# *Ready Jurist One*: Benchmarking Language Agents for Legal Intelligence in Dynamic Environments

**Zheng Jia[1†], Shengbin Yue[1†], Wei Chen[3], Siyuan Wang[4], Yidong Liu[5], Zejun Li[1], Yun Song[6*], Zhongyu Wei[12*]**
[1]Fudan University [2]Shanghai Innovation Institute
[3]Huazhong University of Science and Technology [4]University of Southern California
[5] Midu Technology [6]Northwest University of Political and Law
{zjia24,sbyue23}@m.fudan.edu.cn, zywei@fudan.edu.cn

## Abstract

The gap between existing benchmarks and the dynamic nature of real-world legal practice poses a key barrier to advancing legal intelligence. To this end, we introduce ***J1-ENVS***, the first interactive and dynamic legal environment tailored for LLM-based agents. Guided by legal experts, it comprises six representative scenarios from Chinese legal practices at three levels of environmental complexity. We further introduce ***J1-EVAL***, a dual-metric evaluation framework, designed to assess both task performance and procedural compliance across varying levels of legal proficiency. Extensive experiments on 17 LLM agents reveal that while many models demonstrate solid legal knowledge, they struggle with procedural execution in dynamic settings. Even the SOTA model is below 60% overall performance Ṫhese findings highlight persistent challenges in achieving dynamic legal intelligence and offer valuable insights to guide future research. The detailed resources are available at https://github.com/FudanDISC/J1Bench.

## 1 Introduction

Although large language models (LLMs) have shown remarkable capabilities across conventional legal NLP tasks, *e.g.*, legal information extraction and statute memorization (Yao et al., 2022a; Huang et al., 2021), genuine legal intelligence requires agentic capabilities: procedural awareness, multi-stakeholder interaction, and decision-making in evolving environments. Recent advancements in LLM-based agent frameworks offer a viable solution for the paradigmatic shift from static task completion to dynamic environmental interaction (Yao et al., 2022b; Mou et al., 2025; Yue et al., 2025b), paving promising avenues for the development of *general legal intelligence* (GLI).

| Benchmarks | MT? | Dyn? | Proc? | Type | Evaluation |
|---|---|---|---|---|---|
| LexGLUE | ✗ | ✗ | None | NLP task | Rule |
| LegalBench | ✗ | ✗ | None | NLP task | Rule |
| LawBench | ✗ | ✗ | None | NLP task | Rule |
| LexEval | ✗ | ✗ | None | NLP task | Rule |
| Law-Eval | ✗ | ✗ | None | NLP task | Rule & LLM |
| LAiW | ✗ | ✗ | None | NLP task | Rule & Human |
| AgentCourt | ✓ | ✓ | Brief | Scenario | Rule & Human |
| LegalAgentBench | ✓ | ✓ | Brief | NLP task | Rule |
| MILE | ✓ | ✓ | Complete | Scenario | Rule & LLM |
| ***J1-EVAL*** | ✓ | ✓ | Complete | Scenario | Rule & LLM |

Table 1: Comparison between our ***J1-EVAL*** and existing legal benchmarks, where **MT**, **Dyn**, and **Proc** denote Multi-turn, Dynamic, and Procedural Process, respectively. ***J1-EVAL*** is scenario-based, dynamic, and interactive, aligning better with *real-world legal practices*.

However, the primary obstacle to achieving GLI lies in the lack of reliable experimental platforms and quantitative evaluation tools to expose current deficiencies and guide optimization efforts. Realistic legal services inherently require *multi-turn interactions* to provide step-by-step guidance for the laypeople, alongside a strict adherence to *procedural legality*. In contrast, As shown in Table 1, existing benchmarks are either confined to a **static evaluation paradigm** (Chalkidis et al., 2021; Guha et al., 2023; Fei et al., 2024; Li et al., 2024; Dai et al., 2023) or limited to a few **simplified scenarios** (Chen et al., 2024; Li et al., 2025b; Yue et al., 2025a) with several limitations. **(1) Narrow scenarios**: They are restricted to predefined static environments with limited coverage, failing to support sustained agent-environment interactions. **(2) Over-simplified tasks**: They primarily focus on QA tasks adapted from traditional NLP tasks, lacking the definition of composite targets that reflect real-world legal demands. **(3) Restricted evaluation metrics**: They focus solely on the accuracy of outcomes, neglecting the requirement of procedural justice inherent in the judiciary. To bridge these gaps, we introduce *the first dynamic evaluation benchmark* for legal intelligence including two

[†]These authors contributed equally to this work.
[*]Corresponding author



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10351–10376
July 2-7, 2026 

integrated modules: the interactive environment and the comprehensive assessment system.

***J1-ENVS***: Interactive Legal Environments. We construct an open-ended legal environment in which agents navigate various legal scenarios through dynamic and procedural interactions. As shown in Figure 1, we instantiate heterogeneous agents ranging from layperson to legal professionals, each endowed with distinct behavioral traits and initialized with interrelated factual elements from actual cases to ensure high-fidelity simulation. These agents are then orchestrated into corresponding environments following the structure of legal practice, which are categorized into three hierarchical levels based on increasing environmental complexity: legal consultation (knowledge questioning and consultation), document drafting (complaint and defence drafting), and judicial adjudication (civil and criminal courts).

***J1-EVAL***: Dual-Metric Assessment System. In order to enable comprehensive assessment, we construct 508 specific interactive environment instances from real-world Legal sources (*Chinese Judgment Documents* and *Chinese legal Articles*). Our framework utilizes a dual-metric system consisting of **Outcome-oriented** and **Process-oriented** metrics. The former quantifies the quality of the final target, while the latter scrutinizes adherence to indispensable intermediate constraints, reasoning steps, and procedural norms. Assessment is conducted using rule-based or LLM-based methods, with each instance equipped with **explicit ground truth** for the final objectives, ensuring evaluation reliability and minimizing potential bias introduced by LLM-based evaluators. Together, ***J1-EVAL*** can provide deeper insights into the skill sets required for legal task execution, and present a realistic and reliable overview of agent capabilities, facilitating further exploration into the similarities and differences between LLM-based agents and human learning mechanisms.

We perform an extensive assessment of 17 leading LLM agents, spanning proprietary, open-source, and legal-specific models. Results reveal significant capability gaps in dynamic environments, with performance degrading markedly in complex, multi-agent, and long-horizon settings. Crucially, these failures stem not from knowledge deficits, but from limitations in procedural adherence and role interaction. Our analysis confirms that process-oriented competence is essential for valid outcomes, yet current models frequently misalign procedural execution with final objectives.

**Contributions:** **(A)** We construct ***J1-ENVS***, the first open-ended interactive legal environment, orchestrating heterogeneous agents across three hierarchical stages and enabling dynamic agent–environment interaction. **(B)** We introduce ***J1-EVAL***, a comprehensive assessment system comprising 508 real-world instances, featuring a dual-metric mechanism that integrates both procedural and outcome perspectives. **(C)** Experimental results reveal the gaps between diverse agents and real-world legal demands, offering valuable insights to guide future progress. **(D)** This work introduces a new paradigm for legal intelligence, shifting from static to dynamic settings. Beyond evaluation, ***J1-ENVS*** can be extended to data generation and reinforcement learning training.

## 2 *J1-ENVS*: Interactive Legal Environments

In this section, we detail the construction pipeline of ***J1-ENVS***, including various agents organized in six scenarios of three levels, namely, legal consultation, document drafting, and judicial adjudication. **(A) Role Agent Setting** (§2.1) constructs diverse agents by synthesizing real-world legal sources and personality theories. **(B) Multi-level Environment Construction** (§2.2) integrates these roles into coherent legal environments governed by specific procedures and relationships.

### 2.1 Role Agent Setting

Legal environments inherently involve heterogeneous roles, distinguishing between legal professionals and layperson with different backgrounds, targets, and behaviors. Without specialized grounding, generic LLM-driven simulators are prone to significant legal illusions (Yue et al., 2025a). Therefore, creating realistic individuals presents two challenges: *diversity of layperson* and *legal consistency among roles*. We address this by modeling individuals with the Big Five personality traits and assigning interconnected legal elements, derived from real-world legal practice, to different agents within the shared environment.

**Role Agents Configuration.** We establish two categories of role repositories to populate the environment: (1) *Layperson* includes the general public with a broad background and specific characters with explicit litigation-related needs. The former are assigned topic lists, background, and behav-

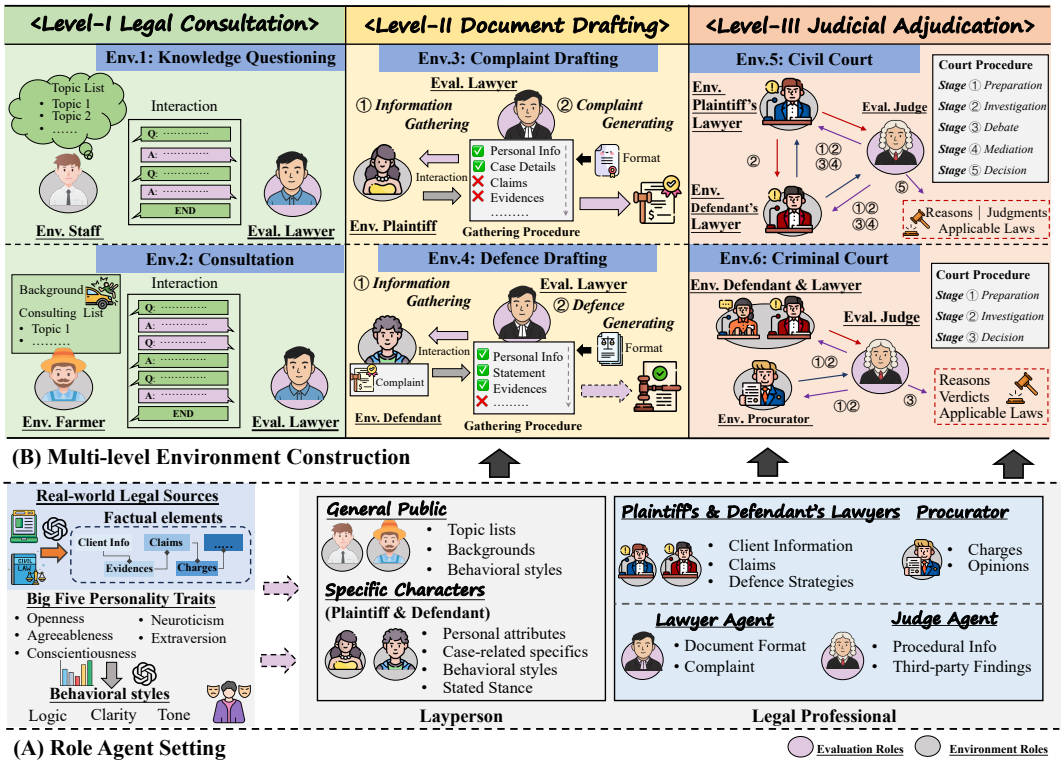


Figure 1: An illustration of ***J1-ENVS*** construction pipeline. (A) Role Agent Setting: We synthesize real-world legal sources and personality theories to construct heterogeneous agents. (B) Multi-level Environment Construction: We structure these roles within specific procedures and relationships to form environments.

ioral styles. The latter, *i.e.,* plaintiffs and defendants, are assigned behavioral styles and individual profiles, including personal attributes (*e.g.*, names, genders, and addresses) and case-related details (*e.g.*, claims, statements of defence, and evidence). (2) *Legal Professionals* include lawyers, prosecutors, and judges who assume formal legal responsibilities. They are assigned authentic legal information corresponding to their current environments. For example, Plaintiff's Lawyers in the civil court possess detailed case information extracted from civil judgment documents, *e.g.*, their client information, case details, claims & defence. In criminal court scenarios, the Procurator agent is assigned to charges and opinions sourced from criminal judgment documents. The methodology for acquiring these interconnected factual elements from real-world legal sources is detailed in Sec. 3.1.

**Setup with Personality Theory.** Behavioral diversity enhances the model to represent complex characters, further aligning with real-world behavior. Guided by social personality theory (Sun et al., 2024), Big-5 Personality Traits are mapped to behavioral styles of non-legal roles (public, the plaintiff, and the defendant). Each trait is assigned one of three levels (high, medium, or low) and GPT-4o is prompted to generate corresponding behavioral descriptions based on the role's profile or background. Details are provided in Appendix A.2.

## 2.2 Multi-level Environment Construction

Building on role repositories, we organize six environments into a three-tiered hierarchy spanning inquiry to adjudication. Below, we detail the participants and interaction logic, with pseudo-code provided in the Appendix A.4.

