AgentsCourt: Building Judicial Decision-Making Agents with Court Debate Simulation and Legal Knowledge Augmentation

Anonymous ACL submission

Abstract

With the development of deep learning, natural language processing technology has effectively improved the efficiency of various aspects of the traditional judicial industry. However, most current efforts focus on tasks within individual judicial stages, making it difficult to handle complex tasks that span multiple stages. As the autonomous agents powered by large language models are becoming increasingly smart and able to make complex decisions in real-world settings, offering new insights for judicial intelligence. In this paper, (1) we propose a novel multi-agent framework, *AgentsCourt*, for judicial decision-making. Our framework follows the classic court trial process, consisting of court debate simulation, legal resources retrieval and decision-making refinement to simulate the decision-making of judge. (2) we introduce *SimuCourt*, a judicial benchmark that encompasses 420 Chinese judgment documents, spanning the three most common types of judicial cases. Furthermore, to support this task, we construct a large-scale legal knowledge base, Legal-KB, with multi-resource legal knowledge. (3) Extensive experiments show that our framework outperforms the existing advanced methods in various aspects, especially in generating legal articles, where our model achieves significant improvements of 8.6% and 9.1% F1 score in the first and second instance settings, respectively.

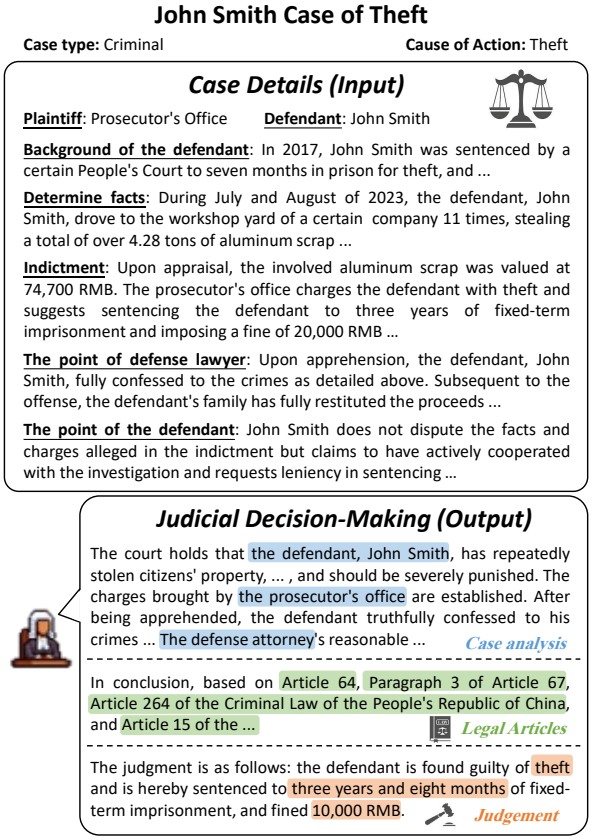

Figure 1: We formulate the Judicial Decision-Making task using the real-world judgement documents: given the case details above, judge agent must 1) conduct a logically clear case analysis; 2) provide precise legal articles; 3) issue a definitive judgement.

1 Introduction

Recent advances in deep learning have significantly impacted the legal domain, with notable achievements in legal question answering (Zhong et al., 2020b; Khazaeli et al., 2021; Cui et al., 2023), legal case retrieval (Sugathadasa et al., 2019; Shao et al., 2020; Li et al., 2023b; Shao et al., 2023) and legal judgment prediction (Xiao et al., 2018; Chalkidis et al., 2019; Wu et al., 2022, 2023b). These developments have effectively alleviated the long-standing issue in the judicial industry of "too many cases, too few legal professionals". However, case trial is a coherent process involving multiple stages such as court debates, case analysis, and legal judgment prediction. The complexity of this process demands close collaboration and interaction between stages. Although current research has made progress in individual stages, it often overlooks the inherent connections between these stages of the trial process. This results in the need to rely on the deep involvement of legal experts when dealing

Framework	AgentsCourt (This work)	LaWGPT (Song et al., 2023)	PLJP (Wu et al., 2023b)	HRN (Lyu et al., 2023)	RLJP (Wu et al., 2022)
Case Analysis	✓	✓	✗	✗	✓
Precedent Retrieval	✓	✗	✓	✗	✗
Web Research	✓	✗	✗	✗	✗
Court Simulation	✓	✗	✗	✗	✗
Judgement Prediction	✓	✓	✓	✓	✓
Legal Articles Generation	**Multiple**	Single	Single	Single	Single
Case Type	**Various**	Various	Crime	Crime	Crime

Table 1: A comparison of our AgentsCourt to notable legal domain frameworks.

with complex judicial decisions. Meanwhile, autonomous agents based on large language models (LLMs) have shown considerable progress in various traditional natural language processing (NLP) tasks (Brown et al., 2020; Wei et al., 2022; Wang et al., 2023; Qian et al., 2023; Wu et al., 2023a) and making decisions in real-world environments (Yao et al., 2023; Richards, 2023; Chen et al., 2023), which offers new insights for judicial intelligence.

However, simulating judicial decision-making is a non-trivial task because agents must navigate complex situations involving multiple stakeholders, understand the subtle nuances of legal provisions, and consider ethical and social justice factors. This presents three unique challenges to the agent system: (1) *Intricate ethical relationships*. In judicial decision, ethical and moral considerations, which are often subtle and multi-faceted, must be taken into account.(2) *Expert knowledge of judicial domain*. Judicial adjudication requires an in-depth understanding and accurate application of specialized knowledge such as laws, regulations and precedents. (3) *Complex and hybrid reasoning*. The agents must be capable of handling a complex amalgamation of logical, factual, and legal reasoning, often interwoven in cases.

