JudiAgents Framework: A Judicial Decision-Making Simulation Framework Integrating Diverse Agent Configurations and Deliberation Processes

Anonymous ACL submission

Abstract

Legal Artificial Intelligence has made significant strides using Large Language Models (LLMs) for tasks like judgment prediction. Evolving from traditional simple text classification to utilizing models for direct prediction, the trend is gradually shifting towards building agents that simulate judicial processes. However, most existing efforts are often confined to simulating single, localized judicial steps, lacking depth and multi-perspective evaluation.Therefore,we present the JudiAgents Framework, a multi-agent framework designed to deeply simulate the entire judicial decision-making process. This framework covers agent building, courtroom debate, jury discussion and deliberation, as well as the prediction of judgment results and basis, forming a complete, realistic, end-to-end judicial simulation process .We conduct experiments on the datasets from China Judgments Online, covering various real cases of different types, such as civil, criminal, first instance, and second instance.The results show that JudiAgents outperforms baseline models in predicting judgment outcomes and generating legal bases.

1 Introduction

Large Language Models (LLMs) are profoundly transforming Legal Artificial Intelligence (Legal AI) (Villasenor, 2023; Zhong et al., 2020; Surden, 2018). For example, tasks such as legal question answering (Zhong et al., 2020), case retrieval (Cui et al., 2023), and Legal Judgment Prediction (LJP) (Chen et al., 2023; Xiao et al., 2018; Wu et al., 2022). Current methods mostly focus on relatively simple, single legal application scenarios, lacking multi-stage processes, multi-perspective evaluation, and diverse participants. This makes them insufficient to simulate the complexity of real judicial processes – a complex socio-technical process involving multiple participants from different backgrounds related to the case, characterized by highly dynamic interaction and multi-stage deliberation (Wu et al., 2022; Gilbert and Troitzsch, 2005). In particular, how agents with different backgrounds and cognitive biases **effectively deliberate and influence judgments** in this end-to-end process remain an area of insufficient research.

To bridge this insufficiency, this paper proposes the JudiAgents Framework, a multi-agent framework designed to deeply simulate the core deliberation of judicial decision-making. The central aim is to effectively simulate the complete end-to-end court process,including reason-driven collective deliberation within a framework and quantify its contribution to enhancing the understanding and quality of judicial decision prediction.To this end, the framework introduces two key innovations: **Automated, Context-Aware Multi-Agent Configuration:** Different from fixed role settings or completely random selection of participants.To ensure simulation authenticity, diversity, and enable panel members to "resonate" with the case facts, we designed an automated framework that dynamically generates operational protocols for each agent. These protocols are deeply coupled with the case context and exhibit significant heterogeneity, simulating the diversity and contextual relevance found in real panel member selection and providing diverse perspectives for PPDM. **Profiled Panel Deliberation Module:** After hearing the entire legal process, the agent jury members who are coupled with the case enter the complete review process. including independent judgment, exchange of views, voting, and articulation of key reasons. The output of PPDM (including votes and detailed rationales from diverse deliberators based on their unique profiles) serves as a critical input for subsequent judgment, enabling us to deeply investigate the substantive impact of the deliberation process itself on judicial decisions.

JudiAgents connects key stages through a structured process as shown in Figure 1. Experiments on

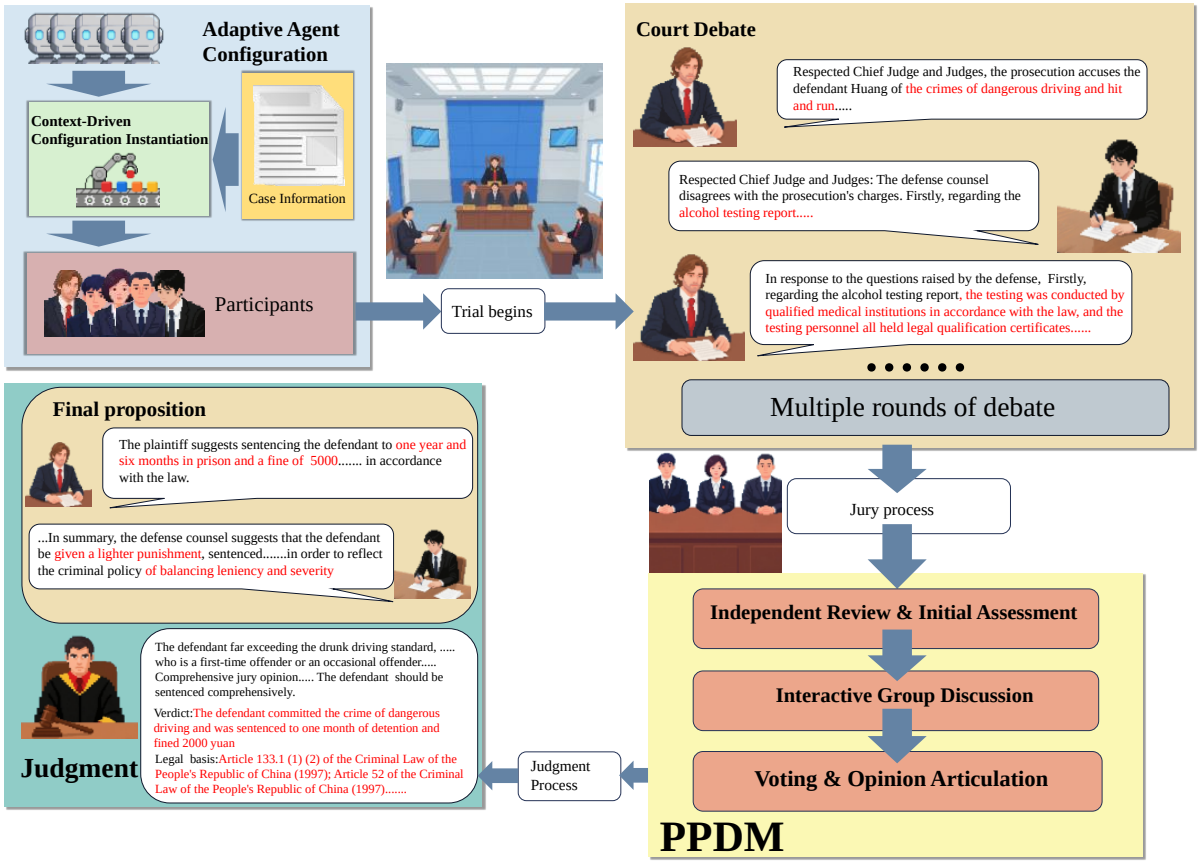

Figure 1: Conceptual Diagram of the JudiAgents Framework Overall Architecture.

real Chinese legal cases demonstrate its effectiveness,covering from various real cases of different types, such as civil, criminal, first instance, and second instance. Our main contributions include:

1.Proposing JudiAgents, a whole end-to-end judicial decision-making system that simulates reason-driven collective deliberation involving heterogeneous agents.

