERTScore (F1)*	*LLMScore*	*BERTScore (F1)*	*LLMScore*
Llama-3.2-1B-Instruct	0.6471††	0.4880††	0.6152	0.4530††	0.6424††	0.4780††	0.6629††	0.5310††
Qwen3-0.6B	0.6647††	0.6280††	0.6353	0.5840††	0.6612††	0.6000††	0.6871††	0.6820††
GLM-4-Flash	0.6857	0.7140††	0.6581	0.6580	0.6842	0.6440††	0.7141	0.7620
GPT-4o-2024-11-20	0.6954	0.7500	0.6553	0.6900	0.6895	0.6960	0.7200	0.8000
DeepSeek-V3-0324	0.6954	0.7780	0.6532	0.6780	0.6832	0.7020	0.7151	0.7940

Table 3: Human-model similarity scores across backbone LLMs. † and †† indicate statistically significant differences from DeepSeek-V3-0324 under a paired sample *t*-test with $p < 0.05$ and $p < 0.01$, respectively.

Stage	Role	Aspect	H	S	T
Begin	Mediator	Fact Summarization	0.05	0.10	**0.85**
		Neutrality Statement	0.00	0.25	**0.75**
		Self Introduction	0.00	0.00	**1.00**
		Turn-Taking Guidance	0.00	0.00	**1.00**
		Preliminary Solution	0.10	**0.50**	0.40
	Parties	Claim Clarity	0.10	0.00	**0.90**
		Fact-Based Claims	0.05	0.05	**0.90**
During	Mediator	Presiding Norms	0.05	0.05	**0.90**
		Debate Direction	0.00	0.10	**0.90**
		Order Control	0.00	0.10	**0.90**
		Fairness Awareness	0.25	0.10	**0.65**
	Parties	Order Compliance	0.00	0.05	**0.95**
		Offense Prohibition	0.05	0.30	**0.65**
		Effective Communication	0.05	**0.50**	0.45
		Focused Debate	0.10	0.05	**0.85**
End	Mediator	Solution Feasibility	0.10	0.00	**0.90**
		Solution Legality	0.05	0.00	**0.95**
		Claim Coverage	0.10	0.05	**0.85**

Table 4: Legal experts' pairwise preferences on 18 aspects across 20 disputes (H=Human, S=Simulation, T=Tie).

Experimental Setup

To test *AgentMediation*, we empirically investigate (1) the reliability of its generated mediation processes, (2) its ability to support analysis of key mediation variables, and (3) the consistency between our automatic evaluation metrics and human judgments. For human evaluations and annotations, we recruit legal experts from prominent law schools. All participants have passed the National Unified Legal Professional Qualification Examination. To ensure fair and motivating compensation, we offer an average hourly wage of $8.25, well above locally mandated minimum wage. To enable controlled experiments, we define a default setting for *AgentMediation*. In this setting, we fix the number of mediators to one and do not assign any predefined behavioral strategies to either disputants or the mediator. Additionally, the mediator has no access to external knowledge sources, and the bargaining stage is limited to 5 dialogue rounds. To ensure reproducibility, the temperature of LLM is set to 0.

Experimental Results

Reliability of Our Simulation (RQ1)

To evaluate the reliability of our simulation, we conduct two human-model comparison experiments under the default setting: utterance-level and case-level.

Utterance-Level. We formulate the task as role-based utterance generation, where both LLM agents and legal experts generate the next utterance for a specified role (mediator or disputant) given the same dialogue context. To ensure diversity in testing scenarios, we uniformly sample 50 dialogue cases (3–15 turns; 28 mediator and 22 disputant roles). For each case, three legal experts independently generate role-specific responses as equally valid references, and the agent responds under the same conditions. Model outputs are evaluated using BERTScore (F1) and LLMScore (a 0–1 semantic similarity score assessed by GPT-4o). We report: (1) the average similarity with each expert, and (2) the average of the highest score per case as an upper bound of human alignment. As shown in Table 3, strong LLMs such as DeepSeek-V3-0324 and GPT-4o achieve BERTScore (LLMScore) values above 0.71 (0.79). As an upper bound for human-human similarity, we take the highest similarity score among the three expert pairs in each case and average over all cases, yielding 0.69 (BERTScore) and 0.76 (LLMScore). These results suggest that our agents exhibit strong semantic alignment with expert responses. We select DeepSeek-V3-0324 as the default backbone for its balance of quality, efficiency, and open-source availability.

