Chinese Court Simulation with LLM-Based Agents System*

Kaiyuan Zhang[1,2†], **Jiaqi Li**[1], **Yueyue Wu**[2,1‡], **Haitao Li**[1], **Cheng Luo**[3]
Shaokun Zou[1], **Yujia Zhou**[1], **Weihang Su**[1], **Yiqun Liu**[1], **Qingyao Ai**[1§]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Quancheng Laboratory, China [3]MegaTech.AI Inc.

Abstract

Mock trial has long served as an important platform for professional legal training and education. Traditional mock trials are difficult to access by the public because they rely on professional tutors and human participants. Fortunately, the rise of large language models (LLMs) provides new opportunities for creating more accessible and scalable court simulations. While promising, existing research ignored the systematic design and procedure evaluation of court simulations, which are critical to the credibility and usage of court simulation in practice. To this end, we propose a novel court simulation paradigm, i.e. SimCourt, based on the real-world procedure structure of Chinese courts, and design a comprehensive evaluation framework focusing on both legal judgment prediction and court procedure analysis. Experiments show that our framework can generate simulated trials that better guide the system in predicting the imprisonment, probation, and fine of each case. Further procedure evaluations show that agents' responses under our simulation framework even outperform judges and lawyers from the real trials in many aspects. These demonstrate the potential of LLM-based court simulation. Codes and datasets are available at: https://github.com/Miracle-2001/SimCourt.

1 Introduction

Mock trial, a widely adopted legal educational method, enables students and professionals to simulate trial proceedings, sharpen courtroom reasoning, and gain hands-on legal practice (Issa et al., 2023; Tang, 2021). Yet, traditional mock trials are heavily based on expert participation, appropriate venues, and offline human participation. Thus, they are time-consuming, expensive, and lack reproducibility (Zhang, 2021), which limits accessibility.

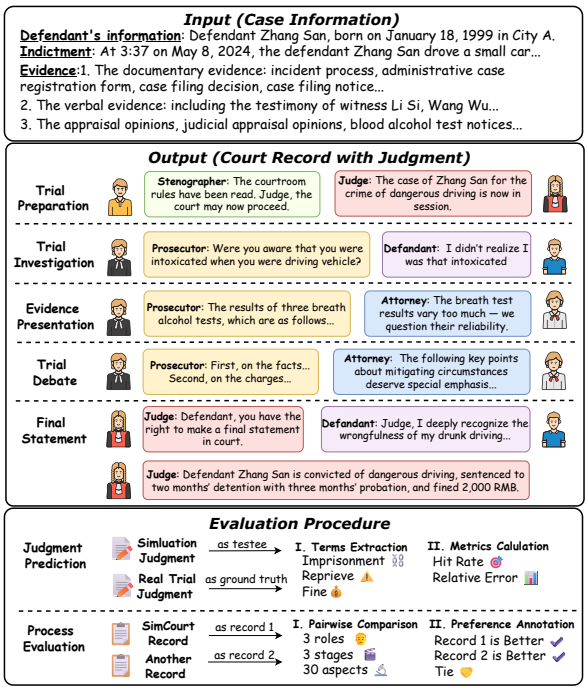

Figure 1: The input of SimCourt consists of 3 components: defendant's information, indictment and evidence. Its output is a 5-stage trial proceeding record, culminating in a final judgment. The evaluation consists of 2 parts: judgment prediction and process evaluation.

Fortunately, recent advances in LLM-based agents provide new opportunities to change this situation. Equipped with modules and tools, LLM-based agents can act as intelligent tools for a specific domain (Xi et al., 2023; Zhang et al., 2024; Cui et al., 2023). These capabilities are highly compatible with the demands of rule adherence, courtroom debate, long-context response, and legal reasoning inherent in the judicial area. With the support of LLM-based agents, researchers can design simulated trial systems capable of automatically generating trial records, providing valuable

*This work is supported by the National Key Research and Development Program of China (Grant No. 2024YFC3307101) and the Research Project of Quancheng Laboratory, China (Grant No.QCL20250105).

†ky-zhang24@mails.tsinghua.edu.cn
‡Corresponding Author: wuyueyue1600@gmail.com
§Corresponding Author: aiqy@tsinghua.edu.cn

Findings of the Association for Computational Linguistics: ACL 2026, pages 8429–8454
July 2-7, 2026

references to law students, legal professionals, and the public.

Despite its potential, existing studies on court simulation remain limited in scope. They mostly focus on simplified debate segments that do not cover the full procedural workflow of a trial, and there is no well-established and rigorous evaluation framework of both the accuracy of the judgment and the quality of the agent interaction in the simulated trials. To generate realistic mock trials, we need to develop a systematic simulation framework strictly following the actual procedure of a trail in the courtroom and build evaluation pipelines focusing on assessing the procedural correctness and downstream usefulness of simulated trial records.

To address these problems, we propose a novel mock trial framework– SimCourt – based on the official procedure of Chinese criminal trials. As illustrated in Figure.1, the input of SimCourt consists of basic case materials collected prior to a trial, including the defendant's information, the indictment, and the corresponding evidence. The simulation output is a full 5-stage trial record with the corresponding judgment document, which contains the final sentence and the explanations created by the judge. In contrast to previous studies that only focus on the Trial Debate stage, we introduce the Trial Investigation stage and the Evidence Presentation stage, and redesign other stages to align the simulated trials with real ones in Chinese courtrooms. In order to achieve professional, coherent, and logical court simulation, we carefully design different courtroom agents with three modules – the Profile, Memory, and Strategy module – as well as external tools to enhance their role maintaining, memorization, and legal reasoning abilities.

To comprehensively evaluate the quality of the simulated mock trial, we develop an evaluation framework and a benchmark based on both the objective goals of the court and a process evaluation guideline widely used by Chinese legal professionals[1]. The framework consists of two parts, namely, the judgment prediction and the process evaluation, which are illustrated in Figure.1. In the judgment prediction, we extract the final judgment from the output of SimCourt and compare it with the real judgment recorded for the same trial. In process evaluation, we summarize 30 aspects commonly used by legal experts to judge the performance of a judge or a lawyer, and conduct pairwise annotations to compare the quality of responses generated by our system and other trials, including those from real human professionals and baseline frameworks. The evaluation results annotated by human legal experts show that SimCourt can generate high-quality trial recordings, and agents in our system even outperform the real human in court from many perspectives.

The main contributions of our work can be summarized as follows.

(1) We propose SimCourt, a simulation framework for Chinese criminal courts based on real trial procedures and a set of courtroom agent frameworks that can simulate professional lawyers or judges and generate high-quality trial records.

(2) We develop a comprehensive evaluation framework as well as a benchmark with human expert annotations for a thorough evaluation of the court simulation system.

(3) The experiment results show that our system can generate responses better than baseline methods and even human expert from many perspectives, demonstrating the potential of LLM-based mock trails for future legal education and practice.

2 Related Work

Leveraging the preliminary understanding of the real world of LLMs, researchers have recently pioneered a new direction: simulating the real world using agents powered by LLMs (Mou et al., 2024). For instance, Wang et.al.(Wang et al., 2023) examine the recommendation system via multi-agent communication on a virtual social media platform. Li et.al. (Li et al., 2024b) construct an entire hospital treatment progress from registration to doctor consultant with LLM-based agent systems. Further researches on software development, gaming, employment consultant and social science area also validate novel application values of LLM-based agent system on real world simulation. (Qian et al., 2023; Xu et al., 2023; Tang et al., 2024; Xie et al., 2024; Jin et al., 2024).

For the court simulation area, Chen et.al. provides a simplified trial debate simulation framework named AgentCourt (Chen et al., 2024), in which lawyer agents can evolve with adversarial debating. He et.al. (He et al., 2024) proposes AgentsCourt, a judicial decision-making agent with simulation of trial debates and enhancement of legal knowledge, which shows the importance of

[1] http://gongbao.court.gov.cn/Details/ee6a5b1d20140c38c800c91c728d63.html

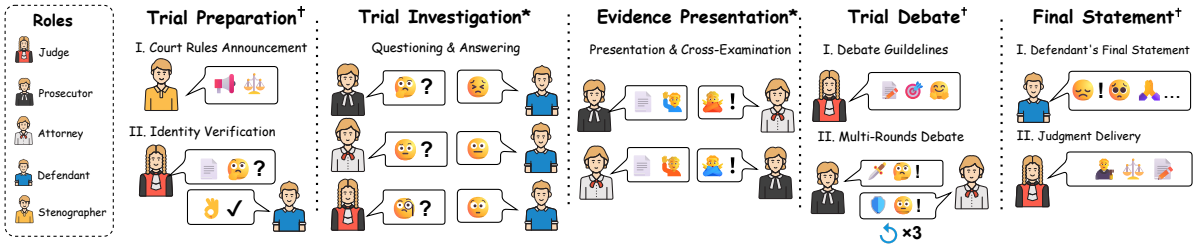

Figure 2: The diagram of the trial procedure in SimCourt, demonstrating the main component of each stage. The asterisk (*) marks courtroom stages newly introduced by SimCourt relative to previous studies, while the dagger (†) marks stages that are further expanded and refined in SimCourt relative to previous studies.

constructing the trial process in the legal judgment prediction (LJP) task. However, compared to SimCourt, these frameworks ignore systematic process design and contain only a simplified simulation evaluation, resulting in restricted application value in legal education and decision support. The detailed comparison between SimCourt and previous work is discussed in Appendix A.

3 Framework and Agent Design

3.1 Overall Procedure and Participants

The main procedure and all participants in SimCourt are presented in Figure.2. The detailed simulation procedure is presented in Appendix B. SimCourt is grounded in the official procedural structure of criminal trials in Chinese courts with 5 stages and 5 critical roles involved. Explanations of each stage are provided in the following.

Trial Preparation Stage represents the start of a trial. Here, Stenographer first announces the rules of the court. Then Judge in turn verifies the identity of parties, advises the Defendant's rights, and makes inquiries about recusal.

Trial Investigation Stage focuses on verifying the basic facts about the case. In this stage, Prosecutor and Attorney in turn examine the Defendant about the case proceedings, outcomes, post-case handling, and other related aspects. Judge may also choose to question the defendant if necessary.

Presentation of Evidence Stage is the time when Prosecutor and Attorney present their evidence in sequence. The evidence is subject to examination by the opposing party.

Trial Debate Stage is one of the most important parts of a trial. Prosecutor and Attorney engage in a comprehensive debate based on their respective positions, covering the nature of the case, the application of laws, and sentencing recommendations. The Debating typically lasts 3 rounds. Judge listens to both parties' viewpoints and may provide guidance at appropriate times.

Final Statement Stage is where Judge grants Defendant the right to make a final allocution for fairness and justice. After that, Judge issues the final judgment documents based on the case details and trial arguments.

