Sim-Court: A Simulation of Court using LLM-based Agents

Kaiyuan Zhang
DCST, Tsinghua University
Beijing,China
ky-zhang24@mails.tsinghua.edu.cn

Zhaoxi Li
DCST, Tsinghua University
Beijing,China
li-zx24@mails.tsinghua.edu.cn

Xuancheng Li
DCST, Tsinghua University
Beijing,China
lixuanch23@mails.tsinghua.edu.cn

Abstract

Court trail has became a useful method to maintain justice and social stability throughout the world. However, the scarce legal resources and huge legal cases put extreme pressure on lawyers, judges, and the whole judicial system. An explainable, portable and efficient autonomous court trial needs to be built. In this paper, we introduce a simulacrum of court called Sim-Court that simulates the entire process of court trial on criminal cases. Our contributions lies in three aspect: (i) An easy-to-use agent framework for court simulation with carefully designed evaluation methods, providing a novel benchmark for LLM under legal scenarios. (ii) A more explicable legal case trial comparing to direct trail prediction methods, showing the advantages and value of this framework. (iii) A list of generated trail information and corresponding court record for facilitating further research on similar areas. Hopefully, Sim-court can paves the way for advancing applications of LLM-powered agent in legal scenarios. The full version is released at: https://github.com/Miracle-2001/Sim-Court.

1 Introduction

After the release of Chatgpt 3.5[10] in late 2022, Large Language Models (LLMs) becomes an extremely hot topic in both research and application areas in the past 2 years. Owing to tremendous training data, numerous model parameters and large calculation center, LLMs gains considerable comprehension to real world and human socialization[15].

With the help of LLMs, many attempts have been made by researchers and engineers who eager to build and release social simulation systems based on LLM-based agents.[3] For example, "Stanford Town"[11] construct a virtual village with multiple agents communicating with each other. Other research such as RecAgent[12], Agent Hospital[8], AgentCourt[2] also publish novel multi-agent simulation systems and unexpected discoveries. These simulation systems not only show the potential power of LLM-based agents but pave another way for application of LLMs.

When it comes to legal area, court trail plays an irreplaceable role in maintaining justice and social stability throughout the world. However, the scarce legal resources and huge legal cases put extreme pressure on lawyers, judges, and the whole judicial system. In 2023, 25.8 million legal cases were accepted in China, while there were only 10,145 court running over the country[6]. Except for the real court shortage, legal education is also hindered due to the limit amount of legal practitioners.[1]

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

If an explainable, portable and efficient autonomous court simulation system were built, this situation is hopefully to reach a remission. In this paper, we proposed a newly defined benchmark and framework for building a more advanced court simulation system via LLM-based agent. The main contribution of this system are listed as followings:

(i) An easy-to-use court simulation framework based on real court progress for criminal cases. Carefully designed evaluation methods are provided to maintain a logical and proper trail process. Experiments are conducted to analyze the role-play ability on legal domain of different LLMs.

(ii) Sim-Court presents a more explicable legal case trial comparing to direct trail prediction methods, showing that with a detailed court trail record, AI can give out more proper judgment.

(iii) 100 generated trail information and corresponding court record are released. These synthetic data can facilitate further research on similar areas.

Hopefully, this system can make contributions to save legal resources and provide an education platform for law students and even real world lawyers and judges.

2 Related Work

2.1 Social Simulation with LLM-based multi-agent system

The research and development of social science area are persistently facing challenges such as limited subjects, expensive but unrepeatable experiments and ethic concerns, etc. Recently, social simulation with LLM-based agent provide a solution for scientists. [9]

LLM-based agent can present human-like conversations according to the given instructions. With few lines of prompt and conversation context, a LLM can act a specific role and give outputs conforming to one's status. When multiple agents are contained in the system, co-operation and social actions can be found.[13] RecAgent[12] examine the recommendation system via multi-agent communication on a virual social media platform. AgentHospital[8] construct an entire treatment progress from registration to doctor consultant. AgentCourt[2] simulated simplified court trials via multi-agent debate. With the help of multi-agent system, tasks can be more easily evaluated and accomplished.