**Level-I: Legal Consultation.** This level involves two participants, the general public and a legal agent, aiming to simulate the response capabilities of the legal trainee in dynamic environments. (1) **Knowledge Questioning** (**KQ**). Through progressive dialogue, the legal agent resolves user confusion, moving from broad themes to specific issues. The public, assigned a topic list and occupation, continues the interaction until all topics are addressed. (2) **Consultation**(**LC**). This presents a complex counseling setting grounded in a specific incident. Since legal agents lack information about the public, they need actively ask for relevant details to provide better responses.

| Scenario | ID | Metric | Definition | Range | Type | Method |
|---|---|---|---|---|---|---|
| Level-I | I-1 | Binary Accuracy | Binary (Yes/No) answer accuracy | [0,0.5,1] | OO | EM |
| | I-2 | Non-binary Score | Score for open-ended answers | [0-1] | OO | LLM |
| Level-II | II-1 | Format-following Score | Document structure alignment | [0-1] | PO | EM |
| | II-2 | Document Score | Average component-level accuracy | [0,1] or [0-1] | OO | EM+LLM |
| Level-III | III-1 | Procedural-following Score | Procedural stage completeness | [0-1] | PO | LLM |
| | III-2 | Judgment Score | Final judgment quality | [0-1] | OO | LLM |
| | III-3 | Crime Accuracy | Charge prediction accuracy | [0-1] | OO | EM |
| | III-4 | Penalty Deviation | Log-distance penalty error | [0-1] (↓) | OO | Dist. |
| | III-5 | Reason Score | Legal reasoning quality | [0-1] | PO | LLM |
| | III-6 | Law Accuracy | Cited law precision | [0, 1] | PO | EM |

Table 2: Metrics in ***J1-ENVS***. OO, PO, EM, and Dist. denote outcome-oriented metrics, process-oriented metrics, exact match, and nlog-distance, respectively. Except for III-1, all metrics are evaluated against explicit ground truth.

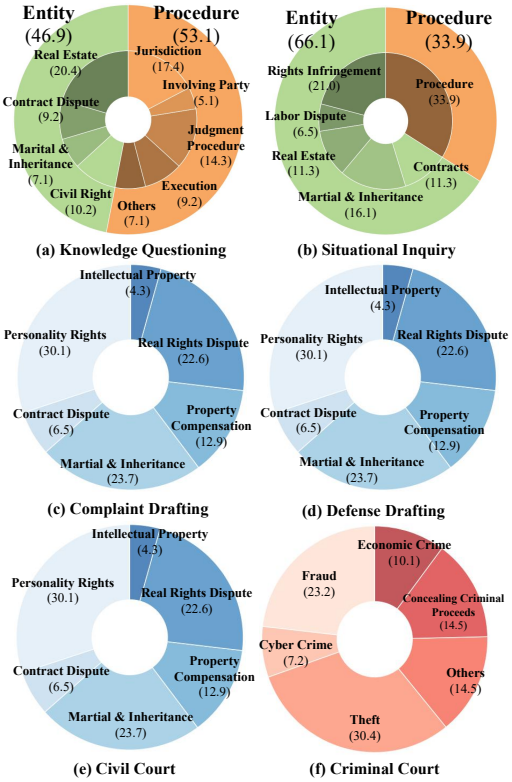


Figure 2: Distribution of legal attributes for six environments in ***J1-EVAL***, showing a wide range of coverage.

**Level II: Document Drafting.** This level involves a litigant with a concrete legal demand and a legal agent responsible for task execution, and is designed to simulate how a practicing lawyer independently gathers relevant information and completes a litigation task according to a given agenda. (1) **Complaint Drafting** (**CD**). Guided by a predefined format, the legal agent gathers information through interaction with the plaintiff, autonomously plans the required content, and generates a complete complaint. (2) **Defence Drafting** (**DD**). Unlike above, this scenario starts from an existing complaint. The legal agent interacts with the defendant, formulates counterarguments based on the complaint, and generates the corresponding defence document.

**Level III: Judicial Adjudication.** This level involves multiple participants and is governed by strict court procedural norms. (1) **Civil Court** (**CI**). The Civil Court addresses disputes concerning rights and obligations between individuals or organizations, involving three participants: the plaintiff, the defendant, and the judge. It includes five stages: court preparation, court investigation, court debate, court mitigation, and court decision. The judge agent oversees the process by following the procedural framework and interacting with both parties' attorneys. In the final stage, the agent renders a judgment in response to the plaintiff's claims. (2) **Criminal Court** (**CR**). The Criminal Court centers on adjudicating criminal charges against individuals or groups, involving four participants: the defendant, the defendant's lawyer, the procurator, and the judge. It includes three stages, court preparation, court investigation, and decision. Compared to civil court, legal agents in criminal proceedings interact with a broader range of participants, eliciting statements from the defendant and the lawyer, requesting evidence from the procurator, and working toward the establishment of criminal liability.

## 3 *J1-Eval*: Dual-metric Assessment System

### 3.1 Setup with Real-world Legal Sources

To ensure authenticity, we assign distinct legal elements from the same event to various roles, drawing from two primary sources: (1) *Chinese Judgment Documents*. For Level II and Level III scenarios, we rely on Chinese court judgments, as

| Model | Level-I | | | | Level-II | | | | Level-III | | | | | | | | | AVG | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | KQ | | LC | | CD | | DD | | CI | | | | CR | | | | | | |
| | I-1 | I-2 | I-1 | I-2 | II-1 | II-2 | II-1 | II-2 | III-1 | III-2 | III-5 | III-6 | III-1 | III-3 | III-4 | III-5 | III-6 | OO | PO |
| *Multilingual LLMs (close source)* | | | | | | | | | | | | | | | | | | | |
| GPT-4o | 69.7 | 62.7 | 55.9 | 45.1 | 58.9 | **89.9** | 45.4 | **84.8** | **87.0** | **28.5** | **50.2** | 13.9 | 49.3 | 60.9 | 38.5 | 42.8 | 27.0 | 62.2 | 46.8 |
| Claude-3.7 | 68.4 | 60.7 | 52.4 | 49.6 | 68.8 | 89.2 | 38.7 | 68.8 | 65.8 | 21.2 | 40.2 | **17.5** | 56.9 | **88.4** | **15.7** | **72.8** | **41.1** | **62.3** | **50.2** |
| *Multilingual LLMs (open source)* | | | | | | | | | | | | | | | | | | | |
| Deepseek-v3 671B | 67.1 | 60.7 | 54.5 | **49.7** | 84.7 | 63.6 | **57.2** | 76.3 | 24.0 | 12.4 | 19.5 | 7.6 | 42.4 | 76.8 | 33.9 | 57.2 | 32.3 | 57.6 | 40.6 |
| Deepseek-r1 671B | **72.5** | 62.8 | 57.2 | 48.3 | 69.9 | 46.4 | 36.3 | 47.9 | - | 0.8 | 1.1 | 0.7 | 0.8 | 1.4 | 98.6 | 1.4 | 0.7 | 42.2 | 13.9 |
| Llama3.3-it. 70B | 64.8 | 50.1 | 43.9 | 23.6 | **97.9** | 84.5 | 32.3 | 62.4 | 61.0 | 12.8 | 25.3 | 4.5 | 50.7 | 81.2 | 25.0 | 50.1 | 17.2 | 52.9 | 42.4 |
| Qwen3 32B | 69.6 | **66.2** | **58.3** | 47.2 | 91.2 | 84.4 | 46.7 | 63.5 | 52.7 | 17.0 | 34.1 | 15.4 | 26.8 | 49.3 | 58.3 | 40.0 | 22.9 | 56.9 | 41.2 |
| Gemma3-it 27B | 64.1 | 49.5 | 54.1 | 26.0 | 96.3 | 84.8 | 37.9 | 68.8 | 61.7 | 15.5 | 32.4 | 6.4 | **60.0** | 49.3 | 22.1 | 36.1 | 15.6 | 51.5 | 43.3 |
| Qwen3 14B | 67.7 | 62.8 | 56.2 | 44.3 | 63.9 | 89.2 | 36.6 | 62.4 | 13.0 | 4.3 | 12.2 | 9.8 | 16.9 | 26.1 | 73.6 | 22.8 | 17.1 | 51.6 | 24.0 |
| Gemma3-it 12B | 62.0 | 42.0 | 47.9 | 19.2 | 51.8 | 84.1 | 35.3 | 63.9 | 39.1 | 12.0 | 23.7 | 3.5 | 41.3 | 58.0 | 40.8 | 38.3 | 8.1 | 48.6 | 30.1 |
| GLM4 9B | 66.0 | 54.9 | 54.0 | 32.4 | 56.6 | 82.2 | 40.4 | 52.2 | 12.7 | 2.3 | 8.0 | 1.5 | 25.1 | 39.1 | 75.0 | 24.8 | 10.1 | 47.9 | 22.4 |
| Qwen3 8B | 66.2 | 57.1 | 55.0 | 36.9 | 74.2 | 72.7 | 17.6 | 22.8 | 9.9 | 5.8 | 14.2 | 7.2 | 16.3 | 26.1 | 67.4 | 19.4 | 16.7 | 42.8 | 21.9 |
| InternLM3-it. 8B | 65.3 | 55.3 | 55.5 | 42.4 | 91.0 | 72.2 | 36.1 | 53.8 | 31.8 | 7.7 | 19.0 | 4.2 | 29.1 | 60.9 | 45.1 | 46.7 | 28.7 | 51.6 | 35.8 |
| Ministral-it. 8B | 59.6 | 37.6 | 51.5 | 14.5 | 60.0 | 2.8 | 34.2 | 14.0 | 22.7 | 5.7 | 12.4 | 0.5 | 2.1 | 4.3 | 98.0 | 2.3 | 0.9 | 23.8 | 16.9 |
| Qwen2.5-it. 7B | 67.2 | 54.9 | 52.6 | 31.3 | 89.5 | 81.4 | 47.1 | 59.7 | 15.6 | 7.3 | 14.1 | 7.5 | 26.4 | 49.3 | 54.5 | 37.5 | 25.4 | 50.5 | 32.9 |
| Qwen3 4B | 62.2 | 55.9 | 50.7 | 37.2 | 51.4 | 83.7 | 33.1 | 51.5 | 28.0 | 11.3 | 18.3 | 7.6 | 20.8 | 44.9 | 57.3 | 32.9 | 18.8 | 49.7 | 26.4 |
| *Legal-specific LLMs* | | | | | | | | | | | | | | | | | | | |
| LawLLM 13B | 56.3 | 47.3 | 44.9 | 20.8 | 12.9 | 2.9 | 0.6 | - | 1.0 | - | 0.3 | - | 12.7 | 17.4 | 87.4 | 10.9 | 2.5 | 23.7 | 5.1 |
| Chatlaw2 7B | 60.6 | 53.1 | 50.4 | 25.0 | 22.4 | 17.0 | 1.3 | 0.9 | - | - | - | - | - | - | - | - | - | 25.9 | 3.0 |

Table 3: Performance among LLM-driven legal agents on ***J1-ENVS***, where the bolded score indicates the best performance, while the underlined score represents the second-best. Note that "-" denotes failure to accomplish the task, where the model achieves a 0% completion rate for the final legal objectives across all samples, and lower **III-4** indicates higher performance.

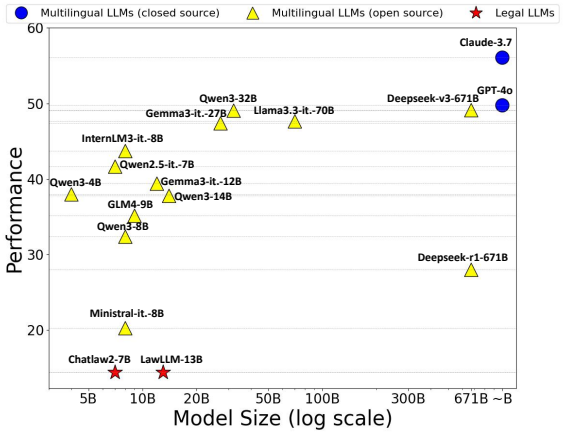


Figure 3: Overall performance ranking across different LLM agent sizes.

they naturally encapsulate the full procedural lifecycle of legal cases, including litigant information, disputes, factual findings, judicial reasoning, rulings, and applicable statutes. We collect both civil and criminal judgments and employ LLMs to extract structured components such as party identities, claims, detailed case facts, and cited legal provisions. (2) *Chinese Legal Articles*. For Level I scenarios, which prioritize progressive inquiry and logical authenticity, we curate expert-authored legal articles from the *HUALV* website [1] and the book *Civil Trial Practice Q&A*. Given that these articles typically feature step-by-step questions or thematic headings, we prompt LLMs to utilize this inherent structure to generate coherent topic lists (classified as binary or open-ended) alongside their corresponding answers. Further data construction details are provided in Appendix A.1.

**Data Details.** All documents are collected from 2025 onward to ensure legal timeliness. After preprocessing, ***J1-EVAL*** consists of 508 instances, including 160 Level I, 186 Level II, and 192 Level III instances. Dataset statistics are reported in Appendix Table 7. Our benchmark spans a diverse and nuanced set of legal attributes aligned with real-world legal practice, derived directly from raw case data. As shown in Figure 2, KQ and LC involve different entities and procedural topics.CD, DD and CI address various types of civil rights disputes. CR covers different categories of criminal offenses. Detailed data analysis are provided in Appendix B.