To tackle the aforementioned challenges, we propose a novel multi-agent framework, **AgentsCourt**, for the Judicial Decision-Making task. As illustrated in Figure 1, given the case details, the task requires the agent to conduct a logically clear *case analysis*, provide precise *legal articles* and issue a definitive *judgement*. AgentsCourt follows the classic court trial process: opening remarks, court debate, precedent retrieval, and judgement, as depicted in Figure 2. Specifically, we first develop a *Court Debate Simulation Module* with three agents, which serves as a platform for all parties involved to present their points to clarify the intricate ethical relationships in the case. One agent serves as the judge to open a court session and announce the

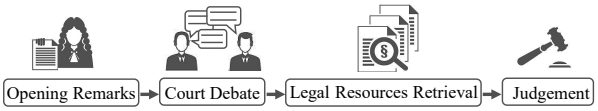

Figure 2: Simplified court trial process.

basic facts of the case. The other two agents are designed as the plaintiff and the defendant respectively, and articulate their points of view during the court debate phase. Then, we devise the *Legal Resources Retrieval Module* to address the inadequacy of expert knowledge. This module employs a judge assistant agent to integrate the most relevant precedents, articles and other information retrieved from the knowledge base we constructed and the internet. Next, we propose the *Decision-Making Refinement Module* to facilitate complex and hybrid reasoning. This module firstly makes a preliminary judgement according to the inherent judicial expertise of the agent elicited by the established facts of current case and the transcripts of court debate, then subsequently refines the judgement using legal information retrieved.

The comparison between our framework and prior works is listed in Table 1. It is worth noting that our framework is not tailored to a specific legal system. AgentsCourt can achieve court simulation, precedent retrieval, judgment prediction, and supports the generation of multiple legal articles for practical judicial practice.

We also introduce **SimuCourt**, a judicial benchmark designed to evaluate Agent-as-Judge across a spectrum of different cases. SimuCourt encompasses 420 Chinese judgement documents, spanning the three most common types of judicial cases — criminal, civil, and administrative — in both first-instance and second-instance (appellate) courts, as well as covering three key societal roles: government agencies, the prosecutor's office, and individuals. Specifically, criminal cases involve acts

that are identified as violations of criminal law, such as theft. Civil cases typically involve disputes between individuals, such as contract disputes or torts. Administrative cases concern disputes between individuals and government agencies. All the cases come from the China Judgements Online [1], which is an official platform established by the Supreme People's Court of China, aimed at publicly releasing the judgement documents of courts at all levels in China. Furthermore, we construct a large-scale legal knowledge base, **Legal-KB**, to support this domain task. It encompasses a variety of legal knowledge, including effective laws and regulations, highly cited judicial papers, and precedents from recent years. The use of real data allows the agents developed on it can be transferred into real applications without any gaps.

We summarize our contributions as follows:

- We propose a novel multi-agent framework AgentsCourt. Given the basic information of a case, our framework can sequentially simulate court debate, retrieve precedents, analyze cases, provide legal articles, and deliver clear judgment. The new judicial paradigm simplifies the process of making judicial decisions, significantly enhancing judicial efficiency.

- We introduce SimuCourt, a judicial benchmark encompasses the three most common types of cases, enabling reliable assessment of the judicial analysis and decision-making power of agents for real judicial practice. Furthermore, we construct a legal knowledge base, Legal-KB, with multi-resource legal knowledge to support this task.

- We perform extensive experiments and ablation studies. The results indicate that our framework outperforms the existing advanced methods in various aspects, especially in generating legal articles, where our system achieves notable improvements of 8.6% and 9.1% F1 score in the first and second instance experimental settings, respectively.

2 Related Work

Legal Artificial Intelligence Legal Artificial Intelligence seeks to improve legal tasks by employing artificial intelligence techniques (Surden, 2019; Zhong et al., 2020a; Katz et al., 2023). With the continuous development of deep learning, the legal field has witnessed the emergence of more intelligent applications across various legal tasks. These tasks span across areas such as legal judgment prediction (LJP) (Xiao et al., 2018; Zhong et al., 2018; Xu et al., 2020; Yue et al., 2021; Wu et al., 2022, 2023b), legal question answering (Zhong et al., 2020b; Cui et al., 2023; Louis et al., 2024; Fei et al., 2023), legal language understanding (Chalkidis et al., 2022; Xiao et al., 2021; Niklaus et al., 2023; Yu et al., 2023), legal case retrieval (Sugathadasa et al., 2019; Shao et al., 2020; Li et al., 2023b; Shao et al., 2023), legal document summarization (Kanapala et al., 2019; Jain et al., 2023, 2024). While these existing efforts have made progress in individual legal tasks, they have overlooked the interconnection between different tasks, resulting in the necessity to heavily rely on the deep involvement of legal experts when dealing with complex judicial decisions. In this work, we focus on completing the entire process of judicial decision-making through multi-agent collaboration.

Multi-agent framework Cooperation among agents like human group dynamics can enhance the efficiency and effectiveness of task accomplishment. Li et al. (2023a) enables two communicative agents to engage in a conversation and cooperate with each other to solve assigned tasks. Park et al. (2023) found social behaviors autonomously emerge within a group of agents. Qian et al. (2023); Hong et al. (2023) present innovative paradigms that leverages LLMs throughout the entire software development process by natural language communication. Du et al. (2023); Zhang et al. (2023); He et al. (2023); Chen et al. (2023); Wu et al. (2023a) further leverage multi-agent cooperation to achieve better performance on multiple tasks.

2.1 Task Formulation

We propose a generative task to evaluate agent as judge. Specifically, as shown in Figure 1, we formulate the Judicial Decision-Making task as given the case details of a case, such as Determine facts, Complaint/Indictment, Statement of the plaintiff and the defendant, the agent system needs to make a complete judicial decision, which includes a clear and reasonable case analysis, rigorous legal articles, and definitive final judgement. SimuCourt encompasses two experimental settings:

First Instance This setting refers to the trial court level, where the judge determines the guilt of the defendant, and assesses whether punitive

