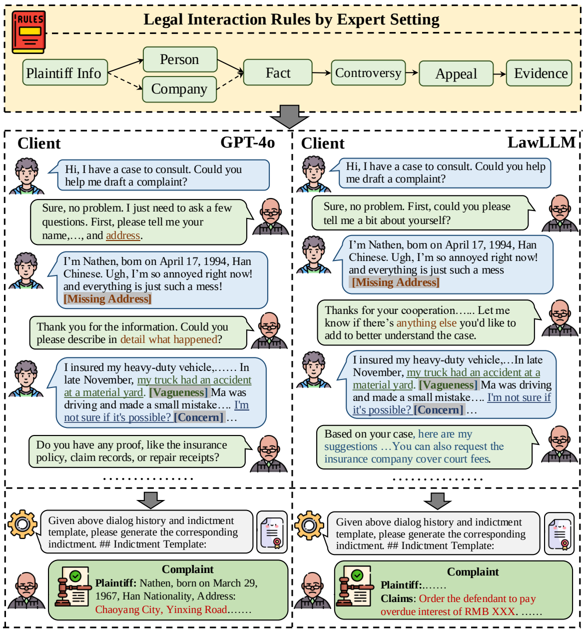
Multi-Agent Simulator Drives Language Models for Legal Intensive Interaction

Shengbin Yue[1†], Ting Huang[1†], Zheng Jia[1†], Siyuan Wang[2], Shujun Liu[1], Yun Song[3*], Xuanjing Huang [1], Zhongyu Wei[1*]
[1]Fudan University, China
[2]University of Southern California, USA
[3]Northwest University of Political and Law, China
{sbyue23,thuang24,zjia24}@m.fudan.edu.cn, zywei@fudan.edu.cn

Abstract

Large Language Models (LLMs) have significantly advanced legal intelligence, but the scarcity of scenario data impedes the progress toward interactive legal scenarios. This paper introduces a Multi-agent Legal Simulation Driver (MASER) to scalably generate synthetic data by simulating interactive legal scenarios. Leveraging real-legal case sources, MASER ensures the consistency of legal attributes between participants and introduces a supervisory mechanism to align participants' characters and behaviors as well as addressing distractions. A Multi-stage Interactive Legal Evaluation (MILE) benchmark is further constructed to evaluate LLMs' performance in dynamic legal scenarios. Extensive experiments confirm the effectiveness of our framework. The detailed resources are available at https://github.com/FudanDISC/MASER.



(a) GPT-4o as Legal Professional **(b) LawLLM as Legal Professional**

Figure 1: Examples of general LLM (*i.e.*, GPT-4o) and legal LLM (*i.e.*, LawLLM) as legal professionals in drafting legal documents. LLMs struggle to maintain flexible interaction patterns under legal agendas.

1 Introduction

The emergence of Legal Artificial Intelligence (Legal AI) (Atkinson et al., 2020; Ge et al., 2021) induces significant transformations within the legal field. Especially, recent advancements in Large Language Models (LLMs) (Achiam et al., 2023; Yao et al., 2024; Yue et al., 2024a), through continuous training on legal knowledge and instructions, have shown remarkable capabilities in various legal tasks, such as legal information extraction (Shen et al., 2024), judgment prediction (He et al., 2024) and instrument drafting (Yue et al., 2023). However, previous endeavors (Fei et al., 2023) predominately concentrated on non-intensive interactive tasks, involving prediction with directly specified elements, thereby restricting the wider real-world application of Legal AI.

The real-world judicial service scenario is highly complex, presenting two major challenges: (1) High professionalism. The rationality of content and the legality of procedures are crucial for legal information services, requiring practitioners to follow a procedural agenda while offering services. (2) Intensive interactivity. Since users often lack basic legal knowledge, legal consultations typically require multiple rounds of interaction, with practitioners guiding users step by step through the process. As shown in Figure 1, drafting a legal complaint is a complex task that demands fulfilling all the listed requirements. However, when the user exhibits incomplete information, vagueness or concerns, both GPT-4o (Achiam et al., 2023) and LawLLM (Yue et al., 2024a) struggle with adaptive handling while simultaneously following agendas. This results in misunderstanding of the user's demands and omission of necessary facts, ultimately

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

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6552–6585
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

inadequately drafting complaints. Current LLMs face challenges in this dynamic legal scenario. A fundamental fact lies in the proprietary nature and scarcity of such data in real legal scenarios, which creates substantial barriers to equipping LLMs with such advanced capabilities.

Therefore, we explore a role-driven scenario simulation framework to construct high-quality data in scale for enhancing legal LLMs. Compared existing role-driven approaches for general contexts (Shao et al., 2023; Yu et al., 2024a,b), two key aspects require consideration to better model real-world legal scenarios. (1) *Character Authenticity*. The legal simulation involves various roles, including legal characters with expert knowledge, solutions and agendas, as well as non-legal characters, who may exhibit diverse behavior styles and backgrounds. Due to LLM's susceptibility to legal illusions, legal configurations of roles (non-legal roles' needs and legal roles' solutions) require consistency with real-word scenarios, ensuring the authenticity and reliability of legal services. (2) *Character-behavior Consistency*. The response behavior of each role should consistently align with their characteristic throughout the interaction. For example, non-legal roles should respond according to their profiles while legal roles should always follow legal knowledge and agendas. Additionally, non-legal roles may present distracting behaviors (forgetfulness or misunderstanding), leading to vague responses, while legal roles must respond and direct professionally.

We present a framework to address these aspects, forming the **Multi-agent Legal Simulation Driver** (**MASER**), where the legal agent (Lawyer) guides and questions the non-legal agent (Client) to complete specific legal tasks in accordance with legal agendas. MASER incorporates two strategies: (1) *Role Agent Presetting*: we map multi-level traits based on a sociological personality theory to ensure diversity in non-legal roles. To ensure legal correspondence, we extracted different elements from judicial documents (which establish a complete logical chain of legal events) to develop the legal traits of the roles. This enables the legal role to have prior knowledge of addressing user demands. (2) *Multi-Agent Legal Simulation*: Beyond the legal agent (Lawyer) and non-legal agent (Client), we introduce a supervisor for character-behavior and distractor alignment at the sentence level, assisting participants in revising their behaviors in each turn. MASER leverages this multi-agent simulator as a synthetic data engine, to drive arbitrary offline LLMs for dynamic legal interaction.

Moreover, we propose a **Multi-Stage Interactive Legal Evaluation** (**MILE**) benchmark, to assess LLM variants' ability to navigate interactive legal tasks. This benchmark is derived from a meticulous collection of high-quality Chinese legal cases, providing real-world and consistent legal configurations. Leveraging GPT-4o to simulate non-legal characters (client), we conduct a thorough evaluation of the performance of LLMs-driven Lawyer within this dynamic and realistic legal interaction environment. A supervisor oversees the client to ensure the consistency of its behavior. The MILE benchmark evaluates the Lawyer along two stages of intensive interaction and goal achievement.

Extensive experiments demonstrate that MASER significantly enhances the performance of arbitrary LLMs on interactive legal tasks. In our framework, trained models surpassed proprietary advanced LLM, *e.g.*, GPT-4o (Achiam et al., 2023), and specialized legal LLM, *e.g.*, LawLLM (Yue et al., 2024a). We envision our framework as a general paradigm that extends more complex and private verticals, bridging the gap between intensive interaction and achieving special goals.

2 Multi-agent Legal Simulation Driver

As shown in Figure 2, MASER achieves legal objectives (*i.e.*, complaint drafting) through intensive interactions, which consist of three agents: a client, a lawyer, and a supervisor. Each character assumes specific roles and responsibilities within the framework. The MASER consists of *Role Agent Presetting* (§2.1) and *Multi-Agent Legal Simulation* (§2.2), ultimately producing synthetic scene data to train arbitrary LLMs (§2.3).

2.1 Role Agent Presetting.

Due to limited expertise, vanilla LLM-driven simulators tend to produce substantial legal illusions, compromising reliability of synthetic data. Hence, the authenticity setting of role agent presents two challenges: the client's diversity and legal consistency among roles. To this end, we develop the individual's features across different dimensions based on the Big-5 Personality Traits. Additionally, we configure various legal elements in real legal events with the same logic chain to different agents.

Setup with Real-world Legal Source. To establish legal correspondence, the intuition is to assign

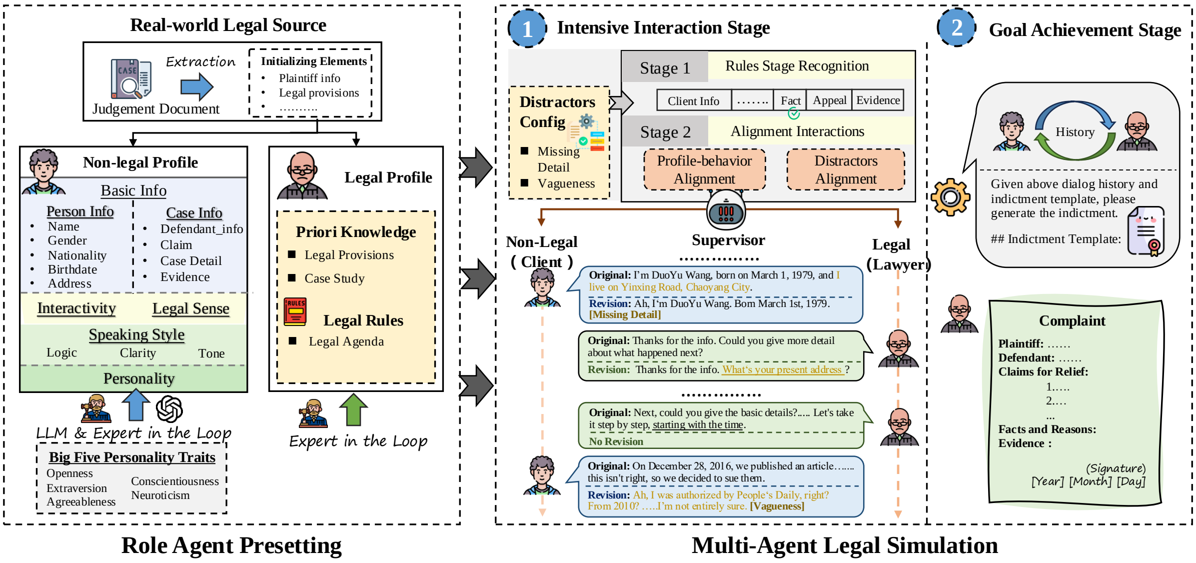


Figure 2: Overview of Multi-agent Legal Simulation Driver (MASER), which consists of role agent presetting and multi-agent legal simulation. Leveraging the MASER, synthesized sentence-level data can drive arbitrary LLMs for legal intensive interaction.

different legal elements of the same legal event to various roles. Chinese judgment documents encompass all the essential components of the entire proceedings of the case, *e.g.*, parties' personal details, legal disputes, facts, evidence, rulings, and applicable laws, which align perfectly with our requirements. We delicately capture these legal elements from documents as legal initialization attributes for lawyers and clients. Specifically, we collect 4,532 civil cases from pre-2021 judgment documents, covering more than 230 civil case types. Then, we prompt GPT-4o to extract six initialization elements from Chinese judgment documents: plaintiff information, defendant information, claim, case details, evidence, case study, and legal provisions. Details are provided in the Appendix A.1.