### 3.2 Evaluation Methods

To comprehensively evaluate, we propose a dual evaluation framework that integrates **outcome-oriented** and **process-oriented** metrics. The former quantifies the quality of the final objectives (e.g., judgment accuracy), whereas the latter assesses the quality of essential intermediate constraints throughout task execution. As shown in Table 2, these metrics are strategically combined based on scenario characteristics to facilitate assessment across different environmental levels. Each metric adopts a score range in **[lower**

[1] https://www.66law.cn

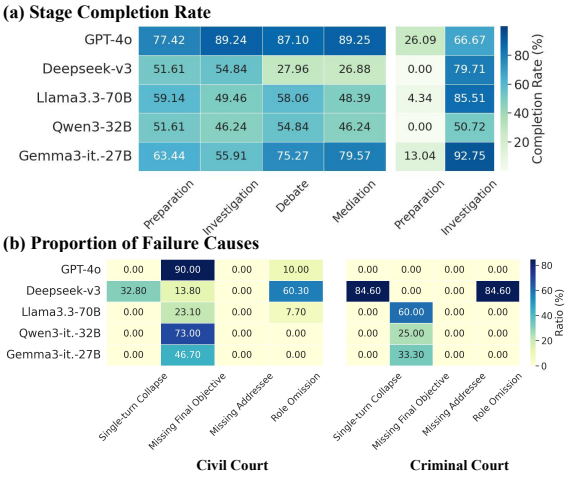


Figure 4: Procedural-following performance of legal agents in civil and criminal courts. (a) Completion rate of court stages in Civil & Criminal Court. (b) Proportion of failed cases attributed to each cause.

**bound–upper bound]** form, utilizing either rule-based or LLM-based automatic methods. Crucially, each instance is equipped with an **explicit ground truth** for the final objective. This approach rigorously mitigates subjective judgment and potential bias, aligning with established methodologies for long-form generation tasks (Li et al., 2025a). See Appendix C for details on evaluation metrics.

**Outcome-oriented metrics.** These metrics evaluate the fidelity of the agent's final deliverables against the ground truth across three-level scenarios. (1) *Legal consultation*. Binary accuracy (**I-1**) evaluates yes/no responses via exact match, while Non-binary score (**I-2**) assesses open-ended answers based on factual accuracy and legal coverage. (2) *Document drafting*. Document score (**II-1**) measures content correctness, with personal information verified by exact match and substantive components (e.g., claims). (3) *Judicial adjudication*. Civil judgment score (**III-1**) and crime accuracy (**III-3**) evaluate civil rulings and criminal charges, respectively. Additionally, Penalty deviation (**III-4**) measures the accuracy of fines and sentences via log distance (Chen et al., 2019), where lower values indicate higher accuracy.

**Process-oriented metrics.** These metrics assess compliance with procedural norms, formatting constraints, and logical reasoning pathways. Format-following Score (**II-1**) evaluates document structure by verifying the presence, order, and labeling of mandatory components. Procedural-following Score (**III-1**) measures trial completeness, where a stage is considered complete only if all predefined actions are correctly executed. Reason Score (**III-5**) evaluates logical coherence of judicial reasoning. Law Accuracy (**III-6**) evaluates correctness of cited statutes via ground-truth matching. Except for **III-1** metric, all process-oriented metrics are evaluated against explicit ground-truth references.

| Agent | Env.GPT-4o | Env.Qwen3 $_{32B}$ |
|---|---|---|
| Evaluator(GPT-4o) | | |
| GPT-4o | 94.22 | 95.49 |
| Deepseek-v3 $_{671B}$ | 90.97 | 94.71 |
| Qwen3 $_{32B}$ | 89.97 | 94.99 |
| Evaluator(Human) | | |
| GPT-4o | 76.25 | 75.63 |
| Deepseek-v3 $_{671B}$ | 74.13 | 75.80 |
| Qwen3 $_{32B}$ | 74.00 | 74.25 |

Table 4: Behavior consistency of environment roles across different legal agents (GPT-4o, Deepseek-v3, Qwen 32B) under Human and LLM evaluation.

## 4 Experiments

### 4.1 Experimental Settings

**Agent Models.** We evaluate legal agents driven by various LLMs in two categories. (1) *General multilingual LLMs*: GPT-4o[2] (Achiam et al., 2023), Claude-3.7[3] (Anthropic, 2023), Deepseek-v3 (Liu et al., 2024), Deepseek-r1 (Guo et al., 2025), InternLM3-instruct 8B (Cai et al., 2024), Qwen2.5-instruct 7B (Yang et al., 2024), Qwen3 4B/8B/14B/32B (Yang et al., 2025), ChatGLM4-chat 9B (GLM et al., 2024), and Ministral-Instruct-2410 8B (Jiang et al., 2023a). (2) *Legal-specific LLMs*: LawLLM 13B (Yue et al., 2024), and Chatlaw2 7B (Cui et al., 2023).

**Implementation Detail.** We utilize arbitrary LLMs, GPT-4o (Achiam et al., 2023) or Qwen3-32B (Yang et al., 2025), to drive the environments. Due to page limitations, details of evaluation settings are provided in the Appendix D.

### 4.2 Performance among Various Legal Agents

In this setting, we leverage GPT-4o to drive our interactive environments.

**Performance at Level I & II.** Table 3 reports our metrics across levels. (1) At *Level I*, both general-purpose and legal-specific models achieve strong performance on KQ, reflecting solid legal knowledge. Performance declines on the more interactive

[2]gpt-4o-2024-11-20
[3]claude-3-7-sonnet-20250219

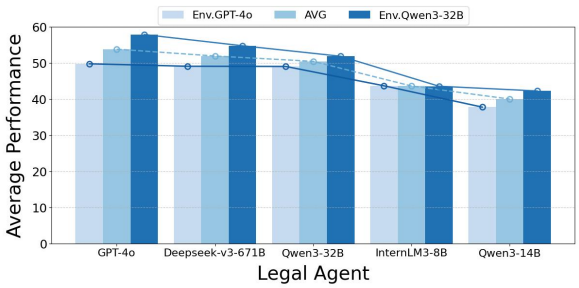


Figure 5: Overall performance of legal agents with ***J1-ENVS*** driven by GPT-4o or Qwen3-Instruct-32B, where **AVG** denotes the agent's average performance.

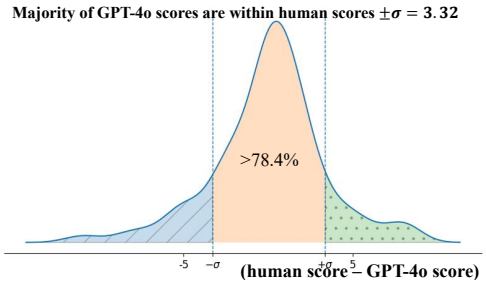


Figure 6: The distribution of the difference between GPT-4o scores and human evaluation

LC task, which demands proactive engagement. (2) At *Level II*, LLaMA-3.3 and Deepseek-v3 achieve the highest scores on **II-1** for CD and DD tasks, demonstrating strong template adherence in the long legal context. Notably, performance in DD scenarios is lower than in CD, as DD settings require models to construct more complex and coherent counterarguments in response to an opposing party's complaint. Detailed results are reported in Tables 8 and 9 of the Appendix.

**Performance at Level III.** Civil and criminal courts involve distinct procedures and judgment forms. The **civil court** comprises five stages, yet most models struggle to complete the full process, particularly in legal-specific models, reasoning models, and smaller models. GPT-4o and Claude-3.7 perform well in **III-1**, **III-2** and **III-5**. The criminal court includes three stages and more participants, where Claude-3.7 achieves leading performance. Additionally, Deepseek-r1's poor performance in executing the court process highlights the limitations of reasoning models in complex environments. Overall, *effective reasoning and adherence to procedural protocols within extended legal contexts are essential for achieving accurate legal judgments*. Detailed judgment component results are reported in Tables 10 and 11.

**Overall Performance.** We compute the average scores across all environments except **III-4**, and rank the models by size. As shown in Figure 3, Claude-3.7 achieves the best performance, showing strong legal intelligence. Notably, although the legal-specific LLMs perform comparably to GPT series on existing legal benchmarks (Fei et al., 2024), they exhibit significantly weaker performance in our setting, falling behind even smaller models. This highlights a key limitation: **despite possessing legal knowledge, the absence of interactive capability hampers their effectiveness in dynamic, realistic environments**.

### 4.3 Further Analysis

**Analysis from process- and outcome-oriented perspectives.** Table 3 presents the average performance of Outcome-oriented (OO) and Process-oriented (PO) metrics. Models with higher PO consistently achieve superior OO, and the OO scores significantly surpass the PO across all models. Based on these, we draw two conclusions: (1) Processal adherence is not a mere formality but the bedrock of logical validity and factual completeness, making it a necessary condition for precise legal judgment. (2) In the legal domain, processal justice is often prioritized over the outcome itself. The lag in PO metrics reveals a **severe deficiency in current models' capabilities to understand, adhere to, and execute legal procedures**.

**Analysis of court completeness.** As shown in Fig. 4, the success rates of the agent in completing all required steps vary across different stages of civil (CI) and criminal (CR) courts. The results indicate that (1) Current legal agents encounter difficulties in the preliminary stages of court proceedings. While these stages are largely formalistic, they are indispensable for ensuring procedural justice in real-world practice. (2) Navigating the criminal court poses greater challenges for legal agents, as the environment contains more roles that increase the complexity of the interaction. In addition, we identify four causes of failure: single-turn collapse, missing final objective, missing addressee, and role omission. Definitions of causes are provided in Appendix E.3. Figure 4 shows that missing the final objective accounts for most failures. These findings underscore the need for enhanced **procedural learning for legal agents**, and the importance of **strengthening their communication**

| Agent | KQ | | LC | | CD | | DD | |
|---|---|---|---|---|---|---|---|---|
| | I-1 | I-2 | I-1 | I-2 | II-1 | II-2 | II-1 | II-2 |
| GPT-4o | 65.7±1.9 | 77.3±0.4 | 48.5±1.3 | 47.8±0.1 | 59.4±0.2 | 84.9±0.9 | 82.1±1.1 | 66.5±0.8 |
| Deepseek-v3 $_{671B}$ | 64.5±1.0 | 77.2±0.1 | 51.7±0.9 | 48.0±0.1 | 98.2±0.7 | 81.7±1.0 | 74.3±0.5 | 51.7±0.5 |
| Qwen3 $_{32B}$ | 68.6±1.9 | 76.7±0.5 | 56.6±2.2 | 47.7±0.2 | 85.3±0.9 | 87.2±0.2 | 89.5±1.3 | 57.1±0.8 |
| Qwen3 $_{14B}$ | 64.5±2.5 | 77.8±0.3 | 54.6±2.0 | 48.0±0.2 | 71.9±1.3 | 61.1±0.4 | 71.9±1.3 | 61.1±0.4 |
| InternLM3-it. $_{8B}$ | 63.0±0.2 | 76.8±0.1 | 51.4±1.1 | 45.6±0.1 | 98.0±0.2 | 58.5±0.1 | 57.1±1.1 | 47.6±0.4 |

Table 5: Performance comparison of legal agents across multiple simulations(KQ, LC, CD, & DD) in ***J1-ENVS*** driven by Qwen3-32B. Scores are reported as average ± standard deviation.

| Agent | CI | | | | CR | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | III-1 | III-2 | III-5 | III-6 | III-1 | III-3 | III-4 | III-5 | III-6 |
| GPT-4o | 80.9±1.3 | 27.3±0.9 | 47.2±1.7 | 20.9±0.9 | 21.1±0.9 | 28.5±1.8 | 74.2±2.8 | 24.1±1.4 | 16.6±1.0 |
| Deepseek-v3 $_{671B}$ | 21.7±1.0 | 10.8±0.8 | 13.7±1.5 | 8.0±1.8 | 45.4±0.3 | 78.3±0.0 | 30.4±3.7 | 62.1±2.0 | 36.6±1.5 |
| Qwen3 $_{32B}$ | 45.9±0.2 | 13.1±1.4 | 29.3±0.9 | 16.2±0.5 | 33.7±0.6 | 53.6±1.2 | 49.7±0.4 | 44.3±1.1 | 26.1±0.7 |
| Qwen3 $_{14B}$ | 17.8±1.6 | 6.6±1.6 | 14.9±2.8 | 8.8±1.7 | 23.2±1.0 | 41.1±0.8 | 67.0±2.7 | 31.3±1.3 | 21.9±0.8 |
| InternLM3-it. $_{8B}$ | 24.1±0.2 | 8.4±1.3 | 18.0±0.7 | 9.2±1.1 | 23.5±0.5 | 51.7±0.7 | 56.2±0.7 | 37.8±0.5 | 26.2±0.2 |

Table 6: Performance comparison of legal agents across multiple simulations(CI & CR) in ***J1-ENVS*** driven by Qwen3-32B. Scores are reported as average ± standard deviation.

