

Article

CogMed: A Multi-Agent Legal Mediation Framework Fusing Cognitive Strategies and Dynamic Beliefs

Jia Chen ¹, Yiheng Ma ² and Shijuan Gao ^{3,*}

¹ School of Law, Xiangtan University, Xiangtan 411105, China; 202331140260@smail.xtu.edu.cn

² School of Law, Hunan University of Technology and Business, Changsha 410083, China; 2408010045@stu.hutb.edu.cn

³ Information & Network Center, Central South University, Changsha 410083, China

* Correspondence: gaoshijuan@csu.edu.cn

Abstract

Legal mediation is an important mechanism for resolving social conflicts and handling disputes. It involves complex interpersonal interactions and unstructured decision-making processes, and therefore holds significant research value as a domain. Leveraging the outstanding logical reasoning capabilities of large language models, multi-agent systems for simulating complex social interactions have become a cutting-edge research direction in artificial intelligence, providing a new supporting vehicle and research pathway for the intelligent study and practical application of legal mediation. However, directly applying general-purpose multi-agent techniques or general-purpose, opaque LLMs to long-horizon, multi-party, and high-conflict professional mediation tasks exposes several deep-seated structural cognitive deficiencies, including a lack of process awareness, insufficient domain-specific intervention capabilities, and limited theory-of-mind reasoning. To address these challenges, this study proposes CogMed, a cognitively enhanced multi-agent framework for legal mediation simulation, which aims to compensate for the limitations of general models in professional strategic interactions through an explicit cognitive architecture. Rather than introducing entirely new individual reasoning modules, the proposed framework focuses on cognitively coordinated integration of process control, strategic intervention, and belief modeling mechanisms under legal mediation settings. CogMed models the mediation process as a Finite State Machine (FSM) to capture macro-level decision logic and introduces a Strategic Toolkit (STK) that serves as a set of action primitives for micro-level interventions. Meanwhile, a Dynamic Belief Tracking (DBT) mechanism is incorporated into party agents to simulate psychological anticipation and strategic reasoning during negotiation. Experimental results demonstrate that CogMed effectively improves both mediation success rates and the quality of negotiated outcomes. Furthermore, the findings suggest a preliminary framework-level compensation pattern under the current experimental setting, where cognitively structured coordination mechanisms may partially enhance the mediation capability of medium-scale models. These preliminary experimental observations suggest that cognitively structured coordination mechanisms may partially compensate for certain limitations associated with model scale under the current controlled mediation setting, thereby offering a potential research direction for cognitively structured legal mediation simulation systems under controlled experimental settings.



Academic Editors: Arkaitz Zubiaga and Katsuhide Fujita

Received: 9 April 2026

Revised: 25 June 2026

Accepted: 26 June 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: legal mediation; multi-agent simulation; large language models; structural cognition; theory of mind

1. Introduction

Within China's rule-of-law framework, mediation has been elevated to a strategic position for resolving social conflicts and advancing the modernization of the national governance system and governance capacity, and has become a core principle guiding judicial practice and grassroots governance. Under this policy context, a large-scale mediation system has been established, ranging from court-led pre-litigation mediation to community-based people's mediation organizations distributed across both urban and rural areas. The effective functioning of this system relies heavily on mediators' professional judgment and communication strategies.

Such a nationwide system places exceptionally high demands on the strategic decision-making and communicative intervention capabilities of millions of full-time and part-time mediators when handling complex disputes. Consequently, systematic analysis and evaluation of mediation strategies hold substantial practical value and application potential for enhancing the effectiveness of grassroots governance in China.

However, mediation is inherently a dynamic and unstructured process of interpersonal strategic interaction. Effective mediation strategies often manifest as tacit knowledge that is difficult to articulate and formalize. This tacit expertise encompasses macro-level awareness of procedural stages, the application of micro-level intervention tactics, and nuanced assessments of disputants' psychological states. As a result, the large-scale replication and transmission of such expertise remain highly challenging.

With the emergence of Large Language Models (LLMs), the development of professional legal agents has become increasingly feasible. Existing legal foundation models, represented by LawGP [1] and ChatLaw [2], primarily focus on tasks involving explicit legal knowledge processing, such as legal information retrieval, summarization, and application. Although these systems demonstrate strong performance in document drafting and legal consultation, mediation, as a multi-party and high-level social cognitive activity, requires far more than legal expertise alone. It demands sophisticated strategic planning, empathy, and conflict resolution capabilities.

To address this complexity, recent research has begun exploring Multi-Agent systems to replicate the mediation process. However, existing multi-agent frameworks—whether simple role-playing-based social simulations or collaboration-oriented architectures such as CAMEL [3] and MetaGPT [4] designed for task decomposition—remain difficult to directly apply to legal mediation. When confronted with long-horizon and high-conflict negotiation scenarios, these approaches exhibit pronounced structural cognitive limitations.

Specifically, three key deficiencies can be observed. First, a lack of process awareness: models often fail to maintain macro-level control over mediation stages and easily lose track of long-term objectives during extended multi-turn interactions. Second, insufficient professional intervention capabilities: when facing negotiation deadlocks, they struggle to proactively employ structured strategies comparable to those used by human experts. Third, limited theory-of-mind reasoning: they lack explicit modeling of opponents' intentions and beliefs during strategic interactions. Simply scaling up model parameters proves insufficient to overcome these higher-order social-cognitive bottlenecks. In this study, these limitations are operationally summarized as cognitive shallowness and strategic drift in long-horizon mediation interactions. Strategic drift refers to the tendency of LLM agents to gradually deviate from the original mediation objective, negotiation strategy, or settlement goal as the dialogue proceeds, often resulting in inconsistent intervention choices, weakened procedural control, or premature convergence to unstable proposals. Cognitive shallowness refers to the tendency of agents to rely primarily on local turn-level responses without maintaining sufficient process awareness, executable intervention planning, or persistent opponent-state tracking. In legal mediation, these issues are particularly problematic

because effective dispute resolution requires sustained coordination across procedural stages, strategic intervention at negotiation deadlocks, and continuous assessment of each party's goals, constraints, and willingness to compromise.

To address the aforementioned challenges, this study proposes CogMed, a cognitively inspired multi-agent simulation framework designed to move beyond the simplistic assumption of “LLM-as-agent”. Instead, CogMed embeds LLMs within an explicit cognitive architecture jointly constrained by process control, professional action mechanisms, and belief modeling. In this manner, high-level strategic reasoning and structured decision-making are achieved through architectural design rather than relying solely on raw model capacity.

Figure 1 illustrates the complete workflow of the CogMed framework, spanning from parameterized scenario configuration to automated evaluation.

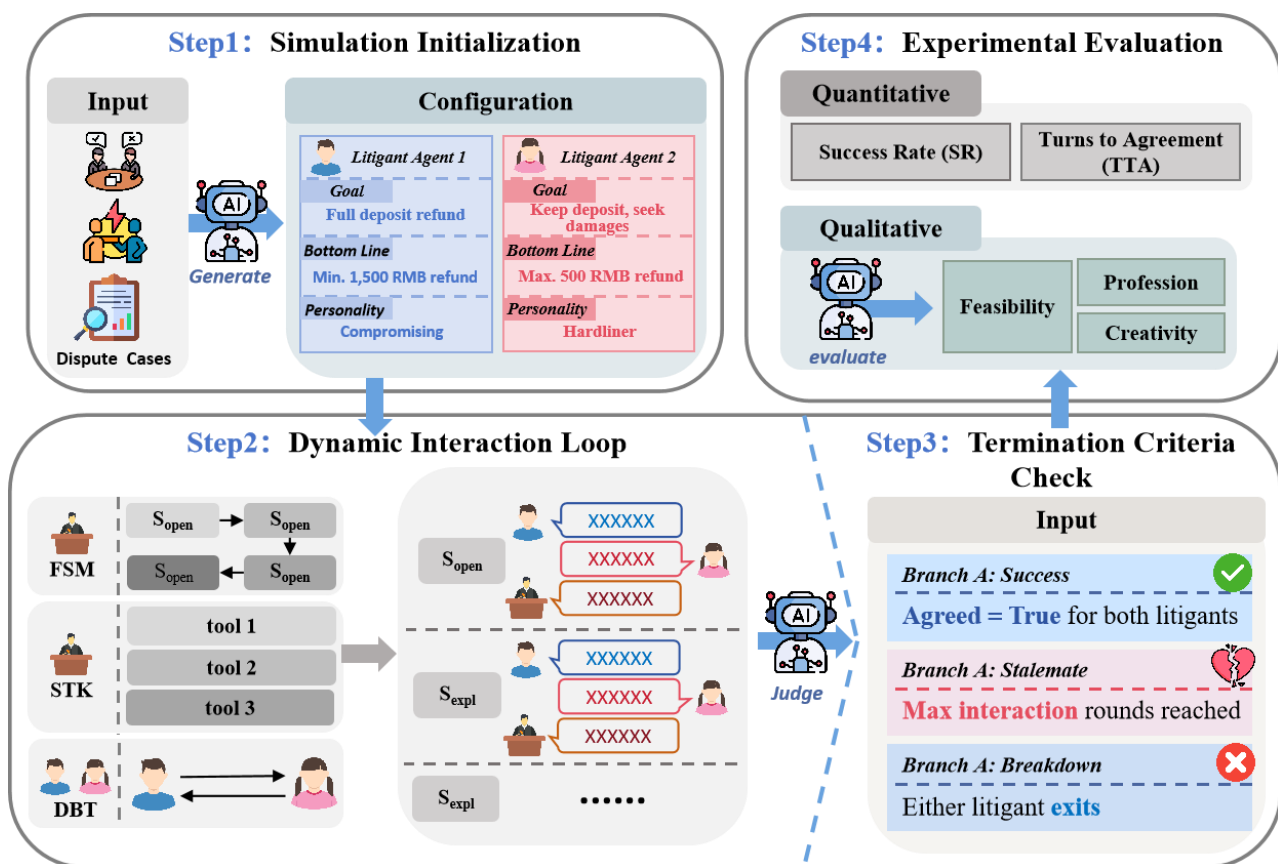


Figure 1. High-level overview of the overall workflow of the CogMed framework, from scenario initialization to automated mediation evaluation.

Specifically, CogMed enhances the cognitive capabilities of the underlying LLMs through three key modules. First, a Finite State Machine (FSM) is employed to model mediation stages, enabling the system to emulate how human experts switch strategic objectives across different phases of the mediation process and thereby maintain macro-level procedural awareness. Second, a Strategic Toolkit (STK)-based decision action space is constructed, which transforms professional intervention experience into computable strategy primitives. This design grounds abstract expertise in executable operations and supports structured, targeted interventions during negotiation. Third, Dynamic Belief Tracking (DBT) is introduced to equip party agents with theory-of-mind capabilities. By continuously inferring opponents' intentions and beliefs in real time, agents are able to engage in more realistic and strategically informed interactions that better approximate human behavioral patterns.