Setup with Big-5 Personality Traits. Behavioral diversity facilitates the model's ability to comprehend and represent the complexity of characters, further improving simulation reliability and generalization. Guided by social personality theory (Sun et al., 2024), Big-5 personality traits are mapped to two dimensions of behavioral attributes: speaking style and interactivity, where speaking style consists of logic, clarity and tone. In our implementation, three levels (high, medium, or low) are assigned to each of the big-5 personality traits, then we prompt GPT-4 to generate personality descriptions and further develop speaking style and interactivity. Details are shown in the Appendix A.3.

Roles Profile Configuration. As shown in Figure 2, these categories of elements from the above steps are assigned to the corresponding agents. Specifically, (1) *Client* is equipped with personal and case information from real-world legal sources, along with personality, speech styles, and interactivity from Big-5 personality traits. Additionally, five levels of legal sense are manually designed by legal experts, aiming to model the level of legal knowledge. (2) *Lawyer* is configured with case analysis and applicable laws from the same legal source, which is the prior knowledge of addressing the client's demands. In addition, legal agendas are manually designed by legal experts. (3) *Supervisor* owns all the information and mounts it on demand, *e.g.*, case information and personality features are configured when supervising Client's behavior.

2.2 Multi-Agent Legal Simulation.

As shown in Figure 2, the lawyer guides the client to collect legal needs, ultimately completing the user's legal task goal (*i.e.*, complaint drafting). The Supervisor oversees interactions between Client and Lawyer at the sentence level, guaranteeing profiles-behavior alignment and distractor consistency. The simulation consists of intensive interaction and goal achievement stages. We leverage powerful LLM (*e.g.*, GPT-4o) to power each agent, enabling them to embody their roles authentically.

Client Agent Behavior. The Client aims to exhibit a set of realistic behavior patterns to enhance the authenticity of the legal simulation: 1) *Cooperation*. The agent should respond to the lawyer's inquiries, and have subjective biases in describing their condition according to setting legal sense; 2) *Communication*. The agent possesses an individualized speaking style, combining logic, clarity, and emotional response; 3) *Curiosity*. The agent should express concerns based on their level of understanding and interactivity, seeking clear explanations from the lawyer to address their doubts; 4) *Distraction behavior*. The agent's response exhibits two types of distractors, missing some important details or vagueness, requiring further confirmation to clarify. The prompt is shown in Figure 12.

Lawyer Agent Behavior. The Lawyer agent aims to emulate a skilled and patient legal service provider in real-world practice: 1) *Agenda compliance*. Throughout the interaction, the agent must adhere to the given legal rule and agenda to guide Client, ensuring the acquisition of all elements necessary to achieve the legal objectives. 2) *Flexible reaction*. The agent uses the configured prior knowledge of the case to address the user's confusion expertly, and when the user exhibits distracting behavior, further inquiry is required to ascertain the facts. The Lawyer's prompt is shown in Figure 14.

Supervisor Agent Behavior. The Supervisor agent oversees multi-turn conversations between the Client and the Lawyer at the sentence level, improving their consistency and interactivity: 1) *Profile-behavior alignment*. The agent assesses the consistency between each participant's profile and their speech, in conjunction with providing advice based on diverse styles of the Client and legal agendas of the Lawyer. 2) *Distractors alignment*. Relying solely on profiles to prompt the Client's distracted behavior and lawyers' flexible reaction is difficult. Thus, the agents use preset distractor configurations to guide participants' interactions under distraction, *e.g.*, the Client missing details and the Lawyer queries again. The Supervisor first determines the current agenda stage by identifying conversations, and subsequently implementing appropriate supervisory measures according to the stage. The supervisor provides feedback and modification suggestions for the Client or Lawyer to adjust their responses with natural language. The prompts are shown in Figure 15, 16 and 18.

Legal Simulation Flow. As shown in Figure 2, our framework simulates a realistic complaint drafting process through a structured multi-turn interaction flow. The Client initiates the conversation based on the configured real-case cause. The lawyer agent acts as the dialogue facilitator, engaging in interactions with the Client directed by agendas. Across interactions, the Supervisor interacts only with the current speaker in each turn. As shown in Algorithm 1, in t-th round, after Client $\mathcal{C}$ (or Lawyer $\mathcal{L}$) generates a response r^c_{before}, the Supervisor $\mathcal{S}$ provides suggestions r^s based on the previous dialogue h_{t-1} and Client's profile P_c. If r^s is deemed "correct," r^c_{before} is added to memory; Otherwise, the Client (or the Lawyer) revises their response r^c_{after} according to the Supervisor's suggestions r^s before adding it to memory. The interaction ends when $\mathcal{L}$ reaches the inquiry end or the predefined maximum turn m. Ultimately, $\mathcal{L}$ generates the complaint y based on the complete history h_m and template y_{temp}, simultaneously up-

Algorithm 1 Simulation Process

Require: Lawyer $\mathcal{L}$, Supervisor $\mathcal{S}$, Client $\mathcal{C}$, Max-turn m, Dialogue history h, Client's response r^c, Lawyer's response r^l, Supervisor's response r^s, Complaint template y_{temp}
Input: Client's profile P_c, Lawyer's profile P_l
Output: History h_e, complaint y

for each t in $0, 1, ..., m$ **do**
  $\mathcal{C}$ generates r^c_{before}
  $\mathcal{S}$ generates r^s given P_c,h_{t-1},r^c_{before}
  if $r^s =$ "correct" **then**
    $\mathcal{C}$ memorizes r^c_{before} in h_{t-1}
  else
    $\mathcal{C}$ generates r^c_{after} given r^s, r^c_{before}
    $\mathcal{C}$ memorizes r^c_{after} in h_{t-1}
  end if
  $\mathcal{L}$ generates r^l_{before} given h_{t-1}
  $\mathcal{S}$ generates r^s given P_c, P_l, h, r^l_{before}
  if $r^s =$ "correct" **then**
    $\mathcal{L}$ memorizes r^l_{before} in h_{t-1}
  else
    $\mathcal{L}$ generates r^l_{after} given r^s, r^l_{before}
    $\mathcal{L}$ memorizes r^l_{after} in h_{t-1}
  end if
  if "Inquiry ends" in r^l **then**
    break
  end if
end for
$\mathcal{L}$ predicts y given h_m, y_{temp}
$h_e = \mathcal{L}$ memorizes y in h_m

dating it into h_m to form scenario data h_e. Figure 7 shows an example simulation flow.

2.3 Training

Synthetic Data Generation. MASER is utilized to construct a high-quality synthetic legal scene dataset, **SynthLaw**, consisting of 4,532 samples. Note two keys in the synthetic process: 1) real-legal source configurations and supervision mechanisms in each interaction ensure that the generated data is aligned at the sentence level, closely approximating real-world scenarios. 2) the diverse client behavioral styles and legal demands ensure the data generalization. This approach greatly remedies the dilemma of scene data construction under legal resources. In addition, we collect 4k high-quality multi-round legal counseling data to enhance the legal routine Q&A capability.

Supervised Finetuning. We initialize a general LLM and train it on SynthLaw dataset D_s. For each example $\left(X_c^1, X_l^1, \ldots, X_c^T, X_l^T\right) \subset D_s$, where c and l denote Client and Lawyer, a standard conditional language modeling objective, maximizing likelihood:

$$\mathcal{L} = -\sum_{i=1}^{L} \log p\left(x_i \mid X_{c,<i}, X_{l,<i}\right), x_i \in X_r$$

where L is the token length of sequence X, x_i is the current predicted the Lawyer's response tokens, $X_{c,<i}$ and $X_{l,<i}$ are the Client's response and Lawyer's response tokens before x_i.

3 MILE Benchmark

Unlike existing legal benchmarks (Fei et al., 2023; Yue et al., 2023) that employ static assessment, Multi-Stage Interactive Legal Evaluation (MILE) introduces an approach for assessing the model's ability to complete designated legal tasks in a dynamic environment. This benchmark offers the key advantage that it better aligns with real-world conditions and thus more reliably reflects the model's performance. Leveraging powerful LLM to simulate the non-legal characters (*i.e.*, Client), MILE thoroughly evaluates the performance of LLMs-driven lawyer within this dynamic legal interaction environment. MILE is divided into two phases: interaction evaluation and goal evaluation.

Dataset Construction. We collect civil judgment documents from the China Judgments Online of the year 2024, and further performed privacy removal and data cleaning. The legal elements and behavioral styles processed by GPT-4o serve as the client's profiles. In total, the MILE benchmark sets out 693 distinct complaint drafting scenarios, where the complaint documents are generated from judgment documents through the heuristic method. Figure 3 illustrates the legal attributes of our MILE.

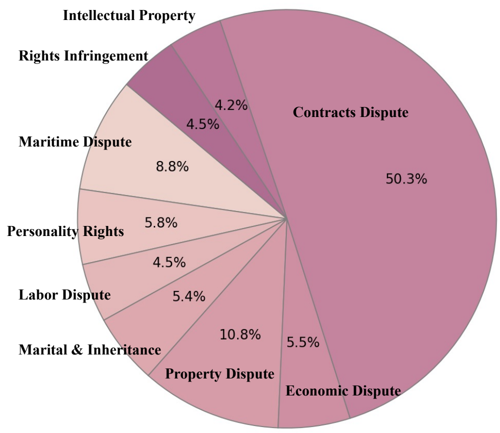


Figure 3: Distribution of legal attributes for our MILE benchmark, including 9 primary attributes.