Besides the accomplishment of applications, theories in sociology and psychology can be validated using LLM-based multi-agent simulations. Jin et.al.[5] examine the influences on acceptance rate by simulating reviewers and chairs with different character, discovering theories such as anchoring bias, echo chamber effects and altruism fatigue which are aligned with sociology findings. Jia et.al.[14] test whether LLM can simulate human trust behavior via classical trust games designed by psychologists. The experimental results also validating phenomena like reciprocity anticipation, risk rerception and prosocial preference.

2.2 Artificial Intelligence in legal area

Legal tasks has been prominent improved by applying Artificial Intelligence technique. The legal field has witnessed an emergence of potential AI applications on various aspects. Tasks such as legal judgement prediction, legal question answering, legal language understanding, legal case retrieval, legal document summarization are ushered to a new research paradigm with the use of AI, especially LLM[7].

For legal simulation topic, Chen et.al. established AgentCourt [2], which provide a simplified court simulation system and each agent can maintain their role to some extent. He et.al. [4] show the vital function of constructing court process in Legal judgement prediction task. However, neither of them released a detailed designed court simulation system for the entire trial process simulation, which may not only provide a useful benchmark and application but also facilitate agent evolution on legal ability in the future.

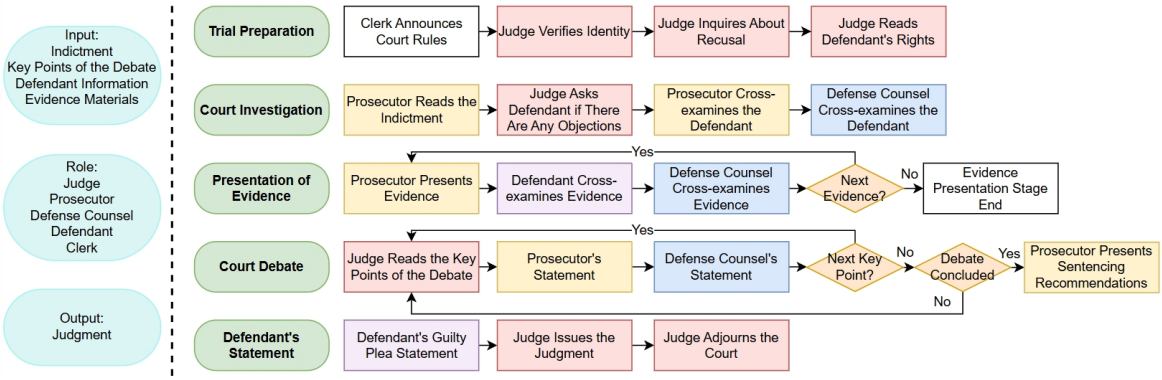

Figure 1: The flowchart of Sim-Court on criminal cases. The left part describe the input information, agent roles, and the expected output. The right part indicate the entire process of court simulation on criminal cases. Stages are designed according to reality.

3 Methods and Framework

3.1 Overall Court Simulation Progress

For a normal process of a court trail on criminal cases, 5 basic agents should be included, which are "Judge", "Prosecution", "Advocate", "Defendant" and "Stenographer". Besides, 5 stages are included in a court trial, which can be described as "Trial Preparation", "Court Investigation", "Presentation of Evidence", "Court Debate" and "Defendant's Statement". Each stage has detail flow path to make it more similar to reality. The flowchart of the whole process is shown as Fig.1

3.2 Role configuration and instructions

Each agents has its profiling and task. Profiles of roles will be restrict by LLM prompts at the beginning of the court and each role has a specific description in each stage, which will help him understand the tasks.

"Prosecution" should discover the truth and punish criminals. "Advocate" have to argue for rights and commutation for his party. These two roles should present and argue in the court. "Judge" should host and conclude the trial as well as give the verdict. "Defendant" is supposed to answer questions honestly to seek mitigation of sentence. "Stenographer" must announce court discipline and write down the court progress. Detailed instruction and prompt design are listed in Tab. 1.

