SimuCourt: Building Judicial Decision-Making Agents with Real-world Judgement Documents

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 solely on individual judicial stage, overlooking cross-stage collaboration. 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 introduce *SimuCourt*, a judicial benchmark that encompasses 420 judgment documents, spanning the three most common types of judicial cases, and a novel task *Judicial Decision Making* to evaluate the judicial analysis and decision-making power of agents. To support this task, we construct a large-scale judicial knowledge base, JudicialKB, with multiple legal knowledge. (2) we propose a novel multi-agent framework, *AgentsCourt*. Our framework follows the real-world classic court trial process, consisting of court debate simulation, legal information retrieval and judgement refinement to simulate the decision making of judge. (3) we perform extensive experiments, the results demonstrate that, compared to the existing advanced methods, our framework outperforms the existing advanced methods in various aspects, especially in generating legal grounds, where our system achieves significant improvements of 8.6% and 9.1% F1 score in the first and second instance settings, respectively.

1 Introduction

Recent advances in deep learning have significantly impacted the legal domain, with notable achievements in legal event detection (Yao et al., 2022a; Fei et al., 2023), legal question answering (Zhong et al., 2020; Khazaeli et al., 2021; Martinez-Gil, 2023), and legal judgment prediction (Chalkidis et al., 2019; Hwang et al., 2022; Fei et al., 2023). 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 precedents retrieval. The complexity of this process demands close collaboration and interaction between stages. Although current research has made progress in individual areas, 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 with complex judicial decisions. Meanwhile, au-

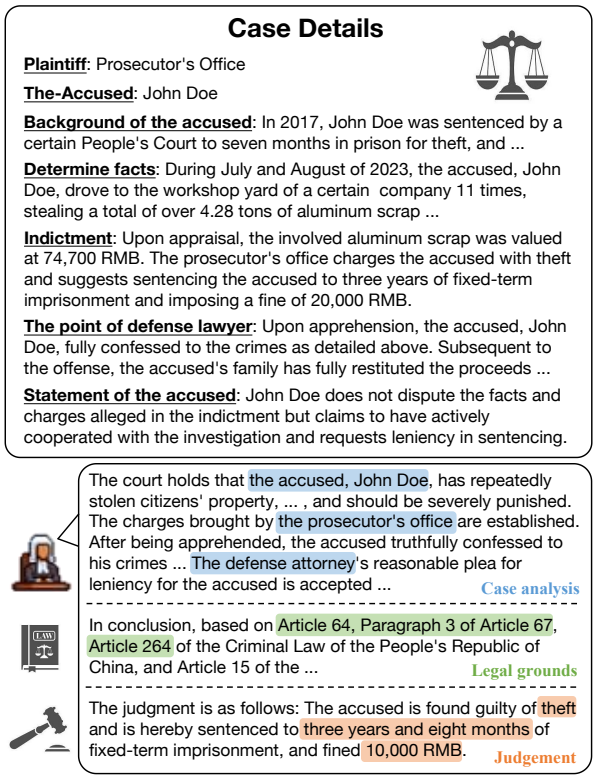

Figure 1: We formulate the Judicial Decision-Making task using the real-world judgement documents: given the case details above, judge agent must 1) conduct a logically clear case analysis; 2) provide precise legal grounds; 3) issue a definitive judgement.

tonomous 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., 2023). An increasing number of agents are being proposed to make 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) *Expert knowledge of judicial domain*. Judicial adjudication requires an in-depth understanding and accurate application of specialized knowledge such as laws, case precedents, and judicial procedures. (2) *Complex and hybrid reasoning*. The agents must be capable of handling a complex amalgamation of logical, factual, and legal reasoning, often interwoven in cases. (3) *Intricate ethical relationships*. In judicial decision, ethical and moral considerations, which are often subtle and multi-faceted, must be taken into account.

In this paper, we introduce **SimuCourt**, a judicial benchmark designed to evaluate Agent-as-Judge across a spectrum of different cases. SimuCourt encompasses 420 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 deal with acts that violate laws. 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. We formulate a novel task **Judicial Decision Making**, as illustrated in Figure 1. Given the case details, agent must conduct a logically clear *case analysis*, provide precise *legal grounds* and issue a definitive *judgement*. Furthermore, we construct a large-scale judicial knowledge base, **JudicialKB**, 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.