To comprehensively evaluate the effectiveness of the proposed framework, systematic experiments were conducted on a real-world civil case dataset. A hybrid evaluation protocol was adopted, combining a quantitative metric—mediation success rate—with qualitative comprehensive scoring based on an LLM-as-a-Judge paradigm.

Experimental results show that CogMed consistently improves both mediation success rates and the quality of negotiated outcomes. More importantly, the findings reveal a key insight: framework intelligence exhibits a compensatory effect with respect to model scale. In particular, a lightweight model empowered by an advanced cognitive framework can achieve professional performance comparable to, or even surpassing, that of a substantially larger model operating without explicit architectural guidance. This result highlights the critical role of structured cognitive design beyond mere parameter scaling. The purpose of CogMed is not to propose a new general-purpose foundation model, nor to claim that FSM, STK, or DBT are entirely new mechanisms in isolation. Rather, the contribution lies in the task-specific coordination of these mechanisms for legal mediation simulation. Although the individual components are inspired by general agent-system design principles, they are instantiated here with distinct operational roles in the mediation process: FSM provides stage-level procedural control, STK provides executable mediation interventions, and DBT provides dynamic opponent-state tracking for litigant agents. Their integration forms a controlled mediation-oriented interaction architecture designed to address long-horizon strategic drift and cognitive shallowness in simulated legal dispute resolution. In summary, the main contribution of this work is not the introduction of entirely novel individual modules, but the construction of a cognitively coordinated multi-agent mediation architecture tailored for long-horizon adversarial legal negotiation scenarios. Specifically, the proposed framework integrates process-level stage control, executable strategic intervention mechanisms, and temporally persistent belief modeling into a unified mediation-oriented interaction system.

Within this architecture, FSM provides explicit procedural coordination for maintaining long-horizon strategic consistency, STK operationalizes mediation expertise into executable intervention primitives, and DBT enables temporally coherent opponent modeling during adversarial negotiation. Through this cognitively constrained integration, the framework aims to improve controllability, strategic coherence, and negotiation stability in legal mediation simulations.

Accordingly, the contributions of this work can be summarized as follows:

- (1) A mediation-oriented multi-agent framework, CogMed, is proposed to coordinate procedural control, strategic intervention, and opponent-state tracking in long-horizon legal mediation simulations.
- (2) A structured interaction mechanism integrating FSM, STK, and DBT is developed to transform selected mediation practices into controllable and executable simulation components.
- (3) Controlled experiments and ablation analyses are conducted to provide preliminary empirical evidence on how structured cognitive coordination affects mediation success, agreement quality, and interaction stability under the current experimental setting.

2. Related Work

2.1. Large Language Models in the Legal Domain

Large Language Models (LLMs) have demonstrated strong capabilities in the legal domain, particularly in text-centric processing of explicit knowledge. Representative domain-specific models, such as LawGPT [1], ChatLaw [2], and LexiLaw [5], are trained on large-scale legal corpora and exhibit proficiency in legal knowledge retrieval, summarization, and application.

Meanwhile, advanced general-purpose foundation models, including GPT-4 [6] and Claude [7], have also achieved competitive performance across a variety of legal benchmark tasks, further demonstrating the broad applicability of large-scale pretrained models to legal reasoning and text-based problem solving.

However, the technical architectures of these models are primarily optimized for single-agent tasks such as knowledge question answering and text generation. Their design objectives generally do not account for the demands of sustained strategic interaction or process-level management.

Entering 2025, research attention has gradually shifted toward the construction of interactive legal scenarios. For instance, the MASER framework and the MILE Benchmark [8] have been introduced to enable large-scale synthetic data generation and multi-stage evaluation of model performance in dynamic and labor-intensive legal interaction tasks. These efforts have advanced the assessment of legal AI systems beyond static benchmarks.

Nevertheless, their core tasks still focus predominantly on explicit legal knowledge processing, achieving strong performance in relatively static settings such as legal consultation, case analysis, and document drafting. In contrast, legal mediation constitutes a long-horizon, adversarial, and multi-party dynamic game, which imposes substantially higher demands on agents' social-cognitive capabilities, including strategic planning, negotiation, and conflict resolution.

Despite the continuous growth in parameter scale, existing legal foundation models generally lack dedicated cognitive architectures tailored for interactive and high-conflict negotiation scenarios. As a result, they struggle to effectively manage the dynamic strategies and complex interpersonal interactions required in real-world mediation processes.

2.2. Multi-Agent Simulation

To model the complex interactions inherent in mediation processes, recent research has increasingly turned to Multi-Agent systems. Driven by the rapid advancement of LLMs, two dominant technical paradigms have emerged.

The first paradigm focuses on social interaction simulation frameworks. A milestone work in this line of research is Generative Agents, proposed by Stanford University, also known as the “Smallville” town simulation Generative Agents [9]. This work demonstrates that LLM-driven agents can exhibit complex and seemingly spontaneous social behaviors in open-ended environments. Subsequent efforts include LegalAgentBench [10], MT-Bench [11] for multi-agent dialogue evaluation, as well as broader attempts to enhance traditional Agent-Based Modeling (ABM) with LLM capabilities.

The technical core of these frameworks typically lies in equipping agents with memory mechanisms (e.g., memory streams), planning modules, and reflective reasoning abilities. More recent studies, such as behavior-alignment evaluations for conflict dialogues, have begun to systematically examine the gap between LLM agents and human participants in terms of strategic behavior and emotional dynamics during adversarial interactions [12]. These findings suggest that even in personality-driven simulations, models still suffer from notable limitations in strategy consistency.

In addition, AgentMediation [13], the first multi-agent system specifically designed for civil dispute mediation, highlights the importance of constructing controllable and observable “social laboratories” for mediation research.

Despite these advances, such frameworks are primarily designed to explore open-ended and goal-agnostic social dynamics, where behavioral emergence is the central objective. This design philosophy fundamentally conflicts with the requirements of legal mediation, which demand structured stage progression and strategic convergence. When

applied to mediation scenarios, the absence of explicit process modeling and executable action primitives often results in agent behaviors that are difficult to constrain and stabilize.

The second paradigm centers on collaborative task-solving frameworks, which emphasize structured interactions among agents to accomplish complex objectives. Early explorations include AutoGPT [14] and BabyAGI [15], both of which attempt to enable autonomous task planning and execution. Subsequent research has introduced more refined collaboration schemes. For example, CAMEL [3] proposes a role-playing-based prompting strategy, while MetaGPT [4] incorporates standard operating procedures (SOPs) and role definitions to emulate the workflow of a software organization. ChatDev [16] further specializes this paradigm for software engineering pipelines, and AutoGen [17] provides a customizable conversational infrastructure that allows both human and AI agents to participate under differentiated roles (e.g., UserProxyAgent and AssistantAgent). In addition, AgentVerse [18] explores parallel execution and coordinated cooperation among multiple agents.

Despite their architectural diversity, these frameworks share a common underlying design philosophy: cooperation. Whether through CAMEL's inception prompting or MetaGPT's meta-programming-style coordination, the primary objective is to efficiently converge toward a shared goal through collaborative decomposition and information exchange. Such cooperative assumptions, however, are fundamentally mismatched with the inherently adversarial nature of many legal scenarios.

To better address legal interactions, LegalSim [19] attempts to formalize litigation as a multi-agent reinforcement learning environment, simulating strategic behavior by exploiting procedural loopholes. Beyond legal-domain simulations, interpretable machine learning methods have also been widely explored for data-driven decision support in other professional domains; for example, Huang et al. [20] integrated an optimized XGBoost model with SHAP analysis to improve both prediction accuracy and interpretability in rock strength estimation. Nevertheless, legal mediation is intrinsically a high-conflict, mixed-motive game rather than a purely cooperative task. Directly applying frameworks designed for collaboration to conflict-driven settings fails to capture the complex strategic confrontation and dynamic belief updating among disputing parties. Consequently, these approaches remain insufficient for realistically modeling mediation processes.

It should be noted that the present study primarily focuses on architecture-level cognitive coordination under legal mediation settings rather than comprehensive benchmarking against all existing negotiation-oriented agent frameworks. More extensive comparative evaluations against advanced agent architectures and real-world mediation systems remain important directions for future research. In addition, many existing multi-agent frameworks differ substantially in task objectives, interaction protocols, implementation availability, and evaluation settings, making direct controlled empirical comparison challenging under a unified mediation simulation environment.

3. CogMed Framework Design

CogMed is designed to address the problem of cognitive shallowness commonly observed in existing multi-agent systems when simulating complex human social interactions. Such systems often lack stable capabilities for process control, action selection, and opponent modeling during long-horizon interactions, resulting in inconsistent strategies and degraded performance over extended negotiations.

Unlike conventional paradigms that treat LLMs as monolithic black-box agents, CogMed adopts a hybrid cognitive architecture. Rather than relying solely on the implicit reasoning ability of large models, the proposed framework explicitly integrates

structured process management, strategic action mechanisms, and belief modeling to provide systematic cognitive support for mediation-oriented decision making.

As illustrated in Figure 2, the proposed architecture presents a structured multi-agent interaction loop driven by a central Controller. The entire simulation process is organized as a closed-loop workflow to ensure stable process management and coordinated decision-making across agents.