**and coordination with different roles**.

### 4.4 Can LLMs drive *J1-ENVS* and *J1-EVAL* ?

This section aims to validate the **stability** and **reliability** of our environment (P1, P2) and evaluation method (P3). we leverage GPT-4o and Qwen3-32B to drive the ***J1-ENVS*** separately. All human evaluation settings are detailed in the Appendix F.

**P1: behavioral consistency analysis of environment roles.** Table 4 presents average consistency scores (1–10) evaluated by GPT-4o (all samples) and humans (20 per scenario). Across different evaluators, environment roles maintain high and stable scores, regardless of the underlying LLM. Furthermore, scores maintain stability across diverse environments for the same legal agent. The result validates the reliability and effectiveness of our environment, laying a solid foundation for assessing agents in interactive legal scenarios. Detailed results are provided in Tables 13 and 14.

**P2: analysis of agent performance across different environments.** Figure 5 shows that all five agents successfully complete their tasks in the Qwen-based environment, with performance rankings consistent with those observed in the GPT-based environment. Although agent performances under the Qwen3-32B setting differ from those in GPT-based environments due to differences in model scale and capability, such variation is expected and acceptable. Importantly, achieving a balance between dynamics and controllability is more critical than strict performance equivalence across models. In addition, Appendix Tables 5 and 6 demonstrate the stability of agent performance across multiple repeats. Results validate the robustness of our framework and support its applicability to diverse deployment scenarios.

**P3: analysis of LLM–human agreement in *J1-EVAL*.** We assess the reliability of LLM-based metrics by measuring agreement between human judgments and model outputs, adopting a unified 10-point Likert scale as the singular scoring standard for both human and machine evaluation. Specifically, human annotators and the LLM-as-Judge assign discrete scores from 1 (completely inconsistent) to 10 (fully consistent). Following the protocol in Appendix F, human annotators are instructed to score 20 cases per scenario. The average pairwise difference between human and model judgments is computed after excluding at most one sample each corresponding to the maximum and minimum differences. As shown in Fig. 6: (1) The average difference is 0.03, indicating **a high level of overall agreement between LLM-based and human assessments**. (2) 78.4% of the cases exhibit a score difference within one standard deviation($\leq 1 * \sigma = 3.32$), suggesting that **evaluation differences are statistically controlled**. Causes of the divergence are analyzed in Appendix F.3. Overall, these results support the reliability of LLM-based evaluation.

**P4: analysis of Cohen's $\kappa$ for the LLM evaluator in *J1-EVAL*.** We further analyze the reliability of LLM-based metrics by computing the quadratically

weighted Cohen's $\kappa$ between human evaluators and the LLM-as-Judge, obtaining a value of 0.6027. For reference, we also compute Cohen's $\kappa$ among human annotators, yielding an average of 0.5855 ± 0.1629, which reflects notable inherent subjectivity among human evaluators. In this context, the LLM–human agreement is not lower than the average human–human agreement, validating the reliability of LLM-based evaluation.

**P5: analysis of confidence intervals.** To validate the robustness of our framework across repeated trials, we conduct simulations three times in environments driven by Qwen3-32B, and analyze the confidence intervals of outcomes. As shown in Table 5 and Table 6, scores across multiple simulations are generally stable, with only minor variations reflected in standard deviations. This indicates that our evaluation framework produces robust and consistent measurements of legal agent performance across repeated runs, validating the reliability of the metrics in ***J1-ENVS***.

## 5 Related Work

**Legal Benchmark.** Existing legal evaluations of LLM are based on static, single-turn settings (Fei et al., 2024; Li et al., 2024; Guha et al., 2023; Dai et al., 2023; Yue et al., 2024). For example, LegalBench (Guha et al., 2023) and LawBench (Fei et al., 2024) assess legal capabilities by adapting existing NLP legal tasks to LLMs, while Law-Eval (Yue et al., 2024) evaluates models using multiple-choice questions from law exams. LexEval (Li et al., 2024), inspired by Bloom's taxonomy, organizes legal abilities into a cognitive hierarchy, and LAiW (Dai et al., 2023) structures legal reasoning into three levels based on syllogistic logic. However, these static, task-centric benchmarks fail to capture the dynamic complexity of real scenarios, rendering them inadequate for agent evaluation.

**LLM-based Simulation.** By emulating human perception and behavior, LLMs can be used to construct role agents with diverse personalities and environmental contexts, ranging from individuals (Shao et al., 2023; Argyle et al., 2023) to demographic groups (Zhang et al., 2025; Jiang et al., 2023b). These agents are capable of coordinating, collaborating, exchanging information, and even engaging in debates, thereby exhibiting social intelligence (Zhou et al., 2023; Mou et al., 2025). Building on such role agents, recent work has explored more complex simulated environments (Stade et al., 2024; Fan et al., 2025). For example, Generative Agents (Park et al., 2023) model a small town with 25 agents, while AI Hospital (Fan et al., 2025) simulates dynamic medical interactions among doctors and NPCs. Overall, the strong role-representation ability of LLMs provides a foundation for constructing complex community-level simulations.

**Legal Intelligence.** Legal AI applies artificial intelligence to enhance or automate various tasks in the legal domain, such as legal information extraction (Yao et al., 2022a; Shen et al., 2024), judgment prediction (Huang et al., 2021), and legal question answering (Kien et al., 2020; Zhong et al., 2020). Early approaches (Zhong et al., 2020; Ji et al., 2021) rely on task-specific models and curated datasets for specific legal tasks. With LLMs' strong generalizability (Brown et al., 2020), they have shown notable capabilities in the legal domain. For example, Deng *et al.* (Deng et al., 2023) introduced a structured prompting method to enhance judgment prediction. ChatLaw (Cui et al., 2023) and LawLLM (Yue et al., 2023, 2024) enhance multitasking capabilities through training on legal knowledge and instructions. Recent explorations (Yue et al., 2025a; Chen et al., 2024) have employed LLM-driven agents to simulate legal scenarios. However, a key challenge remains: the lack of a unified framework for evaluating agent performance and guiding future progress.

## 6 Conclusion

We present the first dynamic benchmark for assessing legal intelligence, powered by ***J1-ENVS***, an environment system designed to simulate agent interactions within authentic legal scenarios, featuring 508 specific environments across three hierarchical levels and six distinct scenarios; and ***J1-EVAL***, which evaluates complex task performance based on both outcome rationality and procedural justice. Through extensive benchmarking of general-purpose and legal-oriented LLMs, we reveal the deficiencies of existing models in realistic, dynamic legal contexts. This work underscores the necessity for agents with robust interactive capabilities and procedural awareness, while positioning ***J1-ENVS*** as a foundational infrastructure for subsequent data construction and agent training.

## Limitations

Our benchmark focuses primarily on the procedural flow of legal tasks, while in real-world applications, agents may require more complex capabilities to complete tasks effectively, such as statutory retrieval or precedent reasoning. The prerequisite for evaluating such capabilities is the establishment of a dynamic interactive environment. Therefore, we expect our pioneering effort to serve as a necessary starting point for future research on legal agents' integration of these advanced capabilities in dynamic environments. In addition, we concentrate on developing dynamic environments within the Chinese legal system, presenting a comprehensive methodology for reconstructing interactive scenarios based on real legal data. The approach is readily adaptable to other legal systems for future work.

## Ethics Statement

All data used in our study were obtained from publicly accessible sources with sensitive information properly anonymized. ***J1-ENVS*** and ***J1-EVAL*** are designed as simulation and evaluation tools to enhance legal professionals' capabilities and advance our understanding of legal AI systems. We acknowledge several important ethical considerations. First, while ***J1-ENVS*** demonstrates promising results in simulating legal scenarios, it is not intended to replace human legal professionals or make actual legal decisions. The system should be used as a supplementary tool for legal training and research purposes only. Second, the legal knowledge and strategies learned by our system should not be misused for generating deceptive legal arguments. Furthermore, we emphasize that the system requires careful review and validation by qualified legal professionals before any practical application. The system's responses should not be considered as formal legal advice without proper human oversight.

## Acknowledgments

We thank the anonymous reviewers and the area chair for their constructive comments. The research is supported by the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization(Grant No.2025-GZL-RGZN-BTBX-02028).The project's computational resources are partially supported by CFFF platform of Fudan University.

## References

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

Anthropic. 2023. The claude 3 model family: Opus, sonnet, haiku.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

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.

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

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

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*.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367.

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*.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2023. Laiw: a chinese legal large language models benchmark. *arXiv preprint arXiv:2310.05620*.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal

judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009, Singapore. Association for Computational Linguistics.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962.

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*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yunyun Huang, Xiaoyu Shen, Chuanyi Li, Jidong Ge, and Bin Luo. 2021. Dependency learning for legal judgment prediction with a unified text-to-text transformer. *arXiv preprint arXiv:2112.06370*.

Lu Ji, Zhongyu Wei, Jing Li, Qi Zhang, and Xuanjing Huang. 2021. Discrete argument representation learning for interactive argument pair identification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5467–5478.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023b. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, and 1 others. 2025b. Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*.

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, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. 2025. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 4975–5001.

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

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Chenchen Shen, Chengwei Ji, Shengbin Yue, Xiaoyu Shen, Yun Song, Xuanjing Huang, and Zhongyu Wei. 2024. Empowering llms for long-text information extraction in chinese legal documents. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 457–469. Springer.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

Libo Sun, Siyuan Wang, Xuanjing Huang, and Zhongyu Wei. 2024. Identity-driven hierarchical role-playing agents. *arXiv preprint arXiv:2407.19412*.

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*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv e-prints*, pages arXiv–2412.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022a. Leven: A large-scale chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and 1 others. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Shengbin Yue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuan-Jing Huang, and Zhongyu Wei. 2025a. Multi-agent simulator drives language models for legal intensive interaction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6537–6570.

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, and 1 others. 2024. Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval. In *International Conference on Database Systems for Advanced Applications*, pages 304–321. Springer.

Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2025b. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25796–25804.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, and 1 others. 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*.

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

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

## Content of Appendix

In this paper, we introduce ***J1-ENVS*** and ***J1-EVAL*** to encourage research on interactive legal intelligence. We showed that our framework can be used to evaluate interactive legal intelligence. In the appendix, we provide the following items for a detailed illustration:



# A Role Setting Details

## A.1 Legal Sources Processing

We utilize two primary legal sources to construct role-specific settings.

**Chinese Judgment Documents** are official court records of judicial proceedings, outcomes, and factual details. We collect both civil and criminal judgment documents, and employ GPT-4o to extract legal elements. For civil judgment documents, the following legal elements are extracted.

- **Personal information** covers both individuals and legal entities. For individuals, we collect name, gender, ethnicity, birthdate, and residential address. For legal entities (corporations), personal information includes the corporate name, registered address, and the identity of the legal representative.
- **Case information** include plaintiff claims, case detail, defence strategies, and evidences.
- **Court Information** includes the final judgment, third-party findings, and legal provisions applied by the court in the case.

For criminal judgment documents, the following legal elements are extracted:

- **Personal information** include the defendant's name, ethnicity, address, gender, birth date, occupation, level of education, and custody status.
- **Plea** captures the stated stance of the defendant and the lawyer towards the accusations.
- **Court information** contains the final verdict, court findings, and legal provisions cited by the judge.
- **Procurator information** comprises formal charges and procurator opinions.

**Chinese Legal Articles** are authored by legal experts to explain legal concepts, discuss controversial topics, and address issues in legal practice.

- ***Civil Trial Practice Q&A*** **literature** is a curated collection of legal articles on civil trials practice, each offering an in-depth analysis of a specific legal issue. Leveraging GPT-4o, we identify lay person with relevant occupations, and construct a topic list in the format of <topic, answer> pair. The topics are then summarized into a shared theme.
- ***HUALV*** **website** includes frontier legal articles, where each article is framed as a specific legal question accompanied by a detailed background description. We leverage GPT-4o to divide the article into two parts: the background part and the explanation part. The explanation part is utilized to construct <topic, answer> pairs.

To ensure diversity, we categorize topics into two categories: Binary(i.e., yes/no) topics and open-ended questions, where responses are expected to be either statutory provisions or legal phrases. Comprehensive statistics related to these topics are provided in Sec. B.

## A.2 Personality Theory Construction

Real-world clients exhibit behavioral styles according to occupation and experiences. Previous research(Yue et al., 2025b; Chen et al., 2024) has demonstrated that LLM-driven agents could effectively follow predefined traits, which motivates our design. Therefore, we model the personality of lay person based on the Big Five Personality Traits theory, including five dimensions: extraversion, neuroticism, openness, agreeableness, and conscientiousness. Each dimension is divided into three levels (high, medium, and low). We leverage GPT-4o to assign a level based on the individual information of lay person. Specifically, in **(1) Knowledge Questioning**, occupations are primarily exploited

| Level | Level I | | Level II | | | Level III |
|---|---|---|---|---|---|---|
| **Env.** | KQ | LC | CD | DD | CI | CR |
| **Count** | 98 | 62 | 93 | 93 | 93 | 69 |
| **Source** | Legal Article | | Civil Documents | | | Criminal Documents |

Table 7: Data statistics of ***J1-EVAL***.

as individual information. For **(2) Consultation**, background descriptions are leveraged to generate behavioral styles. For **(3) Complaint Drafting**, case details are utilized to construct behavioral styles. In **(4) Defence Drafting**, personality modeling is based on the defendant's statement of defence. Dimensions that are initially assigned to the medium level are randomly redistributed across the three levels following a 2:1:2 ratio. Based on the five-dimensional traits, GPT-4o is leveraged to generate consistent behavioral styles.