[1] https://wenshu.court.gov.cn/

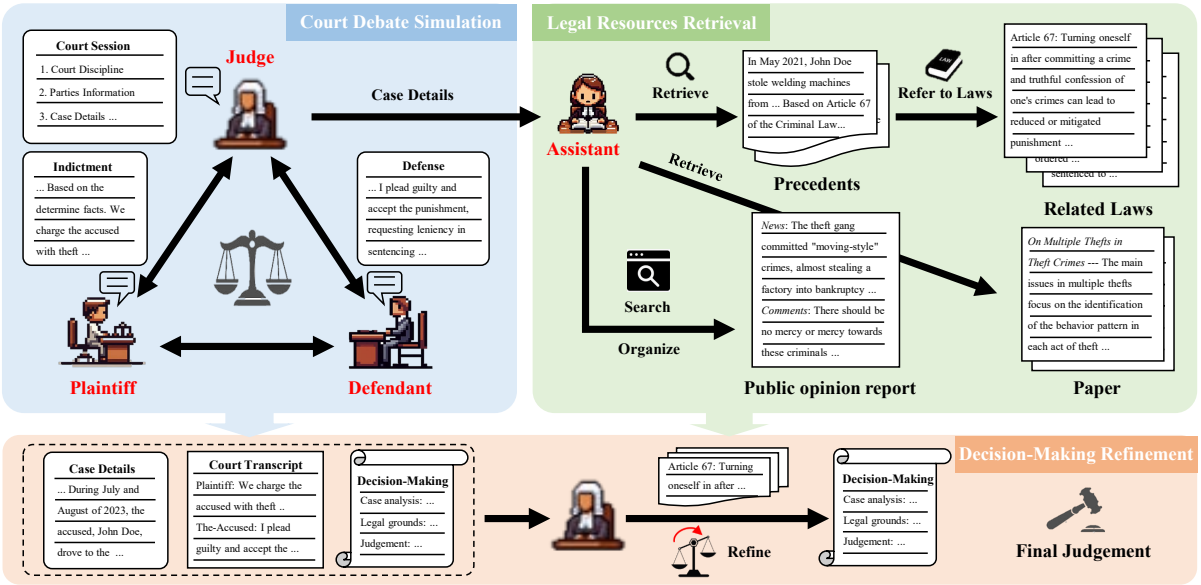

Figure 3: Overview of our multi-agent framework. The Court Debate Simulation Module recreates the court debate process through role-playing, mining different parties' points from limited real records. TheLegal Resources Retrieval Module employ an assistant agent to integrate information retrieved. The Decision-Making Refinement Module exploit the inherent judicial expertise of the judge agent and refines the judgment using information retrieved.

measures are warranted. Within this setting, the primary focus is on evaluating the agent's understanding and analysis of case facts.

Second Instance This setting refers to the appellate court level. During this stage, the judge re-evaluates the case, considering new evidence. The objective at this stage is to ensure the legality and fairness of the initial judgement, identifying legal errors or inappropriate application of regulations from the first instance and demonstrating the capability to effectively handle new evidence.

3 The AgentsCourt Framework

We propose a novel multi-agent framework, as shown in Figure 3. Our framework is based on real-world court trial process and aims to study the collaboration of multiple agents, as well as how they contribute to judicial decision-making.

3.1 Court Debate Simulation

The court debate provides a platform for all parties involved to present their points and arguments comprehensively and fairly, which can significantly influence the judgement of the case.

Court Simulation Due to the majority of judgement documents only recording the key points of the plaintiff's and defendant's statements, obtaining complete court transcripts is challenging. Fortunately, as large language models have shown remarkable ability in role-playing (Li et al., 2023a; Qian et al., 2023; Chen et al., 2023), in this module, we aim to reconstruct the court debate with multiple agents for each case. We set up three agents to play the roles of the judge, plaintiff, and defendant respectively. For each agent, we carefully design an role-playing prompt to build their character personality and use the actual statements from judgment documents as the their starting prompts. It is worth noting that due to the limited record of statements in judgment documents, we combine the plaintiff and their representative, as well as the defendant and their representative, into the plaintiff and defendant, respectively, without setting separate roles for representatives.

Court Debate In this stage, both the plaintiff and the defendant need to present their arguments in line with their interests. The plaintiff should vigorously argue their complaint, articulating their stance and reasoning. Meanwhile, the defendant must defend their actions, aiming to prove their innocence or seek a lighter penalty. During the court session, the judge agent first delivers opening remarks, which include basic information about the plaintiff and the defendant, determination of facts, and so on. Then, the trial moves into the court debate stage and the communication between the agents will be recorded as court transcripts. We present an example of court transcripts in Table 12.

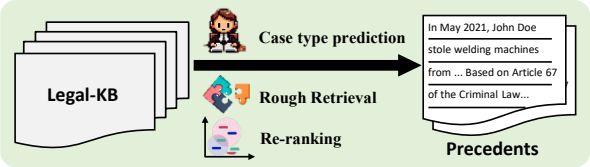

Figure 4: Automatic retrieval of precedents.

3.2 Legal Resources Retrieval

Court debate serves as a platform to thoroughly explore the facts and contentious issues within a case, making the judge better comprehend the complexity of the matter. Furthermore, to make accurate judicial decisions, judges must possess extensive legal knowledge and case information.

Judge Assistant We assign an agent as judge assistant who is responsible for accessing the internet and the knowledge base. In terms of internet information acquisition, the assistant can use web research to seek open information, such as "Does the case have any public opinion?" This aids the judge in understanding the societal impact of the case and potential public perspectives. Ultimately, the agent organizes the retrieved news, comments to the judge, supporting the judge in making rational and well-founded judicial decisions.

Automatic Information Retrieval In terms of knowledge base retrieval, as presented in Figure 4, the assistant first predict the type of case based on the determine facts of the current case. Due to the vast number of documents in the knowledge base, and the fact that cases with the same cause often have more similar keywords, we employs the BM25 model (Lin et al., 2021) for efficient *rough retrieval* to obtain the top 100 documents from the knowledge base. Building on this, we further utilize the BGE-Large model (Xiao et al., 2023) to encode and *re-rank* these retrieved documents and choose the most relevant document to the current case as the optimal precedent. Additionally, to obtain more comprehensive laws and regulations relevant to the current case without introducing additional context, the judge assistant extracts the corresponding legal articles from the top 5 precedents as related legal provisions of current case.

3.3 Judgement Refinement

In this module, we first exploit the inherent judicial expertise of the agent by utilizing determine facts of current case and transcripts of court debate to make a preliminary judgment. Then, the judge agent refines the judgment using information retrieved.

Preliminary Judgement As shown in the bottom of Figure 3, after receiving the determine facts of current case and transcripts of simulated court debate, the judge agent takes the action of analysis, then provides its legal articles and subsequently reaching a preliminary judgement.

Judgement Refinement After obtaining the preliminary judgement which involves analyzing the specific details of the case, the judge agent uses precedent and relevant legal information from the assistant to refine the its judgement and provide the final judgement. This includes but is not limited to analyzing the precedent, referring to legal regulations and considering opinions of public.

4 The SimuCourt Benchmarck

The task, Judicial Decision-Making, requires agents to conduct case analysis, generate legal articles and judgments. However, most existing legal datasets suffer from several limitations when it comes to assessing the Agent-as-Judge paradigm: 1) only contain the factual information of cases; 2) only focus on criminal cases; 3) only evaluate judgments. To this end, we propose SimuCourt, a judicial benchmark for a reliable assessment of the judicial analysis and decision-making power of agents. A comparison between our dataset and previous works is presented in Table 2.