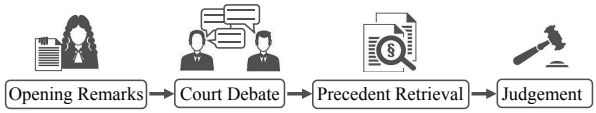

Figure 2: Simplified court trial process.

We also present a novel multi-agent framework, **AgentsCourt**. It follows the classic court trial process and simplifies the process into four phases: opening remarks, court debate, final statement, and judgement, as depicted in Figure 2. Specifically, we first develop a *Court Debate Simulation Module* with three agents. One agent serves as the judge to open a court session and announce the 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. This module provides a platform for all parties involved to present their points and arguments fairly. Then, we devise the *Legal Information Retrieval Module* which 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 *Judgement Refinement Module* which 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, and subsequently refines the judgement using legal information retrieved.

We summarize our contributions as follows:

- 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.

- We propose a novel multi-agent framework AgentsCourt. Given the basic information of a case, AgentCourt can sequentially simulate court debate, retrieve precedents, analyze cases, provide legal grounds, and deliver clear judgment. The new judicial paradigm simpli-

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

fies the process of making judicial decisions, significantly enhancing judicial efficiency.

- We perform experiments and ablation studies to explore factors that impact performance. The results indicate that our framework outperforms the existing advanced methods in various aspects, especially in generating legal grounds, where our system achieves significant improvements of 8.6% and 9.1% F1 score in the first and second instance experimental settings, respectively. We provide our data in the supplementary material for the double-blind review and our code will be open-sourced after the review stage.

2 Related Work

Real-world Agent tasks Agents are becoming smarter and more autonomous, focusing on practical real-world task beyond traditional tasks. Shridhar et al. (2020) introduces a simulator that enables agents to learn abstract, text-based policies in TextWorld(Côté et al., 2019) and then execute house-holding goals. Toyama et al. (2021) proposes the use of an Android simulator to learn phone operations. Yao et al. (2022b); Zhou et al. (2023); Deng et al. (2023) create simulated environments to develop web agents that perform web browsing tasks. Liu et al. (2023) examine agents' abilities to operate on real databases via SQL and evaluate agents in genuine OS' interactive bash environments. Anonymous (2024) introduces a lifelong learning environment for developing autonomous agents capable of performing human-like analysis on societal topics such as economics.

Multi-agent framework Cooperation among agents like human group dynamics can enhance the efficiency and effectiveness of task accomplishment. Li et al. (2023) 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. (2023) further leverage multi-agent cooperation to achieve better performance on multiple tasks.

Agent system for judicial industry The rapid emergence and growing popularity of language models have laid the groundwork for the development of specialized legal language models (Cui et al., 2023; Nguyen, 2023; Huang et al., 2023). Although most existing work primarily uses large models as unstructured knowledge bases for legal question answering. , and the results demonstrate that LLMs have surpassed 90% of human-level performance in the Uniform Bar Exam (Achiam et al., 2023), there is a concern that information retrieval systems might select the correct answer for incorrect reasons (Hubbard et al., 2017; Couch et al., 2018). Furthermore, complex judicial processes, such as judicial decision-making, are still mainly conducted by legal experts. In this paper, we focus on modeling the judicial decision-making process as an agent generative task, which involves analyzing case details, providing judicial grounds, and determining judgement.

3 The SimuCourt Benchmarck

3.1 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 Complaint, Statement of Plaintiff and Defendant, Determine facts, etc. For second-instance cases, each sample contains Petition for appeal, Statement of the appellant and appellee, Determine facts in the first Instance, etc. Detailed list can be found in the Appendix A. Most of cases were released after April 2023. This minimizes the risk of data leakage.[2]

3.2 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 Appendix C, 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

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

Feature	Criminal	Civil	Administrative
# of Cases	140	140	140
# of Causes of action	44	51	33
Avg # of Legal grounds	6.3	3.3	1.6
Max # of Legal grounds	11	10	8
Total # of Legal grounds	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 1: Statistics of SimuCourt.