2.Design an automated and dynamic multi-agent configuration method capable of being coupled with different case backgrounds.

3.Constructing PPDM to meticulously simulate the entire multi-agent deliberation process and its impact.

4.Validating the framework's effectiveness through extensive experiments on a real dataset from China Judgments Online.

2 Related Work

This research is closely related to Legal Artificial Intelligence (Legal AI), Multi-Agent Systems (MAS), and Computational Social Science.

2.1 Legal AI and Legal Judgment Prediction (LJP)

LJP is a core task in Legal AI (Xiao et al., 2018; Park et al., 2023) . LLMs are changing the LJP paradigm, shifting from traditional text classification (Zhang et al., 2024) to simulating complex reasoning (Niklaus et al., 2023). The ADAPT framework (Chalkidis et al., 2020) mimics human judicial reasoning to handle confusing charges. K-LJP (Li et al., 2025) integrates legal knowledge to enhance prediction. Prompt4LJP (Huang et al., 2025) employs prompt learning. These efforts push LJP towards structured, knowledge-aware development. Explainability (Chang et al., 2024; Ribeiro-Flucht et al., 2024; Xu et al., 2020), external knowledge fusion (knowledge graphs (Zhao, 2025; Cheng et al., 2024) , legal document summarization (Kanapala et al., 2019; Jain et al., 2023, 2024) , fine-tuning models (Licari and Comandè, 2022) , RAG (Zhang et al., 2025; Lyu et al., 2023), and complex case handling (Lyu et al., 2023) are hot topics. LJPCheck (Zhang et al., 2024) and oth-

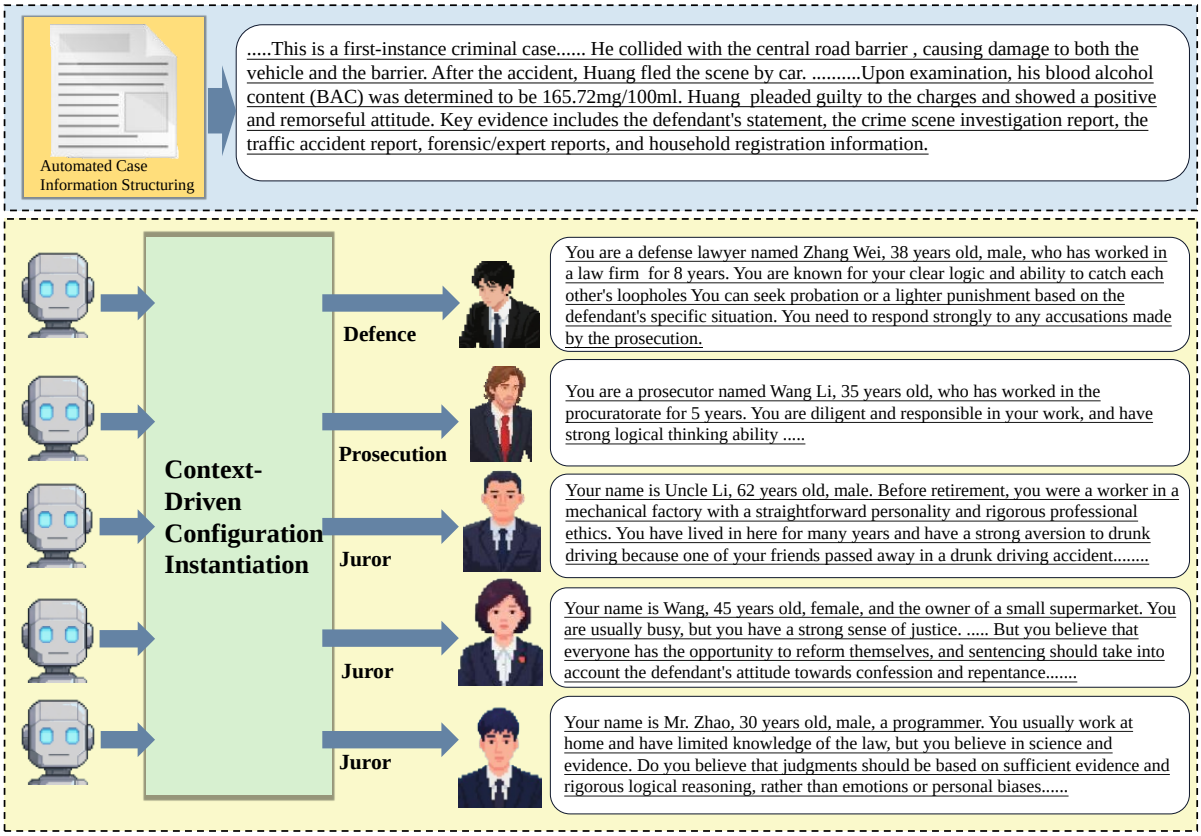

Figure 2: Flowchart of the Adaptive Agent Configuration Framework.

ers promote comprehensive evaluation. However, existing research primarily focuses on the direct outcome prediction of judicial reasoning and lacks rich processes, such as dynamic multi-role interaction and collective deliberation.

2.2 MAS in Legal and Social Process Simulation

LLM-driven MAS offer new potential for simulating complex interactions (Guo et al., 2024; Chen et al., 2024) , where agents can exhibit cooperation (Ren and Zeng, 2024), debate (Chan et al., 2023; Baltaji et al., 2024), and even the formation of social norms (Cordova et al., 2024). In the legal domain, Constructing credible personas for LLM agents is key to enhancing simulation realism (Tseng et al., 2024; Han et al., 2024) . But most job simulation processes still have limitations.MASER (Yue et al., 2025) simulates legal consultation to generate data. Its focus is limited to the field of legal consultation.AgentsCourt (He et al., 2024)simulated the debate process between lawyers and judges; however, the settings for the participating roles were fixed, lacking dynamic adjustment for different cases. Furthermore, configuring a single deliberator for the ruling introduced the risk of bias and limited perspective. AgentsBench (Jiang and Yang, 2024) focuses on collegial panel collaborative deliberation. It provided a certain degree of multi-perspective deliberation. However, the participating roles were randomly selected from predefined settings, similarly lacking dynamic adaptability to different cases and the simulation process was relatively simple.