4.1 Data Collection

We collect 420 real-world cases from the China Judgements Online, which span across two fundamental trial stages: first instance and second instance. These cases encompass three types: criminal, civil, and administrative. For first-instance cases, each sample includes the indictment, the point of the defendant, determine facts, etc. For second-instance cases, each sample contains petition for appeal, the point of the appellant and appellee, etc. Detailed list and data examples can be found in the Appendix C. Most of cases were released after April 2023. This minimizes the risk of data leakage[2]. Detailed data statistics of SimuCourt are shown in Table 3. Furthermore, our dataset undergo rigorous scrutiny, ensuring the accuracy and completeness of the legal texts and information. Details of data collection and quality inspection can be found in Appendix D.

[2]The cutoff date of pretraining data for gpt-3.5-turbo-0613 and gpt-4-1106-preview is officially before April 2023.

Resource	SimuCourt	CAIL	SLJA-SYN
Background of Defendant?	✓	✗	✗
Statement of Different Parties?	✓	✗	✗
Multi-article Scenario?	✓	✗	✗
Case Analysis Evaluation?	✓	✗	✓
Judgement Evaluation	✓	✓	✓
Laws Involved?	**443**	1	1
Case Retrival?	**6.5M**	2.6M	✗
Various Case Types?	**Crime, Civial, Admini.**	Crime	Crime
Different Instances Involved?	**First/Second**	First	First

Table 2: A comparison of our SimuCourt to remarkable legal domain datasets. CAIL (Xiao et al., 2018) is a widely used legal judgment prediction dataset, where each case comes with a fact description; SLJA-SYN (Deng et al., 2023) is a comprehensive legal dataset designed to support multiple tasks such as article retrieval, article interpretation generation, criminal element generation and legal judgment prediction.

Feature	Criminal	Civil	Administrative
# of Cases	140	140	140
# of Causes of action	44	51	33
Avg # of Legal articles	6.3	3.3	1.6
Max # of Legal articles	11	10	8
Total # of Legal articles	198	153	92
Avg. Length of Facts	468.7	487.5	673.3
Avg. Length of Analysis	346.3	486.1	722.7
Avg. Length of Cases	2362.6	2473.8	3315.5

Table 3: Statistics of SimuCourt. Length is measured via the number of words

4.2 Legal Knowledge Base Construction

To make accurate judicial decisions, judges must possess extensive legal knowledge. Furthermore, given the diversity and complexity of human society, each case may involve different facts, parties, and locations. To this end, we construct a large scale legal knowledge base consists of laws, regulations, judicial interpretation, journal articles, and precedents. Detailed data statistics of *Legal-KB* are shown in Table 4.

Laws, Regulations and Judicial interpretations We download various legal documents from the National Laws and Regulations Database of China[3], an authoritative resource for legal information that includes national laws, administrative regulations, local regulations, and judicial interpretations. We remove legal documents that are no longer in effect.

Journal Articles Journal articles, typically authored by legal experts, can provide in-depth analysis and unique perspectives on specific legal issues. We collect highly-cited journal articles from 2010 to 2023 from the Chinese Legal Resources Knowledge Database [4].

[3] https://flk.npc.gov.cn

[4] https://lawnew.cnki.net/

Type	Num	Tokens	Avg. Tokens
Laws and Regulations	9K	66M	7390
Journal Articles	29K	15M	521
Precedents	6.5M	27.1B	4111

Table 4: Statistics of our legal knowledge base.

Precedents We collect all judgement documents of criminal, civil and administrative cases from the China Judgements Online for the years 2017 to 2022. However, as illustrated in Figure 9 in the Appendix, the data exhibits a significant long-tail distribution. To balance the type of case, we limit the number of cases for each cause of action to no more than 20k. For those causes of action with more cases, we retain only the top 20k cases with the longest text as representatives of complex cases.

5 Experiments

5.1 Automatic Evaluation

As example data illustrated in Table 11, the legal articles and judgement are concise and structured. Therefore, we propose corresponding metrics for legal articles and judgement evaluation.

Legal Articles Evaluation The correct legal articles is crucial for a fair judgment. Thus, we employ the strict matching method to assess the legal articles generated by the agent system. Specifically, we calculate the number of entries that match and do not match between the legal articles list of the agent system and the reference legal articles list. These counts are then micro-averaged to determine the overall precision, recall and F1 scores. Details can be found in Table 13.

Judgement Evaluation for Civil and Administrative Cases The judgment of each civil or administrative case may encompass multiple results. While each result typically revolves around a single key point, it may involve specific monetary amounts and interest rate information. Consequently, traditional text matching methods based on similarity struggle to accurately capture these key points. Thus, we employ GPT-4 as an evaluator. Specifically, we separately count the number of matching and non-matching key points in the agent system's judgment results compared to the reference judgment results. The micro-averaged counts are used to calculate the overall precision, recall and F1 scores. Details is presented in Table 14.

Judgement Evaluation for Criminal Cases Different from other cases, the sentence of criminal

	Model	Legal Articles			Judgement Results						Case Analysis		
					Civil and Admini.			Criminal					
		P	R	F	P	R	F	Charge	Prison term	Fine	Correctness	Logicality	Concision
First	**GPT-3.5**	0.127	0.109	0.117	0.367	0.498	0.423	0.822	0.253	0.412	0.466	0.510	0.493
	GPT-4	0.139	0.133	0.136	0.398	0.559	0.465	0.875	0.287	0.462	0.503	0.553	**0.543**
	ReAct	0.161	0.109	0.131	0.387	0.532	0.448	0.866	0.262	0.437	0.516	0.567	0.533
	AutoGPT	0.171	0.123	0.143	0.392	0.543	0.455	0.862	0.275	0.450	0.523	0.576	0.520
	LaWGPT	0.183	0.105	0.133	0.414	0.548	0.471	0.875	0.237	0.425	0.506	0.546	0.533
	AgentsCourt	**0.219**	**0.189**	**0.203**	**0.437**	**0.603**	**0.507**	**0.887**	**0.337**	**0.500**	**0.550**	**0.596**	0.526
Second	**GPT-3.5**	0.206	0.169	0.186	0.317	0.429	0.365	0.716	0.166	0.516	0.496	0.540	0.526
	GPT-4	0.200	0.267	0.228	0.356	0.482	0.409	0.800	0.183	0.533	0.530	0.583	0.576
	ReAct	0.209	0.235	0.221	0.364	0.457	0.405	0.800	0.150	0.516	0.526	0.586	0.570
	AutoGPT	0.217	0.248	0.231	0.371	0.478	0.417	0.816	0.166	0.550	0.540	0.590	0.583
	LaWGPT	0.225	0.231	0.227	0.382	0.472	0.422	0.850	0.133	0.483	0.503	0.553	0.566
	AgentsCourt	**0.271**	**0.284**	**0.277**	**0.400**	**0.528**	**0.456**	**0.833**	**0.200**	**0.583**	**0.583**	**0.633**	**0.593**