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

Table 2: Statistics of our judicial knowledge base.

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 grounds; (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. Detailed data statistics of SimuCourt are shown in Table 1.

3.3 Data Quality

Our selected judgement documents undergo rigorous scrutiny, ensuring the accuracy and completeness of the legal texts and information. The clarity of information in these documents facilitates our efficient and precise data annotation. 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 quality of SimuCourt from various aspects. Detailed data quality inspection can be found in Appendix B.

3.4 Judicial 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 judicial knowledge base consists of laws, regulations, judicial interpretation, journal articles, and precedents. Detailed data statistics of *JudicialKB* are shown in Table 2.

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]. These articles span various legal fields, including but not limited to civil law, commercial law, criminal law and administrative law.

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 7 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. This comprehensive and diverse repository of precedents facilitates the identification of similar cases and provides effective references and guidance for judges when dealing with new cases.

3.5 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 grounds, and definitive final judgement. SimuCourt encompasses two experimental settings:

First Instance This setting refers to the trial court level, where the judge listens to arguments, determines the guilt of the defendant, and assesses whether punitive measures are warranted. Within this setting, the primary focus is on evaluating the agent's understanding of relevant laws and its analysis of case facts.

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

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

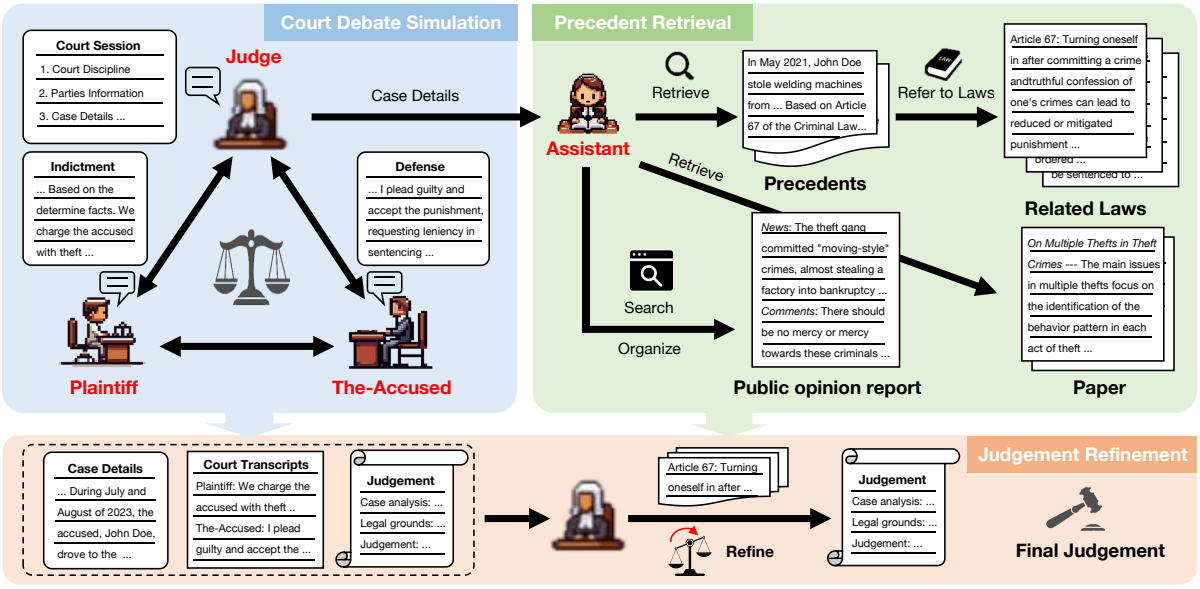

Figure 3: Overview of our multi-agent framework. The Court Debate Simulation Module recreates the court debate process through role-playing, mining different parties' points from limited real records. The Legal Information Retrieval Module employ an assistant agent to integrate information retrieved.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

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. Through these assessments, we aim to comprehensively evaluate the agent's legal intelligence and logical reasoning abilities in judicial practice.