3 Methodology

This section details the overall architecture of the JudiAgents Framework, its adaptive agent configuration framework, and the three core process modules: court debate simulation, PPDM, and judgment generation.

3.1 Adaptive Agent Configuration Framework

To ensure realistic, diverse agent behaviors deeply linked to case facts, we developed an automated, multi-stage configuration framework. It dynamically creates unique configuration for each agent

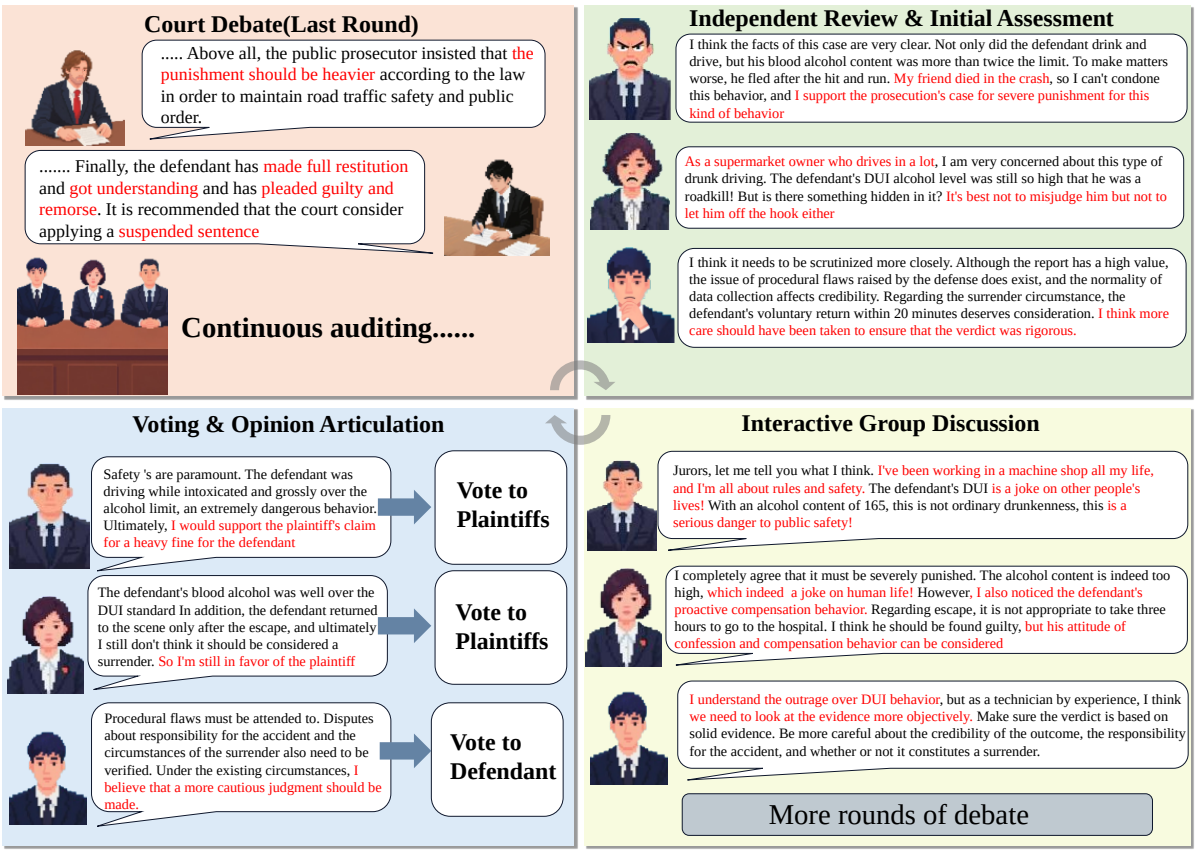

Figure 3: Workflow Diagram of the PPDM Module.

per case. **Stage One: Automated Case Information Structuring.** Automated Case Information Structuring. LLMs extract key case elements from judgment documents, providing structured data for personalized configuration. **Stage Two: Context-driven Configuration Instantiation.** Context-driven Configuration Instantiation in Figure 2. Core to agent profiling and heterogeneity, this stage integrates legal knowledge, social behavior principles, and heterogeneity. Guided by this and Stage One's structured data, an LLM generates unique, case roles for each agent. For panel member heterogeneity, it directs the LLM to combine diverse preset dimensions (e.g., background, cognitive style, values), simulating real-world panel diversity and providing PPDM varied perspectives.

3.2 Court Debate Simulation

Prosecution and defense lawyer agents, presided over by a judge agent, engage in structured multi-round debates. Lawyers argue, present evidence, cross-examine, or rebut based on their operational protocols generated by the adaptive configuration framework (containing case-specific knowledge) and the LLM's internal knowledge. The content of each round of debate is recorded. Debate records form a key PPDM input.The complete simulation process for each case involves approximately 25-40 interaction records.

3.3 Profiled Panel Deliberation Module (PPDM)

PPDM (Figure 3), the framework's core, deeply simulates panel deliberation. It uses panel agents with unique, dynamically generated profiles (via adaptive configuration) to ensure diverse inputs. The protocol involves: **1.Individual Review and Preliminary Judgment:**Each member independently analyzes court records and, using their profile, forms an initial opinion. **2.Interactive Collective Discussion:** Members engage in multi-round dialogues, exchanging views and seeking clarification to simulate idea collision and fusion. **3.Voting and Rationale Articulation:** Post-discussion, members independently vote and must provide detailed textual rationales for their vote. PPDM outputs voting statistics and complete opinion texts from each member, with these detailed rationales