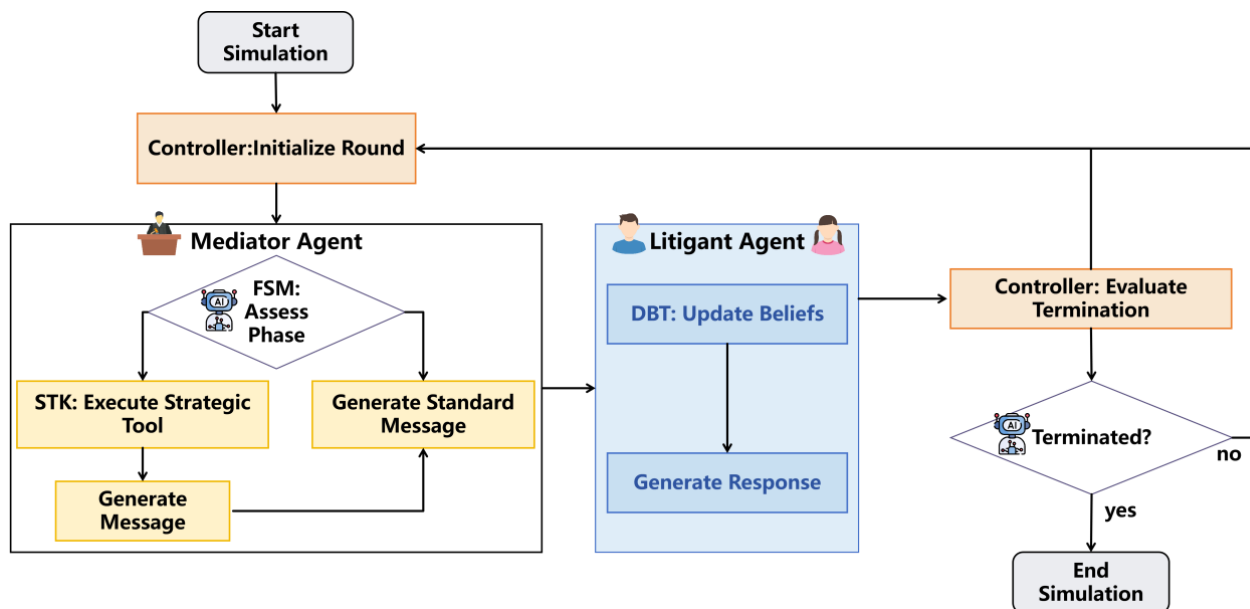


Figure 2. Detailed interaction mechanism of the CogMed framework, illustrating the controller-coordinated mediation loop and the interactions among FSM, STK, and DBT modules.

The simulation begins with round initialization by the Controller. The Mediator Agent, guided at a macro level by a Finite State Machine (FSM), first evaluates the current mediation stage and determines the appropriate course of action. Based on this assessment, it either invokes the Strategic Toolkit (STK) to execute targeted professional interventions or generates standard conversational messages.

Meanwhile, the Litigant Agent, upon receiving incoming information, activates the Dynamic Belief Tracking (DBT) module to update its internal cognitive state and subsequently produces a response that reflects its inferred beliefs and intentions.

At the end of each interaction round, the Controller evaluates whether predefined termination conditions have been satisfied. Depending on the outcome, the simulation either concludes or proceeds to the next round.

By introducing explicit process- and strategy-level constraint modules to guide and regulate the implicit reasoning of LLMs, the proposed mechanism ensures both high fidelity and controllability in legal mediation simulation. This design enables agent behaviors to remain aligned with structured procedural objectives while preserving the generative flexibility of large language models.

The following sections provide a detailed description of the key components of the proposed architecture.

3.1. Simulation Environment and Parameterized Scenario Initialization

To ensure experimental controllability and reproducibility, CogMed constructs the simulation environment in a parameterized manner. Each mediation instance is formalized as a structured configuration, allowing systematic manipulation of contextual and agent-specific attributes while maintaining consistent evaluation conditions.

Formally, a mediation scenario is represented as a tuple $\varepsilon = \langle C, A_M, \{A_{L_i}\}_{i \in \{1,2\}} \rangle$, where C denotes the Case Context, which comprises both unstructured case narratives and structured representations of disputed issues; A_M denotes the Mediator Agent, driven by both macro-level strategic control and micro-level action modules; A_{L_i} denotes the Litigant Agents, consisting of two parties (A_{L_1} and A_{L_2}). Each litigant is assigned a private internal state $S_i = \{G_i, B_i, P_i\}$, corresponding to its goal payoff (Goal), bottom-line constraints (Bottom Line), and personality prior (Personality Prior), respectively.

This parameterized initialization mechanism ensures that agent behaviors are not solely driven by the immediate dialogue history, but are also governed by stable, intrinsic role configurations. Consequently, the resulting interactions more faithfully reflect persistent strategic preferences and individualized behavioral tendencies observed in real-world mediation settings.

3.2. Mediator: Dual-Layer Cognitive Augmentation Architecture

General-purpose LLMs frequently exhibit strategic drift in long-horizon mediation scenarios, where objectives become inconsistent and decision-making degrades over extended interactions. To address this limitation, the Mediator Agent is equipped with a dual-layer cognitive architecture inspired by hierarchical cognition.

Specifically, the proposed design decomposes the mediator's reasoning process into two complementary levels: macro-level planning and micro-level execution. This hierarchical structure enables the agent to maintain a stable strategic direction at the global level while performing flexible, context-sensitive interventions at the local level.

1. Macro-Level: FSM-based phase-aware strategy

When handling long-horizon mediation tasks, general-purpose LLMs often suffer from strategy drift caused by the sliding nature of the context window, gradually losing track of overarching objectives as the dialogue progresses. To mitigate this issue, CogMed introduces an explicit stage-based representation of the mediation process using a Finite State Machine (FSM).

Specifically, the mediation workflow is modeled over a discrete state space $S = \{s_{\text{open}}, s_{\text{expl}}, s_{\text{nego}}, s_{\text{draft}}\}$, corresponding to four key phases: opening, exploration, negotiation, and agreement drafting.

The module operates through an evaluation–execution loop. Prior to each interaction round, a dedicated lightweight evaluator is invoked to approximate the state transition function $\delta : S \times \rightarrow S_{t+1}$, where the current dialogue history H_t is analyzed to determine whether the mediation should transition to the next stage. The resulting state S_{t+1} is then used to dynamically reconstruct the mediator's system prompt, effectively adapting its macro-level strategy according to procedural progress.

For instance, during the exploration stage (s_{expl}), the agent is softly constrained to prioritize information gathering, whereas in the negotiation stage (s_{nego}), more assertive intervention strategies are enabled. By explicitly coupling strategic behavior with procedural states, the mediator's decision-making remains aligned with stage-specific objectives.

This mechanism effectively addresses the loss of focus commonly observed in extended conversations. The explicit stage modeling ensures orderly transitions between potentially conflicting goals—such as information collection and compromise facilitation—thereby maintaining strategic coherence throughout long-horizon negotiations.

2. Micro-Level: strategic toolkit as executable action primitives

To move beyond pure language generation and equip the mediator with professional intervention capabilities, we abstract the tacit expertise of experienced mediators into a Strategic Toolkit (STK). The STK serves as a set of executable action primitives that enable

structured and goal-directed interventions, transforming implicit professional know-how into explicit, callable operations.

Specifically, the toolkit consists of the following actions:

(1) `summarize_and_clarify` (H_t):

This tool invokes the LLM to perform neutral information compression and key-point extraction over recent dialogue history. It is designed to mitigate information overload and topic drift, both of which are common in long-horizon negotiations. By reframing emotionally charged exchanges into concise, fact-oriented summaries, the conversation is re-anchored to a rational and constructive basis.

(2) `perform_reality_check` (A_L , issue):

This tool emulates the perspective of an experienced legal professional by leveraging the model's internal legal knowledge to conduct an objective assessment of risks and costs associated with a litigant's rigid claims. Through this analysis, it corrects overconfidence biases and encourages the party to recalibrate expectations based on a more realistic Best Alternative to a Negotiated Agreement (BATNA), thereby steering the negotiation back toward feasible compromise.

(3) `propose_bridging_solution` (H_t , $\text{Interests}(A, B)$):

This tool employs higher-order Chain-of-Thought reasoning to shift from surface-level positions to deeper underlying interests. Based on this cognitive reframing, it generates creative and integrative proposals aimed at breaking negotiation deadlocks and advancing the mediation process toward constructive solution selection.

Collectively, these tools constitute a finite set of callable action primitives available to the mediator. By explicitly encoding complex professional interventions into controllable decision options, the STK enables structured, reliable, and reproducible strategy execution within the mediation process.

3.3. Litigant: Explicit Theory-of-Mind-Driven Cognitive Modeling

Authentic strategic interaction is driven not only by the maximization of one's own interests, but also by the anticipation of an opponent's intentions. However, general-purpose LLMs typically lack stable Theory-of-Mind (ToM) capabilities, often reacting myopically to local context rather than forming persistent beliefs about their counterparts. As a result, their behaviors in adversarial settings tend to be short-sighted and strategically inconsistent.

To address this limitation, CogMed introduces a Dynamic Belief Tracking (DBT) mechanism that explicitly models the temporal inference process through which disputants reason about their opponents' intentions. Concretely, each litigant agent is required to maintain an explicit belief state $B_t(A_{-i})$, representing its current estimation of the opposing party's goals, constraints, and likely strategies.

Upon receiving a message at time step t , the agent first triggers a belief-update procedure. Importantly, this update process does not aim to perfectly reconstruct the opponent's true psychological state. Instead, its primary purpose is to impose temporal consistency and strategic stability on the agent's behavior. By maintaining a coherent and continuously evolving belief model, the agent's decisions become grounded in accumulated interaction history rather than isolated utterances.

Through this mechanism, DBT enables agents to engage in more realistic and strategically informed negotiations, better approximating the sequential reasoning patterns observed in human adversarial interactions.

Rather than relying on a formally defined analytical update rule, the current DBT mechanism performs opponent-state updating through structured LLM prompting conditioned on the dialogue history, current interaction context, and previously estimated

negotiation attributes. At each interaction round, the mediator agent re-evaluates several dynamically evolving attributes of the opposing party, including perceived goals, emotional tendencies, and flexibility estimates, in order to support subsequent strategic coordination during mediation.

The DBT model explicitly tracks three key latent attributes of the opponent: perceived goal, emotional valence, and flexibility score. Together, these variables characterize not only what the opponent aims to achieve but also their affective stance and willingness to compromise, thereby providing a compact yet expressive representation of their evolving strategic posture.

This belief state is updated recursively as new interaction evidence becomes available, enabling agents to continuously refine their inferences over time. Such a recursive modeling mechanism equips litigant agents with deeper strategic reasoning capabilities, allowing them to adopt more sophisticated countermeasures—such as strategic concession, calibrated signaling of weakness, or tactical bluffing—rather than relying on purely reactive responses.

By grounding decision-making in an explicit and temporally coherent belief model, DBT enhances both the realism and complexity of simulated negotiations, producing behaviors that more closely resemble the adaptive strategies observed in real-world adversarial games.

3.4. Simulation Interaction Cycle

CogMed employs a centralized controller to orchestrate a round-based asynchronous interaction protocol, ensuring that the simulation proceeds in a well-ordered and temporally consistent manner. This design provides explicit control over information flow, decision timing, and termination conditions, thereby preventing the uncontrolled or chaotic exchanges that commonly arise in purely free-form multi-agent conversations.

Under this mechanism, the mediation process is executed as a sequence of discrete interaction rounds, where each agent updates its internal state and generates actions according to the current global context. Such structured temporal coordination guarantees reproducibility and strategic coherence throughout long-horizon negotiations.