4 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.

4.1 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. However, 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., 2023; Qian et al., 2023; Chen et al., 2023), in this module, we aim to reconstruct the court debate with multiple agents for each case.

Simulated Process Since we have already collected determine facts of each case from judgement documents, we simplify the simulated court process into four stages: opening remarks; court debate; final statements; judicial decision. We set up three agents to play the roles of the judge, plaintiff, and defendant respectively. 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.

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. 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

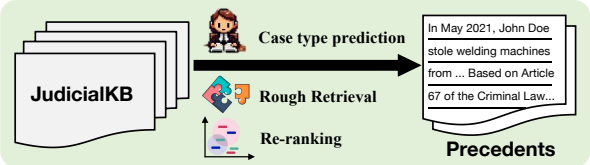

Figure 4: Automatic retrieval of precedents.

the plaintiff and defendant, respectively, without setting separate roles for representatives.

4.2 Legal Information Retrieval

Court debate serves 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 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 similar 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 grounds from the top 5 precedents as related legal provisions of current case.

4.3 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 grounds and subsequently reaching a preliminary judgement. Here is the preliminary judgement of judge regarding the case in Figure 1:

```
Analysis: The court finds John Doe
premeditatedly stole significant assets,
constituting theft. His criminal record
and rapid reoffense warrant a stricter
sentence ...
Legal grounds: Articles 65, Paragraph 3
of Article 67, and Article 264 of the
Criminal Law of the People's Republic of
China.
Judgement: The accused is found guilty of
theft and is hereby sentenced to three
years and six months of fixed-term
imprisonment, and fined 5,000 RMB.
```

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.

5 Experiments

5.1 Automatic Evaluation Metrics

As illustrated in Appendix D, the legal grounds and judgement are concise and structured. Therefore, we propose corresponding automatic evaluation metrics for each and we will make the evaluation code publicly available.

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

	Model	Legal Grounds			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	0.51	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	0.52
	AgentCourt	**0.219**	**0.189**	**0.203**	**0.437**	**0.603**	**0.507**	**0.887**	**0.337**	**0.500**	**0.55**	**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	0.54	0.526
	GPT-4	0.200	0.267	0.228	0.356	0.482	0.409	0.800	0.183	0.533	0.53	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	0.57
	AutoGPT	0.217	0.248	0.231	0.371	0.478	0.417	0.816	0.166	0.550	0.54	0.59	0.583
	AgentCourt	**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 3: Overall performance of our framework and baselines in the first and second instance experimental settings.

Judgement Evaluation for Civil and Administrative Cases The judgment of each civil or administrative case may encompass multiple results, such as the confirmation of legal obligations, compensation orders, and the allocation of litigation costs. 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. 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.

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

We employ manual assessment for the generated case analysis. This is because 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 undergraduate students 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.

5.4 Main Results

As shown in Table 3, our framework outperforms other models in all aspects. For the evaluation on legal grounds, 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

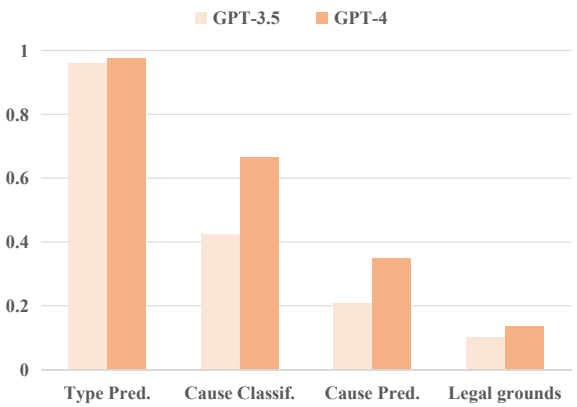

Figure 5: Judicial knowledge evaluation of LLMs.

Figure 6: The absolute difference change.

improvement in terms of correctness and concision.