### A.3 Role Profile Configuration

Based on the functional distinctions, we configure two types of roles, each associated with different sets of legal elements.

***Type I: Lay Person*** represent non-professional participants in legal scenarios, including general public in Level-I, plaintiff & defendant in Level-II, and the defendant in Criminal Court:

- **General Public** denotes clients of various professional and backgrounds seeking legal counseling on a particular theme(Env.1) or a specific case(Env.2). For Env.1, we assign each role agent a topic list together with a shared theme, and equip them with behavioral styles derived from the personality traits described in Sec.A.2. For Env.2, detailed background information is additionally provided.

- **Specific Characters** denote the plaintiff and the defendant in Level-II document drafting scenarios. We provide personal attributes (*e.g.*, name and gender) as well as case-related information. Specifically, plaintiffs are assigned claims and supporting evidence, while defendants are provided with statements of defence and corresponding evidence. Behavioral traits are further incorporated to enhance the realism of simulation.

- **Defendant** denotes the accused in Criminal Court. The role is provided with case details and a stated stance.

***Type II: Legal Professionals*** represent trained practitioners in legal scenarios, including the lawyer agent in Level-II, the plaintiff's & defendant's lawyer, the procurator, and the judge agent in Level-III,

- **Lawyer agent** targets gathering information from the plaintiff/defendant, and drafting complaints/defences accordingly. The agent is provided with predefined legal document formats. Additionally, the lawyer agent in Env.4 is supplemented with complaint information to support defence drafting.

- **Plaintiff's & defendant's lawyers** are provided with client(Plaintiff & defendant) information. The plaintiff's lawyer is equipped with the client's claims, while the defendant's lawyer is assigned defence strategies.

- **Procurator** is assigned charges and opinions.

- **Judge Agent** targets predicting judgments while following predefined court procedures. The agent is configured with procedural information and neutral third-party findings.

### A.4 Pseudo-code

The simulation pseudocodes are described in algorithm 1, algorithm 2, and algorithm 3.

## B Data Statistics

As shown in Table 7, we meticulously select 322 quality data in ***J1-Eval***, with 160 pieces of legal articles, 93 civil judgment documents(shared among CD, DD, and CI), and 69 criminal judgment documents, totaling 508 instances. Legal articles are then utilized for agent configurations in Knowledge Questioning and Legal Consultation, with 98 pieces for Knowledge Questioning and 62 pieces for Legal Consultation. Civil judgment documents are utilized for Complaint & Defence Drafting, as well as the Civil Court. Criminal judgment documents are employed for Criminal Court setups. Therefore, we include 508 cases in ***J1-Eval***.

**Algorithm 1** KQ/LC Scenario Simulation

**Input:** Legal agent $A_l$, General public agent $A_g$ with topic set $T$, Supervisor $S$, maximum dialogue turns $N$
**Output:** Dialogue history $H$

Initialize dialogue history $H \leftarrow \emptyset$
Initialize unaddressed topic list $U \leftarrow T$
Initialize turn counter $k \leftarrow 0$
**while** $k < N$ **and** $U \neq \emptyset$ **do**
    $k \leftarrow k + 1$
    $q_k \leftarrow A_g.\text{ASK}(U)$
    Append $q_k$ to $H$
    $a_k \leftarrow A_l.\text{ANSWER}(q_k)$
    Append $a_k$ to $H$
    $S$ updates $U$ based on $(q_k, a_k)$
**end while**
**return** $H$

**Algorithm 2** CD/DD Scenario Simulation

**Input:** Lawyer agent $A_l$, Specific character agent $A_c$, legal document template $\mathcal{T}$, maximum dialogue turns $N$
**Output:** Generated legal document $D$

Provide template $\mathcal{T}$ to lawyer agent $A_l$
Initialize dialogue history $H \leftarrow \emptyset$
Initialize turn counter $k \leftarrow 0$
**while** $k < N$ **do**
    $k \leftarrow k + 1$
    $q_k \leftarrow A_l.\text{ASK}(H, \mathcal{T})$
    Append $q_k$ to $H$
    **if** $q_k$ indicates end-of-dialogue **then**
        **break**
    **end if**
    $a_k \leftarrow A_c.\text{ANSWER}(q_k)$
    Append $a_k$ to $H$
**end while**
$D \leftarrow A_l.\text{GENERATEDOCUMENT}(\mathcal{T}, H)$
**return** $D$

In addition, for topic lists in Knowledge Questioning, 286 binary topics and 188 open-ended topics are designed, with an average of 4.84 topics for each case. For topic lists in Legal Consultation, we construct 116 binary topics and 107 open-ended topics, with an average of 3.60 topics for each of the cases. The balanced percentage and diversity of different topic types have ensured a comprehensive evaluation of the model-driven agent's performance.

**Algorithm 3** CI/CR Scenario Simulation

**Input:** Judge agent $A_j$, participant agent set $\mathcal{A}$, court procedure $\mathcal{P}$, maximum dialogue turns $N$
**Output:** Dialogue history $H$, predicted judgment $V$

Provide court procedure $\mathcal{P}$ to judge agent $A_j$
Initialize dialogue history $H \leftarrow \emptyset$
Initialize turn counter $k \leftarrow 0$
**while** $k < N$ **do**
    $k \leftarrow k + 1$
    $u_k \leftarrow A_j.\text{SPEAK}(\mathcal{P}, H)$
    Append $u_k$ to $H$
    **if** $u_k$ indicates end-of-trial **then**
        **break**
    **end if**
    Identify addressee role $r_k$ from $u_k$
    Select agent $A_r \in \mathcal{A}$ with role $r_k$
    $v_k \leftarrow A_r.\text{RESPOND}(u_k)$
    Append $v_k$ to $H$
**end while**
$V \leftarrow A_j.\text{PREDICTJUDGMENT}(H)$
**return** $(H, V)$

## C Evaluation Metric

We construct a comprehensive evaluation framework that assesses the performance of the legal agent from multiple perspectives with a set of metrics.

### C.1 Level-I

We leverage an LLM-based check mechanism to label each interaction round with its corresponding topic. As illustrated in Sec.A.1, the topics are divided into binary type and open-ended type, which are evaluated for **Binary accuracy** and **Non-binary score** respectively. To be specific:

- **Binary accuracy** measures accuracy in binary questions via exact matching. By comparing the summarized yes/no answer produced by the checking mechanism with the ground truth, the score is assigned as 1 for correct responses, 0 for incorrect responses, and 0.5 for conditionally acceptable answers.

- **Non-binary score** evaluates open-ended responses of the legal agent. When the ground truth consists of statutory provisions, we apply a matching-based evaluation, assigning a

| Agent | Complaint Drafting | | | | |
|---|---|---|---|---|---|
| | II-2 | | | | |
| | PLA | DEF | CLA | EVI | F&R |
| Multilingual LLMs(close source) | | | | | |
| GPT-4o | 92.7 | **96.3** | 97.3 | 77.7 | 85.3 |
| Claude-3.7 | 92.0 | 93.1 | 96.2 | **85.2** | 79.6 |
| Multilingual LLMs(open source) | | | | | |
| Deepseek-v3 671B | 71.4 | 87.7 | 67.3 | 35.2 | 56.5 |
| Deepseek-r1 671B | 75.3 | 71.0 | 44.3 | 11.7 | 29.8 |
| Llama3.3-it. 70B | 78.9 | 93.1 | 98.0 | 66.7 | **86.0** |
| Qwen3 32B | 74.0 | 94.3 | **98.5** | 76.0 | 79.4 |
| Gemma3-it. 27B | 86.5 | 94.4 | 98.4 | 74.7 | 70.1 |
| Qwen3 14B | **94.0** | 94.5 | 97.1 | 79.7 | 80.5 |
| Gemma3-it. 12B | 80.4 | 91.5 | 96.9 | 73.8 | 78.2 |
| GLM4 9B | 80.8 | 83.9 | 89.4 | 73.6 | 83.8 |
| Qwen3 8B | 76.3 | 77.1 | 81.0 | 60.9 | 68.3 |
| InternLM3-it. 8B | 77.4 | 52.1 | 79.6 | 77.0 | 74.8 |
| Ministral-it. 8B | - | 14.2 | - | - | - |
| Qwen2.5-it. 7B | 84.2 | 94.1 | 91.8 | 52.3 | 84.7 |
| Qwen3 4B | 89.3 | 91.1 | 94.8 | 76.1 | 67.00 |
| Legal-specific LLMs | | | | | |
| LawLLM 13B | 0.2 | 2.1 | 5.0 | 4.0 | 3.0 |
| Chatlaw2 7B | 6.9 | 16.7 | 23.3 | 15.0 | 23.0 |

Table 8: Level II: **II-2** in Complaint Drafting. where PLA, DEF, CLA, EVI and F&R denote defendant information, claims, evidence, and fact & reason respectively.

| Agent | Defence Drafting | | |
|---|---|---|---|
| | II-2 | | |
| | RES | DEF | EVI |
| Multilingual LLMs(close source) | | | |
| GPT-4o | **92.5** | **86.6** | 75.3 |
| Claude-3.7 | 79.9 | 85.3 | 41.1 |
| Multilingual LLMs(open source) | | | |
| Deepseek-v3 671B | 73.9 | 74.3 | **80.5** |
| Deepseek-r1 671B | 73.6 | 53.7 | 16.5 |
| Llama3.3-it. 70B | 71.0 | 71.9 | 44.1 |
| Qwen3 32B | 73.4 | 79.3 | 37.7 |
| Gemma3-it. 27B | 78.9 | 80.2 | 47.3 |
| Qwen3 14B | 77.1 | 76.2 | 33.8 |
| Gemma3-it. 12B | 68.3 | 83.7 | 39.6 |
| GLM4 9B | 69.8 | 61.6 | 25.2 |
| Qwen3 8B | 25.7 | 27.9 | 15.0 |
| InternLM3-it. 8B | 47.4 | 70.9 | 43.2 |
| Ministral-it. 8B | 14.1 | 20.2 | 7.6 |
| Qwen2.5-it. 7B | 69.3 | 73.6 | 36.3 |
| Qwen3 4B | 68.8 | 56.1 | 29.7 |
| Legal-specific LLMs | | | |
| LawLLM 13B | - | - | - |
| Chatlaw2 7B | 1.1 | 1.6 | - |

Table 9: Level II: **II-2** in Defence Drafting, where RES, DEF and EVI denote respondent, defence claims, evidence respectively.

score of 1 if the relevant provisions are correctly cited and 0 otherwise. For ground truths involving legal phrases, we leverage GPT-4o to produce a score ranging from 0 to 10, which is normalized to the interval [0,1].

### C.2 Level-II

We evaluate the legal document drafted by the legal agent from two perspectives.

**(1) Format-following score** measures the alignment with the given format. We break down the format into a set of labels $L = [L_1, L_2, ..., L_k]$, which are organized in a specific sequential order $S$. As shown in Figure 11, the complaint format includes labels of *Plaintiff*, *Defendant*, *Claims*, *Facts and Legal Grounds*, and *evidence and Sources, Names and Addresses of Witness*. As shown in Figure 12, the defence format includes labels of *Defendant*, *defence arguments*, *case id*, *parties and cause of action*, *evidence and Sources, names and addresses of witnesses*. Through matching, we check whether the generated sequential order $S_{agent}$ follows the given order $S$, and score the accuracy of labels by comparing the generated labels $L_{agent}$ with $L$. The final result is calculated with the following equation.

$$FOR = Seq * Label \quad (1)$$

Where: $FOR$ is **II-1**; $Seq$ is 1 if $S_{agent}$ is identical with $S$, and 0 if not; $Label$ is the accuracy of labels $S_{agent}$.

**(2) Document score** evaluates the accuracy of the information in the drafted documents. For the plaintiff and the defendant(*i.e., name, ethnicity, address, birthdate, gender*), we apply the exact match method(1 for correct and 0 for incorrect). For other components(i.e., claims, evidences, statement of defence, and case details), we apply the LLM-as-Judge method.

### C.3 Level-III

We evaluate the legal agents' performance in Civil Court and Criminal Court from four dimensions.