To navigate and complete each stage of the courtroom process smoothly and effectively, each court participant must perform their job accurately and professionally. Specifically:

- **Defendant** needs to protect its own interests within the scope of the case information.
- **Prosecutor** should hold a proper questioning strategy, make logical accusations and effective defense rebuttal.
- **Attorney** should be aware of the Defendant's rights, and propose logical evidence rebuttal and defense arguments.
- **Judge** must ensure the fairness in the courtroom and make interventions or guidance when necessary. Also, it must fully investigate the truth and issue appropriate judgment based on the case details and trial arguments, typically including imprisonment, possible probation, and fines.
- **Stenographer** should record the trial and participants' interactions faithfully and correctly.

3.2 Agent Structure and Implementation

To fulfill the jobs of the participants, the agent we proposed consists of 3 modules and 2 external tools. These components and the agent workflow are shown in Figure.3.

Modules and Tools. To guide the agent in maintaining the characteristics of their roles and following the trial procedure accurately, we introduce

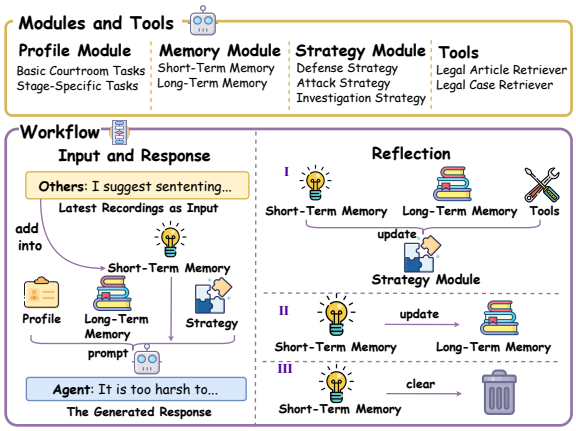

Figure 3: Agent structure and workflow of SimCourt, which consists of Profile module, Memory module, Strategy module and two external tools. The input-response and reflection workflow are also illustrated.

the *Profile module*, which consists of the descriptions of basic tasks for courtroom roles and stage-specific tasks for each specific stage, as illustrated in Appendix G. Moreover, in order to help agents generate coherent responses under the long-context courtroom conversation scenario, agents are provided with a *Memory module*, which is constructed with a short-term memory and a long-term memory. The short-term memory memorizes the latest statements and interactions in the current stage, and the long-term memory stores the summarization of critical statements of previous stages. We also design a *Strategy module* to assist the agents to generate structured, logical, and appropriate statements. Specifically, there are 3 types of strategies, which are Defense strategy, Attack strategy, and Investigation strategy. The Defense strategy focuses on maintaining the relevance and probative value of one's opinions; the Attack strategy focuses on undermining the opposing party's opinions; and the Investigation strategy focuses on clarifying critical issues related to conviction and sentencing.

To compensate general LLMs with legal domain knowledge, we further introduce two useful tools, i.e., a legal article retriever and a legal case retriever, in each courtroom agent. The *legal article retriever* can retrieve articles from a law base consisting of 55,347 Chinese laws, including all articles in the Criminal Law of the People's Republic of China. Here, we adopt the simplest yet effective approach by directly retrieving the content of legal provisions based on their names and numbers. The *legal case retriever* is implemented with a commercial search service named LegalOne[2], which is a LLM-based search engine that can retrieve and summarize judicial cases according to natural language queries. LegalOne is built from a large-scale corpus with more than 2 million legal documents and 200,000 representative cases selected by The Supreme People's Court of China.

Workflow. As shown in Figure.3, agents use the latest recordings as input and generate responses via their Profile, Memory, and Strategy modules. At each stage's end, each agent performs a reflection (also shown in Figure.3) before proceeding. Specifically, the agent first adjusts its Strategy based on short-term memory, long-term memory, and tool retrieval results. It then updates its long-term memory by integrating the current short-term memory. Finally, the short-term memory is cleared, preparing the agent for the next trial stage.

Detailed mechanisms such as agent initialization and tools-calling are demonstrated in Appendix.H

4 Evaluation Framework and Benchmark

For a comprehensive evaluation for the LLM-based mock trial, we develop an evaluation framework that consists of two parts: judgment prediction evaluation and simulation process evaluation, the procedures of which are shown in the bottom of Figure.1. A benchmark based on this evaluation framework is also constructed.

4.1 Judgment Prediction Evaluation

For judgment prediction, we extract the final judgment made by Judge in the simulation and compare it with the ground truth recorded in the original case document on 3 aspects, i.e., imprisonment, probation, and fine, from two perspectives, i.e., categorical accuracy and quantitative error.

Categorical Accuracy focuses on evaluating whether the judgment produced by SimCourt falls into the correct categories supported by law. For *imprisonment* prediction, inspired by the LJP track in CAIL2018 (Xiao et al., 2018), we examine whether the prediction imprisonment number is within the same acceptable range with the ground truth value from the case document, which we define as a "hit". Specifically, if and only if the predicted imprisonment and the actual imprisonment both fall within the same statutory sentencing range, the prediction is considered to be a "hit". Consequently, hit rate

[2] https://legalone.com.cn/

is defined as the proportion of cases where the predicted sentence hits the correct statutory sentencing interval, which can be calculated as follows:

$$HitRate = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[L_i \leq P_i \leq U_i], \quad (1)$$

where N denotes the number of cases. For the ith case, P_i represents the predicted imprisonment while L_i and U_i denote the lower and upper bounds of the sentencing interval of actual imprisonment, respectively. For *probation* prediction, we treat it as a binary classification task. A prediction is considered correct if and only if both the predicted and the actual judgment agree that probation is applicable, or both agree that it is not applicable. The *Fine* prediction is evaluated in the same manner as the probation.

Quantitative Error evaluates the relative error between the predicted judgment by court simulation and the ground truth from the case document, which provides a different angle to understand the performance of judgment prediction. For example, let the ground truth imprisonment for the ith case be T_i. Then the relative error of predicted imprisonment is computed as

$$RelativeError = \frac{1}{N}\sum_{i=1}^{N} \frac{|P_i - T_i|}{T_i}, \quad (2)$$

The lower the relative error is, the better the imprisonment prediction is. The relative error for probation and fine is computed in a similar way. Note that in the above equation, we explicitly exclude cases where T_i is 0 or where T_i represents a non-numeric sentence (e.g., the death penalty) to avoid numerical problems. This exclusion may introduce a certain bias to the metric. Therefore, a better practice is to jointly consider both the categorical accuracy and quantitative error for the evaluation of different models or systems.

4.2 Trial Process Evaluation

Based on the evaluation guideline for Chinese legal professionals[3], we propose a process evaluation framework that is based on 30 different aspects, as shown in Table.3. Detailed explanations are listed in Table.11. Because the Defendant's and Stenographer's statements, as well as the agents' utterances during the Trial Preparation and Final Statement phases, are relatively standardized, we focus only on three main roles and three most important stages of the court, namely Judge, Prosecutor, Attorney, and Trial Investigation, Evidence Presentation, and Trial Debate, respectively. We also examine the overall performance of Judge, Prosecutor, and Attorney. All these 30 aspects have been carefully checked and refined by legal experts.

In practice, quantifying the actual performance of each role in each court stage is usually difficult even for legal experts. Therefore, we conduct pairwise evaluation instead of pointwise evaluation for better reproducibility. Specifically, when comparing two trial records, annotators should decide whether the roles in the first record outperform, underperform, or perform equally to those in the second record on each of the 30 aspects.

4.3 Benchmark Construction

To ensure the reproducibility of our work and support future studies, we construct and release a mock trial benchmark based on our proposed evaluation framework. For evaluation of judgment prediction, we need the description of the case as input and the actual judgment as the ground truth, which can be automatically extracted from the case documents. Thus, we randomly select 200 cases from LeCaRDv2 (Li et al., 2024a), a large-scale Chinese legal case retrieval dataset that contains 55,192 cases, of the 40 most common charges (5 per charge). To avoid unnecessary confusion in case descriptions, all selected cases are first-instance trials involving a single defendant. We use the Deepseek-v3 model (Liu et al., 2024) to extract the defendant's information, indictment, and evidence as input, and refine details with human legal expert review. The statutory sentencing intervals applicable to the actual sentence are also extracted by LLM and checked by legal expert, based on the relevant legal provision and the imposed sentence.

For process evaluation, in order to better stimulate the simulation ability of the system, the cases provided should contain more detailed indictments, more concise evidence, and even interactions of real human litigants as references. To this end, we select and download 50 open trial videos with the full trial records provided on China Court Trial Online[4] and convert the audio into text (with human proofreading). To ensure high-quality conversational records, the majority of the selected cases are drawn from the nationally recognized "Top 100

[3] http://gongbao.court.gov.cn/Details/ee6a5b1d20140c38c800c91c728d63.html

[4] https://tingshen.court.gov.cn

Method	Imprisonment		Probation		Fine	
	Relative Error↓	Hit Rate↑	Relative Error↓	Accuracy↑	Relative Error↓	Accuracy↑
Llama-3-8b	1.403±1.573**	0.845	0.834±1.298**	0.385	3.774±6.792**	0.655
Qwen-3-8b	0.487±0.570**	0.860	0.438±0.601	0.530	1.410±4.440	0.675
GLM-4-9b	1.617±1.533**	0.740	1.028±0.684**	0.625	3.223±8.217*	0.565
GPT-3.5-turbo	1.071±1.181**	0.850	0.856±0.981**	0.440	1.568±3.856*	0.665
GPT-5.2	0.465±0.645*	0.840	0.447±0.420	0.590	1.350±2.538*	0.740
Deepseek-v3	0.445±0.442*	0.860	0.494±0.566	0.470	1.055±1.562	0.660
ReAct & Tool	0.533±0.632**	0.875	0.440±0.532	0.685	1.415±2.676*	0.730
AgentCourt	0.464±0.513*	0.865	0.493±0.496	0.500	1.275±1.923*	0.670
AgentsCourt	0.602±0.748**	0.875	0.601±0.617*	0.700	1.973±5.286*	0.695
PLJP	0.536±0.734**	0.825	0.523±0.605	0.690	1.194±2.366	0.725
SimCourt	**0.350±0.367**	**0.880**	**0.410±0.435**	**0.730**	**0.770±1.210**	**0.835**

Table 1: Legal judgment prediction results comparing with baselines on imprisonment, probation and fine. Relative error (with standard deviation), hit rate and accuracy relative to the real documents are reported. The best-performing methods are highlighted with boldface. */** denotes that SimCourt performs significantly better than baselines at $p < 0.05/0.01$ level.