Table 5: Overall performance of our framework and baselines in the first and second instance experimental settings.

case typically include three core elements: charge, prison term, and fine. The determination of the charge must match the facts of the case. The specific amounts of the prison term and fines are based not only on the facts but also take into account the defendant's performance in court, including their attitude towards the crime and the defense they present for their actions. We calculate the accuracy of the agent system separately for these three items.

5.2 Human Evaluation

The case analysis entails intricate logical reasoning and ethical considerations that are challenging to evaluate through automatic metrics or GPT-4. For each setting, we present a panel of three graduate students majoring in law a random sample of 100 entries from each setting and the following binary True/False criteria guidelines: 1) **Correctness**: Mark true if and only if the analysis is satisfying and considers all parties involved. 2) **Logicality**: Mark false if the analysis contains any illogical or untrue reasoning. 3) **Concision**: Mark true if the analysis covers all necessary information without any extra information.

5.3 Baselines

Vanilla We employ `gpt-3.5-turbo-1106` and `gpt-4-1106-preview` with few-shot as vanilla models. Furthermore, due to limited budget, we only use the `gpt-3.5-turbo-1106` as foundation models of all agent systems.

ReAct (Yao et al., 2023) This system enables the agent to improve its actions based on the outcomes of past activities like searches or tool usage.

AutoGPT (Richards, 2023) This is the most advanced agents framework, incorporating a variety of tools and prompts designed to facilitate the automatic planning and execution of specified tasks.

LaWGPT (Song et al., 2023) This is currently the most popular Chinese legal large language model[5], which has undergone extensive pretraining on Chinese legal corpora and fine-tuning on legal instructions, based on the general Chinese foundation model (Chinese-LLaMA-7B). It possesses strong capabilities in understanding and generating legal content.

5.4 Main Results

As shown in Table 5, our framework outperforms other models in all aspects. For the evaluation on legal articles, our proposed framework achieved performance improvements of 8.6% and 9.1% in the two experimental settings, respectively. In contrast, GPT-4's performance in the first and second instance settings only reach 13.6% and 22.8%, respectively. This not only indicates significant shortcomings in the capabilities of LLMs in sourcing legal provisions, but also reflects the high challenge of our benchmark. In terms of judgment results evaluation, while all models performed well in the conviction of criminal cases, there is still a significant gap in determining prison term and fines compared to standard results. Furthermore, although the analysis of these systems has shown a certain degree of logicality, there is still room for improvement in terms of correctness and concision.

5.5 Discussion and Analysis

Legal Knowledge of LLMs As indicated in Figure 5, all three language models exhibit excellent performance on the simple task of predicting case

[5] https://github.com/pengxiao-song/LaWGPT

Model	Legal Articles	Judgement Results			
		Civil and Admini.	Charge	Prison term	Fine
SimuCourt	0.203	0.507	0.887	0.337	0.500
w/o Court simulation	0.171	0.473	0.875	0.300	0.462
w/o Knowledge base	0.145	0.462	0.850	0.312	0.475
w/o Web search	0.196	0.488	0.865	0.325	0.487

Table 6: Ablation study of our framework in the first instance setting.

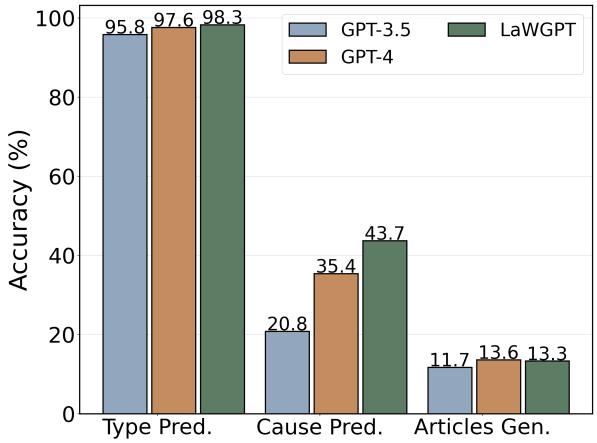

Figure 5: Legal knowledge evaluation of LLMs.

Case type	Precision	Recall	F1 Score
All	0.219	0.189	0.203
Criminal	**0.489**	**0.264**	**0.343**
Civil	0.073	0.063	0.067
Administrative	0.126	0.250	0.167

Table 7: Legal articles evaluation of AgentsCourt.

types. However, their performance is less impressive on the challenging task of predicting case reasons, the GPT-4 model achieves only 35.4% accuracy, while LaWGPT, which has undergone extensive pre-training with professional knowledge, achieves only 43.7%. For the task of article generation, the performance of all models is poor, with LaWGPT sometimes producing garbled output, resulting in even worse performance.

Multi-agent Court Simulation The results of the ablation experiments, as shown in Table 6 in Appendix, demonstrate that our designed court debate simulation module effectively enhances the accuracy of judicial decisions. We further investigate the specific impact of this module on the prison term and fines in criminal case judgements. As depicted in Figure 6, it is evident that the absolute difference in prison term and fines significantly diminishes following the simulation of court debates.

Difficulty of Distinct Types of Cases Table 7 presents the results of our framework in generating legal articles across different types of cases in the first instance setting. The agent system produces more reliable legal articles in criminal cases, while its use and understanding of relevant legal statutes in civil and administrative cases are notably weaker. This observation may be attributed to the fact that the civil and administrative cases involve more complex issues, with multiple vested interests, such as contract disputes, family matters, or government decisions, requiring a deeper understanding of legal and social knowledge.