Case-Level. To further assess simulation reliability at the case level, we employ legal experts to reconstruct the detailed mediation processes of 20 real disputes. Using the same case descriptions, we use *AgentMediation* (based on DeepSeek-V3-0324) to generate simulated mediation processes for these 20 disputes. Legal experts then conduct pairwise preference evaluations between the human reconstructed and system generated versions across 18 fine-grained aspects. Each case is independently scored by three experts, and final judgments are determined via majority vote. As shown in Table 4, our simulations are indistinguishable from real mediation processes across most evaluation criteria, and even exceed them in certain aspects. These results provide further evidence of our simulation's reliability.

The Effect of Different Factors (RQ2)

Having validated the reliability of *AgentMediation*, we analyze how key factors influence outcomes, using over 14,000 simulated processes generated from 330 real disputes.

The Effect of Behavioral Strategies We investigate how disputants' behavioral strategies affect mediation outcomes. Based on the TKI Conflict Mode, we examine five typical strategies: Compromising, Competing, Accommodating, Avoiding, and Collaborating. We evaluate two configura-

Setting	NUM = 1				NUM = ALL			
	Success Rate ($SR\uparrow$)	Satisfaction ($Sat\uparrow$)	Consensus ($Con\uparrow$)	Litigation Risk ($LR\downarrow$)	Success Rate ($SR\uparrow$)	Satisfaction ($Sat\uparrow$)	Consensus ($Con\uparrow$)	Litigation Risk ($LR\downarrow$)
Default Setting	82%	54.77	70.96	24.00	82%	54.77	70.96	24.00
+ Compromising	95%	55.72	72.50	22.50	99%	58.67	75.25	19.25
+ Competing	29%	43.20††	53.25††	40.50††	11%	35.00††	33.00††	64.00††
+ Accommodating	88%	54.21†	71.50	21.00	97%	53.35††	75.50	16.75
+ Avoiding	54%	49.63††	64.75††	26.50††	37%	44.73††	60.75††	25.50††
+ Collaborating	93%	57.71	73.25	21.25	99%	61.10	76.25	19.00

Table 5: Comparison of mediation outcomes across different behavioral strategies, using DeepSeek-V3-0324 as the backbone LLM. *NUM = 1* indicates that one party in the case is randomly replaced with the specific behavioral strategy; *NUM = ALL* means all parties are replaced. † and †† indicate statistically significant differences from the *Compromising* mode under the chi-squared test (Greenwood and Nikulin 1996) with $p < 0.05$ and $p < 0.01$, respectively.

Setting	$SR\uparrow$	$Sat\uparrow$	$Con\uparrow$	$LR\downarrow$
Default Setting	82%	54.77	70.96	24.00
+ Information Conflict	53%	37.55	51.00	45.00
+ Resource Conflict	47%	37.23	49.50	51.00
+ Behavioral Misconduct	36%	31.18††	45.25	53.00†
+ Legal Issue	61%	39.21	57.00†	36.00††
+ External Factors	68%	43.60††	56.75†	39.75

Table 6: Comparison of mediation outcomes across different dispute causes, using DeepSeek-V3-0324 as the backbone LLM. † and †† indicate statistically significant differences from the *Information Conflict* setting under the chi-squared test, with $p < 0.05$ and $p < 0.01$, respectively.

tions: (1) one randomly selected party is replaced with the specific behavioral strategy, and (2) all parties are replaced. As shown in Table 5, we can find the following findings:

Overall Trends Align with Theoretical Expectations. Under the *NUM = 1* setting, the five behavioral strategies already exhibit typical patterns. (1) Competing produces the poorest results: low success (29%), low satisfaction (43.20), low consensus (53.25), and the highest litigation risk (40.50), reflecting its escalation-prone nature. (2) Collaborating shows the opposite extreme, achieving high success (93%), the best satisfaction (57.71), strong consensus (73.25), and the lowest risk among active modes (21.25), consistent with its integrative, win–win orientation. (3) Compromising yields moderate results across all metrics: success 95%, satisfaction 55.72, consensus 72.50, and risk 22.50. (4) Avoiding lowers visible risk (26.50) but also depresses satisfaction (49.63) and consensus (64.75), showing that backing away reduces open conflict yet leaves problems unsolved. (5) Accommodating achieves high consensus (71.50) with low risk (21.00), though satisfaction remains modest (54.21), indicating that harmony may come at the cost of personal needs. These strategy-specific tendencies persist under the *NUM = ALL* setting.

Evidence of Group Polarization (Moscovici and Zavalloni 1969) under Full Replacement. Comparing *NUM = 1* and *NUM = ALL* reveals a clear group polarization effect: when all parties adopt the same behavioral strategy, the collective dynamics become more extreme. In the Competing setting, full replacement significantly reduces performance, with success rate dropping from 29% to 11% and litigation risk rising from 40.50 to 64.00. This suggests that mutual aggressiveness amplifies conflict escalation and breakdown. In contrast, Collaborating under *NUM = ALL* leads to gains across all metrics compared to partial replacement (e.g., consensus increased from 73.25 to 76.25), reflecting a synergistic dynamic when cooperation is mutual. These findings support the theoretical expectation that interaction symmetry can either reinforce cooperation or escalate conflict, depending on the nature of the strategy.

Surface-level Consensus (Kelman 1958; Habermas 1985) in the Accommodating Strategy. The discrepancy between high consensus (75.50) and relatively low satisfaction (53.35) in the Accommodating suggests the presence of a Surface-level Consensus, where disputants agree outwardly but remain internally dissatisfied. This occurs when individuals suppress their own needs to preserve harmony, leading to unspoken tensions or unmet expectations. Compared to the Avoiding strategy, which also yielded low satisfaction but lower consensus, the Accommodating strategy appears more effective in reaching surface-level agreements, but potentially at the cost of long-term resolution quality. This aligns with prior research cautioning against over-reliance on accommodative behaviors, especially when structural issues remain unaddressed.

The Effect of Dispute Causes In this section, we investigate how different dispute causes affect mediation outcomes. While each case in our dataset is associated with a fixed legal dispute type (e.g., housing contract), the underlying causes can vary considerably in practice. To analyze their specific impact, we construct controlled variants of each case by injecting or amplifying a particular dispute cause within the Dispute Brief (DB). We consider five representative dispute causes, defined by the two-level taxonomy introduced earlier: Information Conflict, Resource Conflict, Behavioral Misconduct, Legal Issues, and External Factors. Given a specified top-level cause, we first prompt DeepSeek-V3-0324 to select the most contextually appropriate sub-cause based on the original DB. Then, conditioned on the selected sub-cause, the LLM generate targeted modifications that reinforce the presence of the cause in the DB. As Table 6 shows, our main findings include:

	Different Backbone LLM	Points of Contention		Legal Bases *(Recall ↑)*			
		BERTScore (F1) ↑	*ROUGE-L (F1)* ↑	w/o EK	w/ EK (top-3)	w/ EK (top-5)	w/ EK (top-10)
Open-Source	Llama-3.1-70B-Instruct	0.7647	0.3847	0.0295	0.1151	0.1392	0.1689
	GLM-4-Flash	0.7994	0.4640	0.1742	0.2260	0.2461	0.2718
	DeepSeek-V3-0324	0.8247	0.5404	**0.2999**	**0.3335**	**0.3446**	**0.3594**
	DeepSeek-R1	0.7994	0.4783	0.2988	0.3302	0.3386	0.3542
Closed-Source	GLM-4-Plus	**0.8354**	**0.5581**	0.2771	0.3277	0.3395	0.3593
	GPT-4o-mini-2024-07-18	0.8260	0.5288	0.0624	0.1442	0.1668	0.1939
	Qwen-Plus	0.8231	0.5176	0.2917	0.3301	0.3411	0.3578

Table 7: Performance comparison across different backbone LLMs under different evaluation metrics. EK denotes access to External Knowledge. The top-k represents different retrieval settings, where k represents the number of legal bases retrieved.