Method	Imprisonment		Probation		Fine	
	Relative Error↓	Hit Rate↑	Relative Error↓	Accuracy↑	Relative Error↓	Accuracy↑
SimCourt	**0.350±0.367**	**0.880**	**0.410±0.435**	**0.730**	**0.770±1.210**	**0.835**
w/o Stage 1,2,3,4	0.413±0.450	0.870	0.443±0.526	0.505	0.937±1.344	0.680
w/o Stage 1,2,3	0.449±0.510	0.875	0.467±0.505	0.710	0.788±1.332	0.790
w/o Memory	0.370±0.501	0.850	0.452±0.425	0.695	1.044±2.039	0.830
w/o Strategy	0.424±0.497	0.875	0.426±0.537	0.660	0.860±1.734	0.675

Table 2: Results of ablation study on legal judgment prediction task. Relative error (with standard deviation), hit rate and accuracy relative to the real documents are reported. The best-performing methods are highlighted with boldface. Here, for better describing the ablation settings, the 5 stages are numbered sequentially as 1 to 5.

Outstanding Court Hearings"[5] collections of previous years. The charges of the sampled recordings are also within the same 40-charge set used in the judgment prediction evaluation. The recordings are 57-minutes long and contain 9,782 words on average, and we manually extract the indictment, the defendant's information, and the corresponding evidence to serve as input for court simulation.

Additional details on benchmark construction and specific issues, including privacy concern and data leakage prevention, are provided in the Appendix J.

5 Experiments

5.1 Legal Judgment Prediction

This section presents experiments on legal judgment prediction (LJP) tasks, focusing on the effectiveness of imprisonment, probation, and fine decisions in predicted documents relative to real ones. The temperature in all LLMs is set to 0.7.

Comparison with Baselines. First, we examine the performance of different vanilla models, i.e. Llama-3-8b (Dubey et al., 2024), Qwen-3-8b (Yang et al., 2025), GLM-4-9b (GLM et al., 2024), GPT-3.5-turbo (Brown et al., 2020), GPT-5.2 (Singh et al., 2025), and Deepseek-v3 (Liu et al., 2024), the task of which is directly predicting the judgment based on the case information. Afterwards, we introduce another 3 state-of-the-art baseline methods: the aforementioned AgentCourt (Chen et al., 2024) and AgentsCourt (He et al., 2024) as well as PLJP (Wu et al., 2023), which is a LLM-based LJP method via case retrieval and detailed analysis. We also implemented a common baseline method: ReAct combined with tool calling (Yao et al., 2023). For AgentCourt and AgentsCourt, the debate round is set to 3, which is the same as SimCourt. Besides, for fair comparison, we choose the Deepseek-v3 model as the base model for SimCourt and the 3 baseline methods. We also provide all those baselines with the same tools in SimCourt, i.e. the legal article retriever and LegalOne.

The results are listed in Table.1. SimCourt outperforms all baselines in every aspect and metric, with the lowest error rate and standard deviation, as well as the highest hit rate and accuracy, demonstrating the effectiveness of simulation procedure and courtroom agent design. Besides, we observe

[5] https://www.court.gov.cn/zixun/xiangqing/452001.html

Roles	Stage	Aspect	VS. Real Human			VS. AgentCourt			VS. AgentsCourt		
			Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Judge	Trial Investigation	Clarity of Questioning Structure	**0.64**	0.24	0.12	-	-	-	-	-	-
		Neutrality and Procedural Control	**0.58**	0.20	0.22	-	-	-	-	-	-
		Professional Evidence Examination	**0.66**	0.16	0.18	-	-	-	-	-	-
	Evidence Presentation	Normative Conduct	**0.76**	0.14	0.10	-	-	-	-	-	-
		Cross-Exam Legality Control	**0.62**	0.22	0.16	-	-	-	-	-	-
		Awareness of Fair Trial Safeguards	**0.66**	0.22	0.12	-	-	-	-	-	-
	Trial Debate	Clear Adversarial Framing	**0.88**	0.08	0.04	-	-	-	-	-	-
		Impartial Verbal Interventions	**0.68**	0.18	0.14	-	-	-	-	-	-
		Pace and Order Control	**0.72**	0.10	0.18	-	-	-	-	-	-
	Overall Performance		**0.76**	0.10	0.14	-	-	-	-	-	-
Prosecutor	Trial Investigation	Proper Questioning Strategy	**0.88**	0.08	0.04	-	-	-	-	-	-
		Precise Legal Terminology	**0.92**	0.08	0.00	-	-	-	-	-	-
		Lawful Prosecutorial Questioning	**0.52**	0.26	0.22	-	-	-	-	-	-
	Evidence Presentation	Accuracy in Evidence Presentation	**0.90**	0.10	0.00	-	-	-	-	-	-
		Moderation in Aggressive Advocacy	**0.88**	0.08	0.04	-	-	-	-	-	-
		Proper Response to Objections	**0.96**	0.02	0.02	-	-	-	-	-	-
	Trial Debate	Logical Coherence of Accusation	**0.94**	0.02	0.04	**0.94**	0.06	0.00	**1.00**	0.00	0.00
		Accuracy in Legal Citation	**0.96**	0.00	0.04	**0.50**	0.28	0.22	**0.60**	0.08	0.32
		Effective Defense Rebuttal	**0.92**	0.02	0.06	**0.88**	0.10	0.02	**1.00**	0.00	0.00
	Overall Performance		**0.98**	0.00	0.02	-	-	-	-	-	-
Attorney	Trial Investigation	Relevance and Precision Questioning	**0.84**	0.12	0.04	-	-	-	-	-	-
		Awareness of Fair Trial Safeguards	**0.86**	0.10	0.04	-	-	-	-	-	-
		Awareness of Defendant's Rights	**0.84**	0.10	0.06	-	-	-	-	-	-
	Evidence Presentation	Precision in Challenging Key Issues	**0.92**	0.04	0.04	-	-	-	-	-	-
		Rigor in Evidence Analysis	**0.94**	0.04	0.02	-	-	-	-	-	-
		Effectiveness in Evidence Rebuttal	**0.94**	0.04	0.02	-	-	-	-	-	-
	Trial Debate	Clarity of Defense Arguments	**0.88**	0.02	0.10	**0.48**	0.22	0.30	**1.00**	0.00	0.00
		Logical Rigor in Legal Reasoning	**0.88**	0.04	0.08	**0.38**	0.32	0.30	**1.00**	0.00	0.00
		Balanced Legal and Emotional Appeal	**0.82**	0.10	0.08	**0.58**	0.04	0.38	**0.94**	0.04	0.02
	Overall Performance		**0.96**	0.00	0.04	-	-	-	-	-	-

Table 3: Pairwise evaluation results of trial processes, comparing trial records from SimCourt with those from real human, from AgentCourt and from AgentsCourt. The proportions reflect annotators' preferences over 30 aspects across 50 cases. "Win," "Lose," and "Tie" indicate preference for SimCourt, preference for the compared method, and indistinguishable quality, respectively. "–" indicates comparisons that are not applicable because certain trial stages are not fully implemented in the baseline methods.

that other simplified court simulation methods, i.e., AgentCourt, AgentsCourt, and the non-simulation LJP method, i.e., PLJP and ReAct, fail to consistently outperform vanilla methods across the six metrics. This phenomenon indicates that insufficient court interactions and unprofessional legal agents can instead confuse the judge, leading to greater numerical deviations, further highlighting the effectiveness of our process simulation and agent design. It is also notable that the Deepseek-v3 model performs relative stable among the vanilla models, demonstrating the fundamental capability of the chosen base model.

Ablation Study. To further evaluate the effectiveness of our framework, we conduct ablation study on the trial process and the agent modules, with results shown in Table.2. Specifically, when removing Stage 1 to 4 (only the final statement stage remains), Judge agent will not receive any courtroom record and will make the judgment based solely on the provided tools. In that case, the hit rate and accuracy decrease, while the relative error increases, underscoring the need to simulate judicial proceedings. Particularly, we introduce a setting in which Stage 1,2, and 3 are removed, leaving only the trial debate stage before the final statement. This setting is adapted from the design of AgentCourt and AgentsCourt. We find that performance still experiences a consistent decline, also demonstrating the need to design a complete trial process. Moreover, removing either the Memory module or the Strategy module leads to performance degradation across most metrics, highlighting the necessity of both modules in the system design.

5.2 Trial Process Evaluation

Based on the process evaluation framework and benchmark, we apply the trial process evaluation using human preference annotations. We pairwise compare trial records from SimCourt with those from real human, from AgentCourt and from AgentsCourt, respectively. All the input data for court simulation are derive from those 50 trials as mentioned in the process evaluation benchmark.

Annotation Setup. We invite 3 legal experts to annotate the quality of the simulation process. Each annotator has passed the National Unified Legal Professional Qualification Examination. We provide annotators with the 50 pairs of records from SimCourt and those from the other source, with the order of the two records within each pair randomly shuffled to ensure a blind evaluation setting. For each pair, annotators are asked to indicate their preference across 30 evaluation aspects by selecting which record (the first or the second) is better, or by choosing 'tie' if the two were indistinguishable. The whole guidelines for annotators are presented in Figure.7. Specifically, when compared with AgentCourt and AgentsCourt, since the Trial Investigation stage and the Evidence Presentation stage are not implemented in these frameworks, and their judge does not participate throughout the Trial Debate stage, we ask human annotators to focus solely on the performance of the Prosecutor and the Attorney during the Trial Debate stage.

If annotators judged SimCourt's simulated records as superior or selected "tie", the simulation was considered at least comparable and satisfactory. In this condition, the average Cohen's Kappa among annotators is 0.702, reflecting a high level of agreement and demonstrating the validity and reliability of the annotation.

Afterwards, for each case, we aggregate the 3 annotations on each aspect to derive a final preference. Specifically, we exclude 'tie' annotations and determine the majority preference (i.e., SimCourt or the opposite) for each aspect. If no majority exists, the result is recorded as a 'tie'. Finally, we summarize the result of the evaluation in Table.3.

Results and Analysis. Overall, SimCourt demonstrates strong performance across most evaluation aspects, even relative to human legal professionals. A plausible explanation is that, while the cases are representative, the disputed issues are not overly complex, thereby reducing the difficulty of prosecution and defense. Moreover, LLM-based agents tend to participate more proactively in trial dialogues and, owing to their strong generative capabilities, are able to cover a wider range of prosecutorial and defense issues. This allows LLM-based agents to exhibit relatively strong role-playing abilities. Overall, these findings suggest the potential of LLM-based simulations for courtroom scenarios.

Compared to baseline methods, the extensive presence of "–" symbols clearly reflects the superior completeness of SimCourt's trial procedure design and the preference rate also demonstrates the simulation quality. This advantage is possibly attributable to the more complete framework and broader agent design of SimCourt. We also find that AgentCourt demonstrates stronger competitiveness than AgentsCourt, which may be attributed to its more comprehensive simulation process design.