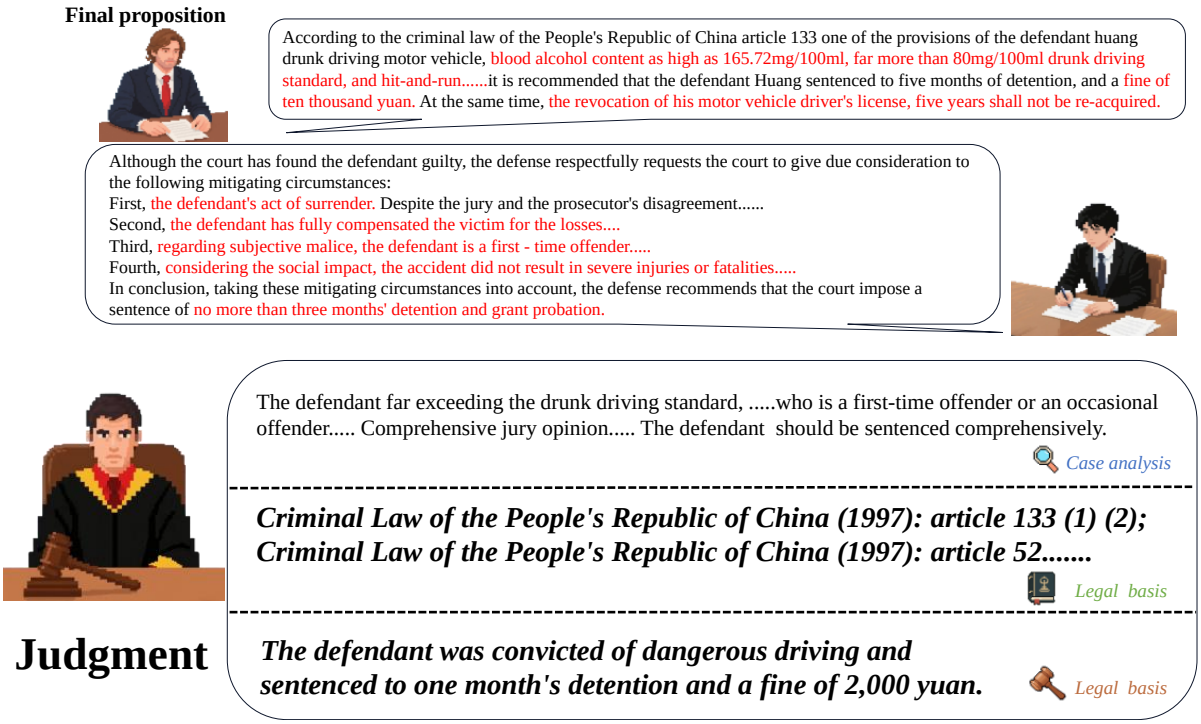

Figure 4: Judgement Generation.

serving as crucial references for the judge's final judgment.

3.4 judgment Generation

The judge agent (Figure 4), after receiving and synthesizing the original case information, court debate records, and the complete output from PPDM (including voting statistics and detailed opinion rationales from each panel member), makes the final decision according to its own operational protocol. Finally, the judge agent generates a structured judgment document containing the verdict, detailed judgment reasons, and cited legal articles.

4 Experimental Setup

This section details the overall architecture of the JudiAgents Framework, its adaptive agent configuration framework, and the three core process modules: court debate simulation, PPDM, and judgment generation.

4.1 Dataset Construction

Our dataset comprises real Chinese legal cases from China Judgments Online (2010-2021), covering diverse civil/criminal and first/second instance proceedings. Cases were randomly selected for representation, then deeply anonymized and structurally processed. Original judgments and legal articles serve as Ground Truth. Appendix A provides detailed statistics.

4.2 Comprehensive Evaluation Metrics

Core Quantitative Evaluation Metrics: For judgment outcome texts and legal basis texts, standard Precision, Recall, F1-score, and Semantic Similarity are used for evaluation. These metrics primarily measure the proximity of predicted text to ground truth text based on semantic or character-level matching.

Auxiliary Quality Evaluation Metrics: Considering the complexity of text generation, we introduced auxiliary evaluation metrics based on GPT-4o to score the **Rationale Logic and Consistency (RLC)**, **Rationale Case-elements Engagement (RCE)**, and **Judgment Support and Coherence (JSC)** on a scale of 1-100. Detailed scoring guidelines are provided in Appendix B.

4.3 Baseline Models

To position the performance of the JudiAgents Framework, we selected a series of general LLMs. These models directly predict judgment outcomes and legal bases from case information, representing the current mainstream level. They include GPT-4o (Achiam et al., 2023), Gemini-2.5, DeepSeek-

Model	Verdict			Legal Basis			RLC	RCE	JSC
	P	R	F	P	R	F			
GPT-4o	0.5969	0.5388	0.5663	0.6182	0.5832	0.6002	71.36	57.59	52.66
Gemini-2.5	0.6124	0.4341	0.5082	0.6667	0.6333	0.6496	70.73	**66.67**	50.00
DeepSeek-v3	0.6735	0.5857	0.6265	0.5109	0.6870	0.5861	76.98	59.91	54.93
Qwen-QwQ	0.6352	0.6563	0.6456	0.6419	**0.7005**	0.6699	78.52	62.94	58.10
Qwen2.5	0.6792	0.6095	0.6425	0.6684	0.6611	0.6647	73.74	59.22	50.92
GLM-4	0.6530	0.5350	0.5883	0.7728	0.6629	0.7136	67.84	54.64	46.87
LLaMA-3.3	0.5828	0.4445	0.5044	0.5437	0.5548	0.5492	62.34	49.22	46.66
Farui-plus	0.3547	0.2901	0.3193	0.5370	0.6198	0.5754	69.18	54.20	52.05
InternLM2.5	0.6138	0.5631	0.5873	0.4752	0.6130	0.5354	72.21	55.70	51.63
JudiAgents	**0.7387**	**0.7517**	**0.7452**	**0.8471**	0.6740	**0.7507**	**81.31**	57.32	**59.19**

Table 1: Overall performance of our framework and baseline in experiments

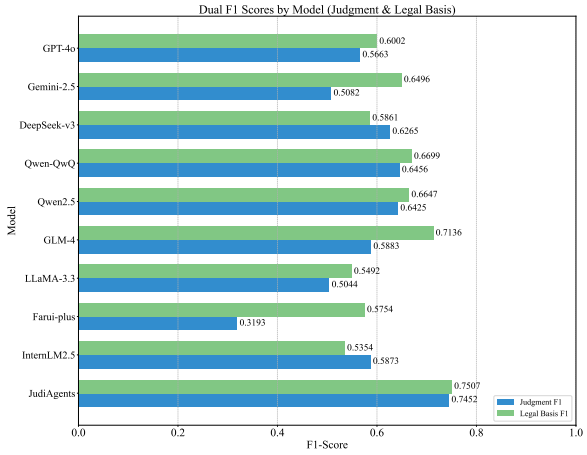

Figure 5: Comparison of F1 Scores for Verdict and Legal Basis Prediction across Different Models.

v3-671B (Liu et al., 2024), Qwen-2.5-72B, Qwen-QWQ-32B (Bai et al., 2023), GLM-4 -32B(GLM et al., 2024), LLaMA-3.3-70B (Grattafiori et al., 2024), Farui-plus, InternLM2.5-20B (Cai et al., 2024). The comparison demonstrates the gains from JudiAgents.Which can be seen in Table 1.