A complete mediation simulation consists of three core stages:

- (1) During the initialization stage, the system first constructs the simulation scenario based on structured dispute issues extracted from real-world cases. An auxiliary LLM, implemented using Qwen-Plus with the temperature fixed at 0.3, is then invoked to generate private role profiles for each litigant agent. Specifically, the auxiliary LLM is prompted with the case summary and disputed issues to produce a structured role profile, including the litigant's initial goal, bottom-line constraint, and personality prior. This profiling process explicitly defines each party's initial goals and bottom-line constraints, while stochastically assigning personality parameters (e.g., hardline or conciliatory types). Such parameterization introduces heterogeneous behavioral tendencies across agents and provides diversified initial conditions for subsequent strategic interactions. By grounding agents in stable and individualized role configurations from the outset, the simulation avoids homogeneous or purely reactive behaviors and more faithfully reflects the variability observed in real-world mediation settings.
- (2) Once the simulation enters the core execution phase, the controller initiates the interaction loop, sequentially activating each agent according to a predefined schedule. This structured activation order ensures deterministic coordination and stable temporal progression throughout the mediation process. Each round begins with the mediator's FSM module evaluating the current dialogue history to update the mediation stage (e.g., transitioning from exploration to negotiation). Based on the strategic

objectives associated with the updated stage, the mediator then either generates direct conversational responses or invokes appropriate STK tools (e.g., risk assessment or strategic intervention) to guide the negotiation. Upon receiving new information, each litigant agent first triggers the DBT module to update its belief state regarding the opponent's intentions. The agent subsequently produces a response conditioned on both its updated beliefs and its predefined role configuration, thereby ensuring behavior consistent with its strategic posture. Meanwhile, the agent updates its internal acceptance state with respect to the current mediation proposal, which in turn influences future decision-making and negotiation dynamics. Through this iterative belief–decision–action cycle, the interaction loop enables coherent, adaptive, and strategically grounded behaviors across extended negotiations.

- (3) At the end of each interaction round, the controller evaluates the global system state to determine whether the simulation should terminate. The mediation process is concluded once any of the following conditions is satisfied. First, if both litigant agents explicitly express agreement within the same round (*Agreed = True*), the case is classified as a successful mediation. Second, if the predefined maximum number of interaction rounds is reached without consensus—set to 10 rounds in our experiments—the process is deemed to have reached a deadlock. This limit allows sufficient opportunity for strategic development while preventing unbounded dialogue that may otherwise lead to strategic drift or unstable behaviors. Third, if either party explicitly withdraws from mediation and opts to pursue litigation, the case is categorized as a mediation breakdown. Together, these termination criteria provide clear and operational stopping rules, ensuring both procedural rigor and consistent evaluation across simulation runs.

The entire interaction process follows a fixed controller-coordinated execution order to ensure procedural consistency across different simulation runs.

4. Experimental Setup

4.1. Dataset and Backbone Models

To ensure the authority and practical relevance of the evaluation, we constructed a test set comprising 50 representative civil dispute cases. This scale balances diversity in case types with the need for controllable and reproducible comparisons of complex negotiation dynamics. It enables systematic assessment of agent performance across heterogeneous yet realistic mediation scenarios.

All cases were curated from the official Multi-Dispute Resolution Case Repository released by the Supreme People's Court of the People's Republic of China. The dataset covers several high-frequency judicial contexts, including contract breaches, tort compensation, labor disputes, neighborhood conflicts, and marriage and family issues. All cases used in this study were obtained from publicly available judicial mediation records released by official sources. During preprocessing, only case summaries, disputed issues, mediation processes, and final settlement information relevant to simulation were retained. Personally identifiable information, including names, addresses, contact information, and other sensitive identifiers, was removed or anonymized. The dataset was used solely for academic research and simulation purposes, and no private or non-public legal documents were involved.

Each case was further cleaned and transformed into a structured format containing three components:

- (1) Case summary, which serves as the initial contextual input for the simulation;
- (2) Disputed issues, which are used to initialize each litigant agent's goals and bottom-line constraints;

- (3) Expert mediation records, including the real-world mediation process and final settlement agreement from the original case.

These records are treated as ground truth references for subsequent LLM-as-a-Judge evaluation, thereby providing a structured reference for comparative evaluation of generated mediation outcomes. To investigate the generalization capability of the CogMed framework under varying computational constraints, three representative models from the Qwen series were selected as the cognitive backbone for the Mediator Agent. Qwen3-8B: a lightweight model representing low-resource settings with limited computational capacity. Qwen3-32B: a mid-scale model representing moderate-resource conditions and serving as a general-purpose baseline. Qwen-Plus: a high-performance, large-scale proprietary model, used as an upper-bound performance reference in our experiments.

This tiered configuration enables systematic evaluation of how cognitive architectural design interacts with model scale, and facilitates analysis of whether the proposed framework can compensate for limited parameter capacity in resource-constrained environments.

4.2. Dataset Construction and Preprocessing

The mediation dataset used in this study was constructed from publicly available civil dispute mediation cases released through the official Multi-Dispute Resolution Case Repository of the Supreme People's Court of the People's Republic of China. To improve diversity of mediation scenarios, the selected cases cover multiple common categories of civil disputes, including contract disputes, labor conflicts, tort compensation, neighborhood disputes, marriage and family disputes, and debt-related mediation cases.

Case selection primarily followed three criteria: (1) the case contained a relatively complete mediation process and final settlement record; (2) the disputed issues and party positions could be clearly identified from the original documentation; and (3) the case involved multi-round negotiation characteristics suitable for mediation simulation. Cases with severely incomplete records, highly ambiguous procedural descriptions, or insufficient mediation interaction details were excluded during preprocessing.

During dataset construction, all cases were transformed into structured mediation scenarios. Specifically, the preprocessing procedure included anonymization of personally identifiable information, extraction of case summaries, identification of disputed issues, reconstruction of litigant goals and bottom-line constraints, and organization of mediation outcomes into standardized evaluation references. The preprocessing process was conducted manually under unified formatting rules to improve consistency across simulation instances. Because the current dataset spans multiple heterogeneous dispute categories with relatively limited sample sizes per category, the present study primarily aims to provide a controlled exploratory evaluation of the proposed framework rather than a definitive large-scale benchmark analysis.

It should be noted that the current dataset size remains relatively limited for large-scale legal AI evaluation. Consequently, the present study is intended primarily as a controlled exploratory evaluation of cognitively structured mediation architectures rather than a comprehensive benchmark study. In addition, because the selected cases originate primarily from publicly available Chinese civil mediation records, potential selection bias and limited cross-jurisdictional generalization may still exist. Future work will therefore incorporate larger and more diverse mediation datasets to further examine robustness, category-level stability, and generalization capability across heterogeneous legal environments.

4.3. Experimental Design and Variable Control

To systematically evaluate the effectiveness of the CogMed framework and disentangle the contributions of its internal mechanisms, this study adopts a strictly controlled single-

factor experimental design, in which only one architectural component is varied at a time while all other conditions remain unchanged. This protocol ensures fair comparison and enables precise attribution of performance differences to specific cognitive modules.

Concretely, the experiment was organized as a controlled component-isolation study. Five mutually exclusive configurations were constructed under the same dataset, controller protocol, interaction limit, and evaluation procedure. First, a ReAct-based baseline agent was implemented as a representative general-purpose LLM agent without explicit cognitive structuring. This baseline relies on ordinary role-based prompting for end-to-end dialogue generation and therefore provides a reference point for evaluating whether additional mediation-oriented structure is beneficial. Second, the CogMed-Full configuration activates all three cognitive modules, namely FSM, STK, and DBT. Third, three ablated variants were constructed by removing one module at a time: CogMed-noFSM, CogMed-noSTK, and CogMed-noDBT. By keeping all other settings unchanged while removing a single component, the ablation design enables a more systematic examination of how procedural control, executable strategic intervention, and opponent-state tracking, respectively, influence mediation performance. This experimental design is therefore intended primarily to evaluate the internal contribution of the proposed cognitively coordinated architecture under controlled simulation settings, rather than to establish an exhaustive benchmark against all existing multi-agent or legal-domain negotiation systems.

To ensure that performance differences can be accurately attributed to variations in the mediator's cognitive architecture, strict variable control protocols were enforced throughout all experiments. In particular, only the mediator agent's cognitive framework configuration and its backbone model (8B/32B/Plus) were varied across settings, while all other components were held constant.

To maintain environmental consistency, the litigant agents, the simulation controller, and the LLM-based judge used for final evaluation were uniformly implemented with the highest-capacity Qwen-Plus model. It should be noted that the current evaluation setting may still introduce potential model-family preference bias because several auxiliary components and the LLM-as-a-Judge evaluator are implemented within the same model family. Accordingly, the reported OQS results should primarily be interpreted as comparative evaluation signals under controlled experimental settings rather than definitive measurements of real-world mediation quality. Their sampling behavior was further stabilized by fixing the temperature parameter to 0.3. This design effectively eliminates potential confounding effects arising from variability in auxiliary agent capabilities, thereby ensuring that the observed performance differences genuinely reflect the efficacy of the mediator's cognitive design and preserving the objectivity and reliability of the evaluation. To further reduce stochastic variability, all experiments were conducted under fixed decoding configurations and standardized execution procedures.

4.4. Evaluation Metrics

To comprehensively assess mediation performance, we adopt a hybrid evaluation framework that integrates both quantitative and qualitative criteria, enabling joint measurement of effectiveness, efficiency, and solution quality.

First, Success Rate (SR) is defined as the proportion of simulations in which both litigant agents reach an agreed = True state within the predefined maximum number of interaction rounds (10 rounds). This metric serves as the primary outcome-oriented indicator, reflecting the framework's ability to successfully guide negotiations toward consensus.

Second, Turns to Agreement (TTA) measures the average number of dialogue rounds required to reach an agreement, computed only over successful cases. This metric captures mediation efficiency by quantifying how quickly consensus can be achieved.

Third, Overall Quality Score (OQS) is introduced to evaluate the substantive quality of mediation outcomes. Following the LLM-as-a-Judge paradigm, an expert-level model—Qwen-Plus—is employed as a neutral evaluator. The judge semantically compares the generated settlement agreements with the corresponding ground-truth agreements from the official case repository of the Supreme People’s Court of the People’s Republic of China. Performance is assessed across three dimensions: solution feasibility, procedural professionalism, and strategic creativity. Each case is rated on a 1–10 scale, and the final OQS is computed as the average of these scores, providing a comparative reference for assessing the relative quality and consistency of generated mediation outcomes across different experimental configurations.

Together, these metrics offer a balanced evaluation of whether the framework not only reaches agreements, but does so efficiently and with high-quality, legally sound solutions.