**(1) Procedural-following score** evaluates the completeness of the procedural stages, each with several mandatory actions. A stage is complete ONLY when all mandatory actions are executed. In the civil court setting, proceedings are divided into five stages(*i.e.*, court preparation, court investigation, court debate, court mitigation, and court decision). Descriptions of stages & actions are provided in Figure 13 and Figure 14. In the criminal court setting, proceedings are divided into three stages(*i.e.*, court preparation, court investigation, and court decision). Descriptions of stages & actions are provided in Figure 15. If a legal agent delivers unilateral speech that executes all procedural stages within a single turn, the performance is

| Agent | Civil Court | | |
|---|---|---|---|
| | III-1 | | |
| | STA | ACT | UNI |
| Multilingual LLMs(close source) | | | |
| GPT-4o | **85.8** | **88.2** | **89.2** |
| Claude-3.7 | 65.3 | 69.1 | 68.8 |
| Multilingual LLMs(open source) | | | |
| Deepseek-v3 671B | 36.6 | 46.5 | 37.6 |
| Deepseek-r1 671B | - | 0.2 | 1.1 |
| Llama3.3-it. 70B | 53.8 | 78.9 | 86.0 |
| Qwen3 32B | 49.7 | 57.5 | 60.2 |
| Gemma3-it. 27B | 68.5 | 75.4 | 67.7 |
| Qwen3 14B | 11.6 | 19.6 | 25.8 |
| Gemma3-it. 12B | 36.8 | 52.2 | 54.8 |
| GLM4 9B | 11.0 | 14.3 | 18.3 |
| Qwen3 8B | 9.1 | 10.6 | 28.0 |
| InternLM3-it. 8B | 25.5 | 38.1 | 47.3 |
| Ministral-it. 8B | 18.8 | 26.6 | 32.3 |
| Qwen2.5-it. 7B | 11.0 | 22.7 | 32.3 |
| Qwen3 4B | 26.6 | 33.4 | 38.7 |
| Legal LLMs | | | |
| LawLLM 13B | 0.5 | 1.5 | 4.3 |
| Chatlaw2 7B | 0.8 | 2.2 | - |

Table 10: Level III: **III-1** in Civil Court, where STA and ACT denote stages and actions respectively, and UNI is the binary indicator.

| Agent | Criminal Court | | | | |
|---|---|---|---|---|---|
| | III-1 | | | III-4 | |
| | STA | ACT | UNI | SEN | FINE |
| Multilingual LLMs (closed source) | | | | | |
| GPT-4o | 46.4 | 52.2 | 66.7 | 38.5 | 38.5 |
| Claude-3.7 | 47.8 | 59.6 | **95.7** | **21.6** | **9.8** |
| Multilingual LLMs (open source) | | | | | |
| Deepseek-v3 671B | 39.1 | 45.6 | 78.3 | 34.9 | 32.9 |
| Deepseek-r1 671B | 2.2 | 2.6 | 1.4 | 98.6 | 98.6 |
| Llama3.3-it. 70B | 44.9 | 56.5 | 85.5 | **21.6** | 28.4 |
| Qwen3 32B | 25.4 | 28.2 | 50.7 | 60.4 | 56.3 |
| Gemma3-it. 27B | **52.9** | **67.1** | 92.8 | 21.9 | 22.3 |
| Qwen3 14B | 15.2 | 18.7 | 29.0 | 73.4 | 73.7 |
| Gemma3-it. 12B | 37.0 | 45.7 | 69.6 | 41.1 | 40.4 |
| GLM4 9B | 21.7 | 28.5 | 39.1 | 76.2 | 73.8 |
| Qwen3 8B | 18.8 | 14.8 | 37.7 | 67.8 | 67.0 |
| InternLM3-it. 8B | 27.5 | 30.8 | 60.9 | 45.4 | 44.9 |
| Ministral-it. 8B | 2.2 | 1.9 | 4.3 | 97.7 | 98.3 |
| Qwen2.5-it. 7B | 21.0 | 31.7 | 52.2 | 56.0 | 53.0 |
| Qwen3 4B | 23.2 | 18.5 | 49.3 | 59.1 | 55.5 |
| Legal LLMs | | | | | |
| LawLLM 13B | 11.6 | 16.1 | 21.7 | 84.3 | 90.6 |
| Chatlaw2 7B | - | - | - | - | - |

Table 11: Level III: **III-2** and **III-4** in Criminal Court, where STA, ACT, SEN, and FINE denote stages, actions, sentences, and fines respectively, and UNI is the binary indicator.

regarded as a failure under our evaluation protocol. The completeness of procedural stages and mandatory actions is evaluated using the LLM-as-Judge method with a checklist-style framework. Specifically, we first segment the dialogue history into procedural units. We then assess each unit with a checklist of binary(Yes/No) questions. To mitigate the risk of keyword hacking, we introduce a general requirement instructing the model to evaluate cases according to their overall semantic content.

**(2) Judgment score** evaluates the quality of the legal agent's judgment. In civil court, we evaluate the judgment based on the semantic alignment with the ground truth. In criminal court, we evaluate the judgment(predicted fines & sentences) with normalized logarithmic deviation. The average of the fine deviation and the sentence deviation is calculated as the final **III-4**.

**(3) Reasoning score** evaluates the quality of the legal agent's reasoning process based on the semantic alignment with the ground truth.

**(4) Law score** evaluates the accuracy of cited legal provisions with exact matching.

## D Implementation Details.

For adapted LLMs, we speed up inference using vllm (Kwon et al., 2023). Greedy decoding was used across the evaluations. For models smaller than 32B, we run evaluations using 8 * RTX 4090 GPUs with 24GB memory each. For models larger than 32B, we run evaluations using 8 * A100 GPUs with 80GB memory. For some metrics calculating in the evaluation, we use GPT-4o [4]. To prevent infinite interaction loops caused by limited model capabilities, we set maximum interaction rounds for each scenario: 15 for knowledge 1 questioning, 10 for legal consultation, 20 for complaint drafting, 15 for defence drafting, 50 for civil court, and 35 for criminal court.

## E Additional Experiment

### E.1 Additional overall performance

We provide additional results of the overall performance among LLM-driven legal agents in ***J1-ENVS***.

- **Complaint & Defence Drafting.** Table 8 and Table 9 report additional **II-2** results in Complaint Drafting and Defence Drafting. Both general and legal-specific models perform relatively well on five sub-metrics of **II-2**, indicating powerful information collection ability through the multi-turn conversation. However, agents driven by smaller or legal-specific models, such as LawLLM, encounter difficulty in accomplishing the assigned tasks, requiring further training.

[4] gpt-4o-2024-11-20

| Model | KQ | LC | CD | DD | CI | CR |
|---|---|---|---|---|---|---|
| *Multilingual LLMs (close source)* | | | | | | |
| GPT-4o | 5.9 | 4.9 | 6.7 | 5.2 | 37.6 | 15.3 |
| Claude-3.7 | 5.9 | 5.0 | 12.8 | 9.2 | 37.5 | 11.6 |
| *Multilingual LLMs (open source)* | | | | | | |
| Deepseek-v3 $_{671B}$ | 6.3 | 4.9 | 13.0 | 4.9 | 11.6 | 7.2 |
| Deepseek-r1 $_{671B}$ | 6.0 | 5.3 | 10.8 | 4.8 | 1.2 | 1.3 |
| Llama3.3-it. $_{70B}$ | 5.9 | 5.2 | 7.0 | 5.0 | 17.5 | 8.9 |
| Qwen3 $_{32B}$ | 5.9 | 4.9 | 9.1 | 7.6 | 33.3 | 21.0 |
| Gemma3-it. $_{27B}$ | 5.9 | 5.1 | 7.7 | 6.1 | 31.5 | 11.4 |
| Qwen3 $_{14B}$ | 6.0 | 5.0 | 6.1 | 6.7 | 32.9 | 16.9 |
| Gemma3-it. $_{12B}$ | 5.9 | 5.2 | 8.7 | 7.1 | 26.5 | 14.5 |
| GLM4 $_{9B}$ | 6.0 | 5.2 | 13.5 | 5.0 | 5.5 | 7.3 |
| Qwen3 $_{8B}$ | 6.1 | 4.8 | 14.3 | 14.1 | 32.3 | 13.1 |
| InternLM3-It. $_{8B}$ | 5.9 | 5.6 | 2.7 | 4.3 | 1.1 | 1.4 |
| Ministral-it. $_{8B}$ | 6.2 | 6.1 | 1.0 | 2.4 | 29.4 | 27.1 |
| Qwen2.5-it. $_{7B}$ | 5.9 | 4.9 | 5.6 | 5.2 | 32.4 | 20.0 |
| Qwen3 $_{4B}$ | 6.0 | 5.2 | 8.1 | 8.2 | 24.2 | 15.5 |
| *Legal-specific LLMs* | | | | | | |
| LawLLM $_{13B}$ | 6.1 | 5.4 | 20.0 | 15.0 | 2.5 | 25.3 |
| Chatlaw2 $_{7B}$ | 7.2 | 7.9 | 6.7 | 6.3 | 21.8 | 15.2 |

Table 12: Average turn of interaction among LLM-driven legal agents on six environments.

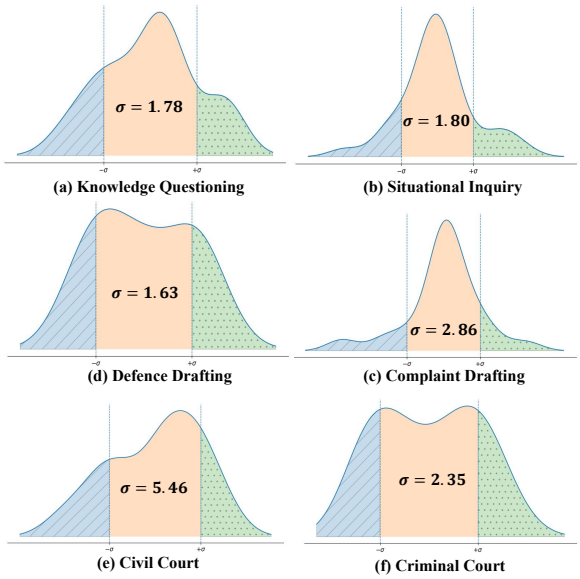


Figure 7: The distribution of the difference between GPT-4o scores and human evaluation across six scenarios

- **Civil Court.** Table 10 reports additional **III-1** results in the Civil Court. Claude-3.7 and Gemma3-Instruct-27B demonstrate remarkable performance in **STA** and **ACT**, indicating strong procedural-following ability of these models, while legal-specific models, the inference model, and smaller models are unable to follow civil court procedures.
- **Criminal Court.** Table 11 reports additional **III-1** and **II-2** results in the Criminal Court.Claude-3.7 score 95.7 in **UNI**. In contrast, legal-specific models fail to navigate the criminal court, resulting lower **UNI** and poorer performance.

### E.2 Analysis of interaction turn.

Table 12 presents the average number of interactions observed across the six environments. Conversely, agents that fail to complete the task tend to have few turns or directly reach the maximum turn limit. For instance, Deepseek-r1, ChatGLM, InternLM3, and LawLLM achieve only 1.2, 5.5, 1.1, and 2.5 turns on average at CI. The results indicate that weak interactive abilities prevent the models from sustaining focused, purposeful dialogues, often leading to off-topic, incomplete, or procedurally flawed interactions.

### E.3 Definition of failed causes

- **Single-turn Collapse**. The judge agent completes all stages & actions within a single turn, which is inconsistent with the purpose of dynamic simulation.
- **Missing Final Objective**. The judge agent fails to generate judgments at the end of the dialogue,
- **Missing Addressee**. The judge agent fails to explicitly contain addressee in an utterance.
- **Role Omission**. The judge agent overlooks one or more environmental roles during the simulation.

## F Human Evaluation

### F.1 Implementation Details

We invite a law professor to assist in recruiting seven evaluators. These evaluators have received systematic legal training and hold master's or higher degrees in law. The human evaluation process is conducted with a dedicated annotation system. To reduce potential bias, all evaluators received identical task instructions and scoring guidelines prior to annotation and completed the evaluation independently. Model identities and experimental settings were concealed, and evaluators made judgments solely based on the content. We randomly sample 20 cases per scenario, and each evaluator independently scored all samples, resulting in a total of 840 human evaluation records. To mitigate the impact of outliers, we discard the maximum and minimum human ratings for each case and compute the average of the remaining scores

| Agent | KQ General Public | LC General Public | CD Plaintiff | DD Defendant |
|---|---|---|---|---|
| Env.GPT-4o | | | | |
| GPT-4o | 86.8 | 80.5 | 90.1 | 86.7 |
| Deepseek-v3 $_{671B}$ | 87.6 | 85.7 | 66.2 | 82.0 |
| Qwen3 $_{32B}$ | 85.1 | 81.5 | 85.2 | 85.2 |
| Env.Qwen3 $_{32B}$ | | | | |
| GPT-4o | 92.6 | 90.3 | 92.8 | 89.1 |
| Deepseek-v3 $_{671B}$ | 94.4 | 89.4 | 86.2 | 88.9 |
| Qwen3 $_{32B}$ | 92.9 | 90.8 | 89.4 | 89.4 |

Table 13: Behavior consistency of various environment roles under GPT evaluation.

| Agent | KQ General Public | LC General Public | CD Plaintiff | DD Defendant |
|---|---|---|---|---|
| Env.GPT-4o | | | | |
| GPT-4o | 77.0 | 73.0 | 78.5 | 76.5 |
| Deepseek-v3 $_{671B}$ | 77.0 | 73.0 | 72.5 | 75.0 |
| Qwen3 $_{32B}$ | 75.0 | 72.0 | 73.5 | 75.5 |
| Env.Qwen3 $_{32B}$ | | | | |
| GPT-4o | 76.0 | 75.0 | 74.0 | 77.0 |
| Deepseek-v3 $_{671B}$ | 75.6 | 75.0 | 74.5 | 79.0 |
| Qwen3 $_{32B}$ | 76.5 | 77.0 | 72.0 | 71.5 |

Table 14: Behavior consistency of various environment roles under human evaluation.

as the final human evaluation score, which is compared against the GPT score.