Farui-plus: One of the current state-of-the-art Chinese legal fine-tuned models from the Qwen team .

4.4 Key Implementation Aspects

All agents in the JudiAgents framework are driven by the DeepSeek-V3 model. Throughout the experiments, we did not perform any fine-tuning on this model. Agent behavior is entirely guided by the specific operational protocols generated by the adaptive agent configuration framework for each case and role.

4.5 Ablation Study Design

To validate the effectiveness of PPDM and the adaptive agent configuration framework, we designed systematic ablation studies:

JudiAgents w/o PPDM: Removes the PPDM, The judge generates the judgment directly based on court debate records and original case information

JudiAgents w/o Profiling: Removes the adaptive agent configuration framework. All agents use a set of generic, non-contextualized operational protocols (see generic configuration examples in Appendix C), instead of personalized protocols dynamically generated for each case. heterogeneity and case relevance.

JudiAgents w/o PPDM & w/o Profiling: Removes both PPDM and adaptive agent configuration. The judge generates the judgment based on court debate (conducted by lawyers with generic configurations), with no deliberation process, and all agents use generic configurations.

5 Results and Analysis

This section will detail the experimental results of the JudiAgents Framework, including performance comparisons with mainstream baseline models, ablation study analysis of core components, quality assessment of judgment rationales, and performance across different case types. We also conduct qualitative case studies to deeply analyze the specific impact of the framework's.

5.1 Overall Performance Comparison

Table 1 shows the comparison of JudiAgents with baseline models. JudiAgents significantly outperforms all baselines in both judgment outcome (F1: 0.7452) and legal basis (F1: 0.7507). Compared to the base model DeepSeek-V3 (judgment F1: 0.6265, legal F1: 0.5861), JudiAgents achieved improvements of 18.94% and 28.08% respectively, demonstrating that integrating multi-agent interaction, configuration, and deliberation mechanisms can effectively overcome the limitations of a single perspective.

5.2 In-depth Analysis of Judgment Rationale Quality

Table 1 compares JudiAgents with key baselines on RLC, RCE, and JSC. JudiAgents excels in RLC (81.31), significantly surpassing the base model DeepSeek-V3 (76.98) , indicating its rationales are logically sound and internally consistent. Its RCE (57.32) is nearly to Qwen-QwQ (62.94), JSC (59.19) is comparable to Qwen-QwQ (58.10) and better than others, showing sufficient argumentation for conclusions. Combined with its leading core metrics, this proves JudiAgents can generate high-quality judgment closely tied to case facts through deep simulation.

5.3 Ablation Study Analysis

w/o PPDM: Removing PPDM led to a significant drop in legal basis F1 and rationale logic (RLC), despite a slight variation in judgment outcome F1. This indicates that while PPDM's diverse deliberation might subtly alter judgment wording, it is crucial for the accuracy of legal grounds and the depth of reasoning, making it indispensable for high-quality judgments.

w/o Adaptive Configuration: Without personalized agent configurations, all metrics declined, particularly RLC. Lacking distinct profiles, panel members' opinions homogenized, preventing PPDM from offering valuable diverse perspectives and potentially amplifying biases. Adaptive configuration is core to PPDM's effective operation and avoiding "echo chambers."

Synergy of PPDM and Adaptive Configuration: Table 2 and Table 3 show that when both components were removed, auxiliary metrics like RLC, RCE, and JSC reached their lowest.This highlights their critical synergy. And the best performance is achieved when PPDM is synergized with adaptive configuration, proving the effectiveness of JudiAgents.

Model Configuration	Verdict F1	Legal Basis F1	Avg F1 (V+L)
JudiAgents	0.7452	**0.7507**	**0.74795**
w/o PPDM	0.7559	0.7024	0.72915
w/o Profiling	0.7256	0.7197	0.72265
w/o Both two	0.7446	0.7187	0.73165

Table 2: Comparison of the main properties of ablation experiments

Model Configuration	RLC	RCE	JSC
JudiAgents	**81.31**	57.32	59.19
w/o PPDM	72.43	57.21	62.91
w/o Profiling	71.26	57.65	52.24
w/o Both two	70.72	46.45	51.21

Table 3: Comparison of auxiliary indicators for ablation experiments

5.4 Qualitative Case Studies

To demonstrate JudiAgents' enhancements to simulation realism and depth, we analyze a criminal case (Huang's Dangerous Driving). Appendices A.1 and C.2 provide case details and sample agent configurations. Figure 6 contrasts the full JudiAgents simulation with versions lacking core components. In the full JudiAgents simulation (Figure 6, "Full"), PPDM and adaptive agent configuration were synergistically crucial. Adaptive configuration produced distinct panel members (e.g., the stern Uncle Li, evidence-focused Mr. Zhao, and tender Ms. Wang). Their unique backgrounds and biases led to varied interpretations and articulated reasons. This heterogeneous deliberation enriched the judge's information and, critically, prompted more prudent, multi-dimensional consideration due to the exchange of diverse perspectives. This aligns with ablation study findings where full JudiAgents excelled in rationale logic (RLC) and legal basis accuracy. Conversely, without adaptive configuration but retaining PPDM (Figure 6, "No Profiling"), panel members' statements became homogenized,