It should be noted that the LLM-as-a-Judge protocol adopted in this study is intended as a scalable auxiliary evaluation approach rather than an authoritative substitute for professional legal experts. Although such automatic evaluation methods provide useful comparative signals under standardized experimental settings, they may not fully capture the nuanced legal reasoning, procedural fairness, and practical mediation expertise required in real-world judicial contexts. Therefore, future work will further incorporate evaluations conducted by professional mediators and legal practitioners to validate the consistency and reliability of the proposed evaluation framework. In addition, the current study primarily relies on a single-model-family LLM-as-a-Judge evaluation protocol, which may introduce potential evaluation preference bias. Although this approach provides scalable comparative assessment under controlled simulation settings, future work will incorporate independent judge models, inter-judge agreement analysis, and professional human expert evaluation to further improve evaluation robustness and practical validity.

4.5. Implementation Details and Reproducibility

To improve methodological transparency and reproducibility, additional implementation details are provided in this study.

All experiments were conducted under fixed decoding configurations. Specifically, the temperature parameter was fixed at 0.3 for all agents and evaluation modules to reduce stochastic variance and maintain behavioral consistency across different experimental settings. The maximum interaction length was limited to 10 rounds under the unified controller framework.

The CogMed framework follows a standardized round-based interaction procedure coordinated by a centralized controller. During each interaction round, the controller sequentially activates the mediator and litigant agents according to a predefined execution order. The mediator first evaluates the current mediation stage through the FSM module and subsequently performs either standard dialogue generation or STK-based strategic intervention. Litigant agents then update their internal belief states through the DBT mechanism before generating responses conditioned on dialogue history and role-specific configurations.

To further improve interpretability, representative interaction examples and mediation dialogue excerpts are additionally provided in Appendix A. Each mediation case is transformed into a structured scenario representation containing case summaries, disputed issues, litigant goals, bottom-line constraints, and personality priors before simulation initialization.

5. Results and Analysis

5.1. Main Results: Framework Effectiveness

To comprehensively evaluate the effectiveness of the proposed CogMed framework, we conducted a series of controlled comparative experiments across three backbone models of different scales. In total, 15 configurations (3×5) were evaluated, corresponding to three model sizes and five architectural settings. Each configuration was tested on 50 standardized mediation cases, resulting in 750 complete simulation runs overall.

Table 1 summarizes the core performance metrics of each backbone model before and after integrating the CogMed framework, providing a direct comparison of baseline and cognitively enhanced agents.

Table 1. Comparative experimental results under different backbone models and framework configurations.

Backbone Model	Configuration	Success Rate (SR)	Overall Quality Score (OQS)	Turns to Agreement (TTA)
Qwen3-8B	Baseline	16%	3.45 (± 0.65)	3.20
	CogMed-Full	46%	4.72 (± 0.95)	6.90
Qwen3-32B	Baseline	42%	4.78 (± 0.85)	6.45
	CogMed-Full	68%	5.55 (± 0.40)	7.20
Qwen-Plus	Baseline	62%	5.42 (± 0.55)	5.80
	CogMed-Full	78%	5.35 (± 0.28)	5.65

Note: Values in parentheses represent the standard deviation of the Overall Quality Score (OQS) across the evaluated mediation cases. The reported “ $\pm$ ” values represent the standard deviation of OQS across the evaluated mediation cases ($N = 50$). No repeated runs or multi-seed averaging were performed in the current study.

From the results reported in Table 1, it can be clearly observed that the CogMed-Full configuration consistently achieves higher success rates (SR) than the corresponding baselines across all backbone model scales, demonstrating the potential effectiveness of the proposed cognitive architecture under the current controlled experimental settings. This trend can be observed not only for high-capacity models but also for lightweight settings, indicating that the benefits of structured cognitive design generalize well across different computational regimes.

It is worth noting that the observed increase in Turns to Agreement (TTA) should not be interpreted as reduced efficiency. Instead, the longer interaction horizon reflects a behavioral shift from premature or rapid failures toward more sustained and constructive negotiation processes. In other words, agents engage in deeper deliberation and extended bargaining before reaching consensus, which ultimately contributes to higher agreement rates and improved solution quality rather than slower performance.

Building upon the quantitative results in Table 1, several key observations can be further highlighted.

First, improvements in SR. For the smaller and mid-scale models, the full CogMed framework yields substantial absolute gains of +30% and +26% for the 8B and 32B backbones, respectively, markedly enhancing their practical usability in real-world mediation tasks. Notably, even for the already strong Qwen-Plus model, the framework continues to provide additional benefits, pushing the success rate to a high level of 78%, which indicates that cognitive augmentation remains effective beyond mere parameter scaling.

Second, improvements in OQS and stability. CogMed also enhances the professional quality of generated agreements for small- and medium-scale models, with OQS increases of +1.20 and +0.75, respectively. More importantly, a reduction in performance variance is observed for the 32B and Plus models, where the standard deviation decreases to ± 0.40 and ± 0.28 . This reduction suggests that the framework not only improves average

performance but also produces more stable and consistent outcomes. Such stabilization indicates that explicit cognitive constraints help regulate agent behavior, leading to more controllable and reliable mediation services.

It should be noted that the current experimental results are primarily intended to provide comparative empirical observations under controlled simulation settings. Due to the stochastic nature of large language model inference, the reported performance differences may still be affected by sampling variability. Although standard deviations are reported to partially reflect output stability, more comprehensive statistical analyses, including repeated experiments with multiple random seeds, confidence interval estimation, significance testing, and sensitivity analysis, will be further explored in future work to strengthen the robustness evaluation of the proposed framework.

5.2. Key Finding: The “Cross-Scale Outperformance” Effect

One notable observation in this study is that cognitively structured coordination mechanisms may partially compensate for limitations associated with model scale under the current experimental setting, giving rise to an observed cross-scale outperformance tendency. As shown in Table 1, the medium-scale Qwen3-32B equipped with the full CogMed architecture achieves an Overall Quality Score (OQS) of 5.55 and a Success Rate (SR) of 68%. These results not only substantially surpass its own baseline configuration, but also consistently outperform the much larger Qwen-Plus baseline model (OQS 5.42, SR 62%), despite the latter having several times more parameters. Interestingly, although CogMed substantially improves the Success Rate (SR) of Qwen-Plus, the corresponding OQS exhibits a slight decrease compared with the unconstrained baseline configuration. One possible explanation is that highly capable large-scale models may already possess strong intrinsic reasoning flexibility, such that additional procedural constraints introduced by FSM, STK, and DBT partially restrict open-ended strategic exploration. This observation suggests that the effectiveness of cognitively constrained architectures may depend on the balance between structural guidance and model autonomy across different model scales.

This observation lends strong support to the hypothesis that architectural constraints can compensate for limited model capacity in specialized domains. However, the current evidence remains limited to a relatively small mediation dataset and a single primary model family under controlled simulation settings. Therefore, broader validation across multiple model architectures, larger mediation corpora, and more comprehensive statistical analyses will be necessary before establishing stronger conclusions regarding the generality of this phenomenon. For complex and long-horizon strategic tasks such as mediation, a moderately sized model endowed with explicit process control, executable intervention mechanisms, and belief modeling capabilities may exhibit competitive or even advantageous mediation performance relative to larger yet structurally unconstrained models under certain controlled mediation settings.

From a practical perspective, this finding highlights a promising alternative to brute-force parameter scaling. Rather than relying solely on increasingly large models, carefully designed cognitive architectures offer a more computationally efficient pathway toward high-performance legal AI systems, suggesting the potential applicability of cognitively structured mediation architectures in resource-constrained simulated mediation environments.

5.3. Ablation Analysis: Deconstructing the Cognitive Modules of CogMed

The ablation analysis was designed to provide a controlled examination of how each architectural component contributes to mediation performance. Specifically, FSM, STK, and DBT were individually removed from the full CogMed configuration while all other settings remained unchanged. This design allows the observed performance changes to be

interpreted as preliminary evidence of the functional role of each component within the current architecture. To precisely quantify the independent contributions of the three core components—FSM, DBT, and STK—we conducted a comprehensive ablation study. By systematically removing each module from the full framework while keeping all other factors constant, the resulting performance variations can be directly attributed to the corresponding cognitive capability.

The experimental results (see Table 2) reveal scale-dependent patterns in cognitive reliance, indicating that models of different sizes exhibit heterogeneous dependencies on specific cognitive mechanisms. This analysis provides deeper insight into how each module contributes to overall mediation performance and clarifies the functional roles of process control, belief modeling, and strategic intervention within the CogMed architecture.

Table 2. Ablation Results of the Full CogMed Model.

Backbone Model	Configuration	Success Rate (SR)	Overall Quality Score (OQS)	Turns to Agreement (TTA)
Qwen3-8B	Baseline	16%	3.45 (± 0.65)	3.20
	CogMed-Full	46%	4.72 (± 0.95)	6.90
	CogMed-noDBT	32%	4.15 (± 1.10)	8.10
	CogMed-noFSM	38%	4.25 (± 1.00)	6.60
	CogMed-noSTK	36%	4.35 (± 1.15)	8.20
Qwen3-32B	Baseline	42%	4.78 (± 0.85)	6.45
	CogMed-Full	68%	5.55 (± 0.40)	7.20
	CogMed-noDBT	48%	4.85 (± 0.92)	8.10
	CogMed-noFSM	56%	5.28 (± 1.15)	6.50
	CogMed-noSTK	28%	4.65 (± 1.25)	7.90
Qwen-Plus	Baseline	62%	5.42 (± 0.55)	5.80
	CogMed-Full	78%	5.35 (± 0.28)	5.65
	CogMed-noDBT	66%	5.38 (± 0.75)	5.30
	CogMed-noFSM	68%	5.05 (± 0.95)	5.10
	CogMed-noSTK	48%	5.25 (± 1.05)	6.40

Note: Values in parentheses represent the standard deviation of the Overall Quality Score (OQS) across the evaluated mediation cases. The noFSM configuration is used to examine the effect of removing explicit procedural stage control. The noSTK configuration is used to examine the effect of removing executable mediation intervention actions. The noDBT configuration is used to examine the effect of removing dynamic opponent-state tracking from litigant agents. Therefore, the ablation study does not merely compare different prompt variants but provides a controlled analysis of how different structural mechanisms affect long-horizon mediation simulation.