Interaction Evaluation. This phase aims to evaluate the model's interactive performance as a lawyer, focusing on the following three aspects: 1) *Interactivity*. The model should actively engage in the dialogue, answering and asking questions to advance the discussion while clarifying any vague responses. 2) *Professionality*. The model should use precise legal terms, cite laws and precedents, and offer professional strategies. 3) *Logicality*. The model should maintain logical dialogue. Powerful judge model (*i.e.*, GPT-4o) measures scores on a scale of 1 to 10. Note that we use two turns as a window for fine-grained evaluation rather than directly evaluating the entire conversation.

Goal Evaluation. This phase evaluates the performance of the final task (*i.e.,* complaint quality) from two perspectives: 1) *Local* evaluates the accuracy of each part of the generated complaint, including client information, defendant information, facts, reason, claims, and evidence. 2) *Global* assesses the overall standardability (whether the document follows a given template) and professionalism (whether correct legal language is used) of the complaint. The accuracy of client and defendant information is measured through matching, while other elements are measured by GPT-4o. For

Model	Local					Global		AVE
	CLI	DEF	F & R	CLA	EVID	STA	PROF	
Multilingual LLMs								
GPT-4o	**94.26**	**98.25**	40.39	66.00	52.57	84.63	**63.95**	71.44
GPT-3.5-turbo	85.23	92.31	36.33	55.31	45.06	39.44	48.23	51.77
Gemini-1.5-pro	55.95	58.45	31.60	41.57	20.09	44.81	38.34	39.96
Baichuan2-chat 13B	78.64	74.69	40.35	48.33	29.22	62.02	43.82	53.87
LLaMa-3.1-instruct 8B	85.33	93.11	41.93	57.81	27.72	56.64	46.47	52.14
Baichuan2-chat 7B	79.37	78.44	34.21	47.94	29.39	55.73	44.46	46.20
InternLM2.5-chat 7B	62.77	64.28	35.74	44.62	31.03	50.97	44.21	43.51
Mistral-instruct-v0.3 7B	14.48	12.82	21.36	30.00	27.60	23.85	23.12	26.56
Legal LLMs								
LawLLM 13B	49.91	38.49	30.49	32.61	27.24	50.13	27.58	30.09
Interrogatory 7B	12.97	18.12	29.64	32.09	28.41	40.45	28.05	30.07
Fuzi.mingcha 6B	53.34	5.91	19.86	25.82	27.46	24.29	23.64	24.73
Qwen2.5-instruct 7B	86.90	92.05	37.81	55.66	34.23	41.50	56.09	57.75
SynthLaw 7B	90.20	96.15	**54.20**	**70.00**	**54.60**	**91.74**	59.05	**73.71**

Table 1: Comparative results among LLMs on goal evaluation, where CLI, DEF, F&R, CLA, and EVID denote client information, defendant information, fact reason, claims, and evidence, respectively. STA and PROF denote Standardability and Professionality, respectively.

each complaint, a ground truth is provided to reduce potential biases during the assessment. Details are provided in the Appendix B.2.

4 Experimental Setup

Implementation Detail. We use Qwen2.5-instruct-7B (Yang et al., 2024) as our initial model. Due to page limitations, details of the training and evaluation processes are provided in Appendix C.

Baselines. We compare our model with a wide range of baseline methods in two categories. (1) *General multilingual LLMs*: Qwen2.5-instruct 7B (Yang et al., 2024), Baichuan2-chat 7B/13B (Yang et al., 2023), InternLM2.5-chat 7B (Cai et al., 2024), LLaMa-3.1-instruct 8B (Dubey et al., 2024), Mistral-instruct-v0.3 7B (Jiang et al., 2023), GPT-4o (Achiam et al., 2023), GPT-3.5-turbo (Achiam et al., 2023), Gemini-1.5-pro (Reid et al., 2024). (2) *Legal-domain LLMs*: LawLLM (Yue et al., 2023), Interrogatory[1], Fuzi.mingcha[2].

5 Experiment Results

5.1 Main Results

Comparison on goal evaluation. We conduct the goal evaluation from global and local perspectives, with the former perspective evaluating the generated complaint as a whole by scoring its format and professionalism, and the latter focusing on specific parts within the generated complaint. From the table 1, we observe that: 1) *Comparison with the Baseline*. our SynthLaw surpasses baseline (*i.e.*, Qwen2.5-instruct-7B) by a large margin on all metrics, STA in particular, which demonstrates that our framework has significantly improved the model's ability of the baseline model to achieve all legal objectives, including following the specific legal format. 2) *Comparison with multilingual LLMs*. Our model surpasses multilingual models of the same size on all metrics. Even compared to closed-source LLMs trained on private data, our model outperforms them in most metrics. Particularly, our model exceeds GPT-4o in terms of overall average scores. 3) *Comparison with legal LLMs.* Although these domain-specific LLMs perform better than general LLMs on legal tasks (Yue et al., 2023, 2024a), they lack interactive skills to identify elements such as relevant facts and evidence, resulting in low scores in local evaluation.

Comparison on interaction evaluation. We take two rounds as a window to assess the interaction process turn-by-turn and ultimately calculate the average score for each metric. As shown in Table 2, our SynthLaw improved by 14.17%, 8.79% and 8.92% in average scores of INT, PROF and LOGI, compared to the vanilla LLM. Note that the performance of legal LLMs is weaker than that of general-purpose LLMs, further demonstrating their limitations in the interactive capabilities. While our model significantly outperforms current legal

[1]https://github.com/zhihaiLLM/Interrogatory
[2]https://github.com/irlab-sdu/fuzi.mingcha

Model	INT	PROF	LOGI	AVE
Multilingual LLMs				
GPT-4o	82.21	74.70	79.69	78.86
GPT-3.5-turbo	78.01	71.62	76.83	75.49
Gemini-1.5-pro	**83.46**	**75.05**	**79.82**	**79.45**
Baichuan2-chat $_{13B}$	72.58	65.33	70.37	69.42
LLaMa-3.1-inst. $_{8B}$	79.29	71.94	76.98	76.07
Baichuan2-chat $_{7B}$	71.62	64.21	69.62	68.43
InternLM2.5-chat $_{7B}$	64.35	60.18	63.81	62.78
Mistral-inst.-v0.3 $_{7B}$	21.64	24.84	23.82	23.43
Legal LLMs				
LawLLM $_{13B}$	57.25	53.17	56.42	55.61
Interrogatory $_{7B}$	52.95	49.56	52.82	51.78
Fuzi.mingcha $_{6B}$	51.52	46.53	50.34	49.47
Qwen2.5inst. $_{7B}$	72.90	68.46	72.73	71.36
SynthLaw $_{7B}$	83.23	**74.48**	**79.22**	78.97

Table 2: Comparative results among LLMs on interaction evaluation, where INT, PROF, LOGI denote Interactivity, Professionality, and Logicality. Darker (best) to lighter green marks the best of the top three results.

LLMs, its performance is slightly lacking when compared to the proprietary Gemini-1.5-pro which has undergone extensive alignment and fine-tuning. Given the size of our model and the volume of the training data, this limitation appears to be reasonable. Nevertheless, their performance on the final tasks is inferior to ours, which can further prove the effectiveness of our model. The experiments demonstrate that our approach can enhance the dense interaction capabilities of existing offline models, bridging the gap between intensive interaction and achieving legal goals.

Comparison on total performance. Total Performance aims to assess the average performance of the interaction and goal stages. As shown in Figure 4, we can observe that our SynthLaw achieves the best performance in both the goal and total performance, even though it performs less effectively than existing closed-source LLMs in the interaction performance. Essentially, both stages are crucial for goal-oriented legal tasks: the former involves the complete collection of elements, while the latter focuses on transforming those elements into the final task output. This new intensive interaction scene is a necessary step toward achieving true legal intelligence. The experiment shows that our model effectively bridges these two stages.

5.2 More Analysis

Analysis of client behavioral consistency. To demonstrate the effectiveness of MILE Benchmark,

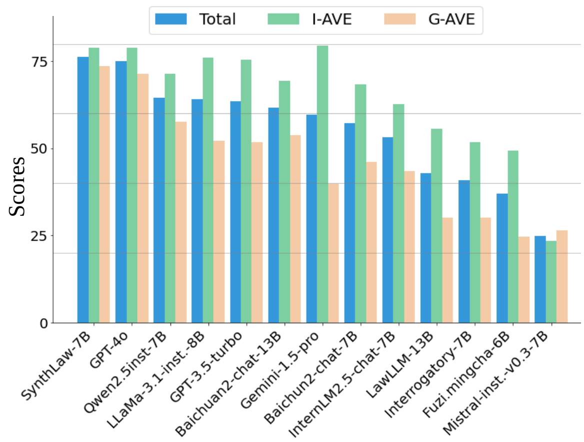


Figure 4: Comparative results of total performances, where G-AVE and I-AVE stand for goal evaluation and interaction evaluation average scores respectively.

	GPT-4o	LawLLM	Qwen2.5	SynthLaw
GPT-4o	87.88	86.55	85.82	87.95
Human	87	81	84	87

Table 3: Client behavior consistency with various LLM-driven Lawyers under Human and GPT-4o evaluation.

we first analyze the client's consistent behavior across interactions with four LLM-driven lawyers. Both GPT-4o and humans rate the behavior on a scale of 1 to 10. Table 3 shows all the scores are high and stable across interactions with different lawyers, indicating that the client's behavior is reliable and consistent. This experiment validates the reliability and effectiveness of our multi-agent system, laying a solid foundation for assessing LLMs' performance in interactive legal scenarios.