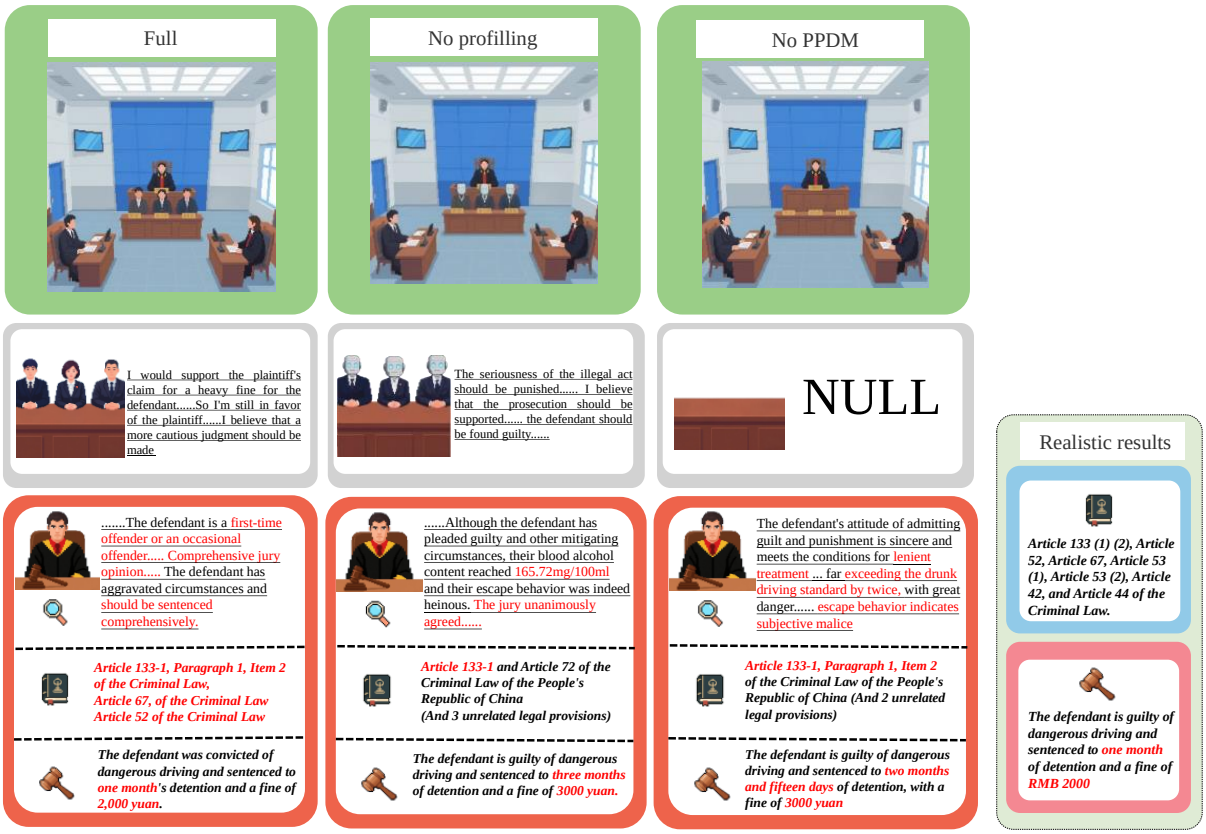

Figure 6: How JudiAgents Influences Case Outcomes.

failing to provide diverse insights. Such "stereotypical" deliberation risked amplifying biases and misleading the judge. When PPDM was removed (Figure 6, "No PPDM"), the judge decided based solely on the court debate. While avoiding "unprofiled deliberation" bias, this lost the multi-perspective input of collective deliberation, leading to potential deficiencies in judgment comprehensiveness and legal rigor, as ablation data suggested.

5.5 Performance by Case Type

Figure 7 shows JudiAgents' performance across different case types compared with different baselines. Judgment outcome metrics for first-instance criminal cases are generally higher than for civil cases, as elements of crime and applicable laws are more defined. The F1 score for legal basis in second-instance cases is prominent, reflecting stricter requirements for legal application in appeals.

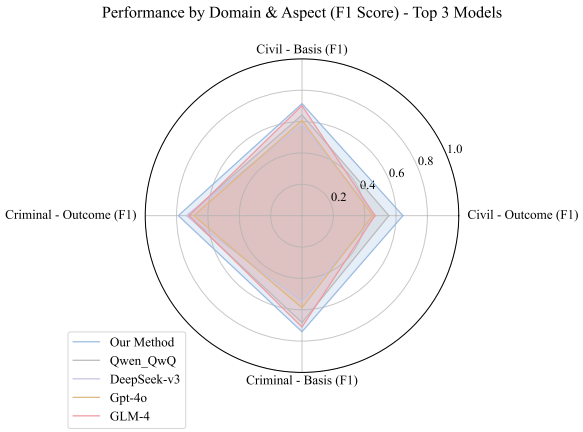

Figure 7: Performance Comparison (F1 Score) of JudiAgents and Baseline Models across Different Case Types.

6 Conclusion

This paper introduced JudiAgents, a multi-agent framework for deeply simulating judicial deliberation. Its automated context-aware agent configuration and the Profiled Panel Deliberation Module notably improve simulation realism, interaction, and decision quality. Experiments show JudiAgents surpasses baselines in judgment and legal article prediction.

Limitations

While promising, this work has limitations. Firstly, our dataset of 481 Chinese legal cases, though detailed, but our study focuses on Chinese only. It could be expanded in scope and jurisdiction for broader generalizability. Secondly, the framework's performance is inherently tied to the capabilities and potential biases of the underlying LLM (DeepSeek-V3), including issues like hallucinations and its "black-box" nature, which can affect simulation fidelity. Thirdly, the agent interaction mechanisms within the Profiled Panel Deliberation Module (PPDM) could be further refined for more nuanced deliberative dynamics. Finally, enhancing the framework by integrating dynamic external legal knowledge bases, beyond the LLM's internal knowledge and injected case specifics, remains an area for future improvement.

Ethical Considerations

We are committed to responsible research in line with ethical guidelines. Key considerations for the JudiAgents Framework include:

Data and Privacy: Case data from China Judgments Online was deeply anonymized to protect privacy, as detailed in appendix A. We acknowledge the sensitivity of using real case data and have sought to minimize associated risks.

Bias and Fairness: The framework relies on LLMs, which may inherit biases. While our diverse agent configuration aims to mitigate this, ongoing vigilance and future auditing are crucial to ensure fairness and prevent the perpetuation of societal biases in simulation outcomes.

Potential for Misuse: JudiAgents is intended as a research tool for understanding judicial processes, not for replacing human judgment or for applications that could be socially harmful. We advocate for its responsible development and circumspect use.

Transparency: The PPDM enhances transparency by simulating deliberation. However, improving the interpretability of LLM-driven components remains an ongoing goal to build trust and understanding.

We aim for our work to contribute positively to Legal AI by fostering deeper insights into judicial decision-making.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Razan Baltaji, Babak Hemmatian, and Lav Varshney. 2024. Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent llm collaboration. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 17–31.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert:" preparing the muppets for court'". In *EMNLP (Findings)*, pages 2898–2904.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Kai-Wei Chang, Anaelia Ovalle, Jieyu Zhao, Yang Trista Cao, Ninareh Mehrabi, Aram Galstyan, Jwala Dhamala, Anoop Kumar, and Rahul Gupta. 2024. Proceedings of the 4th workshop on trustworthy natural language processing (trustnlp 2024). In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*.

Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. 2024. A survey on multi-generative agent system: Recent advances and new frontiers. *arXiv preprint arXiv:2412.17481*.