Legal knowledge base With the support of an external knowledge base, the performance of agent system in judicial reasoning improved significantly, with an increase of up to 6.2%. The achievements are also attributed to our designed automatic retrieval module. As shown in Table 8 in Appendix A, through the rough retrieval, the most similar cases only have a 62% consistency in the cause of action with the current cases. However, after the documents re-ranking, the consistency of the cause of action between retrieved cases and the current cases increased to 85%. This improvement proves the effectiveness of our retrieval module.

6 Conclusion

We propose a novel multi-agent framework AgentsCourt, which can sequentially simulate court debate, retrieve precedents, analyze cases, provide legal articles, and deliver clear judgment. Furthermore, we introduce SimuCourt, a judicial benchmark to evaluate the judicial analysis and decision-making power of agents. Then, we perform experiments to analyze different modules. The new judicial paradigm we presented effectively simulates the judicial decision-making with multi-agent, which significantly enhances judicial efficiency.

7 Limitation

In this paper, we introduce a novel judicial benchmark SimuCourt. After thorough analysis, our work still presents the following limitations:

- Our data only includes Chinese documents from "China Judgments Online." Despite our framewok AgentsCourt not being specifically designed for the civil law system, testing the agent system with real data from different legal systems is important.
- The judgement documents cover the three most common types of cases: criminal, civil, and administrative. Including a broader range of case types in the future would evaluate the judicial analysis and decision-making power of agents more comprehensively.
- Although our database contains a large number of precedents and legal resources, experimental results have shown that overall performance of agent systems is still unsatisfactory.

We look forward to further exploring the potential of the legal knowledge base in future studies.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng Li, Mengshu Sun, and Jun Zhao. 2023. Lego: A multi-agent collaborative framework with role-playing and iterative feedback for causality explanation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9142–9163.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

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.

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

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Soha Khazaeli, Janardhana Punuru, Chad Morris, Sanjay Sharma, Bert Staub, Michael Cole, Sunny Chiu-Webster, and Dhruv Sakalley. 2021. A free format legal question answering system. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 107–113.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023b. Sailer: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1035–1044.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

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 C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Toran Bruce Richards. 2023. Autogpt - the next evolution of data driven chat ai. https://auto-gpt.ai/.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In *IJCAI*, pages 3501–3507.

Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, and Shaoping Ma. 2023. Understanding relevance judgments in legal case retrieval. *ACM Transactions on Information Systems*, 41(3):1–32.

Pengxiao Song, Yixuan Jin, and Zhi Zhou. 2023. LaWGPT. https://github.com/pengxiao-song/LaWGPT.

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2019. Legal document retrieval using document vector embeddings and deep learning. In *Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 2*, pages 160–175. Springer.

Harry Surden. 2019. Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35:19–22.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023a. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

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.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023b. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. *arXiv preprint arXiv:2310.09241*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

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*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, Toronto, Canada. Association for Computational Linguistics.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2023. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3540–3549.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. Jecqa: a legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.

A Retrieval Module

As shown in Table 8, through the rough retrieval and documents re-ranking, the consistency of the cause of action between retrieved cases and the current cases increased to 85%.

B Example of Court Transcript

We present an example of court transcript simulated by multi-agent debate in Table 12.

C Data Demonstration

The detailed list is presented in Table 10. Furthermore, we show examples of the first-instance stage in Figure 7 and second-instance stage in Figure 8, respectively.

D Data Analysis

D.1 Data Description

Our choice of cases is driven by three reasons: (1) **Diversity of causes of action**. Based on our statistical analysis of data from the China Judgements Online over the past few years, we observed a significant long-tail distribution in various types of cases. For example, as shown in Figure 9, in the total civil cases of 2022, the top 15 causes of action accounted for 66% of the total number of cases. To reflect a broader spectrum of legal practice, we focus on maintaining diversity in the types of causes of action; (2) **Clarity of case analysis and facts**. We have meticulously selected judgement documents that provide detailed case analysis and clear determine facts for annotation. This aim is to enhance the quality and accuracy of data annotation while aiding agents in better understanding the judicial reasoning and legal articles; (3) **Uniqueness and accuracy of judgements**. We prioritize cases that are not overturned in appellate review. This ensures the consistency of our evaluation, as these cases have already undergone a rigorous litigation process and the judgements are fair.

D.2 Data Quality Inspection

We first process the privacy information of all documents. Specifically, We have meticulously anonymized sensitive information in the judgement documents. Then, After completing data annotation and handling private information, we manually inspect the data quality from various aspects.

Privacy Information Processing: We have meticulously anonymized sensitive information in the judgement documents. In addition to replacing personal names, place names, and institution names with generic terms, we also anonymize other details that could potentially disclose personal privacy, such as ID numbers, phone numbers, and addresses, to ensure the safety of personal privacy.

Manual Inspection: After completing data annotation and handling private information, we manually inspect the quality of SimuCourt: (1) *Case Meeting Standards*. The selected samples need to include clear case analysis and facts and have not been overturned in the appellate stage. (2) *Accurate Information Annotation*. Annotation should ensure the accurate and error-free extraction of key information from the original legal documents, including case analysis, legal articles, and judgement. (3) *Privacy Information Security*.In order to safe-

Precedents	Rough retrieval	+ Re-ranking
Top1	62%	85%
Top2	60%	82%
Top3	61%	80%

Table 8: Cause of action matching

Criteria	Pass Rate
Case Meeting Standards	98.6%
Accurate Information Extraction	95.8%
Privacy Information Security	100%
Average	98.1%

Table 9: Data quality analysis.

guard individual privacy and security, it is crucial to ensure that each data entry does not contain any content that could potentially disclose sensitive information about the parties involved. We employ three graduate students to manually review all 420 annotated cases. By carefully scrutinizing, our dataset exhibits a high level of quality. Specific quality metrics and analysis results are shown in Table 9.

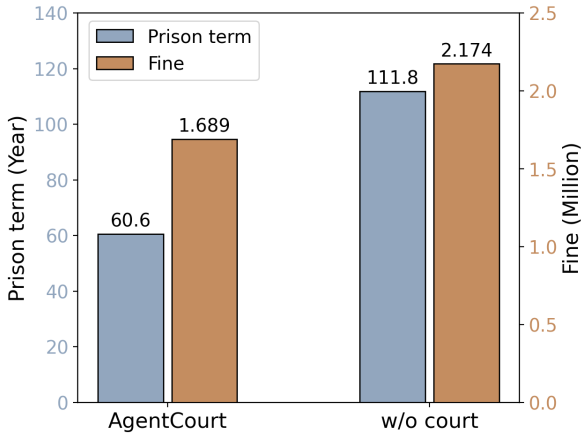

Figure 6: The absolute difference change.