5.5 Discussion and Analysis

Judicial Knowledge of LLMs As indicated in Figure 5 and presented in Table 11 in Appendix, the LLMs demonstrates a high accuracy rate of 96.6% in predicting case types based on limited basic case information. However, its performance in predicting causes of action is less impressive, with only a 35.0% accuracy rate, which increased to just 66.6% even when provided with a list of potential causes. The F1 score of legal grounds generated by LLMs is lower, only 13.6%. This highlights the limitations of LLMs in judicial knowledge.

Difficulty of Distinct Types of Cases Table 4 presents the results of our framework in generating legal grounds across different types of cases in the first instance setting. The agent system produces more reliable legal grounds 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 in criminal cases, the nature of the offense is typically clearer, allowing the agent to more easily apply relevant legal statutes to specific situations. In contrast, civil and administrative cases often 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.

Multi-agent Court Simulation The results of the ablation experiments, as shown in Table 12 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

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 4: Legal grounds evaluation of AgentCourt.

depicted in Figure 6, it is evident that the absolute difference in prison term and fines significantly diminishes following the simulation of court debates.

Judicial 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 5 in Appendix E, 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 introduce SimuCourt, a judicial benchmark to evaluate the judicial analysis and decision-making power of agents. Furthermore, we propose a novel multi-agent framework AgentsCourt, which can sequentially simulate court debate, retrieve precedents, analyze cases, provide legal grounds, and deliver clear judgment. 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: (1) Our data only includes Chinese documents from "China Judgments Online." Despite our framewok AgentCourt not being specifically designed for the civil law system, testing the agent system with real data from different legal systems is important; (2) Our dataset only covers 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; (3) 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 judicial knowledge base in future studies.

References

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

Anonymous. 2024. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *The Twelfth International Conference on Learning Representations*.

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*.

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*.

Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2019. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pages 41–75. Springer.

Brian A Couch, Joanna K Hubbard, and Chad E Brassil. 2018. Multiple–true–false questions reveal the limits of the multiple–choice format for detecting students with incomplete understandings. *BioScience*, 68(6):455–463.

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*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*.

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*.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Joanna K Hubbard, Macy A Potts, and Brian A Couch. 2017. How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats. *CBE—Life Sciences Education*, 16(2):ar26.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

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. 2023. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

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.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *arXiv preprint arXiv:2302.05729*.

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.

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/.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Daniel Toyama, Philippe Hamel, Anita Gergely, Gheorghe Comanici, Amelia Glaese, Zafarali Ahmed, Tyler Jackson, Shibl Mourad, and Doina Precup. 2021. Androidenv: A reinforcement learning platform for android. *arXiv preprint arXiv:2105.13231*.

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. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

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

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022a. Leven: A large-scale chinese legal event detection dataset. *arXiv preprint arXiv:2203.08556*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022b. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

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)*.

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

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

A List of Information

The Detailed List is presented in Table 6

B 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.

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

Table 5: Cause of action matching

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 grounds, and judgement. (3) *Privacy Information Security*.In order to safeguard 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 10.

C Data analysis

The cause of action of civil cases statistics in 2022 is shown in Figure 7

D Data example

We list several cases of criminal, civil and administrative in Table 7-9.

E Retrieval module

As shown in Table 5, 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%.

First instance	Second instance
Case Category	Case Category
Cause of Action	Cause of Action
Plaintiff	Appellant
Defendant/The accused	Appellee
Background information of the defendant	Background information of the appellant
Complaint/Indictment	Petition for appeal
Statement of the plaintiff	Statement of the appellant
Statement of the defendant/the accused	Statement of the appellee
Determine facts	Determine facts in the first Instance
	Judicial analysis in the first Instance
	legal grounds of the first Instance
	Judgement of the first Instance
	Determine facts in the second Instance

Table 6: Information list of different trial stages.