Specifically, Judges in SimCourt are preferred in most cases, especially in adversarial framing (88%), reflecting effective courtroom guidance. Prosecutor receives strong preference, with 96% in objection handling and over 94% in logical accusation, indicating coherent and reasonable charge construction. Attorney also performs well, particularly during the Evidence Presentation stage, achieving 90% preference across all aspects, highlighting its effectiveness in evidence rebuttal. These strengths benefit from the well-structured simulation process and agent framework.

However, certain limitations remain. For Judge, preference is relatively lower in cross-examination legality control (62%) and neutrality control (58%), possibly due to limited flexibility in managing adversarial exchanges and regulating rhetorical tone in dynamic courtroom settings. Prosecutor also perform less favorably in lawful questioning, with only 52% of cases rated above human performance, suggesting insufficient sensitivity to procedural constraints such as restrictions on leading questions, evidence sequencing, and the protection of defendants' rights. Moreover, both Prosecutor and Attorney exhibit unstable performance in legal citation and reasoning, indicating that even with access to legal tools, prompt-driven agents alone struggle to substantially enhance legal understanding. This points to important directions for future work.

Additional analyses, including a case study and stage-wise and role-wise performance discussions, are presented in Appendices C and D. We also summarize the most frequent reasons for role preferences in Appendix E. We further conduct an ex-

ploratory LLM-based evaluation of the simulated trial records; however, due to its low agreement with human expert annotations, these results are excluded from the primary analysis and instead reported in Appendix F.

6 Conclusion

In this paper, we introduce SimCourt, a court simulation framework based on real Chinese trial procedures. A specially designed agent framework is proposed, helping agents simulate professional practitioners. Evaluation framework and benchmark on both legal judgment prediction and simulation process evaluation are developed. Experiments including process comparison with human legal professionals demonstrate the potential application value of our framework and highlight directions for future improvement.

7 Limitations

1. For now, SimCourt only supports court simulation and evaluation on criminal cases. In the future, we hope to expand the framework for application on civil cases and administrative cases.

2. Compared to real-world trials, the procedure of SimCourt is relatively fixed. Handling other scenarios—such as pretrial conferences or cases with multiple defendants—would require further extensions to the current framework.

3. Currently, while prompt-based approaches and legal knowledge enhancement methods can partially enhance the performance of court agents, they remain limited in their ability to thoroughly understand and flexibly apply legal knowledge, suggesting that further parametric fine-tuning may be necessary in future work.

8 Ethical Considerations

All criminal cases and videos are collected from open-access resources and websites, and the basic information of the defendant is anonymized. All annotators signed informed consent forms and received an average hourly wage of 75 CNY, which is significantly higher than the minimum wage requirements in Beijing.

SimCourt may involve potential risks. For example, SimCourt could be misused to fabricate trial records and spread them maliciously, potentially undermining public trust in real judicial decisions. We therefore encourage future users to comply with relevant laws and regulations, respect real-world judicial rulings, use SimCourt responsibly, and safeguard the privacy and security of all involved parties.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (Grant No. 2024YFC3307101) and the Research Project of Quancheng Laboratory, China (Grant No.QCL20250105).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024. 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.

Hamzeh Abu Issa, Thair Kaddumi, and Naji Alwerikat. 2023. The impact of moot courts on the quality of legal education: Students of the faculty of law at the applied science private university as a model. *Journal of Higher Education Theory and Practice*, 23(19):266–270.

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

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024a. Lecardv2: A large-scale chinese legal case retrieval dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260.

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

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

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and 1 others. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, and 1 others. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, and 1 others. 2024. Gensim: A general social simulation platform with large language model based agents. *arXiv preprint arXiv:2410.04360*.

Lang Tang. 2021. The effective application of" moot court" in law teaching. In *2021 2nd International Conference on Computers, Information Processing and Advanced Education*, pages 240–243.

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

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

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

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

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

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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)*.

Hong Zhang. 2021. Construction of digital moot court teaching practice platform for law major. In *2021 4th International Conference on Information Systems and Computer Aided Education*, pages 189–193.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

A Comparison between SimCourt and Existing Works

Table.7 presents the general comparison between SimCourt and previous work on framework and evaluation design.

For framework design, SimCourt introduces and realizes all the 5 stages in criminal court, while AgentCourt and AgentsCourt contain only a few simplified stages. Specifically, in AgentsCourt, there are only 3 rounds of debate carried by Prosecutor and Attorney, while Judge does not participate in the discussion. In AgentCourt, Judge has no option to make guidance during Trial Debate

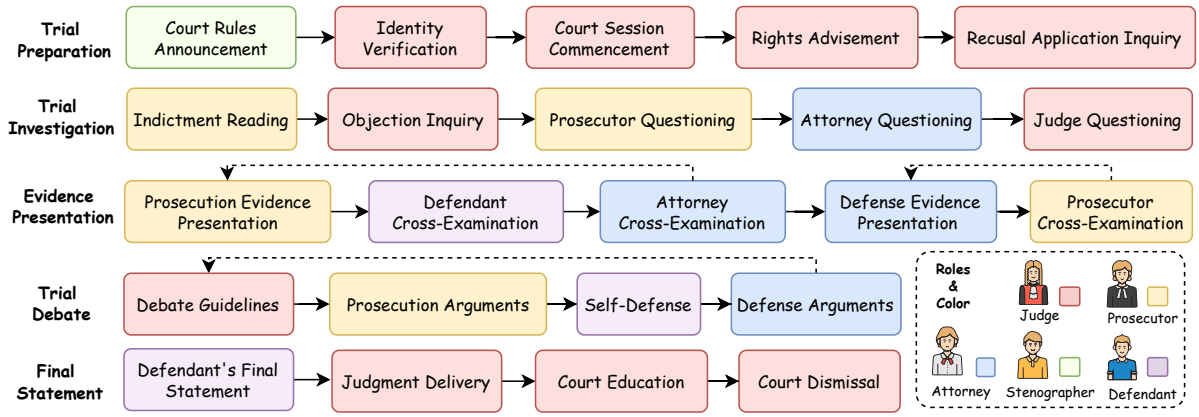

Figure 4: The flowchart of the trial procedure in SimCourt, which is grounded on real trial procedures. Colors of the boxes represent the roles that lead the respective phase. The correspondence between roles and colors can be found at the bottom right corner of the chart. In the diagram, the dashed arrows represent optional backtracking.

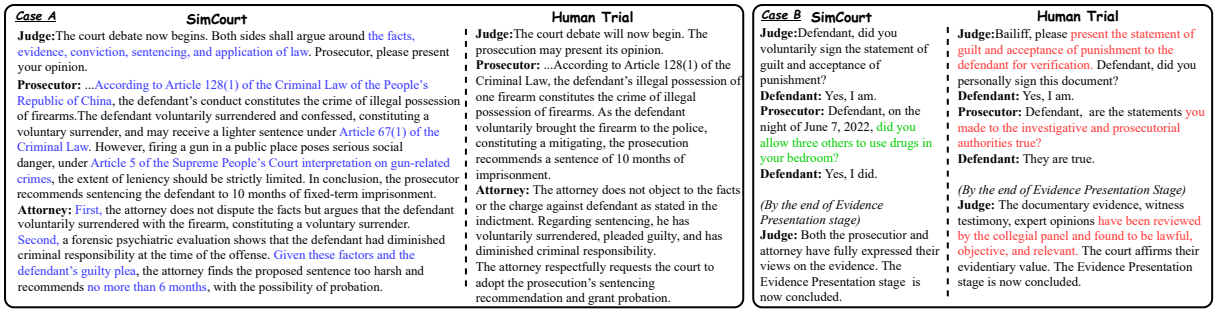

Figure 5: Records comparison between SimCourt and human trial. Case A showcases typical strengths of SimCourt, with effective expressions highlighted in blue. Case B illustrates typical weaknesses, where inappropriate statements are highlighted in green, and well-performed statements from the human trial are highlighted in red.

stage and Defendant also has no option to make final statement during Final Statement stage. Both AgentCourt and AentsCourt completely ignore the implementation of Trial Investigation stage and Evidence Presentation stage, which are crucial stages for investigating the truth. The absence of these components substantially degrades the quality of the generated simulated trial records and, in turn, adversely affects the accuracy of judgment prediction.

For evaluation, AgentsCourt applies the analysis of legal judgment prediction but completely ignores the evaluation of the simulation process. AgentCourt does not introduce the judgment prediction evaluation nor pay enough attention to process evaluation. In detail, AgentCourt introduces three coarse-grained dimensions, i.e., Cognitive Agility, Professional Knowledge, and Logical Rigor. However, in a comprehensive court simulation framework, court agents each have a specific task in each stage, so a fine-grained evaluation methods specifically designed for each role and each stage can better evaluate the performance of the courtroom agents. For this purpose, SimCourt introduces an evaluation framework that covers 30 different aspects, focusing on 3 main roles (Judge, Prosecutor, and Advocate) and 3 main stages (Trial Investigation stage, Evidence Presentation stage, and Trial Debate stage).

B Overall Procedure of SimCourt

Figure.4 presents the detailed flowchart of SimCourt. Colors of the boxes represent the roles that lead the respective phase. Compared to the former framework, SimCourt is designed with an extra Trial Investigation stage, an extra Evidence Presentation stage, and a more comprehensive procedure of the other 3 stages, i.e. Trial Preparation stage, Trial Debate stage and Final Statement stage. Experiments on both judgment prediction (including comparison among baselines and ablation study) and process evaluation illustrate the effectiveness of the simulation in these stages.

C Case Study

To directly present the strengths and weaknesses of the simulation process, we selected two typical cases, as shown in Figure.5. Case A shows typical strengths of SimCourt while Case B demonstrates typical weaknesses.

In *Case A*, the simulated judge emphasized that the debate should focus on key issues such as sentencing, offering a clearer adversarial framing than the human judge. The simulated prosecutor also demonstrated stronger legal grounding by citing multiple legal articles with detailed explanations, whereas the human prosecutor referenced only one article with minimal elaboration. Additionally, the SimCourt attorney presented a logical and well-reasoned defense, outperforming the human attorney who made a relative simpler defense statement with no sentence recommendation.

In *Case B*, the human judge instructed the bailiff to return the statement to the defendant for verification and formally acknowledged the evidentiary validity, demonstrating proper legal procedure. In contrast, the simulated judge failed to recognize or apply basic principles of legality control. Furthermore, the human prosecutor questioned the defendant regarding the authenticity of the statements, whereas the simulated prosecutor posed leading questions, resulting in legally flawed prosecutorial conduct. Hopefully, through training from real courtroom records, simulated agents may further improve their legal knowledge and procedural awareness, helping to mitigate current limitations.

Complete generated trials of SimCourt can be found at the github pages[6].

D Stage-wise and Role-wise Process Performance

We calculated the average preference proportion of SimCourt compared with real human trials over each stage and each role, as shown in Figure.6.