E Details of Automatic Evaluation

E.1 Legal Articles Evaluation

We start by using pattern matching to parse the free text, followed by a hard match against specific legal provisions. For example, as shown in Table 13. Then, with TP (True Positives) = 2, FP (False Positives) = 1, FN (False Negatives) = 2, the corresponding ***Precision*** = 2/3, and ***Recall*** = 2/4.

E.2 Judgement Evaluation for Civil and Administrative Cases

We utilize GPT-4 to assess the judgment results generated by the model in civil and administrative cases. As shown in Table x, we present an evaluation example, which is also a prompt demonstration for GPT-4.

John Smith Case of Theft

Case type: Criminal **Cause of Action:** Theft

Case Details (First Instance)

Plaintiff: Prosecutor's Office

Defendant: John Doe

Background of the defendant: In January 2017, John Doe was sentenced by a certain People's Court to seven months in prison for theft; In October 2017, John Doe was sentenced to seven ...

Determine facts: During July and August of 2023, the defendant, John Doe, drove to the workshop yard of a certain company 11 times, stealing a total of over 4.28 tons of aluminum scrap, and later sold the stolen goods for over 53,000 yuan ...

Indictment: Upon appraisal, the involved aluminum scrap was valued at 74,700 RMB. The prosecutor's office charges the defendant with theft and suggests sentencing the defendant to three years of fixed-term imprisonment and imposing a fine of 20,000 RMB …

The point of defense lawyer: Upon apprehension, the defendant, John Doe, fully confessed to the crimes as detailed above. Subsequent to the offense, the defendant's family has fully restituted the proceeds ...

Statement of the defendant: John Doe does not dispute the facts and charges alleged in the indictment but claims to have actively cooperated with the investigation and requests leniency in sentencing …

Judicial Decision-Making (Output)

The court holds that the defendant, John Doe, has repeatedly stolen citizens' property, ... , and should be severely punished. The charges brought by the prosecutor's office are established. After being apprehended, the defendant truthfully confessed to his crimes ... The defense attorney's reasonable plea for leniency for the defendant is accepted ...

Case analysis

In conclusion, based on Article 64, Paragraph 3 of Article 67, Article 264 of the Criminal Law of the People's Republic of China, and Article 15 of the ...

Legal Articles

The judgment is as follows: the defendant is found guilty of theft and is hereby sentenced to three years and eight months of fixed-term imprisonment, and fined 10,000 RMB.

Judgement

Figure 7: An example case of first-instance stage (translated from Chinese).

John Smith Case of Intentional Injury

Case type: Criminal **Cause of Action:** Intentional injury

Case Details (First Instance)

Appellant: John Smith (original defendant)

Appellee: Prosecutor's Office

Background of the appellant: In July 2010, John Smith was sentenced to two years in prison for robbery and was released upon completing his sentence in December 2012.

Determine facts in the first instance: The defendant John Smith, in August 2022 had a verbal dispute with Emily Taylor (the victim, female, 52 years old) over debt issues. John Smith dragged Emily Taylor to the ground. According to judicial appraisal, Emily Taylor suffered …

Judicial analysis in the first instance: The first-instance court determined that the defendant John Smith intentionally caused bodily harm to another person, resulting in minor injuries …

Legal articles of the first instance: Article 234, paragraph 1, Article 67, paragraph 1, Article 45, and Article 47 of the Criminal Law of the People's Republic of China.

Judgement of the first instance: The defendant John Smith was convicted of the crime of intentional injury and sentenced to six months of imprisonment.

Petition for appeal: The original defendant John Smith appealed. The appellant John Smith's grounds for appeal are that during the debt collection process, the victim Emily Taylor tore and insulted the appellant, showing fault. According to the law, the appellant's criminal responsibility should be mitigated. Since the first instance did not recognize this, the appellant requests …

The point of the appellant: 1. The appellant does not dispute the criminal facts and charges determined in the first instance, and voluntarily admits guilt; 2. The company operated by the appellant John Smith pays an annual tax of over 3 million yuan, which has stimulated local economic development … ; 3. The incident in this case occurred suddenly and was not the intention of the appellant. The victim demanded wages, and the appellant has already compensated the victim with 80,000 yuan, obtaining the victim's forgiveness. Therefore, the appellant requests that the second instance change the judgment to probation for the appellant.

The point of the appellee: The first-instance court found that the facts of John Smith's intentional injury were clear, the evidence was solid and sufficient, the application of the law was correct, and the conviction and sentencing were accurate. The appellant's grounds for appeal lack factual and legal basis. It is recommended that the second-instance court uphold …

Determine facts in the second instance: Upon review during the second instance, it was confirmed that the facts of John Smith's intentional injury, as determined in the original verdict, were clear. During the second instance, the appellant's defense counsel provided a "petition" from company employees, demonstrating John Smith's good behavior on normal occasions.

Judicial Decision-Making (Output)

The court holds that John Smith intentionally caused bodily harm to another person, ... The facts as determined in the first instance were clear, ... After the incident, the appellant John Smith voluntarily surrendered and ... The materials submitted during the second instance demonstrated that the enterprise he operated encountered operational difficulties, resulting in workers' incomes being … *Case analysis*

In conclusion, based on Article 236, paragraph 1, item (ii) of the Criminal Procedure Law of the People's Republic of China, Article 234, … *Legal Articles*

The judgment is as follows: the appellant was convicted of the crime of intentional injury and sentenced to six months' imprisonment, suspended for one year.

Judgement

Figure 8: An example case of second-instance stage (translated from Chinese).

First instance	Second instance
Case type	Case type
Cause of Action	Cause of Action
Plaintiff	Appellant
Defendant	Appellee
Background information of the defendant	Background information of the appellant
Indictment	Petition for appeal
The point of defense lawyer	The point of the appellant
The point of the defendant	The point of the appellee
Determine facts	Determine facts in the first instance
Case analysis	Judicial analysis in the first instance
Legal Articles	Legal articles of the first instance
Judgement	Judgement of the first instance
	Determine facts in the second instance
	Case analysis
	Legal Articles
	Judgement

Table 10: Information list of different trial stages.