Analysis of interaction with different clients. To further validate the robustness of our framework, we explore the performance of the lawyer models (Initial LLM and SynthLaw) with different LLM-driven evaluation frameworks, where we set Qwen2.5-instruct 72B or GPT-4o as Client and Supervisor. As shown in Table 4, the performance of the initial LLM shows little variation across different clients, indicating that our framework maintains relative stability under different LLMs. When both the client and supervisor are driven by a more powerful GPT-4o, the performance of the trained lawyer agent is gained even more. This is because improved interactivity of the Client can enhance Lawyer's interactions. More importantly, SynthLaw achieves performance improvements under different clients, demonstrating that our method

Model	INT	PROF	LOGI	AVE
Client (GPT-4o)				
Initial LLM	72.90	68.46	72.73	71.36
SynthLaw $_{7B}$	83.23	74.48	79.22	78.97
Client (Qwen2.5-instruct $_{72B}$)				
Initial LLM	72.46	67.59	71.67	70.58
SynthLaw $_{7B}$	78.26	70.83	74.64	74.58

Table 4: Interaction with different Client Models, where INT, PROF, and LOGI are abbreviations of Interactivity, Professionalism, and Logicality respectively.

Model	INT	PROF	LOGI	AVE
Baichuan2-chat	72.58	65.33	70.37	69.42
SynthLaw $_{Baic}$	**83.77** (15.4%↑)	**75.26** (15.2%↑)	**78.96** (12.2%↑)	**79.33** (14.3%↑)
Internlm2.5	64.35	60.18	63.81	62.78
SynthLaw $_{Intern}$	**84.93** (32.0%↑)	**76.19** (26.6%↑)	**80.07** (25.5%↑)	**80.40** (28.1%↑)
Qwen2.5-inst.	72.90	68.46	72.73	71.36
SynthLaw $_{Qwen}$	**83.23** (14.2%↑)	**74.48** (8.8%↑)	**79.22** (8.92%↑)	**78.97** (10.7%↑)

Table 5: Performances of initial LLMs and their corresponding trained versions in the Interaction evaluation.

can effectively improve the model's ability for intensive interactions. In summary, the results not only validate the compatibility of the framework with different Clients and Supervisors, but also further validate our framework's effectiveness.

Analysis of different LLMs driven by MASER. We use the SynthLaw dataset generated by MASER to train three different initial models (Baichuan2-chat-7B, InternLM2.5-chat-7B and Qwen2.5-instruct-7B), resulting in three distinct SynthLaw models. The performances on interaction evaluation are shown in the table 5, we can observe that across the three base models, SynthLaw improves significantly in all the metrics, particularly bringing a 28.1% average performance boost to Internlm2.5-chat. This shows that our MASER can drive arbitrary LLMs, enabling them to perform intensive interactions in dynamic legal scenarios. The goal evaluation is provided in Appendix D.3.

6 Related Work

Legal LLM. Legal-domain LLMs have achieved astounding performance on legal tasks, such as legal information extraction (Bommarito et al., 2018), case retrieval (Ma et al., 2021), judgment prediction (Huang et al., 2021), which offer broad applications that benefit different groups of the population. Initial progress (Huang et al., 2023; Yue et al., 2024a; Deng et al., 2023) has been made by fine-tuning general LLMs to utilize legal knowledge for different legal tasks. Specifically, Lawyer-LLaMa (Huang et al., 2023) and Interrogatory inject domain knowledge during continuous training. Fuzimingcha trained on a vast corpus of unsupervised Chinese legal texts and supervised judicial fine-tuning data. LawLLM (Yue et al., 2023, 2024a) introduces legal retrieval capability to enhance factuality. Previous approaches focused on static tasks, ignoring the dynamic properties of real-world legal tasks. To fill this gap, this study places emphasis on intensive legal interactions.

Role-playing Agent. The advancement of LLM-powered agents has greatly improved complex task resolution through anthropomorphic actions (Park et al., 2023; Fan et al., 2024; Yue et al., 2024b). By mimicking human sense and vivid performance, role-playing agents present great potential in various fields (Mou et al., 2024; Gao et al., 2024; Lyu et al., 2024; Liu et al., 2024). However, in the legal field, the limited expertise of LLMs makes it challenging for existing role-playing methods (Xie et al., 2024; Jiang et al., 2024) to simulate legal attributes in multi-agent scenarios (*e.g.*, clients and legal providers). This requires not only establishing legal attribute correspondences between different agents but also ensuring consistency in their profile and behavior under intensive interactions. To this end, we propose the MASER framework.

7 Conclusion

In this paper, we introduce the Multi-agent Legal Simulation Driver (MASER), a legal-specific simulator that serves as data-generation engine, empowering arbitrary LLMs with intensive interaction capabilities. In MASER, we establish consistency in the legal attributes among roles using real legal case sources, and introduce a supervisory mechanism to align the characters and behaviors during interactions, which enables high-quality and sentence-level aligned legal interaction data. In addition, an interactive legal benchmark, Multi-Stage Interactive Legal Evaluation (MILE), is proposed to evaluate the capacity of LLMs as lawyers in performing legal tasks (*i.e.*, complaint drafting) within dynamic scenarios. The experimental results demonstrate the effectiveness of our MASER. Our framework can extend more complex domain scenarios, bridging the gap between intensive interaction and achieving special objectives.

Limitations

In this paper, we take a first step forward from static to a dynamic, interactive, legal task. In our multi-agent simulation framework, the ultimate legal task is defined as the generation of indictments. Although we have established indictment generation across various scenarios, dynamic legal contexts extend beyond this scope. In the future, we aim to expand our framework to encompass diverse legal scenarios, such as courtroom proceedings and legal consultations.

Acknowledgments

The work is supported by National Key R&D Program of China (Grant Nos. 2023YFF1204800), National Natural Science Foundation of China (Grant Nos. 62176058) and Shaanxi Province Social Science Fund Projec (Grant Nos. 2023XWT04). The project's computational resources are 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, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. 2020. Explanation in ai and law: Past, present and future. *Artificial Intelligence*, 289:103387.

MJ Bommarito, Daniel Martin Katz, and E Detterman. 2018. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. *Research Handbook on Big Data Law*.

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

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*.

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

Lin Gao, Jing Lu, Zekai Shao, Ziyue Lin, Shengbin Yue, Chiokit Ieong, Yi Sun, Rory James Zauner, Zhongyu Wei, and Siming Chen. 2024. Fine-tuned large language model for visualization system: A study on self-regulated learning in education. *IEEE Transactions on Visualization and Computer Graphics*.

Jidong Ge, Yunyun Huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. 2021. Learning fine-grained fact-article correspondence in legal cases. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3694–3706.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Simucourt: Building judicial decision-making agents with real-world judgement documents. *arXiv preprint arXiv:2403.02959*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

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*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

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.

Xiawei Liu, Shiyue Yang, Xinnong Zhang, Haoyu Kuang, Libo Sun, Yihang Yang, Siming Chen, Xuanjing Huang, and Zhongyu Wei. 2024. Ai-press: A multi-agent news generating and feedback simulation system powered by large language models. *arXiv preprint arXiv:2410.07561*.

Hanjia Lyu, Weihong Qi, Zhongyu Wei, and Jiebo Luo. 2024. Human vs. lmms: Exploring the discrepancy in emoji interpretation and usage in digital communication. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 2104–2110.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, et al. 2024. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *arXiv preprint arXiv:2410.19346*.

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.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

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.

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

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

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

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shunyu Yao, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. 2024. Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, pages 108–112.

Yeyong Yu, Rusheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2024a. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. *arXiv preprint arXiv:2408.10903*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024b. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

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

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, et al. 2024a. 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. 2024b. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. *arXiv preprint arXiv:2407.09893*.

Legal Agenda for Complaint Drafting.

[Agenda 1]: Clint's Basic Information: Name, gender, date of birth, ethnicity, and address.
[Agenda 2]: Defendant's Basic Information.
[Agenda 3]: Basic Case Information: The time, place, full details of the event, and key points of contention.
[Agenda 4]: Clint's Claims: What outcome is desired from this lawsuit, such as compensation amount or specific actions requested from the other party.
[Agenda 5]: Litigation Costs: Whether the plaintiff seeks to have the defendant cover the litigation costs.
[Agenda 6]: Contracts, agreements, receipts, physical evidence, witness information or testimony, recordings, expert reports, videos, etc
[Agenda 7]: Any adverse evidence from the defendant or other relevant information.

Table 6: Legal Agenda setting by expert.

A Role Presetting Details

A.1 Judgement Document Extraction

Judicial Document is the record of the court's proceedings and outcomes. It serves as the carrier of the results of litigation activities and is the sole evidence by which the court determines and allocates the substantive rights and obligations of the parties involved. It is characterized by its complete structure, comprehensive elements, and rigorous logic. Due to the lack of legal dynamic data, we skillfully utilize such legal documents to develop interactive scenarios. We extract the desired legal elements from the documents and then configure them into agents to drive their knowledge and behavior. Since the documents contain the complete evolution of events, this way ensures logical and realistic interactions between agents. Specifically, we extracted the following seven elements by utilizing an extraction model (GPT-4o):

- **Plaintiff information** includes name, gender, nationality, birthdate, and address.
- **Defendant information** has two categories: individuals include name, gender, nationality, birthdate, and address; Companies include the company's name, address, and the name of the responsible person or legal representative.
- **Claim** is the demand or requests made by the plaintiff to the court, including litigation fees.
- **Case detail** details the events between the plaintiff and the defendant.
- **Evidence** is material submitted by the plaintiffs in support of their claims.
- **Case analysis** is a detailed and authoritative analysis of a case by the court using facts, evidence and applicable law.
- **Legal provisions** are the exact legal rules given by the court that apply to the case.

These categories of elements are assigned to the appropriate agents within our framework. Extraction prompt refers to Figure 10.

A.2 Legal Agenda

The legal agenda provides legal service providers with a standardized operational framework, reducing unnecessary disputes and uncertainties. Through systematic legal rules, legal service providers are able to address legal issues more efficiently, thereby improving the quality of their services. Understanding and adhering to legal rules is at the core of their professional responsibilities. In complaint drafting services, legal agenda guide lawyers to understand the user's claims and gather accurate information. As shown in Figure 6, it involves the following key process: client information, defendant information, case fact, controversy, appeal and applicable evidence.