For role-wise performances, it can be found that Prosecutor and Attorney shows similar performance (about 88.0% preference), likely due to their logical statement and active rebuttal. However, Judge shows the most modest preference (68.9%), which partially lies in the relative poor ability in fairness and legitimacy control. For stage-wise performances, the proportion of preference in the Trial Investigation stage is relatively low (74.9%), which also derive from the unawareness of legality issues. Generally, future refinement can focus on the performance of Judge agent and the performance on Trial Investigation stage.

[6] https://github.com/Miracle-2001/SimCourt/tree/main/experiments/process_evaluation/Trials/SimCourt

E Most Frequent Reasons for Role Preferences of SimCourt

To intuitively investigate why SimCourt receives higher preference rates, we analyze the high-frequency reasons cited when SimCourt roles are preferred.

Specifically, for the three roles in SimCourt—Judge, Prosecutor, and Attorney—we collect all annotated reasons associated with preference judgments. We then perform a frequency-based analysis using tokenization and ranking. To better capture preference-indicative expressions, we focus on bigrams and further extract those that explicitly convey positive preference. For each role, we retain the top three most frequently occurring phrases.

We find that annotators more frequently use phrases such as "procedural compliance," "adequate protection," and "issue guidance" to express preference for the presiding judge; phrases such as "clear logic," "targeted arguments," and "comprehensive multi-perspective coverage" to express preference for the prosecutor; and phrases such as "protection of party rights," "rigorous reasoning," and "evidence analysis" to express preference for the defense attorney. These observations indicate that agents in SimCourt exhibit notable strengths in guiding disputed issues, articulating logical arguments, and proactively safeguarding the rights and interests of litigants.

Notably, these findings are largely consistent with the conclusions from the process-level evaluation, providing complementary evidence for the stability of the system and the reliability of the experimental results.

F Process Evaluation by LLM Evaluators

In addition to human annotators, we also conduct an exploratory evaluation using LLM annotators to assess the simulation process, i.e. the LLM-as-Judge method. This experiment was intended to examine whether LLM-based judgments could serve as a scalable proxy for human expert evaluation in this legally specialized setting.

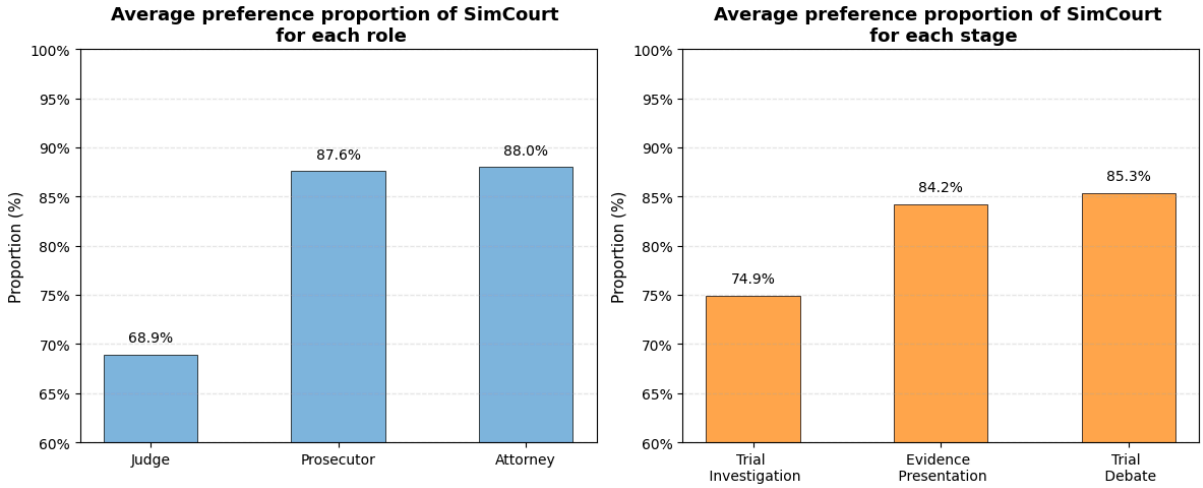

Figure 6: The left part presents the average preference proportion of SimCourt of each stage. The right part presents the average preference proportion of SimCourt of each role.

Roles	Stage	Aspect	VS. Real Human			VS. AgentCourt			VS. AgentsCourt		
			Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Judge	Trial Investigation	Clarity of Questioning Structure	**0.86**	0.02	0.12	-	-	-	-	-	-
		Neutrality and Procedural Control	**0.86**	0.00	0.14	-	-	-	-	-	-
		Professional Evidence Examination	**0.72**	0.18	0.10	-	-	-	-	-	-
	Evidence Presentation	Normative Conduct	**0.82**	0.04	0.14	-	-	-	-	-	-
		Cross-Exam Legality Control	**0.92**	0.04	0.04	-	-	-	-	-	-
		Awareness of Fair Trial Safeguards	**0.90**	0.04	0.06	-	-	-	-	-	-
	Trial Debate	Clear Adversarial Framing	**0.94**	0.02	0.04	-	-	-	-	-	-
		Impartial Verbal Interventions	**0.64**	0.28	0.08	-	-	-	-	-	-
		Pace and Order Control	**0.70**	0.10	0.20	-	-	-	-	-	-
	Overall Performance		**0.60**	0.00	0.40	-	-	-	-	-	-
Prosecutor	Trial Investigation	Proper Questioning Strategy	**0.92**	0.02	0.06	-	-	-	-	-	-
		Precise Legal Terminology	**0.92**	0.02	0.06	-	-	-	-	-	-
		Lawful Prosecutorial Questioning	**0.92**	0.06	0.02	-	-	-	-	-	-
	Evidence Presentation	Accuracy in Evidence Presentation	0.28	**0.68**	0.04	-	-	-	-	-	-
		Moderation in Aggressive Advocacy	0.36	**0.60**	0.04	-	-	-	-	-	-
		Proper Response to Objections	**0.90**	0.06	0.04	-	-	-	-	-	-
	Trial Debate	Logical Coherence of Accusation	**0.92**	0.00	0.08	**0.88**	0.00	0.12	**0.96**	0.00	0.04
		Accuracy in Legal Citation	**0.98**	0.02	0.00	**0.88**	0.00	0.12	**1.00**	0.00	0.00
		Effective Defense Rebuttal	**1.00**	0.00	0.00	**0.92**	0.00	0.08	**0.88**	0.00	0.12
	Overall Performance		**0.82**	0.04	0.14	-	-	-	-	-	-
Attorney	Trial Investigation	Relevance and Precision Questioning	**0.88**	0.06	0.06	-	-	-	-	-	-
		Awareness of Fair Trial Safeguards	0.20	**0.78**	0.02	-	-	-	-	-	-
		Awareness of Defendant's Rights	**0.74**	0.20	0.06	-	-	-	-	-	-
	Evidence Presentation	Precision in Challenging Key Issues	**0.90**	0.06	0.04	-	-	-	-	-	-
		Rigor in Evidence Analysis	**0.90**	0.06	0.04	-	-	-	-	-	-
		Effectiveness in Evidence Rebuttal	**0.90**	0.06	0.04	-	-	-	-	-	-
	Trial Debate	Clarity of Defense Arguments	**0.92**	0.00	0.08	**0.84**	0.00	0.16	**0.98**	0.00	0.02
		Logical Rigor in Legal Reasoning	**0.94**	0.00	0.06	**0.88**	0.00	0.12	**1.00**	0.00	0.00
		Balanced Legal and Emotional Appeal	**0.44**	0.16	0.40	**0.84**	0.12	0.04	**0.48**	0.04	**0.48**
	Overall Performance		**0.78**	0.16	0.06	-	-	-	-	-	-

Table 4: Pairwise evaluation results of trial processes by the LLM annotator, comparing trial records from SimCourt with those from real human, from AgentCourt and from AgentsCourt. The proportions reflect the annotator's preferences over 30 aspects across 50 cases. "Win," "Lose," and "Tie" indicate preference for SimCourt, preference for the compared method, and indistinguishable quality, respectively. "-" indicates comparisons that are not applicable because certain trial stages are not fully implemented in the baseline methods.

Specifically, we build a prompt-driven LLM annotator based on the Deepseek-V3 model and ask it to annotate the preference of the two given trial records on those evaluation aspects. Detailed explanations of each aspect are provided to the annotator. To avoid position bias, we randomly shuffle each pair as the settings in human annotations. Figure.8 presents the annotation guidelines for the LLM annotator and Table.4 presents the annotation results.

According to those tables, the LLM annotator also generally prefers the trial records generated by SimCourt, which is the same as the case for expert human annotators. However, we observe that the agreement between LLM-based scores and human expert annotations is low, **with an average Cohen's Kappa coefficient of approximately 0.3, indicating only a slight agreement;** This substantial discrepancy suggests that LLM cannot reliably align with expert judgments on nuanced legal and procedural criteria.

We attribute this gap to the inherently normative and context-sensitive nature of courtroom evaluation, which requires a fine-grained understanding of legal standards, procedural legitimacy, and role-specific responsibilities. Although the LLM demonstrates strong surface-level coherence and human-like generation abilities, it still struggles to consistently capture expert-level legal reasoning and evaluative standards, even when guided by detailed prompts and structured rubrics.

Consequently, we do not rely on the LLM-based evaluation for our primary conclusions. The results are reported in this appendix solely for completeness and transparency, highlighting both the potential and current limitations of LLM-as-a-judge approaches in complex legal evaluation tasks.

G Task Descriptions in the Profile Module

This section presents task descriptions in the profile module of each agent. Table.8 presents the basic tasks and instructions of each agent, which is fixed during the court trial. Table.9 presents the stage-specific tasks settings, which represents as a supplementary description of tasks on the specific stage.

H Agent Initialization and Tools Calling

Before the commencement of the court, each agent receive their basic courtroom tasks and fill their Profile module, which would be kept thorough the whole trial process. Subsequently, each agent initialize its own strategies. Prosecutor and Attonery initialize their Attack strategy and Defense strategy, while Defendant designs his Defense strategy and Judge formulates an Investigation strategy as guidance to investigate the details of the case. Specifically, we let each agent generate queries related to the application of law based on the case information, and then submit them to LegalOne to retrieve similar cases. We then let agents compile a list of relevant legal article titles and obtain their full content through the legal article retriever. The final Strategy module of an agent is built with the case information, its courtroom tasks, the retrieved legal provisions and similar cases, and the strategy types (i.e., Attack, Defense, and Investigation) assigned to its role.

Detailed prompts templates and an exemplar of tool calling is further illustrated in Appendix L. Other prompts templates are demonstrated in Appendix M and Appendix N.

I Details of Process Evaluation

The illustrations of each aspect is presented in Table.11. These illustrations are also provided to annotators to make sure they fully understand all the aspects.