The cross-model comparison further uncovers several important cognitive mechanisms underlying the effectiveness of CogMed. Across all backbone models, the ablation results consistently identify the Strategic Toolkit (STK) as the most decisive component. Removing STK leads to substantial and systematic performance degradation, with success rates dropping by 40% and 30% for Qwen3-32B and Qwen-Plus, respectively. This pattern highlights a fundamental principle in professional negotiation tasks: possessing strong reasoning capabilities does not necessarily translate into effective outcomes. In complex adversarial settings, knowing what to do is not equivalent to being able to execute it successfully. By encapsulating expert interventions—such as Best Alternative to a Negotiated Agreement (BATNA) analysis—into atomic and executable actions, STK bridges the gap between high-level strategy generation and concrete operational behavior, thereby converting abstract intent into reliable and actionable decisions. It should also be noted that the ablation results should not be interpreted as proving that the three modules contribute independently in a strictly additive manner. In particular, the noSTK configuration indicates that executable intervention actions may function as an enabling component for the interaction between FSM-based procedural control and DBT-based opponent-state tracking.

Without STK, the system may still maintain procedural stages and belief updates, but lack concrete mediation actions to advance the negotiation, which can lead to repetitive or stalled interaction patterns.

At the same time, Dynamic Belief Tracking (DBT) functions as a general-purpose stabilizer of social cognition rather than a model-specific enhancement. Although its quantitative impact appears most pronounced for the medium-scale model (with an approximate 20% performance drop upon removal), disabling DBT consistently increases conversational instability across all settings, frequently manifesting as abrupt negotiation breakdowns. This observation suggests that explicit Theory-of-Mind modeling plays a crucial role in maintaining temporal coherence and strategic stability in long-horizon, non-cooperative interactions, enabling agents to reason about opponents in a more persistent and structured manner.

Finally, an intriguing scale–structure trade-off emerges. As backbone model size increases, the marginal gains introduced by the cognitive architecture gradually diminish and may even exhibit mild antagonistic effects on certain indicators. This phenomenon does not contradict the general advantages of large-scale models; rather, it underscores that, for process-intensive and strategically constrained professional tasks, structured cognitive design is often more critical for behavioral controllability than further parameter expansion. Together, these findings reinforce the view that architectural intelligence, rather than sheer scale alone, constitutes the primary driver of robust performance in complex mediation environments.

With respect to the Overall Quality Score (OQS), a differentiated trend can be observed across model scales. The lightweight Qwen3-8B exhibits a substantial improvement of +1.27, while the mid-scale Qwen3-32B also benefits from a clear gain of +0.77, indicating that the cognitive architecture consistently enhances the professionalism and coherence of solutions generated by small- and medium-capacity models. In contrast, the large-scale Qwen-Plus shows a slight fluctuation after integration of the framework, with OQS decreasing marginally from 5.42 to 5.35. This phenomenon reflects an inherent trade-off between structural constraints and generative flexibility: for highly capable foundation models, enforced procedural control via FSM and predefined STK actions may modestly restrict the upper bound of free-form generation, leading to a small reduction in subjective quality scores.

Nevertheless, this minor trade-off does not undermine the overall benefit of the framework. By imposing explicit strategic structure on the interaction process, CogMed still robustly improves the Success Rate (SR) of the large model from 62% to 78%, demonstrating more reliable and goal-directed outcomes. These findings suggest that, in the era of large-scale models, the role of cognitive architectures gradually shifts from capability augmentation to behavioral regulation. Rather than compensating for insufficient reasoning capacity, the framework primarily serves to stabilize, constrain, and guide decision-making, ensuring consistent performance in complex, long-horizon mediation tasks.

5.4. Qualitative Case Analysis

To intuitively demonstrate how the CogMed framework converts the tacit knowledge embedded in legal mediation into explicit and actionable interventions, this section presents an in-depth analysis of a representative case from the experimental set, namely the private lending dispute between Jing, Zhang, and Meng.

(1) Scenario configuration

① Case focus

The dispute concerns the recovery of an outstanding principal of RMB 120,000 together with an additional RMB 79,000 in accrued interest, while the guarantor, Meng, refuses to assume liability by declining to sign the agreement.

② Party positions

Jing insists on a monthly interest rate of 1.5% and demands joint and several liability, whereas Zhang, citing financial hardship, requests a full interest waiver and proposes repayment over 40 installments.

③ Negotiation status

By the end of the third interaction round, both parties remain firmly entrenched in their positions. Disagreements persist over the repayment term (24 vs. 40 installments) and the degree of interest reduction, causing the dialogue to fall into a repetitive and stagnant loop.

(2) Mediation process trace

As shown in Table 3, the state transitions and intervention effects of the CogMed framework throughout the complete simulation cycle. It records how the mediation process evolves across different procedural stages and illustrates how the mediator dynamically applies structured cognitive modules—such as FSM-guided stage control, STK-based strategic interventions, and DBT-driven belief updates—to regulate negotiation progress and resolve deadlocks.

Table 3. Representative Mediation Case: Private Lending Dispute between Jing, Zhang, and Meng.

Round	Mediation Stage	Mediator Intervention Logic	Dialogue Effect
Round2	Exploration	summarize_and_clarify invoked. The mediator proactively consolidates three core disputed issues—interest rate, repayment schedule, and guarantor liability—instead of passively responding.	From emotion to facts: scattered arguments are anchored to three negotiable parameters.
Round3	Negotiation	Deadlock detected. Jing insists on 24 installments, while Zhang demands 40; the parties' bottom lines do not overlap.	Identification of a potential Zone of Possible Agreement (ZOPA).
Round4	Negotiation	propose_bridging_solution invoked. The mediator introduces a compromise package: (1) 30 installments; (2) 60% settlement of historical interest; (3) key innovation—vehicle collateral replacing guarantor liability.	Breaking the zero-sum structure: the introduction of vehicle collateral bypasses the impasse over Meng's responsibility.
Agreement reached at Round 6	Agreement Drafting	FSM state locked. Upon detecting convergence of intentions, the system enforces commitment to settlement terms to prevent backtracking.	Consensus reached: Zhang accepts the 60% interest settlement, Jing accepts the collateral arrangement; mediation successfully concludes.

(3) Mediation outcome analysis

At the early stage of the mediation, both parties exhibited severe positional confrontation. The lender, Jing, insisted on a monthly interest rate of 1.5% and demanded that Meng assume joint and several liability, whereas the borrower, Zhang, citing financial hardship, requested a complete interest waiver together with an extended repayment schedule. Under such polarized demands, the DBT (Dynamic Belief Tracking) module of

the litigant agents recorded a low level of perceived flexibility on both sides, indicating limited willingness to compromise and a high risk of negotiation breakdown.

Guided by the FSM (Finite State Machine), the mediator first entered the exploration stage to stabilize the interaction. Rather than responding reactively to emotional accusations, the mediator proactively elicited detailed information regarding Zhang's financial capacity and Jing's minimum acceptable terms. This structured information-gathering process prevented the dialogue from devolving into affect-driven exchanges and instead anchored the discussion to two negotiable core variables—repayment feasibility and interest standards—thereby laying a rational foundation for subsequent bargaining.

The turning point of the mediation emerged in Round 4. Confronted with the deadlock over the repayment schedule (24 versus 40 installments) and guarantor liability, the mediator invoked the `propose_bridging_solution` module from the Strategic Toolkit (STK). Instead of passively waiting for unilateral concessions, the mediator proactively introduced a creative compromise by proposing the use of vehicle collateral to replace the guarantor's responsibility, while suggesting that historical interest be settled at 60%. This intervention effectively expanded the solution space beyond the parties' original positions, transforming the negotiation from a rigid distributive conflict into a more flexible, integrative bargaining process.

Subsequently, the mediator employed `perform_reality_check` to guide both parties in reassessing litigation risks, expected recovery time, and potential legal costs. This structured reality-based evaluation encouraged more rational expectations and directly facilitated updates in the agents' belief states, as reflected in the DBT module's increased flexibility scores.

Finally, under the guidance of the FSM-controlled agreement drafting stage, the system detected convergence in the parties' intentions and stabilized the negotiation toward commitment. An agreement was reached in Round 6, achieving a smooth transition from positional confrontation to interest alignment. This trajectory clearly illustrates how explicit cognitive interventions can transform an initially adversarial stalemate into a cooperative and mutually acceptable settlement.

(4) Summary

Under the guidance of the cognitive framework, the mediator demonstrates markedly stronger strategic coherence than a baseline LLM, exhibiting consistent improvements in reasoning stability, intervention effectiveness, and long-horizon control.

① From positions to interests

Through the DBT module's continuous tracking of each party's perceived risks and intentions, the mediator is able to identify shifts in flexibility and strategically apply `perform_reality_check` to recalibrate expectations. This mechanism steers the parties away from rigid positional bargaining and back toward an interest-based negotiation space grounded in realistic assessments.

② Structured intervention capability

The STK transforms tacit judicial expertise—such as debt substitution, collateralization, and other practical mediation tactics—into explicit, executable Chain-of-Thought (CoT) reasoning procedures. By operationalizing these professional strategies as callable actions, the framework resolves a key limitation of baseline models, which often “know that mediation is needed” but lack concrete means to break deadlocks.

③ Stage-level coherence

The FSM mechanism maintains macro-level procedural consistency throughout multi-round interactions, ensuring that the mediator persistently advances toward agreement formation. This explicit process control effectively mitigates the common issue of strategy

drift in long-context dialogues and preserves goal-directed behavior across the entire negotiation trajectory.

Collectively, these effects demonstrate that structured cognitive constraints not only enhance decision quality but also provide stable and controllable mediation dynamics, highlighting the practical value of architectural guidance beyond raw language modeling capacity.

Although the proposed framework generally improves mediation stability, several failure patterns can still be observed in complex scenarios involving persistent adversarial behaviors, emotional escalation, or ambiguous compromise spaces. In such cases, agents may exhibit repetitive strategic loops, ineffective intervention sequences, or prolonged deadlock states, further reflecting the limitations discussed in Section 8.

6. Ethical and Legal Considerations

Because legal mediation involves sensitive social conflicts and potentially consequential legal interests, the ethical and legal implications of automated mediation systems require careful consideration. First, although the cases used in this study are public and anonymized, case repositories may still reflect institutional, regional, or historical biases. Models trained or evaluated on such data may therefore reproduce biased assumptions regarding dispute types, settlement preferences, or acceptable compromise patterns.

Second, the proposed framework should not be interpreted as an autonomous legal decision-making system. CogMed is intended to support simulation, training, and decision assistance rather than replace qualified mediators, lawyers, or judges. Final mediation decisions should remain under human professional supervision, especially when legal rights, vulnerable parties, or high-stakes disputes are involved.

It is also important to distinguish AI-generated mediation simulations from real mediation practice. The dialogues, intervention strategies, and settlement proposals generated by CogMed should be understood as hypothetical simulation outputs rather than verified legal facts, binding mediation records, or authoritative legal recommendations. Because large language models may generate plausible but factually inaccurate or legally unsupported content, there is a risk that AI-generated mediation processes may create an illusion of procedural validity while containing fabricated reasoning, incomplete evidence assessment, or misleading settlement logic.