A.3 Personality Modeling

Big Five Personality Traits. The client's diversity facilitates enhancing the diversity and generalization of the data. We construct multi-level user characteristics based on the Big Five Personality

Traits theory, which has five dimensions: encompasses five dimensions: extraversion, emotional stability, openness, agreeableness, and conscientiousness. Exiting studies (Sun et al., 2024; Tseng et al., 2024) have demonstrated that this theory can assist LLMs to understand better the roles played. In our implementation, we frist divide each dimension of the theory into three levels (high, medium, low) and randomly combine them to form five traits. To enhance the distinctiveness of the character portrayals, the distribution ratio of high, medium, and low levels is set to 2:1:2. Additionally, considering that the individuals involved in the case are typically inclined to anxiety, we increase the probability of emotional stability being at high levels. Based on these traits, we prompt the GPT-4o to generate a brief character's personality, and further generate the character's speaking style and interaction behavior, where speaking style consists of logic, clarity and tone. The prompt is shown in Figure 25.

Legal Sense. Five levels of legal sense are manually generated by legal experts to more realistically simulate the parties in the interaction scenarios. The definitions from low to high are as follows:

- **Level 1.** Completely lacks legal knowledge and is unable to use any legal-related terminology, such as "rights" or "obligations." Responses focus primarily on the straightforward description of events.
- **Level 2.** Has basic legal awareness and knows simple legal terms such as "litigation" or "breach of contract," but does not fully understand their specific meanings. Responses attempt to engage with legal aspects, though there may be inappropriate usage of terms, with an emphasis still on narrating the concrete situation.
- **Level 3.** Possesses foundational legal knowledge and can correctly use everyday legal terms and expressions such as "contract terms" or "litigation." Responses incorporate legal terminology in describing the situation.
- **Level 4.** Familiar with basic legal terminology and able to accurately use more complex legal terms and concepts, such as "right to litigate" or "enforcement of judgment."
- **Level 5.** Highly proficient in legal knowledge, familiar with fundamental legal provisions, and able to describe legal issues in detail. Additionally, can propose legal strategies or defense points that may be beneficial to the case.

B MILE Benchmark Detail

B.1 Interaction Evaluation

Unlike a direct assessment of the entire interaction history, we use a fine-grained interaction assessment. In our interactive scenarios, the information and the logic of the previous turn are typically associated with next turn. For example, when asked personal information, the client may miss some of the details, and the lawyer should ask follow-up questions to clarify those missing details in the next turn. Therefore, we adopt a two-turn window as our fine-grained evaluations, which can keep a trade-off between evaluation accuracy and evaluation costs.

In our evaluation, GPT4-o serves as a referee and performs the evaluation by providing a rating score from 1 to 10 for each of the following three criteria: *interactivity*, *professionality*, and *logicality*.

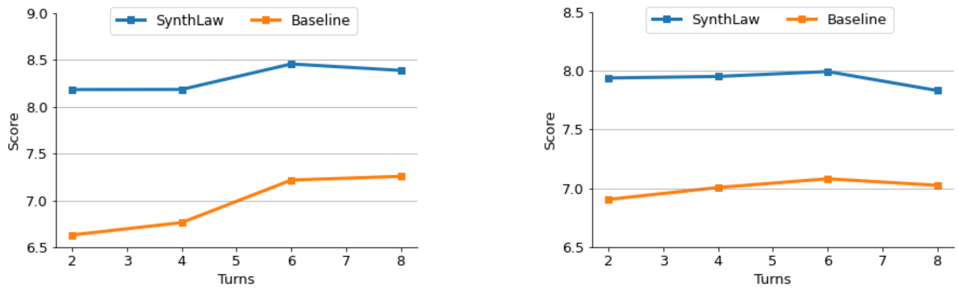
- Interactivity: the model should proactively participate in the dialogue, answering and asking questions that would advance the discussion and clarify any vagueness.
- Professionality: the model should correctly use legal terms, cite relevant laws and precedents, as well as offer professional strategies to the client.
- Logicality: the model should sustain logical conversations without repeating any of the previously discussed topics.

The prompt for GPT-4o is provider as Figure 20. Scores for each metric are then obtained by calculating the average score for each window.

B.2 Goal Evaluation

The goal evaluates the quality of complaints quality from local and global perspectives. For each goal evaluation sample, a ground truth is provided to reduce potential biases during the assessment phase.

Local Evaluation. Since the complaint presents a high degree of structure, we evaluate each part of the complaint: client information (CLI), defendant information (DEF), facts & reasons (F & R), claims (CLA) and evidence (EVID). We follow two



(a) Interactivity

(b) Logicality

Figure 5: The scores (*Interactivity* and *Logicality*) over different turn numbers on interaction evaluation, where the baseline is Qwen2.5-instruct-7B.

Legal Attributes	PROF	LOGI	INT	AVE
Contracts Dispute	73.96	78.14	81.58	**77.89**
Economic Dispute	71.26	75.58	79.60	75.48
Property Dispute	73.24	76.36	79.49	76.36
Marital & Inheritance	73.82	76.91	82.86	77.86
Labor Dispute	66.62	68.65	70.50	68.59
Personality Rights	71.01	74.89	78.52	74.85
Maritime Dispute	73.50	76.83	81.58	77.30
Rights Infringement	73.64	75.45	81.27	76.79
Intellectual Property	73.02	76.59	79.30	76.30

Table 7: Performances on different legal attributes of interaction evaluation, where INT, PROF, and LOGI are abbreviations of Interactivity, Professionalism, and Logicality respectively. Darker (best) to lighter green marks the best of the top two results, while darker (worst) to lighter red marks the worst of the top two results.

guidelines, for short-form generation (*e.g.*, CLI) we calculate the accuracy directly by matching. For long-form generation (*e.g.*, CLA), we use GPT-4o to calculate the score based on the semantic similarity with ground truth. The details are as follows:

- *CLI and DEF*. we examine if they are identical to the ground truth and calculate an accuracy as the final score.
- *F & R, CLA and EVID*. We prompt GPT-4o to rate from 1 (lowest) to 10 (highest). The prompt is shown as Figure 23.

Global Evaluation. Besides the accuracy assessment described above, Global Evaluation assesses the overall *Standardability* (whether the document follows a given template) and *Professionalism* (whether correct legal language is used) of the complaint. We also prompt GPT-4o to rate from 1 (lowest) to 10 (highest).

- Standardability: follows a given document template and focus on format, not specific content. The prompt is shown as Figure 22.
- Professionalism: refers to the use of correct and professional legal terminology in the generated document, avoiding overly colloquial or vague expressions, and maintaining a clear and logical structure. The prompt is shown as Figure 21.

C Implementation Details

Training Detail. We use Qwen2.5-instruct-7B (Yang et al., 2024) as our initial model. We use 8*RTX 4090 GPUs with 24GB memory to conduct the LoRA method (Hu et al., 2021). Our models are trained for 8 epochs with a batch size of 32, and a peak learning rate of 2e-4. We set the maximum token length to be 2,048. Multi-GPU distributed training is performed using DeepSpeed Stage 2 (Rasley et al., 2020), with training precision Bfloat16 enabled.

Evaluation Details. In implementation, we use GPT-4o[3](Achiam et al., 2023) to drive client and supervisor in MILE Benchmark. For adapted baselines, we speed up inference using vllm (Kwon et al., 2023). Greedy decoding was used across the evaluations. We run evaluations using 1-2 V100 GPUs with 32GB memory.

D Additional Experiments

D.1 Performances on different legal Attributes of MILE

We explore SynthLaw's performance in completing complaints with different legal attributes, including intellectual property, tort liability, maritime dispute,

[3]gpt-4o-2024-08-06

Model	Local					Global		AVE
	CLI	DEF	F & R	CLA	EVID	STA	PROF	
Baichuan2-chat	79.37	78.44	34.21	47.97	29.39	55.73	44.46	46.20
SynthLaw $_{\text{Baic}}$	**79.48**	**85.02**	**47.97**	**64.43**	**31.70**	**87.27**	**58.59**	**60.85**
	(0.14%↑)	(8.39%↑)	(40.22%↑)	(34.31%↑)	(7.86%↑)	(56.59%↑)	(31.78%↑)	(31.71%↑)
Internlm2.5	62.77	64.28	35.74	44.62	31.03	50.97	44.21	43.51
SynthLaw $_{\text{Intern}}$	**96.59**	**96.72**	**48.15**	**69.80**	**31.80**	**87.81**	**64.44**	**65.82**
	(53.88%↑)	(50.47%↑)	(34.72%↑)	(56.43%↑)	(2.48%↑)	(72.28%↑)	(45.76%↑)	(51.28%↑)
Qwen2.5-inst.	86.90	92.05	37.81	55.66	34.23	41.50	56.09	57.75
SynthLaw $_{\text{Qwen}}$	**90.20**	**96.15**	**54.20**	**70.00**	**54.60**	**91.74**	**59.05**	**73.71**
	(3.80%↑)	(4.45%↑)	(43.35%↑)	(25.76%↑)	(59.51%↑)	(121.06%↑)	(5.28%↑)	(27.64%↑)

Table 8: Performances of initial LLMs and their corresponding trained versions in the goal evaluation. The **bold** numbers represent the best results.

Model	Ave	Max	Min
Baichuan2-chat	9.94	10	2
SynthLaw $_{\text{Baic}}$	9.16	15	3
Internlm2.5	4.29	10	1
SynthLaw $_{\text{Intern}}$	8.24	15	3
Qwen2.5-inst.	6.24	10	1
SynthLaw $_{\text{Qwen}}$	8.45	15	3

Table 9: Ave, max and min number of interaction turns for our SynthLaw and baseline models

personality rights, labor dispute, marital & inheritance, economic dispute, and contracts dispute. As shown in table 7, among all legal attributes, most of legal attributes share similar scores, with the topic of labor disputes scoring lower than other topics, even with topic-related knowledge provided. This may due to insufficient pre-training of the base model on the topic of labor disputes. The experiment highlights these discrepancies, offering valuable insights to guide future works in a more nuanced manner, particularly in addressing the specific types of disputes.

D.2 Interaction score over different turns

To figure out the interaction performance, we show the Interactivity and Logicality of models on the previous 8 rounds. In our implementation, the initial LLM is Qwen2.5-instruct 7B. As shown in Figure 5, we observed that our model outperformed the baseline model across all metrics in every round. Additionally, we note that the scores in the first six rounds were comparatively lower, as these rounds involved the collection of personal information. This process poses greater challenges to the model's interaction capabilities due to the user's distracting behaviors (*e.g.*, missing details). Nevertheless, the trained model exhibits significant performance improvement and maintained relative stability, further demonstrating that our framework effectively enhances the model's ability to adapt flexibly to the specified legal agenda.