J Details of Benchmark Construction and Specific Claims

To better reconstruct the case circumstances, in addition to extract the basic information of the case (i.e., Defendant's information, indictment and evidence), we also extract the opinions and requests of defendant if they are mentioned in the original judgment documents and add them to 'Defendant's testimony' in the evidence materials. All the information are extracted by the Deepseek-V3 model and checked by human experts.

For privacy concern, we claim that all the cases are derived from open source datasets and open websites, and all private information, including names, identification numbers, and residential addresses, has been either replaced or anonymized.

For possible data leakage and unfair comparison risks, we carefully check each of the cases to make sure that:

1. None of the selected cases is included in the corpus of LegalOne, which means that no model can directly get the answers from retrieved case documents.

2. The selected cases are drawn from the same

time period (within five years), and all involved legal provisions are consistent and not outdated.

3. For the 50 cases used in the process evaluation, all evidence examined in court in the original trial recordings was presented in paper form, with no multimedia materials played during the proceedings. Potential issues related to speech-to-text transcription errors or omitted evidence information are not applicable.

K Annotation Guidelines for Human Annotators

The guidelines provided to the human annotators are presented in Figure.7

L Prompts Templates of Tool Calling

Tools including a legal article retriever and a case retriever named LegalOne are provided. During the initialization, reflection and final judgment, agents may generate queries to these tools for legal knowledge and advice. During the tool-calling, agents first define wether use the tools and then generate queries. The template for tool calling is shown in Figure.9, and a detailed exemplar is presented in Figure.10.

M Prompts Templates for Final Judgment

For an accurate judgment, we provide Judge agent with all the information collected before and during the court trial. A specific designed prompt template are also provided to allow Judge generating a professional judgment document. Note that Judgment may also launch tool callings, seeking for references before final judgment. Prompts templates for final judgment is shown in Figure.15.

N Other Prompts Templates for Agents

Prompts templates for initialization, reflection, and action of agents are listed in this section. All the prompts are designed and checked by legal experts.

For initialization, agents may first generate a tool-calling queries, the prompts of which is illustrated in the last section. Afterwards, the responses made by the tools are performed as 'Reference Materials' and are provided in the initialization prompts. During the initialization, different agents are provided with different instructions, which are shown in Figure.11.

For reflection, the prompts templates are divide into 2 parts: reflection prompts on planning module and reflection prompts on memory module. The templates for planning module reflection is shown in Figure.12 where different agents may be provided with different instructions. Note that agents may also choose to call the tools for legal knowledge enhancement before updating their strategies. The templates for memory module reflection for all the agents is shown in Figure.13, and the summary will be then added into the long-term memory.

For action, agents will utilize all three modules (profile module, memory module, planning module), along with current instructions. The prompts templates of action is shown in Figure.14.

O Significant Testing of the Simulation Process Evaluation

We adopt a two-sided binomial test. In the 50 paired trial records used for the process-level evaluation, let n_S denote the number of pairs where SimCourt is preferred, n_H the number where Human is preferred, and n_T the number of ties, with $n_S + n_H + n_T = 50$.

Under the null hypothesis that annotators have no systematic preference, the probability of preferring SimCourt or Human among non-tied pairs should each be 0.5. Therefore, taking the number of non-tied pairs $n = n_S + n_H$ as the effective sample size, we compute the two-sided probability of observing n_S or a more extreme outcome:

$$p = 2 \times \sum_{k=n_S}^{n} \binom{n}{k} (0.5)^n \quad (\text{if } n_S > n/2), \tag{3}$$

If $p < 0.01$, the difference in preferences is considered statistically significant.

We note that the number of ties reported in Table 3 is relatively small, which makes this testing procedure applicable. Following this approach, we computed the binomial significance test for all 30 evaluation aspects in Table.3.

We observe that the average p-value of all the 30 evaluation aspects is below 0.01, showing a statistically significant preference in favor of SimCourt. This indicates that annotators' preference for the simulated trial results generated by SimCourt is statistically significant, and supports the reliability of our experimental conclusions.

P Potential Distribution Shift between Judgment Prediction and Process Evaluation

The use of two separate datasets is inherent to the nature of the tasks. The judgment prediction task requires large volumes of data including indictments, evidence, and final judgments, which can be extracted from the LeCaRDv2 dataset. The process evaluation task, however, requires complete trial transcripts and detailed evidence materials, which are only available from China Court Trial Online. Many of these cases were not adjudicated on the day of trial, making their judgment outcomes unavailable and thus unsuitable for judgment prediction.

Generally, the two datasets necessarily serve different objectives — one focused on outcome, the other on process — and this divergence is precisely the span we intended to cover in designing the simulated courtroom system. The distributional difference between the two datasets is therefore both expected and justifiable.

We now provide a more detailed analysis of the distribution shift between the two datasets. First, both datasets draw exclusively from the 40 most common charge types, and both consist solely of first-instance cases with a single defendant, which provides a macro-level guarantee against excessive distributional divergence. We further computed the mean, standard deviation, and median of both datasets across the indictment and evidence dimensions, and applied the Mann-Whitney U test for significance testing. In addition, we extracted text embeddings using Qwen3-32B, computed the Maximum Mean Discrepancy (MMD) with an RBF kernel, and assessed significance via permutation testing. We found that (PED refers to the Process Evaluation Dataset, and JPD refers to the Judgment Prediction Dataset):

Regarding indictments: character count and sentence count show no significant difference ($p > 0.05$), indicating comparable overall length. However, the number of legal provisions cited differs significantly ($p = 0.0026$), with PED averaging slightly more citations (7.0) than JPD (5.3). The MMD test further reveals a significant difference in semantic distribution ($p = 0.0099$), indicating divergence in high-dimensional semantic space.

Regarding evidence: both character count and sentence count show highly significant differences ($p < 0.001$), with PED containing substantially longer evidence texts (mean 3,255 characters, 71 sentences) compared to JPD (mean 1,621 characters, 42 sentences). The number of legal provisions cited is low in both datasets with a median of 0, showing no significant difference ($p = 0.8624$). The MMD test is again significant ($p = 0.0099$), indicating meaningful semantic divergence in evidence content as well.

In summary, under the shared macro-level constraints of drawing from the 40 most common charges, first-instance cases, and single-defendant cases, PED involves more complex legal citations and more extensive evidence, which poses greater challenges for the simulated system on the process evaluation task. We argue that this distribution shift is fully consistent with — and indeed expected by — the design intent of the process evaluation component.

Hallucination Type	Total Section	Hallucination Section	Hallucination Rate
Case-level	49851	54	0.11%
Legal-level	150	2	1.33%

Table 5: Legal Hallucination rate among case-level and legal-level hallucinations.

Q Legal Hallucinations Error Analysis

We would like to clarify how SimCourt currently handles hallucinations. During the generation of each statement, the system first produces a draft, then performs an internal self-check to detect any fabricated legal provisions or case facts. If any are identified, the draft is revised before the final statement is produced.

Following your suggestion, we define "Legal Hallucinations" in two categories: **case-level hallucinations** and **law-level hallucinations**.

Case-level hallucinations refer to errors where the agent fabricates case details or facts. Law-level hallucinations refer to errors involving fabricated legal provisions or invented sub-clauses.

Based on this definition, we conducted the following additional analysis:

For case-level hallucinations, we used large language models combined with manual verification to inspect every agent utterance in both the judgment prediction and process evaluation tasks. All fabricated case facts were annotated. The dataset includes a total of 49,851 utterances.

For law-level hallucinations, we examined the annotation rationales provided by three human an-

notators for the "Accuracy of Legal Citation" criterion. If any rationale mentioned that the SimCourt-generated transcript contained fabricated laws or sub-clauses, it was marked as a law-level hallucination. A total of 150 evaluations regarding legal citations were analyzed. The results of this hallucination analysis are summarized in Table.5.

Overall, we observe that the incidence of both case-level and law-level hallucinations is relatively low, indicating that SimCourt demonstrates a notable ability to mitigate hallucinations.

We note that the rate of law-level hallucinations is higher than that of case-level hallucinations. A possible explanation is that case facts are inherently more objective, making them harder for the model to confuse or fabricate. In contrast, while legal provisions themselves are objective, their proper application requires deeper understanding. When the model cannot reliably link retrieved statutes to its utterance strategy, it may generate fabricated legal provisions.

R Computational cost and Latency

Regarding the token consumption and simulation duration of SimCourt, we conducted the following supplementary experiments:

We randomly selected 20 cases from the dataset and tested them using the backbone models reported in the paper, Deepseek-v3 and Qwen3-4B. For each case, we measured the token usage, the average number of dialogue turns, the average simulation duration, and the maximum token consumption per single model call. The results are summarized in Table.6:

From the above cost analysis, we observe that the average processing time exceeds 20 minutes, with token usage approaching 500,000, and a single instance consuming over 20,000 tokens. This indicates that prompt engineering indeed poses significant challenges to the model's contextual capacity. Future improvements will focus on maintaining efficiency while reducing token and time costs. We plan to explore parameterized training approaches, such as supervised fine-tuning or reinforcement learning, to reduce reliance on prompts.

We also note that using Deepseek-v3 as the backbone incurs higher costs than Qwen-3-4B. This is likely because Deepseek-v3 has stronger generative capabilities and produces more detailed discussions per case, resulting in more dialogue rounds and higher token consumption. Additionally, API calls are more time-consuming than using a locally deployed SGlang server.

It is worth emphasizing that even with a single simulation costing $0.16 and taking 26 minutes, it can generate a high-quality trial simulation record, which remains acceptable. After streamlining redundant steps, this approach still holds potential for practical commercial deployment.

Backbone Model	Driver	Conversation Rounds	Elapsed Time (mins)	Token Cost	Money Cost ($)	Max tokens per turn
Deepseek-v3	API	188.30	26.78	481156	0.161	21288
Qwen3-4B	SGlang	177.85	21.95	459770	NA	22353

Table 6: Computational cost and latency of SimCourt.

Framework & Evaluation	SimCourt (ours)	AgentCourt	AgentsCourt
Trial Preparation Stage	✓	✓	✗
Trial Investigation Sage	✓	✗	✗
Evidence Presentation Stage	✓	✗	✗
Trial Debate Stage	✓	△	△
Final Statement Stage	✓	△	✗
Legal Judgement Prediction	✓	✗	✓
Aspects of Process Evaluation	30	3	0

Table 7: Comparison between SimCourt and former court simulation framework. ✓ denotes full implementation, △ denotes partial implementation, and ✗ denotes no implementation.

Role	Basic Courtroom Task
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.
Prosecutior	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.
Attorney	You are an experienced attorney. 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. Your goal is to protect your own interests within the scope of the information you know. At the same time, you are required to cooperate with the court investigation and answer questions truthfully.
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 8: Basic courtroom tasks of each agent, which are stored in the profile module. The content is designed after the Criminal Procedure Law of the People's Republic of China.