In addition, historical mediation data may contain unequal bargaining patterns, institutional preferences, or socially biased assumptions. If such patterns are directly reproduced by an automated system, the model may cascade biased or inequitable practices by normalizing compromise outcomes that disadvantage weaker parties. For example, an AI mediator may overemphasize settlement efficiency, pressure vulnerable parties to accept unfair concessions, or treat historically common but normatively problematic outcomes as desirable references. Therefore, mediation success should not be evaluated solely by agreement formation, but also by procedural fairness, voluntariness, transparency, and protection of legitimate rights.

To mitigate these risks, any practical use of CogMed should incorporate strict safeguards, including human professional review, verification of factual claims, independent legal assessment of proposed settlements, transparency regarding AI involvement, traceability of generated recommendations, and bias-oriented auditing of mediation outcomes. The system should not be allowed to independently determine legal responsibility, pressure parties into settlement, or replace qualified mediators. Its outputs should remain subject to human oversight and institutional accountability.

Third, transparency is essential for responsible use. Users should be informed when AI-generated mediation suggestions are involved, and the reasoning process, data sources,

and limitations of the system should be made clear. In addition, human mediators should retain the authority to review, reject, or revise any system-generated recommendation.

Fourth, potential misuse should be carefully prevented. Automated mediation tools may be misused to pressure parties into settlement, generate biased negotiation strategies, or create an illusion of legal authority. Therefore, deployment of such systems should be accompanied by clear accountability mechanisms, professional oversight, data governance procedures, and safeguards against coercive or deceptive use.

Overall, the proposed framework should be regarded as a research-oriented, human-in-the-loop mediation simulation and decision-support tool. Its outputs should not be treated as real mediation records, legal advice, or binding settlement recommendations. Any transition from controlled simulation to real-world legal practice would require professional validation, ethical review, bias auditing, and institutional regulation.

7. Application Scenarios and Future Prospects

Targeting the inherently complex and unstructured strategic nature of legal mediation, this study proposes CogMed, a cognitively enhanced multi-agent simulation framework. By integrating explicit process control, structured interventions, and belief modeling into LLM-driven agents, the framework demonstrates substantial generalizability beyond the experimental setting. Both its architectural design philosophy and modular technical components exhibit broad practical potential, particularly in the context of the ongoing digital transformation of judicial systems and grassroots governance. In these domains, CogMed offers a scalable and intelligent solution for supporting diverse mediation tasks and improving the consistency, efficiency, and professionalism of dispute resolution services.

Specifically, the potential application scenarios of CogMed include, but are not limited to, the following four categories:

(1) Judicial pre-litigation mediation assistance and training

In court-led pre-litigation mediation settings, the CogMed framework may provide a conceptual reference for future integration into intelligent judicial mediation support systems as an auxiliary virtual mediation support tool, working alongside human judges or mediators rather than replacing them. By simulating the dynamic strategic interactions between disputing parties, the system can rapidly generate multiple potential negotiation trajectories and breakthrough solutions, enabling mediators to anticipate likely deadlocks, identify underlying interests, and select appropriate intervention strategies—such as reality checks or bridging proposals—at critical moments.

Beyond real-time assistance, the framework also holds significant value for professional training. Through high-fidelity case replay and structured strategy simulation, CogMed can serve as a practical sandbox environment for novice mediators, allowing them to experiment with different tactics, observe negotiation consequences, and internalize expert decision-making patterns. Such experiential learning accelerates the accumulation and formalization of tacit knowledge, thereby shortening the learning curve and improving the overall quality and consistency of mediation practice.

(2) Intelligent pre-resolution of grassroots disputes

Within grassroots mediation settings—such as neighborhood offices, community service centers, and local people’s mediation committees—practitioners frequently confront heavy caseloads, diverse dispute types, and uneven professional expertise. In such environments, CogMed can be deployed as a lightweight, localized system capable of operating efficiently on small- and medium-scale backbone models, such as Qwen3-8B and Qwen3-32B. This deployment strategy aligns closely with the framework’s experimentally validated

compensation effect, whereby structured cognitive design offsets limited model capacity while maintaining strong performance.

By inputting the basic facts and key points of contention, the system can conduct multi-round negotiation simulations and automatically generate preliminary mediation strategies, risk assessments, and communication guidance. Such decision support empowers frontline mediators to handle disputes more efficiently and consistently, improving both response speed and solution quality. This capability is particularly valuable for high-frequency community conflicts, including neighborhood disputes, property management disagreements, and family or marital issues, where rapid and pragmatic resolution is essential for maintaining local social stability.

(3) Negotiation simulation and strategy optimization for commercial disputes

In the domain of commercial mediation, enterprises frequently encounter complex strategic conflicts, including contract breaches, equity disputes, and intellectual property controversies, all of which involve prolonged negotiations and high economic stakes. In such scenarios, the Dynamic Belief Tracking (DBT) and Strategic Toolkit (STK) modules of CogMed can provide corporate legal teams and negotiation units with a dedicated sandbox environment for scenario-based rehearsal and strategy testing.

By explicitly modeling opponents' potential strategies, intentions, and psychological states, the system enables multi-round simulation of negotiation dynamics under varying assumptions. As a decision-support tool, CogMed allows teams to preview negotiation trajectories, evaluate the likely outcomes of different intervention tactics, and stress-test alternative settlement plans before entering real discussions. This process facilitates more informed and data-driven decision-making, helping organizations refine their bargaining strategies and increasing the probability of achieving favorable and sustainable agreements.

(4) Strategy adaptation for cross-cultural and cross-jurisdictional disputes

In international commercial mediation and cross-border civil disputes, where parties often operate under different cultural norms and legal systems, successful resolution frequently depends on accurately accounting for variations in communication styles, legal expectations, and decision-making conventions. In such contexts, CogMed provides a flexible and extensible framework for strategy adaptation across heterogeneous environments.

Specifically, the architecture supports the incorporation of cultural dimensions and jurisdictional characteristics through the expansion of agents' personality priors and belief models. By parameterizing these factors, the system can simulate culturally grounded behavioral patterns, potential misunderstandings, and divergent legal interpretations that may arise during negotiation. This capability enables mediators to anticipate cross-cultural friction points and proactively tailor intervention strategies to local expectations. As a result, CogMed assists practitioners in designing solutions that are both culturally sensitive and legally compatible, thereby reducing the likelihood of negotiation breakdowns caused by cognitive or normative mismatches.

In summary, the proposed CogMed multi-agent simulation framework operationalizes a hybrid cognitive architecture composed of Finite State Machines (FSMs), Strategic Toolkits (STK), and Dynamic Belief Tracking (DBT), through which the tacit expertise of human mediators is systematically transformed into explicit and computable modules. By embedding structured process control, executable interventions, and belief modeling into LLM-driven agents, CogMed bridges the gap between intuitive professional judgment and algorithmic decision-making, enabling complex mediation strategies to be replicated in a stable and reproducible manner.

Extensive empirical evaluations not only demonstrate consistent improvements in mediation success rates and solution quality, but also uncover a key phenomenon referred to as

the framework intelligence compensation effect: carefully designed cognitive architectures can enable small- and medium-scale models to match or even surpass the performance of substantially larger models in specialized mediation tasks. This finding underscores that architectural design, rather than sheer parameter scaling alone, plays a decisive role in professional, long-horizon negotiation settings.

Taken together, these results provide both theoretical justification and practical guidance for the low-cost, high-efficiency deployment of legal AI systems in resource-constrained grassroots judicial environments, offering a scalable pathway toward intelligent and trustworthy digital mediation services.

8. Limitations and Future Directions

Although CogMed demonstrates strong performance in simulating legal mediation, several limitations remain, particularly in handling highly irrational behaviors and adapting across heterogeneous legal systems. The current framework is primarily grounded in a bounded rationality paradigm, where agents are modeled as strategic decision-makers operating under structured incentives and logical reasoning. While this assumption aligns well with most institutional mediation settings, it may limit simulation fidelity in cases dominated by extreme emotional reactions or non-logical, stochastic behaviors. In such scenarios, where decisions are driven more by impulsive or affective factors than by strategic calculation, the current modeling approach may not fully capture the complexity of real-world dynamics. Incorporating richer psychological or affective models therefore constitutes an important direction for future work.

In addition, the experimental evaluation in this study is largely conducted within the context of the Chinese civil law system. As a result, the generalizability of the proposed strategies to other legal traditions—such as common law jurisdictions—and to multilingual or cross-cultural environments remains to be systematically validated. Future research should leverage larger and more diverse international case repositories to examine the transferability and robustness of the framework across different legal norms, procedural practices, and communication styles, thereby ensuring broader applicability and practical relevance.

In addition, the deployment of AI-driven mediation systems raises profound challenges related to legal ethics, algorithmic transparency, and accountability. As such systems transition from controlled simulation environments to real-world judicial assistance, critical questions emerge regarding how to ensure that AI-generated intervention strategies remain unbiased and procedurally fair, and how responsibility should be attributed when mediation outcomes result in inequitable or adverse consequences. These concerns extend beyond technical performance and directly touch upon the normative foundations of justice and due process.

Future research will therefore place greater emphasis on establishing comprehensive ethical governance and oversight mechanisms. This includes the development of transparent auditing procedures, interpretable decision processes, and clearly defined responsibility frameworks, as well as the incorporation of a human-in-the-loop evaluation paradigm to maintain meaningful human supervision over critical decisions. By balancing cognitive augmentation with institutional safeguards, the goal is to enhance mediation effectiveness while ensuring that technological deployment remains firmly aligned with the principles of judicial fairness and legal integrity. In addition, the current study remains primarily grounded in controlled simulation settings and standardized experimental configurations. Although the proposed framework demonstrates promising performance in simulated legal mediation tasks, the reported findings should be interpreted within the scope of the experimental environment rather than as definitive evidence of real-world deploy-

ment readiness. Several important challenges therefore remain open, including statistical robustness under stochastic LLM inference, broader comparisons against advanced multi-agent frameworks, validation through professional human experts, and adaptation to more diverse legal-cultural contexts.

Future work will further investigate dynamically reconfigurable strategy generation, richer emotional-social cognition modeling, repeated multi-seed experimental evaluation, human-centered assessment protocols, and broader real-world validation under heterogeneous mediation environments. By integrating cognitive augmentation with ethical governance and institutional safeguards, future research aims to improve both the practical reliability and responsible deployment of AI-assisted mediation systems. Future work will also investigate cross-model evaluation consistency and multi-judge assessment protocols to reduce potential model-family bias in automatic mediation quality evaluation. Future work will further incorporate broader comparative evaluation against role-playing multi-agent systems, memory-augmented negotiation agents, legal-domain mediation frameworks, and advanced prompt-engineering baselines to more comprehensively evaluate the generality and robustness of the proposed architecture.