Moral Emotions Theory (Haidt et al. 2003) suggests that perceiving others' misconduct, such as deception or intentional rule-breaking, can elicit moral emotions like anger, contempt, and disgust. These emotions often lead to punitive or distancing responses, making compromise more difficult in mediation. This pattern aligns with our empirical observations: cases involving Behavioral Misconduct exhibit the highest litigation risk (53.00), the lowest satisfaction (31.18), and the lowest resolution success rate (36%).

Realistic Conflict Theory (Sherif et al. 1961) argues that disputes perceived as zero-sum competition over limited resources lead to increased hostility. Our Resource Conflict cases support this view, with elevated risk (51.00), low satisfaction (37.23), and a reduced success rate of 47%.

The Common In-Group Identity Model (Gaertner et al. 1993) states that facing a common external threat can blur group boundaries and encourage cooperation. Consistent with this view, disputes driven by External Factors (e.g., policy shocks or natural disasters) and Legal Issues achieve the highest consensus scores (56.75 and 57.00) and elevated success rates (68% and 61%). Litigation risk is also lower for Legal Issues (36.00), suggesting that when parties frame the problem as procedural rather than personal, they are more willing to compromise and avoid escalation.

The Effect of Mediator Expertise This section examines how mediator expertise influences mediation outcomes. As shown in Table 7, DeepSeek-V3-0324 outperforms all open-source LLMs and rivals closed-source LLMs, achieving strong results in contention identification (BERTScore 0.8247, ROUGE-L 0.5404) and the highest recall in legal citation. These results make it a highly competitive backbone LLM in our *AgentMediation*. Moreover, all LLMs benefit from External Knowledge. For instance, DeepSeek-V3-0324's recall improves to 0.3594 with top-10 retrieval, showing that retrieval-augmented generation (RAG) can effectively enhance domain-specific performance.

Overall, the above experiments addressing RQ2 show that *AgentMediation* can effectively explore the impact of key mediation variables and uncover phenomena consistent with established social theories, highlighting its ability to capture legal-social dynamics and its practical value for advancing legal research and intelligent mediation.

Reliability of Our LLM-as-a-Judge (RQ3)

To assess the reliability of our LLM-as-a-judge, we conduct a human-model agreement study on the three key outcome metrics: Satisfaction, Consensus, and Litigation Risk. For each, we randomly sample 50 representative mediation cases. Nine legal experts are recruited, with three assigned per metric to independently score all cases. We calculate agreement between model predictions and the majority vote of the three expert annotations per case. Using DeepSeek-V3-0324 as the judge, we observe Cohen's Kappa scores of 0.358 for Satisfaction, 0.519 for Consensus, and 0.672 for Litigation Risk. These results suggest that while Satisfaction remains the most subjective metric (even inter-annotator agreement is low, at 0.344), both the LLM-as-a-judge's predictions for Consensus and Litigation Risk show strong correlation with human annotations, providing empirical support for leveraging LLM-as-a-judge in large-scale mediation simulations. Additional details are in Appendix Section 3.

Ablation Studies and Discussions

Beyond the main experiments, we conducted a broad range of in-depth follow-up studies. We performed ablations under the default setting to assess the contribution of individual mediation stages and examined the impact of varying bargaining rounds. Results show that removing any single stage leads to a noticeable decline in mediation performance, while increasing the number of bargaining turns yields only marginal gains. The more detailed results are provided in the appendix.

Conclusion

We present *AgentMediation*, the first LLM-based agent framework for simulating dispute mediation. Grounded in real-world disputes, it produces realistic mediation processes and serves as an extensible platform for controlled studies on key factors. Our simulations align with established sociological theories and offer insights into mediation dynamics. Looking forward, our framework provides a foundation for developing intelligent mediation systems.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (Grant No. 2024YFC3307102).