Stage	Role	Stage-Specific Task
Trial Preparation	Judge	This is the preparation stage. As the judge, your task is to verify the defendant's identity, ask whether the defendant requests any recusals, and inform the defendant of their legal rights.
	Prosecutor	This is the preparation stage. As the prosecutor, you are not required to speak during this stage.
	Defendant	This is the preparation stage. As the defendant, your task is to respond to the judge's inquiries.
	Stenographer	This is the preparation stage. As the Stenographer, your task is to call in the relevant parties and announce the court rules.
	Attorney	This is the preparation stage. As the attorney, you are not required to speak during this stage.
Court Investigation	Judge	This is the court investigation stage. As the judge, your task is to ask the defendant whether they object to the indictment.
	Prosecutor	This is the court investigation stage. As the prosecutor, your task is to read the indictment aloud and question the defendant.
	Defendant	This is the court investigation stage. As the defendant, your task is to answer questions from the judge and the prosecutor.
	Stenographer	This is the court investigation stage. As the Stenographer, you are not required to speak during this stage.
	Attorney	This is the court investigation stage. As the attorney, your task is to question the defendant.
Evidence Presentation	Judge	This is the evidence presentation stage. As the judge, your task is to ask the defendant for their opinion on each piece of evidence.
	Prosecutor	This is the evidence presentation stage. As the prosecutor, your task is to read each piece of evidence in turn.
	Defendant	This is the evidence presentation stage. As the defendant, your task is to express your opinion on each piece of evidence.
	Stenographer	This is the evidence presentation stage. As the Stenographer, you are not required to speak during this stage.
	Attorney	This is the evidence presentation stage. As the attorney, your task is to express your opinion on each piece of evidence.
Court Debate	Judge	This is the court debate stage. As the judge, your task is to present each debate topic in order and decide whether to proceed to the next one.
	Prosecutor	This is the court debate stage. As the prosecutor, your task is to present your opinion on each debate topic. Focus on sentencing and conviction, respond to prior arguments logically, stay focused, and avoid trivial issues. When citing laws, recite the articles explicitly.
	Defendant	This is the court debate stage. As the defendant, your attorney will speak on your behalf. If you are questioned, you must answer truthfully.
	Stenographer	This is the court debate stage. As the Stenographer, you are not required to speak during this stage.
	Attorney	This is the court debate stage. As the attorney, your task is to defend the defendant on each debate topic. Focus on sentencing and conviction, respond to prior accusations logically, stay focused, and avoid trivial issues. When citing laws, recite the articles explicitly.
Final Statement	Judge	This is the defendant's final statement stage. As the judge, your task is to listen to the defendant's statement and make a fair judgment based on the full court record, evidence, indictment, and defendant's background.
	Prosecutor	This is the defendant's final statement stage. As the prosecutor, you are not required to speak during this stage.
	Defendant	This is the defendant's final statement stage. As the defendant, your task is to make your final statement.
	Stenographer	This is the defendant's final statement stage. As the Stenographer, your task is to arrange the departure of relevant parties.
	Attorney	This is the defendant's final statement stage. As the attorney, you are not required to speak during this stage.

Table 9: The stage-specific tasks of each stage for each agent, which are stored in the profile module. The content is designed after the Criminal Procedure Law of the People's Republic of China.

No.	Charge	No.	Charge
1	Dangerous Driving	21	Gathering to Fight
2	Theft	22	Bribery
3	Intentional Injury	23	Extortion
4	Traffic Accident Causing Injury or Death	24	Illegal Absorption of Public Deposits
5	Drug Smuggling, Trafficking, Transporting, or Manufacturing	25	Embezzlement
6	Fraud	26	Inducing, Sheltering, or Introducing Prostitution
7	Provoking Troubles	27	Intentional Homicide
8	Allowing Others to Take Drugs	28	Snatching
9	Operating Gambling Establishments	29	Occupational Embezzlement
10	Obstructing Official Duties	30	Illegal Manufacturing, Trading, Transporting, Mailing, or Storing of Firearms, Ammunition, or Explosives
11	Robbery	31	Rape
12	Illegal Possession or Concealment of Firearms or Ammunition	32	Illegal Occupation of Farmland
13	Illegal Detention	33	Refusing to Enforce Court Judgments or Rulings
14	Illegal Possession of Drugs	34	Falsely Issuing Special VAT Invoices or for Tax Fraud
15	Credit Card Fraud	35	Producing, Selling, or Providing Counterfeit Drugs
16	Gambling	36	Negligent Homicide
17	Illegal Logging	37	Forging, Altering, or Trading in Official Documents, Certificates, or Seals
18	Intentional Destruction of Property	38	Producing or Selling Toxic or Harmful Food
19	Illegal Business Operations	39	Offering Bribes
20	Contract Fraud	40	Illegal Fishing

Table 10: List of the most common 40 Criminal Offenses

Annotation guidelines for human annotators.

Task Overview
You will be given **50 pairs of trial records**, with **two records per case**. For each pair, conduct a **pairwise preference comparison** across multiple aspects related to **courtroom roles** and **trial stages**.
You do **not** need to assign numerical scores. For each aspect, indicate **which record performs better** and provide a **brief reason**.

Data Description
- Each trial record is approximately **10,000 words**.
- Each record involves **five roles**: Court Clerk, Judge, Prosecutor, Defense Attorney, and Defendant.
- The trial procedure includes five stages: Court Preparation, Trial Investigation, Evidence Presentation and Examination, Trial Debate, and Defendant's Statement.

Evaluation Focus
- **Roles**: Judge, Prosecutor, Defense Attorney
- **Stages**: Trial Investigation, Evidence Presentation and Examination, Trial Debate
Each role may have **multiple evaluation aspects** within each stage. Please provide a **preference judgment** for each aspect, as well as an **overall preference** for each role. Detailed definitions of the **30 evaluation dimensions** are provided in the attached table.

Annotation Labels
- **A**: the first record performs better
- **B**: the second record performs better
- **C**: indistinguishable

Notes
- Ignore minor stylistic issues (e.g., repetitions or verbal habits).
- Focus on substantive content rather than surface style.
- Apply **consistent evaluation standards** across all cases.

Figure 7: Annotation guidelines for human annotators

Evaluation Aspects	Illustrations
Clarity of Questioning Structure	Whether the judge questioning in a clear, logical, and progressive manner.
Neutrality and Procedural Control	Whether the judge maintains impartiality and ensures a smooth procedural flow.
Professional Evidence Examination	Whether the judge handles evidentiary issues with legal expertise
Reasonableness of Interrogation	Whether the prosecutor uses a legally sound and goal-oriented strategy in questioning.
Focus and Professional Language	Whether the prosecutor employs accurate and professional legal language.
Evidence Introduction Legitimacy	Whether the prosecutor's questioning complies with evidentiary and procedural rules.
Specificity of Questioning	Whether the attorney asks focused and legally relevant questions.
Sensitivity to Legal Procedure	Whether the attorney demonstrates sensitivity to procedural legality and judicial norms.
Defendant's Rights Protection	Whether the attorney actively safeguards and invokes the defendant's rights.
Normative Conduct	Whether the judge moderates proceedings according to legal standards and decorum.
Legality Control of Cross-Exam	Whether the judge ensures cross-examinations comply with legal and evidentiary rules.
Awareness of Fair Trial Safeguards	Whether the judge upholds fairness and equality between prosecution and defense.
Accuracy in Evidence Presentation	Whether the prosecutor presents evidence clearly, accurately, and without distortion.
Appropriateness of Aggressiveness	Whether the prosecutor maintains persuasion without undue hostility or improper pressure.
Proper Response to Objections	Whether the prosecutor responds to objections with legal and procedural propriety.
Relevance in Raising Challenges	Whether the attorney identifies and questions the core issues in the prosecution's case.
Rigor in Evidence Analysis	Whether the attorney provides logically structured and thorough analysis of the evidence.
Effectiveness of Responding	Whether the attorney persuasively challenges or neutralizes the prosecution's evidence.
Clarity in Guiding Debate Focus	Whether the judge appropriately identifies and frames the points of legal contention.
Impartial Verbal Interventions	Whether the judge's verbal involvement maintains neutrality and procedural fairness.
Pace and Order Control	Whether the judge effectively manages the tempo and discipline of courtroom proceedings.
Logical Coherence of Accusation	Whether the prosecutorial argument is internally consistent and legally structured.
Accuracy in Legal Citation	Whether legal authorities are cited correctly and relevantly to support the argument.
Effectiveness of Responding	Whether the prosecutor addresses and counters the defense's claims with clarity and force.
Clarity of Defense Arguments	Whether the attorney arguments are clearly articulated and logically developed.
Logical Rigor in Legal Reasoning	Whether legal reasoning is precise, internally consistent, and legally sound.
Human Reasoning Expression	Whether the argument balances doctrinal reasoning with appropriate emotional resonance.
Overall Performance of Judge	The overall performance of Judge
Overall Performance of Prosecutor	The overall performance of Prosecutor
Overall Performance of Attorney	The overall performance of Attorney

Table 11: Detailed illustrations of the 30 process evaluation aspects. These illustrations are also provided to annotators to make sure they fully understand all the aspects.

Annotation guidelines for the LLM annotator

You are a professional courtroom performance evaluation expert. You are now required to conduct a comparative evaluation of the performance of the **{role}** during the **{stage_name}** stage based on two trial transcripts.

Description of the responsibilities of {role} during the {stage_name} stage:

{role_description}

Evaluation requirements:

1. Please ignore verbal habits such as filler words or limited local repetitions, and focus primarily on the substantive content.
2. Conduct evaluations based on specific aspects; avoid high-level or overly general assessments.
3. For each evaluation dimension, carefully compare the two transcripts and determine which one demonstrates better performance.
4. Explanation of evaluation outcomes:
   - If Transcript A (the first transcript) performs better, output: **A**
   - If Transcript B (the second transcript) performs better, output: **B**
   - If the performance cannot be distinguished or the two are equivalent, output: **C**
5. If one transcript does not exhibit the relevant content while the other does, the latter should be considered superior.
6. If neither transcript exhibits the relevant content, output **C**.
7. For each evaluation dimension, provide a detailed justification with a moderate level of elaboration.

Contents of the {stage_name} stage from the two trial transcripts:

===== Transcript A =====
{stage_content_a}

===== Transcript B =====
{stage_content_b}

Evaluation dimensions and detailed descriptions
Please evaluate the following **{len(items)}** dimensions individually:
{items_description}

Output format requirements:
Please strictly follow the JSON format below and do not include any additional text:

```
{"{item_name}": {"option": "A/B/C", "text": "Brief justification"}}
```

Please output the JSON directly without any additional explanation.