3.3 Evaluation Methods

Evaluation methods varies on different roles. Roughly, the evaluation focus on three aspects: task completion quality, content and expression and spot performance as listed in Tab. 2. Specifically, "Judge", as the symbol of justice, is expected to present a trial documents with proper format and exact measurement of penalty. Besides, the hosting ability will be considered for the "Judge" agent. For "Advocate", this agent is supposed to take the defendant to prison, so his goal is to persuade "Judge" to produce a high penalty. For "Prosecution", the less penalty the defendant bear, the more successful the agent is. What's more, agents should maintain his role during conversation, and the expression should be concise and logical, without repetition.

Particularly, the tasks for "Defendant" and "Stenographer" are simple, so there is no need to evaluate their performances.

For evaluation technique, we apply autonomous evaluation by LLMs. We compare LLM generated trial documents with real ones to decide the judgment quality and use GPT-4 model to evaluate other aspects of each agent.

Table 1: Basic instruction of each agent. The content is designed after Criminal Law of the People's Republic of China.

Role	Basic Instruction
Judge	You are the presiding judge in this case, presiding over the trial. You have a thorough understanding of the relevant processes in the field of criminal procedure. To ensure the fairness of trial procedures, you should protect the right of the defendants and other participants in the proceedings.
Prosecution	You are an experienced prosecutor, specializing in the field of criminal litigation. Your task is to ensure that the facts of a crime are accurately and promptly identified, that the law is correctly applied, that criminals are punished, and that the innocent are protected from criminal prosecution.
Advocate	You are an experienced advocate. The responsibility of a defender is to present materials and opinions on the defendant's innocence, mitigation, or exemption from criminal responsibility in light of the facts and the law, and to safeguard the litigation rights and other lawful rights and interests of the suspect or defendant.
Defendant	You are the defendant in this case, you committed a crime, the trial will try your charges. The evidence in the case will be public, and you will be questioned by the judge and the prosecutor. You want to plead guilty in order to have your sentence reduced.
Stenographer	As a court stenographer, you are responsible for recording court proceedings, managing court documents, assisting judges, and ensuring the smooth running of court proceedings.

Table 2: Evaluation aspects for different roles. The evaluation focus on three aspects: task completion quality, content and expression and spot performance.

Role	Task Completion Quality	Content and Expression	Spot Performance
Judge	Judgment accuracy and tendency	Logic, clarity, role consistency, document format	Control ability
Prosecution	The degree of the bias between the penalty on indictment and the simulation result. The less, the better.	Logic, clarity, role consistency	Calmness and politeness
Advocate	The degree of penalty. The less, the better.	Logic, clarity, role consistency	Calmness and politeness

4 Experiment

4.1 Dataset

We collected 100 real judgement documents happening in 2018 Jiangsu Province. Afterwards, we used ERNIE-Speed-128K model to extract information including the information of defendant, statement of prosecution, evidence and debate focus. These 100 legal cases and corresponding court simulation records are currently released at https://github.com/Miracle-2001/Sim-Court

4.2 Settings

To find out more suitable language models as the agent, we test different LLM in different cases. We import multiple LLM models including ERNIE-Speed-128K, Chatglm-4-air, Claude3-sConnet, GPT3.5-turbo, GPT4o-mini, GPT-o1-mini, GPT-4. Since time is limited, we only examine the first 10 cases in the dataset.

Table 3: Overall result of different LLM under different evaluation aspects. 'Jud.','Pro.' and 'Adv.' represents as 'Judge','Prosecution' and 'Advocate', respectively.

LLM	Task Completion Quality			Content and Expression			Spot Performance		
	Jud.	Pro.	Adv.	Jud.	Pro.	Adv.	Jud.	Pro.	Adv.
Chatglm-4-air	-0.21	1.1	3.2	6.9	5.9	6.2	9.1	9.7	9.6
ERNIE-Speed-128K	-0.49	0.9	3.4	6.3	6.4	6.2	9.2	9.9	9.7
Claude3-sonnet	-0.34	1.0	4.1	6.8	6.1	6.7	8.9	9.8	9.6
GPT3.5-turbo	+0.21	1.2	4.2	7.0	6.4	6.2	8.6	9.6	9.6
GPT4o-mini	+0.24	0.7	3.1	7.4	7.2	7.1	8.7	9.8	9.8
GPT-4	+0.37	0.6	3.2	7.5	6.9	7.2	8.7	9.8	9.8
GPT-o1-mini	+0.22	0.6	2.4	7.2	7.0	7.4	8.4	9.9	9.7
Average	0	0.9	3.0	7.0	6.6	6.7	8.8	9.8	9.7