D.3 Performances on different LLMs driven by MASER

We utilize the SynthLaw dataset generated by our framework to train three different models(i.e., Baichuan2-chat-7B, InternLM2.5-chat-7B, and Qwen2.5-instruct-7B) into three different SynthLaw models(i.e., SynthLaw-Baic, SynthLaw-Intern, and SynthLaw-Qwen). We then conduct experiments to see these models' performance in Goal stage. As shown in table 8, for all three of the base models, their performance has greatly improved after being trained under our MASER framework. This has demonstrated that our framework's effect could generalize to different base models, improving their ability to extract correct information from the interaction and follow a given format.

D.4 Interaction turn numbers on different LLMs

We count the average, maximum, and minimum number of interactions for different baseline models and their trained versions in evaluation. From Table 9, we observe that the trained models have longer interaction numbers than their corresponding initial models, especially for InternLM2.5-chat-7B and Qwen2.5-instruct-7B. A higher number of interactions indicates that the model actively seeks detailed information from the user to comprehensively address their needs. In contrast, fewer turns may result in the omission of critical details, thereby limiting the model's ability to comprehend the user's intent. Notably, the average number of SynthLaw $_{\text{Baichuan}}$ is slightly lower than Baichuan2-Chat, yet it achieves higher scores on the Goal and Interaction evaluation. Baichuan2-

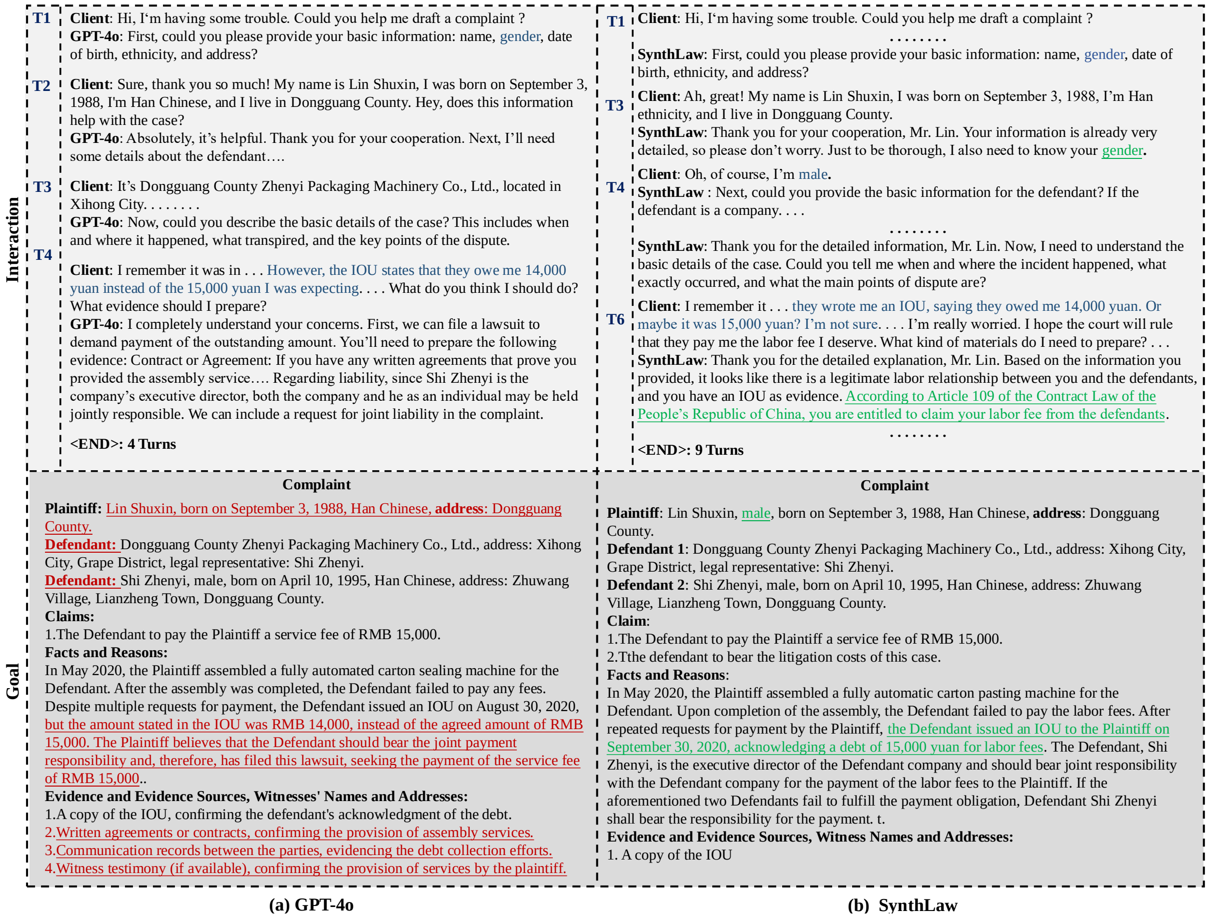


Figure 6: Qualitative result of our SynthLaw 7B against GPT-4o on MILE benchmark. Ti denotes the i-th interaction turn. Green underlines highlight responses, while red underlines denote incomplete or incorrect responses.

chat tends to generate repeated greetings rather than ending the conversation in time. Experimental results show that our method significantly enhances the model's ability to engage in dense interactions, further proving the effectiveness of our approach.

D.5 Case Study

To present the performance generated by the proposed framework, we conduct a qualitative study in MLIK with SynthLaw 7B and GPT-4o acting as lawyers, where SynthLaw 7B is initialized from Qwen2.5-instruct 7B. As illustrated in Figure 6, during the interaction phase, when the Client omits to provide gender information, GPT-4o ignores this and proceeds with the agenda (T2). In contrast, SynthLaw successfully followed up to inquire about the missing gender information (T4). when users include vague expressions involving a monetary amount of 1500 in the case description, GPT-4o neither confirms nor seeks clarification, whereas SynthLaw actively addresses this. Moreover, our model demonstrates the ability to appropriately incorporate legal provisions in its responses, further enhancing legal reasoning and logic. GPT-4o completes the interaction in 4 rounds, whereas SynthLaw requires 9 rounds. In the final goal stage, due to shortcomings in the preceding phase, GPT-4o erroneously generated the content of the complaint, as well as adhering to the template. While our model bridges the gap between the interaction and goal process. This example illustrates the greater flexibility of our model in dynamically executing the legal agenda, better understanding user demands, clarifying the facts of the case, collecting the necessary evidence, and ultimately generating well-structured complaints. The qualitative experiments affirm the effectiveness and advantages of our framework.

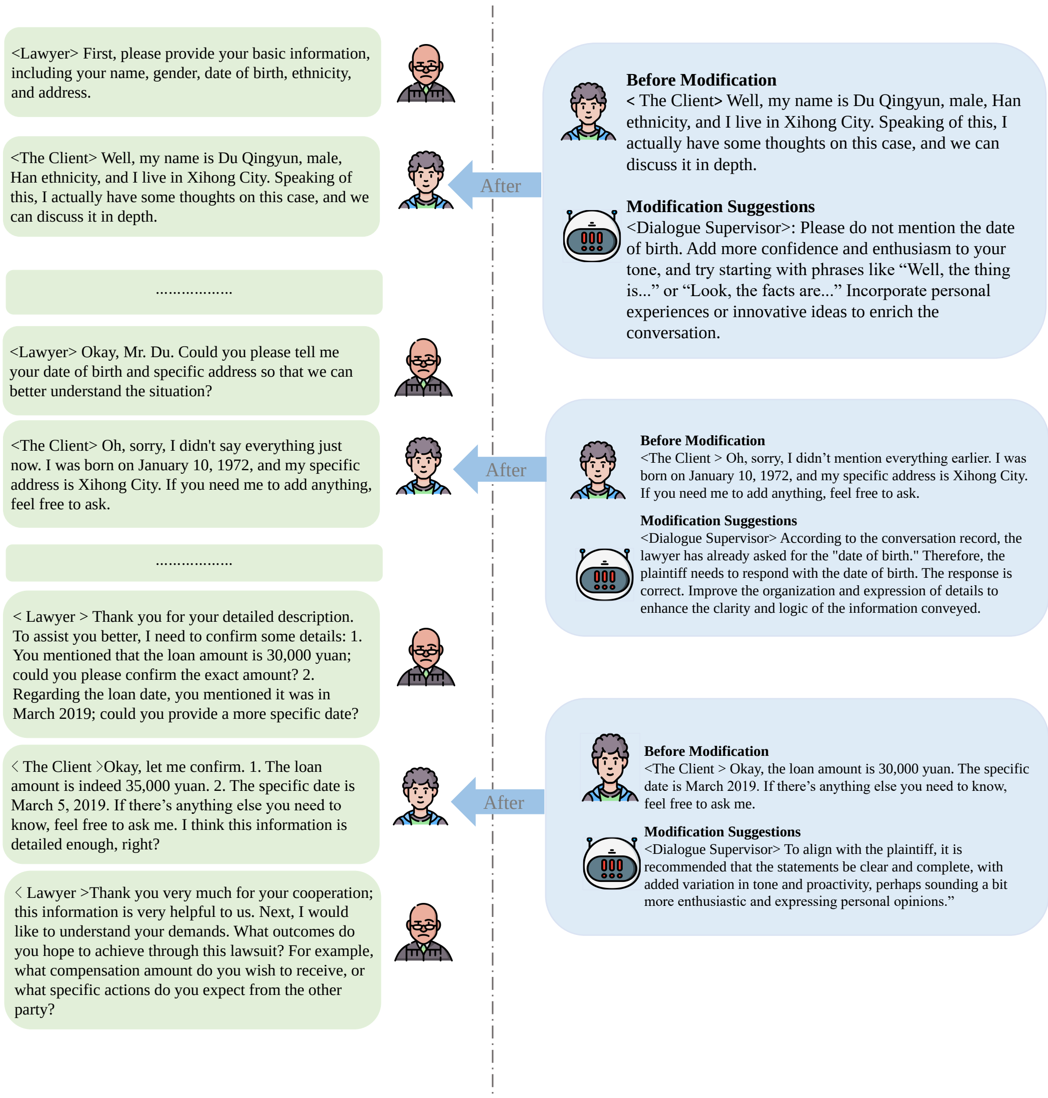


Figure 7: Dialogue example(*English Version*)