Figure 8: Annotation guidelines for the LLM annotator.

Prompt 1: Prompts templates of tool calling for agents.

Now, in order to better complete your task, two tools are available for you to use:
1. Legal Article Retriever: This tool contains all laws of the People's Republic of China, including the Criminal Law, Civil Code, Civil Procedure Law, etc. It returns the full content of legal provisions based on your query.
2. Legal Analysis Assistant ("LegalOne"): By inputting legal questions, facts, points of dispute, or case information, this tool provides relevant case references and expert legal analysis.

You should decide whether to use these two tools based on your task objective and the current information. Also, determine what specific queries you want to ask. You may submit multiple queries at once by placing them in a list. When using the Legal Article Retriever, the full name of the law must be provided—no abbreviations allowed. For example, you must write: "Article 1 of the Criminal Law of the People's Republic of China," not just "Criminal Law Article 1."

Your response must strictly follow the format below (do not include any extra text or explanations!):

Legal Articles:
Enabled: 0 or 1,
Queries: [Your query 1, Your query 2]
LegalOne:
Enabled: 0 or 1,
Queries: [Your query 1, Your query 2]

Figure 9: Prompts templates of tool calling for agents.

Exemplar 1: Exemplars of tool calling and corresponding responses.

The tool calling by agents:

Legal Articles:
Enabled: 1,
Queries: ["Article 264 of the Criminal Law of the People's Republic of China"]
LegalOne:
Enabled: 1,
Queries: ["Case examples of aggravated punishment for recidivism in theft offenses"]

The response by the legal article retriever:

Whoever steals public or private property, if the amount is relatively large, or if the theft is committed repeatedly, as a burglary, with a lethal weapon, or as pickpocketing, shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention, or public surveillance, and shall also, or shall only, be fined; if the amount is huge, or if there are other serious circumstances, the offender shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years and shall also be fined; if the amount is especially huge, or if there are other especially serious circumstances, the offender shall be sentenced to fixed-term imprisonment of not less than ten years or life imprisonment, and shall also be fined or have property confiscated.

The response by the legal case retriever (LegalOne):

Conclusion
The key to imposing aggravated punishment for recidivism in theft offenses lies in determining whether the defendant committed another crime punishable by fixed-term imprisonment or more severe penalty within five years after completing the sentence for the prior offense.
Case Analysis
Judicial practice demonstrates the following key considerations:
1.Recidivism Determination. Representative cases: (2023) Yue 0112 Crinimal Case No. 1177, (2023) Yue 1803 Crinimal Case No. 300. A brief introduction of the cases are...
2.Prior Convictions and Remorse. Representative cases: (2023) Yue 1803 Crinimal Case No. 300, (2022) Min 0623 Crinimal Case No. 74. A brief introduction of the cases are...
3.Attempted Crime and Voluntary Confession. Representative cases: (2018) Ji 0202 Crinimal Case No. 204, (2019) Su 02 Xing Zhong No. 81. A brief introduction of the cases are...

Figure 10: Exemplars of tool calling for agents, including the tool calling queries and corresponding responses made by the legal article retriever and legal case retriever.

Prompt 2: Initialization prompts of agents.

For all agents:
Indictment: {indictment}
Defendant Information: {defendant_information}
Evidence: {evidence}
Reference Materials: {references_provided_by_tools}

For Judge:
(Generating the investigation strategy)
As the presiding judge, you need to ascertain the facts of the case. Therefore, the key investigation issues should also cover the following aspects:
1. Whether the defendant is guilty or not, and the determination of liability (e.g., whether the age meets the criteria, primary or secondary liability, etc.).
2. Sentencing considerations (e.g., mitigating or aggravating circumstances, whether the defendant voluntarily surrendered, whether the victim's forgiveness was obtained).
3. Other unresolved but relevant issues that may impact conviction and sentencing, including any valuable questions you deem necessary for the case.
Number the investigation issues starting from 1.

For Prosecutor:
(Generating the attack strategy and the defense strategy)
Offensive Strategy: The Prosecutor must develop an offensive strategy, including establishing a chain of evidence proving the defendant's guilt through precise interpretation of evidence and legal provisions. This involves challenging the relevance and probative value of the defense's evidence (e.g., questioning the authenticity of purported voluntary surrender claims) and justifying the sentencing recommendations in the indictment based on factors such as the degree of social harm and the defendant's potential for rehabilitation.
Defensive Strategy: The Prosecutor shall formulate a defense strategy to safeguard the relevance and probative value of their evidence against potential challenges from the defense or defendant.
Prior to trial commencement, the Prosecutor shall develop both offensive and defensive strategies based on the indictment, defendant information, and evidence list.
Number the attack strategy and defense strategy starting from 1.

For Attorney:
(Generating the attack strategy and the defense strategy)
Attack Strategy: The Attorney must develop a strategy to challenge the Prosecutor's evidence, such as demonstrating its irrelevance to the case or insufficient probative value, and point out deficiencies in evidence or unreasonable inferences.
Defense Strategy: The Attorney shall formulate a defense strategy to mitigate the defendant's liability. This includes gathering exculpatory evidence, constructing a plausible alternative narrative, providing precise interpretations of legal provisions, emphasizing case-specific circumstances, and highlighting the defendant's demonstrated cooperation and remorse. Additionally, the Attorney must maintain evidentiary relevance between defense materials and case facts to counter potential challenges from the Prosecutor.
Prior to trial commencement, the Attorney shall develop both offensive and defensive strategies based on the indictment, defendant information, and evidence list.
Number the attack strategy and defense strategy starting from 1.

For Defendant:
(Generating the defense strategy)
Defense Strategy: The defendant must develop a defense strategy to mitigate culpability. If the objective is sentence reduction, the defendant should demonstrate remorse by admitting guilt and accepting punishment. If the objective is to deny the charges and prove innocence, the defendant must assert their innocence unequivocally.
Prior to trial commencement, the defendant shall formulate this defense strategy based on the indictment, personal information, and evidence list.
Number the defense strategy starting from 1.

Figure 11: Initialization prompts templates of agents.

Prompt 3: Reflection prompts of the planning module for different agents.

For all agents:
Summary of Court Statements: {long_term_memory}
Latest Trial Record: {short_term_memory}
Current Evidence: {evidence_pool}
Overall Requirements: {former_prompts_during_initialization}
Reference Materials: {references_provided_by_tools}

For Judge:
Current Investigation Issues and Findings: {investigation_strategy}
As the presiding judge, you must ascertain the facts of the case. Your tasks are:
Adjust/update investigation issues based on input information, current investigation issues, and findings.
Summarize the findings for each issue – including overviews of prosecution/defense arguments, discussion adequacy, and your judicial assessment. Be concise.

For Prosecutor and Attorney:
Current Offensive/Defensive Strategies: {attack_strategy}, {defense_strategy}
Adjust or update strategies based on input information.
Attention:
Strategies must be fact-based and realistic.
Return strictly in specified format without extraneous content.

Figure 12: Reflection prompts of the planning module for different agents, including the refinement and update for the memory module and planning module. Note that to simplify the process, Defendant agent will not change his strategy during the court.

Prompt 4: Reflection prompts of the memory module for agents

Latest court record : {short_term_memory}
Based on the latest court hearing records, you need to write a summary of this part of the trial.
The trial summary should include: a chronological summary of events that occurred, and may affect the final conviction and sentencing.
Procedural, repetitive, irrelevant to the case, and utterances without substantive content such as judges interrupting to control the proceedings should be omitted!
For example:
During the trial preparation phase, the presiding judge questioned the defendant xxx, and the defendant stated xx.
During the court investigation phase, the prosecutor questioned xxx in turn, the defendant replied xxx, the defense lawyer questioned xxx, and the defendant replied xxx.
Note: You only need to return the summary of this phase.

Figure 13: Reflection prompts of the memory module for agents. The summary will be then added into the long-term memory.

Prompt 5: Prompts of action for agents.

For Judge:
Your basic courtroom task: {basic_courtroom_tasks}
Stage-specific tasks: {stage-specific_tasks}
Current strategy: {investigation_strategy}
Summary former stages: {long_term_memory}
Latest court record: {short_term_memory}
Current instruction: {instruction}

For Prosecutor and Attorney:
Your basic courtroom task: {basic_courtroom_tasks}
Stage-specific tasks: {stage-specific_tasks}
Current strategy: {attack_strategy}, {defense_strategy}
Summary former stages: {long_term_memory}
Latest court record: {short_term_memory}
Current instruction: {instruction}

Figure 14: Prompts of action for agents. All the three modules as well as the current instruction are used.

Prompt 6: Prompts templates for final judgment.

Basic case information:

Accused crime: {charge}
Defendant information: {defendant_information}
Indictment: {indictment}
Evidence : {evidence}
Investigation issues and findings: {investigation_strategy}
Summary of court statements: {long_term_memory}
Reference Materials: {references_provided_by_tools}

The trial is about to end. As the presiding judge, you need to give a fair judgment based on the trial records, evidence, indictment, defendant information, debate focus, and facts found. Please pay close attention to the debate focus and the results of the facts found! Please distinguish for yourself the relevance and applicability of the statements made by the defendant, prosecutor, and defense lawyer to the real situation of the case. Please think carefully and punish cautiously!

#####

Response requirements

Your response should start with "The judgment is as follows" and then state the judgment.
(1) For the first, if convicted and sentenced, it should be stated as: "1. Defendant ××× is guilty of the crime of ×××, and is sentenced to... (specify the main penalty and additional penalties); 2. Defendant ×××... (specify the decision on confiscation, compensation, or confiscation of property, as well as the types and amount of these properties. If there are none, this item is not written)."
(2) For the second, if convicted but exempted from punishment, it should be stated as: "Defendant ××× is guilty of the crime of ×××, exempted from criminal punishment (if there is confiscation, compensation, or confiscation of property, continue to write as the second item)."
(3) For the third, if declared not guilty, it should be stated as: "Defendant ××× is not guilty."
Specifically, when convicted and sentenced, it should include a real sentence, which may include probation, fines. For example: The judgment is as follows: Defendant Zhang San is guilty of the crime of xxx, sentenced to x months of detention, with a suspended sentence of x months, and fined RMB 2,000.

#####
Attention

1. Provide a specific crime, followed by the sentence. If necessary, include probation and compensation, fine amount.
2. Note! The indictment mentions real sentences, suspended sentences, fines, etc., but whether there is a specific crime, whether suspended sentences apply, and whether fines are imposed, you need to judge for yourself! If you believe that no real sentence is needed, suspended sentences do not apply, or no fines are imposed, then do not mention real sentences, suspended sentences, or fines. Do not strictly follow the indictment
3. When responding, directly give your judgment, do not say extra words.

Figure 15: Prompts templates for final judgment. Note that Judgment may also make tool calling for references.