Weitao Chen, Hongbin Xu, Zhipeng Zhou, Yang Liu, Baigui Sun, Wenxiong Kang, and Xuansong Xie. 2023. Costformer: Cost transformer for cost aggregation in multi-view stereo. *arXiv preprint arXiv:2305.10320*.

Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024. Unified active retrieval for retrieval augmented generation. *arXiv preprint arXiv:2406.12534*.

Carmengelys Cordova, Joaquin Taverner, Elena Del Val, and Estefania Argente. 2024. A systematic review of norm emergence in multi-agent systems. *arXiv preprint arXiv:2412.10609*.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2023. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv preprint arXiv:2306.16092*.

Nigel Gilbert and Klaus Troitzsch. 2005. *Simulation for the social scientist*. McGraw-Hill Education (UK).

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Agentscourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. *arXiv preprint arXiv:2403.02959*.

Qiongyan Huang, Yuhan Xia, Yunfei Long, Hui Fang, Ruiwei Liang, Yin Guan, and Ge Xu. 2025. Prompt4ljp: prompt learning for legal judgment prediction. *The Journal of Supercomputing*, 81(2):420.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2023. Bayesian optimization based score fusion of linguistic approaches for improving legal document summarization. *Knowledge-Based Systems*, 264:110336.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2024. A sentence is known by the company it keeps: improving legal document summarization using deep clustering. *Artificial Intelligence and Law*, 32(1):165–200.

Cong Jiang and Xiaolei Yang. 2024. Agents on the bench: Large language model based multi agent framework for trustworthy digital justice. *arXiv preprint arXiv:2412.18697*.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51:371–402.

Ang Li, Yiquan Wu, Ming Cai, Adam Jatowt, Xiang Zhou, Weiming Lu, Changlong Sun, Fei Wu, and Kun Kuang. 2025. Legal judgment prediction based on knowledge-enhanced multi-task and multi-label text classification. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6957–6970.

Daniele Licari and Giovanni Comandè. 2022. Italian-legal-bert: A pre-trained transformer language model for italian law. *EKAW (Companion)*, 3256.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang, and Zhaochun Ren. 2023. Multi-defendant legal judgment prediction via hierarchical reasoning. *arXiv preprint arXiv:2312.05762*.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Tianyu Ren and Xiao-Jun Zeng. 2024. Enhancing cooperation through selective interaction and long-term experiences in multi-agent reinforcement learning. *arXiv preprint arXiv:2405.02654*.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2024. Explainable ai in language learning: Linking empirical evidence and theoretical concepts in proficiency and readability modeling of portuguese. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 199–209.

Harry Surden. 2018. Artificial intelligence and law: An overview. *Ga. St. UL Rev.*, 35:1305.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

John Villasenor. 2023. Generative artificial intelligence and the practice of law: impact, opportunities, and risks. *Minn. JL Sci. & Tech.*, 25:25.

Yiquan Wu, Yifei Liu, Weiming Lu, Yating Zhang, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2022. Towards interactivity and interpretability: A rationale-based legal judgment prediction framework. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 4787–4799.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and 1 others. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. *arXiv preprint arXiv:2004.02557*.

Shengbin Yue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuanjing Huang, and Zhongyu Wei. 2025. Multi-agent simulator drives language models for legal intensive interaction. *arXiv preprint arXiv:2502.06882*.

Yuan Zhang, Wanhong Huang, Yi Feng, Chuanyi Li, Zhiwei Fei, Jidong Ge, Bin Luo, and Vincent Ng. 2024. Ljpcheck: Functional tests for legal judgment prediction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5878–5894.

Zhuocheng Zhang, Yang Feng, and Min Zhang. 2025. Levelrag: Enhancing retrieval-augmented generation with multi-hop logic planning over rewriting augmented searchers. *arXiv preprint arXiv:2502.18139*.

Qihui Zhao. 2025. Legal judgment prediction via legal knowledge extraction and fusion. *Journal of King Saud University Computer and Information Sciences*, 37(3):1–16.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.

A Dataset Details

The dataset used in this study comprises 481 real Chinese legal cases, all sourced from the publicly available China Judgments Online. In selecting these cases, we considered the diversity of case types (covering civil and criminal domains, first and second instance proceedings) and the completeness of the judgment documents to ensure sufficient contextual information for simulation and evaluation. All case data underwent rigorous deep anonymization before being used for research(For example, "Huang" in the example case is a pseudonym), removing all sensitive information that could point to real individuals or entities, to comply with data privacy and ethical requirements.Data acquisition and usage comply with the terms of service of China Judgments Online, which permits public access for non-commercial research purposes.

Data Type	Criminal	Civil	Total
First Instance	200	200	400
Second Instance	31	50	81
Total	231	250	481

Table 4: Detailed Composition Statistics of the Dataset

During the structural processing phase, we extracted the following key information from the original judgment documents as input for simulation and ground truth for evaluation:

Basic Case Information: Including case name, court of trial, case type, trial procedure, judgment date, etc. Party Information: Anonymized basic role information of the plaintiff/prosecutor, defendant/accused, and related agents or defense lawyers.

Core Case Facts: Structured description of case facts, time of occurrence, location, key actions, etc. Claims/Charges: Plaintiff's claims or the prosecution's charges.

Main Evidence (Summary): Key types of evidence mentioned in the original document.

Ground Truth Verdict: The main text of the judgment from the original document.

Ground Truth Legal Basis: Main legal articles cited in the original judgment document.

The detailed composition of the dataset is shown in Table 4, ensuring coverage across different case types and trial procedures.

A.1 Basic Case Facts Summary

Case of Huang for Dangerous Driving:

Defendant: Huang (Pseudonym)

Core Facts: defendant Huang drove a small ordinary passenger car under the influence of alcohol and collided with the central road barrier, causing damage to the vehicle and barrier. Huang fled the scene after the accident. He was apprehended by police officers in the early hours of the next day. Upon testing, Huang's blood alcohol content was 165.72mg/100ml.

Original Judgment Outcome (Ground Truth): Defendant Huang committed the crime of

dangerous driving and was sentenced to one month of criminal detention and fined RMB 2,000.

Original Legal Basis (Ground Truth): Article 133-1, Paragraph 1, Item 2; Article 42; Article 44; Article 52; Article 53, Paragraph 1; Article 53, Paragraph 2; Article 67 of the "Criminal Law of the People's Republic of China."