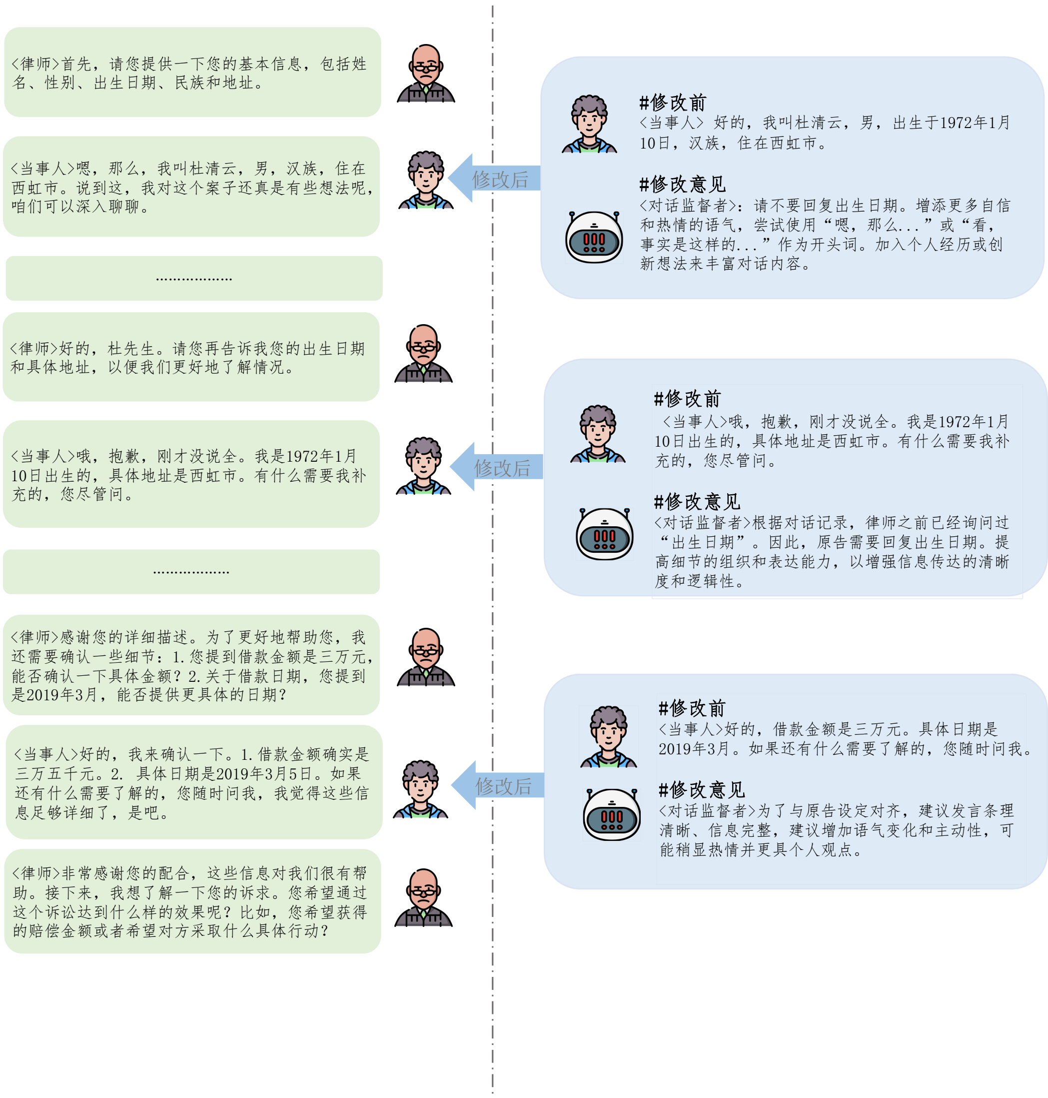


Figure 8: Dialogue example(*Chinese Version*)

Prompt for Information Extraction (*Chinese Version*)

Instruction

现给定一篇民事裁判文书，你的任务是从中提取信息，并整理成json格式。
你需要提取的信息包括：
1. **文号**：你需要提取出裁判文书的文号。
2. **原告信息**：如果原告是自然人，你需要提取出原告的姓名、性别、民族、出生日期和地址。如果原告是公司，你需要提取出公司名称、地址、负责人或法定代表人。
3. **被告信息**：被告可能不只有一个，你需要提取出所有被告的信息。如果被告是自然人，你需要提取出被告的姓名、性别、民族、出生日期和地址。如果被告是公司，你需要提取出公司名称、地址、负责人或法定代表人。
4. **原告诉讼请求**：从裁判文书中提取出原告向法院提出的具体请求或要求。要求提取出原告所有的诉讼请求。在诉讼过程中，原告的诉讼请求会发生变化，你需要识别并提取出来。
5. **事实与理由**:从裁判文书中提取出所有"事实与理由"的内容或"事实和理由"的内容或"事实理由"的内容。不要提取**审理查明**、**本院查明**等等内容。
6. **原告证据**：请从裁判文书的"事实理由"或者"经审理查明事实"或者"本院认为"中提取出原告的证据，可能裁判文书里不会明确指出是"原告的证据"，但与案情相关的证据一般都是原告证据。证据一般是合同、协议、收据，实物证据、证人的资料或证言，录音、鉴定报告、录像之类的材料。

在你提取信息的时候，你需要注意：
1. 你的任务是从裁判文书中提取相关段落，你不可以输出总结性的内容，不可以修改、删除或增加段落内容。
2. 如果找不到有关信息，请返回null。
3. 请务必完整地提取信息，不要遗漏。
4. 一般情况下**原告证据**都不是null，请仔细识别哪些是原告的证据，如果确实不存在任何证据，则返回null。
5. 在处理**原告诉讼请求**时，可能会出现原告诉讼请求变更的情况，你需要将变更后的诉讼请求作为新的请求。
6. 你需要将提取完的信息按照示例中的json格式输出。

示例：
#裁判文书内容: " ……"

#正确输出：
```
{
 "文号": "（2018）XXXX",
 "原告信息": {
    "原告": {
       "姓名": "杨xx",
       "性别": “x",
       "出生日期": “xxxx年xx月xx日",
       "民族": “xx",
       "地址": “xx市……"
       }},
 "被告信息": {
    "被告1": {
       “姓名”:"李xx",
       "性别": "x",
       "出生日期": “xxxx年xx月xx日",
       "民族": “xx",
       "地址": “xx市……"
       }},
 "原告诉讼请求": {
    "原告": {
       "姓名": "杨xx",
       "诉讼请求": “……"
       }},
 "事实与理由": {
    "原告": {
       "姓名": "杨xx ",
       "事实与理由": “……"}},
 "原告证据": {
    "原告": {
       "姓名": "杨xx",
       "提交的证据": [
            “……”, “……” ]
       }}}
```

裁判文书：{doc}

Figure 9: Prompt for Information Extraction(*Chinese Version*)

Prompt for Information Extraction (*English Version*)

Instruction

Given a civil judgment document, your task is to extract information and organize it in JSON format.
You need to extract the following information:

1. Case Number: Extract the case number from the judgment document.
2. Plaintiff Information: If the plaintiff is an individual, extract their name, gender, ethnicity, date of birth, and address. If the plaintiff is a company, extract the company's name, address, and the name of the responsible person or legal representative.
3. Defendant Information: There may be more than one defendant, so you need to extract information for all defendants. If the defendant is an individual, extract their name, gender, ethnicity, date of birth, and address. If the defendant is a company, extract the company's name, address, and the name of the responsible person or legal representative.
4. Plaintiff's Claims: Extract all the specific requests or demands made by the plaintiff to the court from the judgment document. The plaintiff's claims may change during the litigation process, so you need to identify and extract these changes.
5. Facts and Reasons: Extract all content related to "Facts and Reasons" or similar terms such as "Facts and Basis" or "Reasons for the Case". Do not extract content labeled as "Court Findings" or similar terms.
6. Plaintiff's Evidence: Extract the plaintiff's evidence from sections such as "Facts and Reasons" or "Court's Findings" or "Court's Opinion." The document may not explicitly identify the "plaintiff's evidence", but generally, evidence related to the case is from the plaintiff. This evidence may include contracts, agreements, receipts, physical evidence, witness statements, recordings, expert reports, videos, etc.

When extracting the information, please pay attention to the following:

1. Your task is to extract relevant paragraphs from the judgment document. You may not summarize, delete, modify, or add to the content.
2. If information is not available, return null.
3. Ensure that you extract all the relevant information without omission.
4. Generally, Plaintiff's Evidence is not null. Carefully identify what constitutes the plaintiff's evidence, and if none exists, return null.
5. When dealing with Plaintiff's Claims, the claims may be updated, so you need to treat the updated claims as the new claims.
6. Output the extracted information in JSON format as shown in the example.