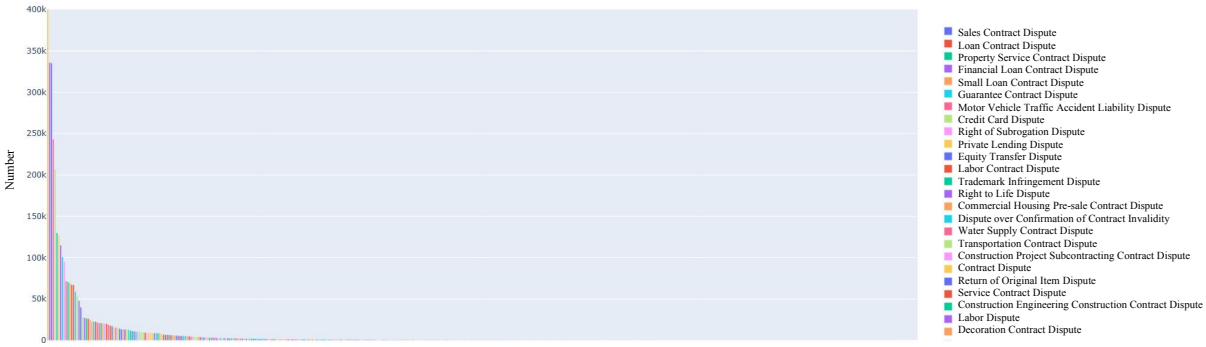

Figure 7: Cause of action of civil cases statistics in 2022

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 grounds	*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 7: Criminal case example.

Cause of action	Item	Content
Private lending dispute	Case analysis	The court holds that legal private lending is protected by law. The mortgage loan contract between the plaintiff and defendant is lawful and valid, obliging all parties to fully comply. After the plaintiff lent the money, the defendant mortgaged their property as collateral and registered this mortgage. The defendant must repay the principal and interest as agreed or bear the breach of contract responsibilities, including the plaintiff's legal fees and preservation guarantee fees incurred for debt collection. The court supports the plaintiff's claim for legal and preservation fees, as stipulated in the contract and evidenced by corresponding receipts.
	Legal grounds	*Article 389 of the Civil Code of the People's Republic of China*; *Article 394 of the Civil Code of the People's Republic of China*; *Article 395 of the Civil Code of the People's Republic of China*; *Article 400 of the Civil Code of the People's Republic of China*; *Article 407 of the Civil Code of the People's Republic of China*; *Paragraph 1 of Article 509 of the Civil Code of the People's Republic of China*; *Article 577 of the Civil Code of the People's Republic of China*; *Article 675 of the Civil Code of the People's Republic of China*; *Article 676 of the Civil Code of the People's Republic of China*; *Article 67 of the Civil Procedure Law of the People's Republic of China*.
	Judgement	*Result 1*: The defendant shall return the principal amount of 800,000 yuan to the plaintiff within ten days from the effective date of this judgment, and pay interest based on the unpaid principal; *Result 2*: The defendant shall pay the plaintiff's attorney fees of 49,000 yuan and the preservation guarantee fee of 1,800 yuan within ten days from the effective date of this judgment.

Table 8: Civil case example.

Cause of action	Item	Content
Labor and Social Security Administration	Case analysis	The court finds that the plaintiff, a company, and the third party, Wang, have a clear labor contract relationship with defined rights and obligations. The fact that Wang was injured in an accident during working hours and at the workplace due to work-related reasons is clear and well-evidenced. The local authority of Gangcheng District, upon receiving the company's application for Wang's work-related injury recognition and legally reviewing the relevant materials, made a decision on the work-related injury recognition within 60 days and delivered the decision document, in compliance with Articles 14(1), 19, and 20 of the Work Injury Insurance Regulations, with legal procedures followed.
	Legal grounds	*Article 69 of the Administrative Litigation Law of the People's Republic of China*; *Paragraph 1 of Article 14 of the Work Injury Insurance Regulations*; *Article 19 of the Work Injury Insurance Regulations*; *Article 20 of the Work Injury Insurance Regulations*.
	Judgement	*Result*: Dismiss the plaintiff's claim.

Table 9: Administrative case example.

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

Table 10: Data quality analysis.

Model	Type Pred.	Cause Classif.	Cause Pred.	Legal grounds
GPT-3.5	96.0	20.8	42.5	11.7
GPT-4	97.6	35.0	66.6	13.6

Table 11: Judicial knowledge evaluation of LLMs

Model	Legal Grounds	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 12: Ablation study