### F.2 Analysis of behavioral consistency

We perform a comprehensive analysis of the behavioral consistency of environment roles during interactions with three legal agents, leveraging GPT-4o and human evaluators to rate the alignment on a 10-point scale. GPT-4o assessment is performed across all samples, while human evaluators assess 20 samples per scenario. As shown in Table 13 and Table 14, in interactions with various legal agents(GPT-4o, Deepseek-v3, and Qwen3-14B), the performance of all environmental roles remains consistently high and stable. The experiment demonstrates the reliability and effectiveness of environmental roles.

### F.3 Analysis of differences between human evaluation and LLM-as-Judge

We analyze the discrepancy between human scores and GPT-4o scores. As shown in Figure 7, the standard deviations in five scenarios are all below 3.0, while the civil court scenario exhibits a higher standard deviation of 5.46, resulting in higher the overall difference. This larger variance is due to the civil court scenario containing more evaluation metrics and longer textual content, which increases the likelihood of divergence between human judgments and GPT-4o's assessments.

### F.4 Ethics Statement

This work involves human annotators who manually score the outputs of legal agents. We carefully considered ethical issues related to human participation, data usage, and potential harm. All annotators were recruited voluntarily and were informed of the purpose of the study, the nature of the annotation tasks, and the expected time commitment in advance. They participated with informed consent and were allowed to withdraw at any time. The annotation tasks involved evaluating model-generated legal outputs and did not require annotators to provide personal information or disclose sensitive data. Annotators were fairly compensated for their work in accordance with local standards. No vulnerable populations were involved in the study.

## G Case Study

To present the performance of legal agents, we conduct a qualitative study in J1-Envs and J1-Eval, with GPT-4o acting as the legal agent.

- **An example of CI**. As illustrated in Figure 8, GPT-4o utilizes the control sentence(highlighted in blue) to designate the target of each dialogue turn while executing procedural steps (highlighted in red). After 32 turns, the legal agent concludes the court session and deliver the final decision. Despite having accomplished most stages and actions, GPT-4o occasionally fails to generate a proper control sentence(T6), resulting in a loop.

- **Evaluation of the model judgment**. We further evaluate the quality of the model judgment in Figure 8. As shown in Figure 9, compared to the ground truth, GPT-4o cites only basic provisions of the Civil Code while omitting relevant interpretive clauses issued by the Supreme Court, which offer detailed guidance, leading to misjudgments.

- **Evaluation of the model-drafted complaint**. As shown in Figure 10, GPT-4o often rigidly reproduces format labels without proper modification. In 31.18% of Complaint Drafting cases, GPT-4o preserve the content within parenthese. In 6.45% of cases, GPT-4o incorrectly split the format label "Evidence and Source of Evidence; Witness and Their Addresses" into two distinct components.

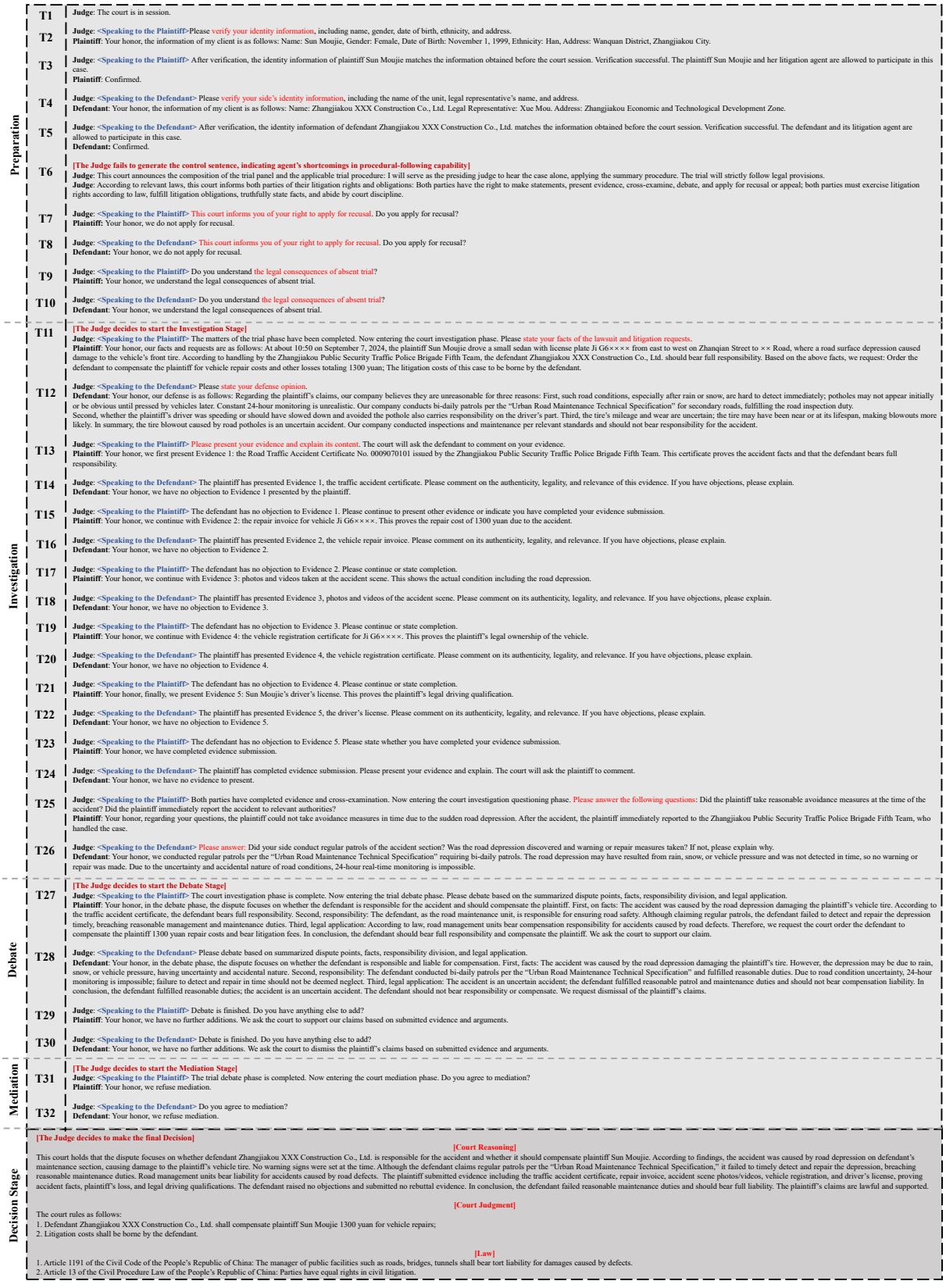


Figure 8: Qualitative result of GPT-4o as the judge in Civil Court. T$i$ denotes the $i$-th interaction turn. Blue denotes the control sentence utilized by the Judge to denote the target, while red denotes preset procedures

| Ground Truth | GPT-4o Judgment |
|---|---|
| **[Court Reasoning]**<br>Where a person infringes upon the civil rights and interests of another due to fault, they shall bear tort liability. If the infringed party is also at fault for the occurrence or aggravation of the damage, the tortfeasor's liability may be mitigated accordingly. In the present case, the defendant, as the entity responsible for the management and maintenance of the road section where the incident occurred, was obligated to fulfill its duty to ensure that the road remained in a condition suitable for normal traffic. However, no warning signs were placed at the accident site, resulting in the plaintiff's vehicle's right front wheel hitting a depression in the road surface, which led to a tire blowout. This indicates that the defendant failed to perform its management and maintenance obligations. Therefore, the defendant shall bear liability for the plaintiff's legitimate and reasonable losses. Nonetheless, the plaintiff also bears certain fault for the occurrence of the accident, as they failed to exercise reasonable caution, and should accordingly assume a proportion of responsibility. Regarding the defendant's claim that the blowout might have been caused by factors such as the mileage or wear of the plaintiff's tire, since the defendant failed to provide sufficient evidence in support of this claim, the court does not accept this line of defense.<br>**[Court Judgment]**<br>The court rules as follows:<br>1. The defendant, Zhangjiakou XX Construction Co., Ltd., shall, within five days from the effective date of this judgment, compensate the plaintiff, Sun Moujie, in the amount of **RMB 1,040** for tire repair expenses.<br>2. **If the monetary obligation is not fulfilled within the period specified in this judgment, interest on the delayed payment shall be doubled for the period of delay, in accordance with Article 264 of the Civil Procedure Law of the People's Republic of China.**<br>3. The case acceptance fee of RMB 50 shall be halved, totaling RMB 25, which shall be borne by the defendant, Zhangjiakou XX Construction Co., Ltd.<br>4. **This judgment is final.**<br>**[Law]**<br>1. **Article 1165, 1173, 1184** of the Civil Code of the People's Republic of China<br>2. **Article 90 of the SPC Interpretation on the Application of the Civil Procedure Law** | **[Court Reasoning]**<br>This court holds that the dispute focuses on whether defendant Zhangjiakou XXX Construction Co., Ltd. is responsible for the accident and whether it should compensate plaintiff Sun Moujie. According to findings, the accident was caused by road depression on defendant's maintenance section, causing damage to the plaintiff's vehicle tire. No warning signs were set at the time. Although the defendant claims regular patrols per the "Urban Road Maintenance Technical Specification," it failed to timely detect and repair the depression, breaching reasonable maintenance duties. Road management units bear liability for accidents caused by road defects. The plaintiff submitted evidence including the traffic accident certificate, repair invoice, accident scene photos/videos, vehicle registration, and driver's license, proving accident facts, plaintiff's loss, and legal driving qualifications. The defendant raised no objections and submitted no rebuttal evidence. In conclusion, the defendant failed reasonable maintenance duties and should bear full liability. The plaintiff's claims are lawful and supported.<br>**[Court Judgment]**<br>The court rules as follows:<br>1. Defendant Zhangjiakou XXX Construction Co., Ltd. shall compensate plaintiff Sun Moujie 1300 yuan**[Error: wrong compensation sum]** for vehicle repairs;<br>2. Litigation costs shall be borne by the defendant.<br>**[Error: omit other judgments]**<br>**[Law]**<br>1. Article 1191**[Error: wrong article]** of the Civil Code of the People's Republic of China: The manager of public facilities such as roads, bridges, tunnels shall bear tort liability for damages caused by defects.<br>2. Article 13 of the Civil Procedure Law of the People's Republic of China**[Error: Incorrect Law]**: Parties have equal rights in civil litigation.<br>**[Error: fails to cite the SPC Interpretation]** |