References

Abdelnabi, S.; Gomaa, A.; Sivaprasad, S.; Schönherr, L.; and Fritz, M. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37: 83548–83599.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ashery, A. F.; Aiello, L. M.; and Baronchelli, A. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20): eadu9368.

Chawla, K.; Ramirez, J.; Clever, R.; Lucas, G.; May, J.; and Gratch, J. 2021. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. *arXiv preprint arXiv:2103.15721*.

Gaertner, S. L.; Dovidio, J. F.; Anastasio, P. A.; Bachman, B. A.; and Rust, M. C. 1993. The common ingroup identity model: Recategorization and the reduction of intergroup bias. *European review of social psychology*, 4(1): 1–26.

Greenwood, P. E.; and Nikulin, M. S. 1996. *A guide to chi-squared testing*. John Wiley & Sons.

Habermas, J. 1985. *The theory of communicative action: Volume 1: Reason and the rationalization of society*, volume 1. Beacon press.

Haidt, J.; et al. 2003. The moral emotions. *Handbook of affective sciences*, 11(2003): 852–870.

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

He, Z.; Cao, P.; Wang, C.; Jin, Z.; Chen, Y.; Xu, J.; Li, H.; Jiang, X.; Liu, K.; and Zhao, J. 2024. Agentscourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. *arXiv preprint arXiv:2403.02959*.

Hsieh, H.-P.; Jiang, J.; Yang, T.-H.; Hu, R.; and Wu, C.-L. 2022. Predicting the success of mediation requests using case properties and textual information for reducing the burden on the court. *Digital Government: Research and Practice*, 2(4): 1–18.

Hua, W.; Fan, L.; Li, L.; Mei, K.; Ji, J.; Ge, Y.; Hemphill, L.; and Zhang, Y. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.

Jebb, A. T.; Ng, V.; and Tay, L. 2021. A review of key Likert scale development advances: 1995–2019. *Frontiers in psychology*, 12: 637547.

Jin, Y.; Zhao, Q.; Wang, Y.; Chen, H.; Zhu, K.; Xiao, Y.; and Wang, J. 2024. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.

Kelman, H. C. 1958. Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*, 2(1): 51–60.

Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; and Liu, Y. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Liu, X.; Lei, X.; Wang, S.; Huang, Y.; Feng, Z.; Wen, B.; Cheng, J.; Ke, P.; Xu, Y.; Tam, W. L.; et al. 2023. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Mnookin, R. H. 1998. *Alternative dispute resolution*. Harvard Law School.

Moffitt, M. L.; and Bordone, R. C. 2012. *The handbook of dispute resolution*. John Wiley & Sons.

Moscovici, S.; and Zavalloni, M. 1969. The group as a polarizer of attitudes. *Journal of personality and social psychology*, 12(2): 125.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Priya, P.; Chigrupaatii, R.; Firdaus, M.; and Ekbal, A. 2025. GENTEEL-NEGOTIATOR: LLM-Enhanced Mixture-of-Expert-Based Reinforcement Learning Approach for Polite Negotiation Dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 23, 25010–25018.

Sherif, M.; Harvey, O. J.; White, B. J.; Hood, W. R.; and Sherif, C. W. 1961. *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*. Norman, OK: University Book Exchange.

Sherman, N.; and Momani, B. T. 2025. Alternative dispute resolution: Mediation as a model. *F1000Research*, 13: 778.

Tan, J.; Westermann, H.; Pottanigari, N. R.; Šavelka, J.; Meeùs, S.; Godet, M.; and Benyekhlef, K. 2024. Robots in the Middle: Evaluating LLMs in Dispute Resolution. In *Legal Knowledge and Information Systems*, 168–179. IOS Press.

Thomas, K. W. 2008. Thomas-kilmann conflict mode. *TKI Profile and Interpretive Report*, 1(11).

Vilendrer Law, P. 2023. Five Main Causes of Conflict and How Mediation Can Resolve Them. Accessed: 2025-05-06.

Wang, J.; Mo, F.; Ma, W.; Sun, P.; Zhang, M.; and Nie, J.-Y. 2024. A User-Centric Multi-Intent Benchmark for Evaluating Large Language Models. *arXiv preprint arXiv:2404.13940*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.