Author Contributions: Conceptualization, J.C. and S.G.; methodology, J.C. and Y.M.; software, J.C.; validation, J.C., Y.M. and S.G.; formal analysis, J.C.; investigation, J.C. and Y.M.; resources, S.G.; data curation, J.C.; writing—original draft preparation, J.C.; writing—review and editing, Y.M. and S.G.; visualization, J.C.; supervision, S.G.; project administration, S.G.; funding acquisition, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hunan Provincial Education Science Planning Project, grant number XJK22QXX001. The APC was funded by Central South University.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request. The dataset was constructed from publicly available judicial mediation cases and further processed into structured mediation scenarios through manual annotation and reconstruction. To ensure the integrity and consistency of the research dataset, the processed data are not publicly available but may be shared for academic research purposes upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Representative Case and Experimental Process

Appendix A.1. Case Introduction

This paper selects a housing rental deposit dispute as a representative case to demonstrate the operation process of the CogMed multi-agent legal mediation simulation framework. In this case, the tenant Zhang San and the landlord Li Si signed a one-year housing lease contract. After the lease term expired, the landlord Li Si withheld the entire deposit of RMB 5000 on the grounds that “there were stains on the walls and minor wear on the furniture.” Zhang San believed that these conditions constituted normal wear and tear resulting from ordinary use and that the property had been cleaned before moving out. Therefore, Zhang San requested the full return of the deposit and an apology for the unreasonable withholding. Li Si, however, believed that the property had indeed been damaged and that the tenant had not used it properly, and therefore was only willing to return part of the deposit.

The core disputes in this case mainly focus on three aspects. First, how much of the deposit should be returned? Zhang San requested the return of the full RMB 5000 deposit, while Li Si was initially willing to return only RMB 1000. Second, a point of focus was whether the wall stains and furniture wear constituted normal wear and tear or damage caused by improper use by the tenant. The third point of focus was whether the landlord should apologize to the tenant. It can be seen that although the facts of this case are

not complicated, it simultaneously involves monetary disputes, liability determination, and emotional demands, making it highly representative of mediation scenarios.

During the simulated mediation process, the system configured three agent roles. Zhang San acted as the tenant and primarily sought to protect his right to the return of the deposit. Li Si acted as the landlord and emphasized his property losses. Judge Wang acted as a neutral mediator responsible for facilitating communication between the parties, identifying the key points of dispute, and proposing mediation solutions. At the beginning of the mediation, both parties maintained relatively firm positions. Zhang San insisted on a full refund of the deposit and an apology, while Li Si insisted on returning only a small portion of the deposit and required Zhang San to acknowledge improper use of the property.

As the mediation progressed, the mediator first summarized the views of both parties and clarified that the deposit amount, liability determination, and apology issue were the primary disagreements. Subsequently, the mediator conducted a reality check regarding the landlord's firm position, pointing out that if the dispute proceeded to litigation, sufficient evidence would be required to prove that the property damage had indeed been caused by the tenant; otherwise, the landlord might face the risk of losing the case and incurring additional costs. The mediator then proposed a compromise solution and guided both parties away from the opposing positions of "whether the full deposit should be returned" and "whether an apology must be given" toward a more practical and actionable resolution.

Ultimately, after multiple rounds of negotiation, the parties reached an agreement. The landlord Li Si agreed to return RMB 3500 of the deposit to Zhang San and provide a neutral statement stating that "both parties regret the dispute that occurred during the lease period." Zhang San accepted this proposal and waived any further claims. This case demonstrates that the CogMed framework can gradually guide originally emotional and confrontational disputes toward negotiable and executable mediation outcomes through strategies such as summarization and clarification, reality checking, and compromise solution generation.

Appendix A.2. Experimental Process Introduction

The experiments in this paper mainly revolve around the CogMed multi-agent legal mediation simulation framework. The basic idea of the experiments is as follows: first, case backgrounds, disputed issues, and actual mediation outcomes are extracted from real civil dispute cases; these pieces of information are then input into the multi-agent simulation system, where different agents play the roles of mediator and litigants, automatically generating multi-round mediation dialogues; finally, the mediation outcomes are evaluated.

During the data preparation stage, this paper constructed a case dataset containing typical civil disputes. Each case generally includes a case summary, disputed issues, handling method, handling result, dispute resolution basis, and dispute resolution essentials. On the one hand, this structured information is used to initialize simulation scenarios so that the agents can understand the case background. On the other hand, it serves as a reference standard for subsequent evaluation, allowing assessment of whether the simulated mediation outcomes are close to the actual mediation solutions.

In terms of agent configuration, the experiments mainly include mediator agents and litigant agents. The mediator agent is responsible for conducting the mediation, controlling the dialogue flow, and proposing intervention strategies. The litigant agents represent the two parties to the dispute and respond according to their own demands, bottom lines, and personality settings. To make the simulation process closer to real-world mediation, CogMed introduces a finite-state machine and a strategic toolkit on the mediator side, and a dynamic belief tracking mechanism on the litigant side. The finite-state machine is used to determine whether the mediation is currently in the opening, exploration, nego-

tiation, or agreement drafting stage. The strategic toolkit is used to perform professional interventions such as summarization and clarification, reality checking, and proposing bridging solutions. The dynamic belief tracking mechanism is used to simulate each party's judgment regarding the other party's intentions and willingness to compromise.

In the experimental design, a baseline group and a full CogMed framework group were established. The baseline group used only ordinary large language model role-playing for mediation and did not include any additional cognitive enhancement modules. The full CogMed framework group simultaneously enabled the finite-state machine, strategic toolkit, and dynamic belief tracking modules. To further analyze the actual contribution of each module, ablation groups were also established in which FSM, STK, or DBT was removed, respectively, in order to observe the impact of different modules on mediation success rates and solution quality.

The experimental evaluation combined quantitative and qualitative approaches. The quantitative metrics mainly included mediation success rate and average agreement round. The mediation success rate was used to measure whether the parties could reach an agreement within the specified number of rounds, while the average agreement round was used to evaluate mediation efficiency. The qualitative metric involved scoring the final mediation solution using a large language model judge, with a focus on feasibility, professionalism, and innovativeness. Through this evaluation approach, the experiments not only assessed whether mediation was successful but also examined whether the generated mediation solutions were reasonable, stable, and of practical reference value.

The experimental results show that the full CogMed framework can effectively improve the mediation success rate compared with the baseline group, particularly for small- and medium-scale models. This indicates that in long-horizon, multi-party, and highly conflictual tasks such as legal mediation, relying solely on the free generation capability of large language models can easily lead to strategy drift or weak mediation performance, whereas the introduction of structured cognitive modules enables the system to guide dialogues more steadily toward convergent solutions. Overall, the experiments verify the effectiveness of the CogMed framework in legal mediation simulation and also demonstrate that explicit cognitive frameworks can improve model performance in professional mediation tasks to a certain extent.

References

1. Zhou, Z.; Shi, J.-X.; Song, P.-X.; Yang, X.-W.; Jin, Y.-X.; Guo, L.-Z.; Li, Y.-F. Lawgpt: A chinese legal knowledge-enhanced large language model. *arXiv* **2024**, arXiv:2406.04614.
2. Cui, J.; Li, Z.; Yan, Y.; Chen, B.; Yuan, L. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv* **2023**, arXiv:2306.16092.
3. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. Camel: Communicative agents for “mind” exploration of large language model society. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 51991–52008. [CrossRef]
4. Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Zhang, C.; Wang, J.; Wang, Z.; Yau, S.K.S.; Lin, Z.; et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
5. Li, H.; Ai, Q.; Dong, Q.; Liu, Y. Lexilaw: A Scalable Legal Language Model for Comprehensive Legal Understanding. Technical Report. 2024. Available online: <https://github.com/CSHaitao/LexiLaw> (accessed on 25 June 2026).
6. 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* **2023**, arXiv:2303.08774.
7. Maxim, E.; Hopkins, M. From llm to nmt: Advancing low-resource machine translation with claude. *arXiv* **2024**, arXiv:2404.13813.
8. Yue, S.; Huang, T.; Jia, Z.; Wang, S.; Liu, S.; Song, Y.; Huang, X.; Wei, Z. Multi-agent simulator drives language models for legal intensive interaction. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, NM, USA, 29 April–4 May 2025.

9. Park, J.S.; O'Brien, J.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023.
10. Li, H.; Chen, J.; Yang, J.; Ai, Q.; Jia, W.; Liu, Y.; Lin, K.; Wu, Y.; Yuan, G.; Hu, Y.; et al. Legalagentbench: Evaluating llm agents in legal domain. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria, 27 July–1 August 2025.
11. Bai, G.; Liu, J.; Bu, X.; He, Y.; Liu, J.; Zhou, Z.; Lin, Z.; Su, W.; Ge, T.; Zheng, B.; et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv* **2024**, arXiv:2402.14762.
12. Kwon, D.; Shrestha, K.; Han, B.; Lee, E.H.; Lucas, G. Evaluating Behavioral Alignment in Conflict Dialogue: A Multi-Dimensional Comparison of LLM Agents and Humans. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025.
13. Chen, J.; Li, H.; Qin, M.; Zhou, Y.; Ren, Y.; Wang, W.; Liu, Y.; Wu, Y.; Ai, Q. Simulating Dispute Mediation with LLM-Based Agents for Legal Research. *arXiv* **2025**, arXiv:2509.06586.
14. Firat, M.; Kuleli, S. What if GPT4 became autonomous: The Auto-GPT project and use cases. *J. Emerg. Comput. Technol.* **2024**, *3*, 1–6. [\[CrossRef\]](#)
15. Yashar, T.; Nadiri, A. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv* **2023**, arXiv:2306.03314.
16. Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; et al. Chatdev: Communicative agents for software development. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024.
17. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In Proceedings of the ICLR 2024 Workshop on LLM Agents, Vienna, Austria, 11 May 2024.
18. Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Chan, C.-M.; Yu, H.; Lu, Y.; Hung, Y.-H.; Qian, C.; et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *arXiv* **2023**, arXiv:2308.10848.
19. Badhe, S. LegalSim: Multi-Agent Simulation of Legal Systems for Discovering Procedural Exploits. In Proceedings of the Natural Legal Language Processing Workshop 2025, Suzhou, China, 8 November 2025.
20. Huang, S.; Yong, W.; Wang, L.; Zhou, J. From measurement while drilling data to rock strength: A robust LMWOA-XGBoost model for predicting uniaxial compressive strength with SHAP interpretability. *Expert Syst. Appl.* **2026**, *325*, 132535. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.