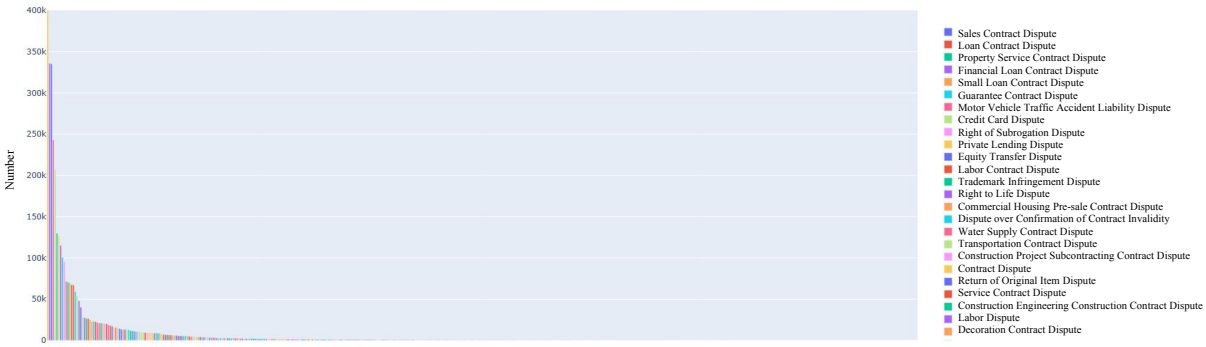

Figure 9: Cause of action of civil cases statistics in 2022

Cause of action	Item	Content
Theft	Case analysis	The court holds that the accused, John Doe, has repeatedly stolen citizens' property, constituting theft, and should be severely punished. The charges brought by the prosecutor's office are established. After being apprehended, the accused truthfully confessed to his crimes, voluntarily pleaded guilty, and returned part of the stolen goods, thus is eligible for a lighter punishment according to law. The defense attorney's reasonable plea for leniency for the accused is accepted.
	Legal articles	*Article 64 of the Criminal Law of the People's Republic of China*; *Paragraph 3 of Article 67 of the Criminal Law of the People's Republic of China*; *Article 264 of the Criminal Law of the People's Republic of China*; *Article 15 of the Criminal Procedure Law of the People's Republic of China*.
	Judgement	*Charge*: The defendant is convicted of theft; *Prison term*: Sentenced to three years and eight months in prison; *Fine*: Fined ten thousand yuan.

Table 11: An example output of criminal case.

Role	Record
Judge	(1)Determine facts: On October 10, 2022, the defendant John Smith noticed the victim, David Lee, wearing a pair of gold earrings near a hotel. He then followed David Lee into an alley, and taking advantage of David Lee's unawareness, he snatched the pair of gold earrings from behind. After successfully obtaining them, the defendant John Smith fled the scene... (2) Background of the defendant: John Smith was sentenced to six months' imprisonment by a certain People's Court in June 2009 for the crime of theft; John Smith was sentenced to two years' imprisonment by a certain People's Court in December 2010 for the crime of robbery... (3) Court Discipline...The plaintiff and the defendant may now make their statements.
Plaintiff	The defendant, John Smith, committed the crime of robbery with the purpose of unlawfully obtaining property from others, involving a considerable amount. His actions violate Article 267 of the Criminal Law of the People's Republic of China. The criminal facts are clear, and the evidence is indeed substantial and sufficient. Therefore, he should be held criminally liable for the crime of robbery.
Defendant	I admit that the robbery occurred on October 10, 2022, which was an act of impulse. After the incident, I have truthfully confessed my crime. I hope the judge can give me a lenient punishment.
Plaintiff	The defendant has previously been sentenced to a term of imprisonment for intentionally committing a crime. If he commits another crime within five years of completing the sentence, he should be sentenced to a term of imprisonment exceeding the original sentence, which constitutes recidivism.
Defendant	I deeply acknowledge that my actions have caused adverse effects on society and irreversible losses to the victims. However, I assure the court that I will actively compensate the victims for their losses. I sincerely repent and am willing to accept the court's judgment.

Table 12: An example of court transcript simulated by multi-agent debate (translated from Chinese).

Reference legal articles	Generated legal articles
Article 67, Section 1 of the Criminal Law of the People's Republic of China	Article 67 of the Criminal Law of the People's Republic of China
Article 52 of Criminal Law of the People's Republic of China	Article 53 of the Criminal Law of the People's Republic of China
Article 53 of the Criminal Law of the People's Republic of China	Article 52 of Criminal Law of the People's Republic of China
Article 15 of the Criminal Procedure Law of the People's Republic of China	

Table 13: A comparison between the reference legal articles and the legal articles generated by the model (translated from Chinese).

Reference judgment	Generated judgment
Result1: The defendant, Company XYZ, shall return the plaintiff's rental deposit of 6000 yuan and the remaining rent of 17025 yuan within five days from the effective date of this judgment.	**Result1**: The defendant shall return the deposit of 6000 yuan to the plaintiff.
Result2: The defendant, Company XYZ, shall pay the plaintiff a penalty of 5950 yuan within five days from the effective date of this judgment.	**Result2**: The defendant shall return the remaining rent of 17025 yuan to the plaintiff.
Result3: The plaintiff's other claims are dismissed	
The analysis by GPT-4:	
Key Point Match Evaluation	The standard answer contains 4 key points: the return of the rental deposit of 6000 yuan, the return of the remaining rent of 17025 yuan, the payment of a penalty of 5950 yuan, and the dismissal of other claims. The generated answer's "Result 1" matches the first key point (the return of the rental deposit of 6000 yuan), and "Result 2" matches the second key point (the return of the remaining rent of 17025 yuan). Therefore, the score for key point match is 2.
Answer Accuracy Evaluation	The generated answer consists of 2 results. The "Result 1" of the generated answer accurately matches one key point without including incorrect information, thus scoring 1 point. Similarly, the "Result 2" of the generated answer accurately matches another key point without including incorrect information, also scoring 1 point. Therefore, the score for accuracy is 2.
Output Result	Standard Key Points Count: 4, Key Point Match Score: 2, Generated Answer Results Count: 2, Answer Accuracy Score: 2
Precision	Answer Accuracy Score / Generated Answer Results Count = 2/2
Recall	Key Point Match Score / Standard Key Points Count = 2/4

Table 14: A prompt demonstration for GPT-4 evaluator (translated from Chinese).