B Experimental Setup and Evaluation Details

B.1 Implementation Details

Core LLM: All agents in the JudiAgents framework are driven by the DeepSeek-V3. This model was not fine-tuned for this task to the experiments.

PPDM Settings: The number of panel members in the PPDM module and the rounds of court debate can be configured according to experimental needs. In the main experiments of this study, PPDM included three panel members, and their deliberation process involved one round of independent analysis and two rounds of cross-discussion.

Court Debate: In the main experiments of this study, the court debate proceeded for three rounds.

B.2 Scoring Guidelines for Auxiliary Quality Metrics

The RLC, RCE, and JSC metrics are scored by GPT-4o on a scale of 1-100 based on the following criteria:

RLC (Rationale Logic and Consistency):

0-40 points: Rationale is chaotic, illogical, self-contradictory, or completely lacks meaningful argumentation.

40-70 points: Rationale is generally understandable, but the logical chain has clear flaws, and some arguments are insufficient or inconsistent.

70-90 points: Rationale is logically clear, argumentation is relatively sufficient, internally mostly consistent, and can support the judgment conclusion well.

90-100 points: Rationale is logically rigorous, argumentation is sufficient and powerful, highly internally consistent, analysis and balancing of complex situations are reasonable, demonstrating excellent legal reasoning ability.

RCE (Rationale Case-elements Engagement):

0-40 points: Rationale barely mentions or responds to key facts, evidence, or points of contention in the case.

40-70 points: Rationale responds to some case elements but improperly handles, omits, or misunderstands some important facts, evidence, or points of contention.

70-90 points: Rationale adequately responds to the main facts, evidence, and points of contention in the case, and conducts reasonable analysis and adoption.

90-100 points: Rationale comprehensively and deeply responds to all key elements of the case, including complex or subtle points, and clearly articulates how these elements affect the judgment.

JSC (Judgment Support and Coherence):

0-40 points: Judgment conclusion lacks effective argumentative support, or there is a serious disconnect between argumentation and conclusion.

40-70 points: Argumentation for the judgment conclusion has some deficiencies; supporting reasons for some key links are not sufficient or persuasive.

70-90 points: Judgment conclusion is supported by relatively sufficient and reasonable argumentation; the connection between reasons and conclusion is clear.

90-100 points: Judgment conclusion is supported by comprehensive, powerful, and logically rigorous argumentation, convincingly deriving the conclusion from the reasons.

C Agent Configuration Prompt Examples

The adaptive agent configuration framework generates unique system prompts for each agent in each case. Below are excerpts of configuration examples generated for Case (Huang's Dangerous Driving Case):

C.1 Generic Agent Configuration Example (used in w/o Profiling mode)

Generic Lawyer Prompt:

You are a lawyer. Your duty is to represent your client (plaintiff or defendant) in court by making statements, presenting evidence, cross-examining, and debating to protect their legal rights. You need to conduct professional legal argumentation based on facts and law. Get a favorable verdict for your side.

Generic Juror/Panel Member Prompt:

You are a panel member. You need to listen carefully to both sides' testimonies and debates, actively ask questions to clarify doubts, and scrutinize all evidence. You should make a judgment based on facts and law, not just listen to statements. Please note, you need to make judgments based on the

debates of the prosecution and defense, and the submitted evidence, not unilateral statements.Actively participate in questioning to ensure you fully understand the case.

C.2 Adaptive Configuration Example for Huang's Dangerous Driving Case

Prosecutor (Wang Li, representing the procuratorate, diligent and responsible):

You are a prosecutor named Wang Li, 35 years old, working at the procuratorate for 5 years. You are diligent, responsible, and possess strong logical thinking skills. Your goal is to clearly present evidence to prove defendant Huang's guilt, refute the defense lawyer's challenges, and request the court to issue a judgment. During the debate, actively respond to the defense's challenges and uphold the prosecution's charges to maintain public safety and social order.

Defense Lawyer (Zhang Wei, logical and adept at finding loopholes):

You are a defense lawyer named Zhang Wei, 38 years old, male, working at a law firm for 8 years. You are known for your clear logic and ability to find loopholes in the opposing party's arguments. You may seek probation or a lighter punishment based on the defendant's specific circumstances. You need to strongly respond to any charges raised by the prosecution. Your goal is to question the prosecution's evidence through debate, find contradictions and loopholes in the prosecution's statements and evidence, and strive for the lightest possible sentence or even acquittal for your client.

Jury Member 1 (Uncle Li, a retired worker who despises drunk driving):

Your name is Uncle Li, 62 years old, male, a retired factory worker, Before retirement, you were a worker in a mechanical factory with a straightforward personality and rigorous professional ethics. upright and meticulous. You have lived in here for many years and despise drunk driving because a friend of yours died in a drunk driving accident. You believe drunk driving is irresponsible towards others' lives and must be severely punished. You need to listen carefully to both sides' testimonies and debates, actively ask questions to clarify doubts, and scrutinize all evidence. You should make a judgment based on facts and law, not just listen to statements. Please note, you need to make judgments based on the debates of the prosecution and defense, and the submitted evidence, not unilateral statements.Actively participate in questioning to ensure you fully understand the case.

Jury Member 2 (Ms. Wang, a gentle supermarket owner):

Your name is Ms. Wang, 45 years old, female, owner of a small supermarket. You often drive in to stock up,and have a strong sense of justice. You believe everyone is equal before the law, and no crime should be condoned for any reason. You often drive and detest drunk driving, considering it irresponsible towards one's own life and others'. However, you believe everyone has a chance to repent and reform, and sentencing should consider the defendant's attitude of admitting guilt and showing remorse. You need to listen carefully to both sides' testimonies and debates, actively ask questions to clarify doubts, and scrutinize all evidence. You should make a judgment based on facts and law, not just listen to statements. Please note, you need to make judgments based on the debates of the prosecution and defense, and the submitted evidence, not unilateral statements.Actively participate in questioning to ensure you fully understand the case.

Jury Member 3 (Mr. Zhao, a cautious programmer with a technical background):

Your name is Mr. Zhao, 30 years old, male, a programmer. You are usually a homebody and don't know much about law, but you believe in science and evidence. You think judgments should be based on sufficient evidence and rigorous logical reasoning, not emotions or personal biases. You need to listen carefully to both sides' testimonies and debates, actively ask questions to clarify doubts, and scrutinize all evidence. You should make a judgment based on facts and law, not just listen to statements. Please note, you need to make judgments based on the debates of the prosecution and defense, and the submitted evidence, not unilateral statements.Actively participate in questioning to ensure you fully understand the case.