4.3 Overall Results

First, we ask the judge model to mark between 1 to 10 for each evaluation aspect of each agent and mean score is calculated among the 10 testing cases. For the 'task completion quality' aspect, we fixed the other two agents as GPT-4 and only changed the testee agent. In addition, for evaluating judgment tendency, we report penalty bias (with year as unit) between the testee LLM and the average results. The general results are shown on Tab. 3.

For "Task Completion Quality", models such as chatglm-4-air and Claude3-sonnet tend to present mild judgment when playing the "Judge" role, while the GPT series shows the opposite phenomenon. When it comes to "Prosecution" agent, GPT series manage to persuade the "Judge" making prison terms more close to indictment than other LLMs. For 'Advocate', GPT-o1-mini performs better, with a shorter prison term in the final judgment. What is more, the variation among LLMs when acting as 'Prosecution' and 'Advocate' is obviously more gentle than that in 'Judge'. This is expected because the judge own the final decision right.

For "Content and Expression", models gain higher points when representing as 'Judge'. One possible explanation is 'Prosecution' and 'Advocate' face more long-context situation than 'Judge', which result in a more frequent failure in role consistency.

For Spot Performance, all LLMs receive a better score than former aspects. That is because LLMs tends to be polite and calm even when the situation is in emergency. Another finding is that 'Judge' is marked lower points as a result of failing to control the court when repetition and meaningless questions occur.

4.4 Bad Case Analysis

Some typical bad cases are presented and analyzed in this section.

Example 1: *... (When debating whether it is a crime of intentional injury) Advocate: Why did you choose to throw the glass bottle* ***instead of solving the problem in other ways?****...*

The advocate's question should follow the debate topic.

Example 2: *... (to Defendant) Prosecution: Elaborate on the circumstances and motives. ...*

The inquiries from prosecution should be exact. General questions have to be avoid.

Example 3: *... Prosecution: Is the* ***prosecution*** *has any other crimes to question? ...*

The prosecution agent failed to maintain his role and ask questions to himself.

These cases reveal that more precise prompt design or model finetuning need to be applied.

5 Future work

1. Exploring and enhancing multi-round conversations and long context retrieval ability. A real court process consists of many rounds of conversations, so agents should remember prior ones and give logical output. Techniques such as retrieval augmented generation (RAG) can make a difference.

2. Profile design and tasks understanding ability. In real world, both judge and lawyer are sophisticated practitioner. However, currently, agents in Sim-Court may forget his profile when facing long context situation. A useful finetuning methods under real court data should be employed.

3. By now, this system only cover criminal cases and 5 agents, civil cases and more agents will be taken into consideration in the future.

6 Conclusion

In this paper, we introduce Sim-Court, a court simulation system with detailedly designed process. Experiments on different LLMs and agents have been carried out, which present the potential to be a useful benchmark and application. Bad cases and future work are discussed by the end of the paper. Hopefully, more insights and further research questions can be discovered in the future.

References

[1] Earl L Carl. The shortage of negro lawyers: Pluralistic legal education and legal services for the poor. *J. Legal Educ.*, 20:21, 1967.

[2] Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *arXiv preprint arXiv:2408.08089*, 2024.

[3] Önder Gürcan. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 134–144, 2024.

[4] Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. Agentscourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416, 2024.

[5] Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*, 2024.

[6] William A Joseph. *Politics in China: an introduction*. Oxford University Press, 2024.

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

[8] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

[9] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*, 2024.

[10] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

[12] Lei Wang, Jingsen Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, and Ji-Rong Wen. Recagent: A novel simulation paradigm for recommender systems. *arXiv preprint arXiv:2306.02552*, 2023.

[13] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[14] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*, 2024.

[15] Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre LS Filipowicz. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, pages 368–370, 2023.