Figure 9: Qualitative result of GPT-4o judgment in comparison with the ground truth, where errors are highlighted in red, and corresponding correct answers are highlighted in green

| Ground Truth | GPT-4o Judgment |
|---|---|
| **Civil Complaint**<br>**Plaintiff:** Zhang, female, Han ethnicity, residing in Tiexi District, Anshan City.<br>**Defendant:** Zhang, female, Han ethnicity, residing in Tiexi District, Anshan City.<br>**Claims:**<br>1. The plaintiff respectfully requests that the People's Court order the defendant to return the 24K gold necklace, 24K gold bracelet, and 24K gold pendant purchased by the plaintiff, with a total value of RMB 16,328.4.<br>2. The plaintiff further requests that the defendant bear all litigation costs, attorney drafting fees, and photocopying expenses incurred in this case.<br>**Facts and Reasons:**<br>The plaintiff and defendant are friends. On January 2, 2023, the plaintiff purchased a 24K gold necklace, 24K gold bracelet, and 24K gold pendant at Liuguifu Jewelry Store for her granddaughter, with a total value of RMB 16,328.4. After the purchase, the plaintiff asked the defendant to check whether the items looked good and if her granddaughter would like them. The defendant, after inspection, said that these styles were only worn by elderly women and that young people did not wear such designs. The defendant then offered to help the plaintiff exchange them for different styles and took away the necklace, bracelet, pendant, and the invoice. Subsequently, when the plaintiff inquired several times about the exchange, the defendant repeatedly evaded with various excuses. Later, the plaintiff approached the defendant multiple times, but the defendant refused to return the items and repeatedly insulted the plaintiff. The plaintiff was left with no choice but to bring the matter to court, requesting that the court lawfully support the plaintiff's claims.<br>**Evidence and Source of Evidence; Witnesses and Their Addresses:**<br>1. One audio recording.<br>2. one screenshot of the purchase details. | **Civil Complaint**<br>**Plaintiff (Natural Person) [Error: parenthetical should be removed]:** Zhang, female, Han ethnicity, residing in Tiexi District, Anshan City.<br>**Defendant (Natural Person) [Error: parenthetical should be removed]:** Zhang, female, Han ethnicity, residing in Tiexi District, Anshan City.<br>**Claims:**<br>1. The plaintiff requests the court to order the defendant to return the jewelry purchased by the plaintiff, specifically one pure gold necklace, one pure gold bracelet, and one pure gold pendant, with a total value of RMB 16,328.4.<br>2. The plaintiff requests that all litigation costs, attorney's fees (for legal document drafting), and photocopying expenses incurred in this case be borne by the defendant.<br>**Facts and Reasons:**<br>The plaintiff and the defendant are friends. On January 2, 2023, the plaintiff purchased a pure gold necklace, bracelet, and pendant from Liuguifu Jewelry Store, totaling RMB 16,328.4. The plaintiff handed the jewelry to the defendant to examine whether the style was suitable. The defendant claimed she could assist in exchanging the items and took both the jewelry and the receipt. However, to date, the defendant has failed to return the jewelry, repeatedly delaying and even verbally abusing the plaintiff. The plaintiff has made multiple demands for return of the items, all of which have been refused. The plaintiff is therefore filing this lawsuit, requesting the court to safeguard her legitimate rights and interests in accordance with the law.<br>**Evidence and Source of Evidence; Witnesses and Their Addresses:**<br>1. Audio recording: A conversation between the plaintiff and the defendant, demonstrating that the defendant acknowledged possession of the jewelry.<br>2. Screenshot of purchase details: Proof of the plaintiff's jewelry purchase at Liuguifu Jewelry Store, showing the value and ownership of the items.<br>3. No witnesses. **[Error: no need to emphasize]** |

Figure 10: Qualitative result of GPT-4o drafted complaint in CD, where highlighted bold red boxes denote format errors, and green denotes corresponding correct answers

**Complaint Format(*English Version*)**

Civil Complaint

Plaintiff (if an individual): XXX, male/female, born on XXXX-XX-XX, ethnicity: X, residing at XXXXXXX.
Plaintiff (if a legal entity): XXX, Legal Representative: XXX, Address: XXXXXXX.

Defendant (if an individual): XXX, male/female, born on XXXX-XX-XX, ethnicity: X, residing at XXXXXXX.
Defendant (if a legal entity): XXX, Legal Representative: XXX, Address: XXXXXXX.

Claims:
1. Claim 1
2. Claim 2
...

Facts and Legal Grounds:
...

Evidence and Sources, Names and Addresses of Witnesses:
...

**Complaint Format(*Chinese Version*)**

民事起诉状

原告（如果是自然人）：XXX，男/女，XXXX年XX月XX日生，X族，住XXXXXX。
原告（如果是法人）：XXX，法定代表人：XXX，住所：XXXXXX。

被告（如果是自然人）：XXX，男/女，XXXX年XX月XX日生，X族，住XXXXXX。
被告（如果是法人）：XXX，法定代表人：XXX，住所：XXXXXX。

诉讼请求：
1. 诉讼请求1.
2. 诉讼请求2.
...
事实和理由：
...
证据和证据来源，证人姓名和住所：
...

Figure 11: Complaint Format

**Defence Format(*English Version*)**

Civil Defence

Defendant (if a natural person): XXX, male/female, born on XXXX-XX-XX, of X ethnicity, residing at XXXXXXX.
Defendant (if a legal entity): XXX, legal representative: XXX, address: XXXXXXX.

In response to the lawsuit filed with the XXX People's Court, case no. (XXXX) … Min Chu …, involving (parties and cause of action), the defence is as follows:
…… (state defense arguments)

Evidence and sources, names and addresses of witnesses:
……

**Defence Format(*Chinese Version*)**

民事答辩状

答辩人（如果是自然人）：XXX，男/女，XXXX年XX月XX日生，X族，住XXXXXX。
答辩人（如果是法人）：XXX，法定代表人：XXX，住所：XXXXXX。

对XXXX人民法院（XXXX）...民初...号...（写明当事人和案由）一案的起诉，答辩如下：
.....（写明答辩意见）
证据和证据来源，证人姓名和住所：
.....

Figure 12: Defence Format

**Civil Court Stages(*English Version*)**

**Opening of the Hearing(Do not include numerical sequence indicators. Each item must be handled in a separate round. Do not merge multiple items into a single step.)**

1. You shall verify the identity of the plaintiff. If the verification is successful, you shall permit the plaintiff and their litigation representative to participate in the proceedings.
2. You shall verify the identity of the defendant. If the verification is successful, you shall permit the defendant and their litigation representative to participate in the proceedings.
3. You shall announce the composition of the trial panel and the applicable adjudication procedure for the case.
4. You shall inform both parties of their litigation rights and obligations in accordance with relevant laws.
5. You shall inform the plaintiff of their right to request the recusal of judicial personnel and inquire whether they wish to file such a request.
6. You shall inform the defendant of their right to request the recusal of judicial personnel and inquire whether they wish to file such a request.
7. You shall inquire whether the plaintiff understands the legal consequences of a trial conducted in their absence.
8. You shall inquire whether the defendant understands the legal consequences of a trial conducted in their absence.

**Court Investigation(Do not include numerical sequence indicators. Each item must be handled in a separate round. Do not merge multiple items into a single step.)**

1. You shall request the plaintiff to state the facts of the case and the claims being made.
2. You shall request the defendant to present their defense arguments.
3. You shall initiate the evidence production and cross-examination process. Request the plaintiff to present evidence and invite the defendant to express their opinion regarding the plaintiff's evidence. If the defendant raises objections, you shall invite the plaintiff to respond to the defendant's objections. Repeat this step until the plaintiff has presented all evidence. Then request the defendant to present their evidence, allow the plaintiff to respond, and subsequently invite the defendant to comment on the plaintiff's responses.
4. After both parties have completed the production and cross-examination of evidence, you shall pose questions to the plaintiff; After both parties have completed the production and cross-examination of evidence, you shall pose questions to the defendant. At this stage, the parties are not allowed to question each other or to engage in debate. Do not confuse court investigation with court debate.

**Court Debate(Do not include numerical sequence indicators. Each item must be handled in a separate round. Do not merge multiple items into a single step.)**

1. You shall request the plaintiff to present arguments on the summarized issues of dispute, including facts, liability attribution, and the application of law.
2. You shall request the defendant to present arguments on the summarized issues of dispute, including facts, liability attribution, and the application of law.
3. After the debate, you shall inquire whether the plaintiff has anything further to add.
4. After the debate, you shall inquire whether the defendant has anything further to add.

**Court Mediation(Do not include numerical sequence indicators. Each item must be handled in a separate round. Do not merge multiple items into a single step.)**

1. You shall inquire whether the plaintiff agrees to mediation.
2. You shall inquire whether the defendant agrees to mediation.
3. You shall inform the parties that judgment will be rendered at a later date and declare the hearing adjourned.

**Final Judgment(Do not include numerical sequence indicators. Each item must be handled in a separate round. Do not merge multiple items into a single step.)**

1. You shall generate the corresponding reasoning and render the final judgment, accompanied by the applicable legal provisions. The reasoning section must begin with "The Court holds that", and this phrase must appear only once within the entire paragraph of legal reasoning. Conclude your statement with: <End of hearing>

Figure 13: Civil Court Stages and Actions(*English Version*)

**Civil Court Stages(*Chinese Version*)**

一、<开庭审理>阶段，你应依次完成以下事项（不要提及数字序列号，每项单独成轮，不能合并）：1. 你核实原告当事人身份，若核实成功则准予原告当事人及诉讼代理人参与诉讼；2. 你核实被告当事人身份，若核实成功则准予被告当事人及诉讼代理人参与诉讼；3. 宣布审判庭的构成以及本案适用的审判程序；4. 依照有关法律告知双方当事人诉讼权利和诉讼义务；5. 告知原告当事人回避的权利并询问是否申请回避；6. 告知被告当事人回避的权利并询问是否申请回避；7. 询问原告是否知晓缺席审理后果；8. 询问被告是否知晓缺席审理后果。

二、<法庭调查>阶段，你应依次完成以下事项（不要提及数字序列号，每项单独成轮，不能合并）：1. 你要求原告陈述起诉事实和诉讼请求；2. 你要求被告陈述答辩意见；3. 你开始举证质证，请原告出示证据并请被告就原告证据发表自己看法，如果被告提出质证意见，需要请原告就被告质证意见发表看法。重复此步骤直至原告出示完所有证据后，再请被告出示证据并由原告质证，之后在请被告就原告质证意见发表看法；4. 双方举证质证后，由你向原告发问；5. 双方举证质证后，由你向被告发问。此阶段不允许双方互问，也不允许双方辩论。不要混淆法庭调查和法庭辩论。

三、<庭审辩论>阶段，你应依次完成以下事项（不要提及数字序列号，每项单独成轮，不能合并）：1. 你要求原告根据归纳好的争议焦点、事实、责任划分和法律适用等展开法庭辩论；2. 你要求被告根据归纳好的争议焦点、事实、责任划分和法律适用等展开法庭辩论；3. 辩论完后，你询问原告是否还有其他想要说的；4. 辩论完后，你询问被告是否还有其他想要说的。

四、<法庭调解>阶段，你应依次完成以下事项（不要提及数字序列号，每项单独成轮，不能合并）：1. 询问原告是否同意调解；2. 询问被告是否同意调解；3. 告知择期宣判，并宣布闭庭。

五、<最终宣判>阶段，你应依次完成以下事项：生成相应的说理内容，给出最终判决，判决后附依据的法条。说理部分应以“本院认为”开头，并且“本院认为”仅在整段推理中出现一次。在最后以“<结束庭审>”结束你的发言。

Figure 14: Civil Court Stages and Actions(*Chinese Version*)

**Criminal Court Stages(*English Version*)**

**(1) <Opening of the Hearing> stage, you shall complete the following items in sequence (do not include numerical sequence indicators, handle each item in a separate round, and do not merge multiple items):**

1. You verify the identity information of the defendant. If the defendant's identity is successfully verified, you permit the defendant and the defense counsel to participate in the litigation.
2. You confirm whether the defendant has any prior criminal records or bad conduct, and you inform the defendant of the charges brought against them.
3. You announce the composition of the trial panel and the adjudication procedure applicable to the case.
4. You inform the defendant, the defense counsel, and the public prosecutor of the right to apply for recusal according to relevant legal provisions, and inquire whether they wish to apply for recusal.
5. You ask the defendant whether they plead guilty and accept punishment. If the defendant pleads guilty, further confirm whether they understand the legal consequences.

**(2) <Court Investigation> stage, you shall complete the following items in sequence (do not include numerical sequence indicators, handle each item in a separate round, and do not merge multiple items):**

1. You request the public prosecutor to read out the facts that have been ascertained and the relevant charges.
2. You hear the defendant's statement on the content of the indictment.
3. You hear the defense counsel's opinion on the content of the indictment.
4. You question the defendant.
5. You request the public prosecutor to present evidence.

**(3) <Final Judgment> stage, you shall complete the following items in sequence:**

1. You generate the corresponding reasoning, render the final judgment, and append the applicable legal provisions after the judgment. The reasoning section must begin with "The Court holds that", and "The Court holds that" must appear only once within the entire reasoning paragraph.

End your statement with "<End of hearing>".

**Criminal Court Stages(*Chinese Version*)**

一、<开庭审理>阶段，你应依次完成以下事项（不要提及数字序列号，每项单独成轮，不能合并）：1. 你核实被告当事人身份信息，若被告当事人身份核实无误，你准予被告当事人与被告辩护人参与诉讼；2. 你明确被告当事人是否有前科劣迹，你告知被指控的罪名；3. 你宣布审判庭的组成及本案适用的审判程序；4. 你依照有关法律规定告知被告当事人、被告辩护人与公诉机关依法申请回避的权利，并询问是否申请回避；5. 你询问被告当事人是否认罪认罚，若认罪，进一步确认其是否清楚法律后果。

二、<法庭调查>阶段，你应依次完成以下事项（不要提及数字序列号，每项单独成轮，不能合并）：1. 你要求公诉机关宣读查明事实、有关指控；2. 你听取被告当事人对起诉内容的陈述；3. 你听取被告辩护人对起诉内容的意见；4. 你对被告当事人进行讯问；5. 你要求公诉机关出示证据。

三、<最终宣判>阶段，你应依次完成以下事项：你生成相应的说理内容，给出最终判决，判决后附依据的法条。说理部分应以"本院认为"开头，并且"本院认为"仅在整段推理中出现依次。在最后以"<结束庭审>"结尾。

Figure 15: Criminal Court Stages and Actions