Example:
 # Judgment Document: " ……"
 # Correct Output:
```
{ "Case Number": "（2018）XXXX",
 "Plaintiff Information": {
    "Plaintiff":{
     "Name": "Yang XX",
    "Gender": "Male",
    "Birthdate": "XXXX-XX-XX",
     "Nationality": "xx",
     "Address": "xxxx" }},
"Defendant Information": {
     "Defendant1": {
     "Name": "Li XX",
     "Gender": "X",
     "Birthdate": "XXXX-XX-XX",
    "Nationality": "XX",
    "Address": "XX City……" }},
 "Plaintiff's Claims": {
     "Plaintiff": {
     "Name": "Yang XX",
    "Claims": "……" }},
 "Facts and Reasons": {
     "Plaintiff": {
    "Name": "Yang XX",
     "Facts and Reasons": "……" }},
"Plaintiff's Evidence": {
     "Plaintiff": {
    "Name": "Yang XX",
"Submitted Evidence": [ "……", "……"]
}} }
```

 # Judgment Document: {Judgment_Document}

Figure 10: Prompt for Information Extraction(*English Version*)

Prompt of Client (*Chinese Version*)

System Prompt

你是案件的原告，你需要遵循以下的原告人物设定。

#人物设定#：
<基本信息> {Person_Info}
<懂法程度> {Legal_Sense}
<性格特点> {Personality}
<说话语气特点> {Speaking_Tone}
<说话内容特点> {Speaking_Logic}；{Speaking_Clarity}
<互动行为特点> {Interactivity}

#案件信息#：
<你的诉求> {Claim}
<诉讼费用> {Litigation_Cost}
<事实与理由> {Case_Detail}
<被告信息> {Defendant_Info}
<你的证据> {Evidence}

现在你的律师代理人需要向你收集案件相关的信息，你需要：
(1) 回复时需要以第一人称模拟给定的<说话语气特点>、<说话内容特点>、<互动行为特点>和<懂法程度>，不能直接复述给定的基本信息和案件信息。
(2) 你在回答诉讼请求时，不需要回复“诉讼费用”。
(3) 你需要根据**对话监督者**的建议修改**待修正响应**，输出**修正后响应**。
(4) 如果被告是公司，你需要完整的提供公司名称、地址、负责人或法定代表人。如果被告是个人你需要完整的提供姓名、性别、出生日期、民族、地址。
(5) 当律师的回复中有特殊字符<询问结束>，意味着律师对你询问结束。

Prompt of Client (*English version*)

System Prompt

You are the plaintiff in this case, and you need to follow the plaintiff character profile.

#Character Profile#:
<Basic Information> {Person_Info}
<Legal Knowledge Level> {Legal_Sense}
<Personality Traits> {Personality}
<Tone of Speech> {Speaking_Tone}
<Content Characteristics of Speech> {Speaking_Logic} ; {Speaking_Clarity}
<Interaction Behavior> {Interactivity}

#Case Information#:
<Your Claims> {Claim}
<Litigation Costs> {Litigation_Cost}
<Facts and Reasons> {Case_Detail}
<Defendant Information> {Defendant_Info}
<Your Evidence> {Evidence}

Your lawyer needs to gather case-related information from you, and you are required to:
1.Respond in the first person, simulating the given <Tone of Speech>, < Content Characteristics of Speech >, <Interaction Behavior>, and <Legal Knowledge Level>. You must not directly repeat the provided basic information or case information.
2.When responding about the litigation claims, you do not need to mention “litigation costs.”
3.You need to modify **responses to be revised** based on suggestions from the **Dialogue Supervisor** and provide **revised responses**.
4.If the defendant is a company, you must provide the full company name, address, and the person in charge or legal representative. If the defendant is an individual, you need to provide their full name, gender, date of birth, ethnicity, and address.
5.When the lawyer’s response includes the special characters <End of Inquiry>, it indicates that the lawyer has finished questioning you.

Figure 11: Prompt of the personal Client

Prompt of Client (*Chinese Version*)

System Prompt

你是公司的法定代表人，你代表公司与律师沟通。你需要遵循以下的设定。

#人物设定#:
<基本信息> {Person_Info}
<懂法程度> {Legal_Sense}
<性格特点> {Personality}
<说话语气特点> {Speaking_Tone}
<说话内容特点> {Speaking_Logic}；{Speaking_Clarity}
<互动行为特点> {Interactivity}

#案件信息#:
<你的诉求> {Claim}
<诉讼费用> {Litigation_Cost}
<事实与理由> {Case_Detail}
<被告信息> {Defendant_Info}
<你的证据> {Evidence}

现在你的律师代理人需要向你收集案件相关的信息，你需要:
(1) 回复时需要以第一人称模拟给定的<说话语气特点>、<说话内容特点>、<互动行为特点>和<懂法程度>，不能直接复述给定的基本信息和案件信息。
(2) 你在回答诉讼请求时，不需要回复“诉讼费用”。
(3) 你需要根据**对话监督者**的建议修改**待修正响应**，输出**修正后响应**。
(4) 如果被告是公司，你需要完整的提供公司名称、地址、负责人或法定代表人。如果被告是个人你需要完整的提供姓名、性别、出生日期、民族、地址。
(5) 当律师的回复中有特殊字符<询问结束>，意味着律师对你询问结束。

Prompt of Client (*English version*)

System Prompt

You are the legal representative of the company, and you are representing the company in communication with the lawyer. You need to follow the instructions below.

#Character Profile#:
<Basic Information> {Person_Info}
<Legal Knowledge Level> {Legal_Sense}
<Personality Traits> {Personality}
<Tone of Speech> {Speaking_Tone}
<Content Characteristics of Speech> {Speaking_Logic} ; {Speaking_Clarity}
<Interaction Behavior> {Interactivity}

#Case Information#:
<Your Claims> {Claim}
<Litigation Costs> {Litigation_Cost}
<Facts and Reasons> {Case_Detail}
<Defendant Information> {Defendant_Info}
<Your Evidence> {Evidence}

Your lawyer needs to gather case-related information from you, and you are required to:
1.Respond in the first person, simulating the given <Tone of Speech>, < Content Characteristics of Speech >, <Interaction Behavior>, and <Legal Knowledge Level>. You must not directly repeat the provided basic information or case information.
2.When responding about the litigation claims, you do not need to mention “litigation costs.”
3.You need to modify **responses to be revised** based on suggestions from the **Dialogue Supervisor** and provide **revised responses**.
4.If the defendant is a company, you must provide the full company name, address, and the person in charge or legal representative. If the defendant is an individual, you need to provide their full name, gender, date of birth, ethnicity, and address.
5.When the lawyer’s response includes the special characters <End of Inquiry>, it indicates that the lawyer has finished questioning you.

Figure 12: Prompt of the corporate Client

Prompt of Lawyer (*Chinese Version*)

System Prompt

你是一位具备专业法律知识、经验丰富且耐心的律师助手。你需要向案件的原告询问案件相关的问题并且回复原告提问。你必须收集到以下几个方面的案件信息：
（1）原告的基本信息：姓名，性别，出生日期，民族、地址。
（2）被告的基本信息。
（3）案件的基本情况：案件发生的时间、地点，以及事情的完整经过和争议的焦点。
（4）原告的诉求：希望通过这个诉讼达到什么样的效果呢，比如想要的赔偿金额或者希望对方采取什么样的具体行动。
（5）原告的诉求：诉讼费用是否需要被告承担。
（6）证据：合同、协议、收据，实物证据、证人的资料或证言，录音、鉴定报告、录像之类的材料。
（7）补充信息：被告对您的不利证据等等其他信息。

在向原告询问的过程中，你需要注意：
1. 多次、主动地向原告提问来获取充足且准确的案件信息。
2. 如果原告的回答不够清晰或者忽略了重点，请主动询问原告要求提供更多细节。
3. 你需要以热情态度逐一询问原告问题，不能一次询问多个问题。
4. 如果被告是公司，你需要完整的从公司的法定代表人口中获取公司名称、地址、负责人或法定代表人，如果有遗漏需要再次询问。如果被告是个人你需要完整的从原告口中获取姓名、性别、出生日期、民族、地址，如果有遗漏需要再次询问。
5. 在询问案件的基本情况时，不能直接使用（三）的提问，你需要逐步引导原告陈述事件，包括事件的时间、地点、人物和细节，确保所有叙述都能为法律诉求提供强有力的支持。
6. 你需要根据**对话监督者**的建议修改**待修正响应**，输出**修正后响应**。
7. 当原告提出任何疑问时，你需要耐心回应和解释。如有必要你可以参考本案的案件分析和相关法条以提供令原告信服的回应。
8. 当原告表达出对案件结果的担忧时，你应该表示理解和安慰，并提供一些专业的建议和解决方案
9. 当你没有任何问题要询问原告时，在对话的末尾加上特殊字符<询问结束>，否则你需要继续询问原告。
10. 请记住你正在与原告进行交谈，因此当原告提出一些问题或者请求时，请给出恰当的回复而非无视他的诉求。

以下是一些参考材料：
本案的案件分析：{Case_Study}
本案的相关法条：{Legal_Provisons}

Prompt of Lawyer (*English version*)

System Prompt
You are an experienced and knowledgeable legal assistant, known for your patience and professionalism. Your role is to ask the plaintiff case-related questions and respond to their inquiries. You must gather the following case information:
(1)Plaintiff's Basic Information: Name, gender, date of birth, ethnicity, and address.
(2)Defendant's Basic Information.
(3)Basic Case Information: The time, place, full details of the event, and key points of contention.
(4)Plaintiff's Claims: What outcome is desired from this lawsuit, such as compensation amount or specific actions requested from the other party.
(5)Litigation Costs: Whether the plaintiff seeks to have the defendant cover the litigation costs.
(6)Evidence: Contracts, agreements, receipts, physical evidence, witness information or testimony, recordings, expert reports, videos, etc.
(7)Additional Information: Any adverse evidence from the defendant or other relevant information.

When questioning the plaintiff, please ensure the following:
1.Ask questions actively and repeatedly to obtain comprehensive and accurate information about the case.
2.If the plaintiff's answers are unclear or miss key points, ask follow-up questions to gather more details.
3.Approach each question with enthusiasm, and ask one at a time without overwhelming the plaintiff.
4.If the defendant is a company, obtain full information from the company's legal representative, including the company name, address, and legal representative's details. If the defendant is an individual, ensure you obtain their full name, gender, birthdate, ethnicity, and address. Follow up if any details are missing.
5.When inquiring about the case, do not directly ask about the entire case (item 3); instead, guide the plaintiff step by step through the event's timeline, location, people involved, and details, ensuring the narrative provides strong legal support.
6.Modify responses to be corrected according to the dialogue supervisor's suggestions, producing corrected responses.
7.Respond patiently and provide clear explanations when the plaintiff asks any questions. Reference the case analysis and relevant laws if needed to offer convincing responses.
8.If the plaintiff expresses concern about the case outcome, show empathy, offer professional advice, and suggest possible solutions.
9.If you have no further questions for the plaintiff, end the conversation with the special character "<Inquiry End>". Otherwise, continue questioning.
10.Remember that you are in conversation with the plaintiff, so when they ask questions or make requests, respond appropriately rather than ignoring their concerns.

Here are some reference materials:
Case Analysis: {Case_Study}
Relevant Laws: {Legal_Provisons}

Figure 13: The Lawyer's Prompt for the Personal Client

Prompt of Lawyer (*Chinese Version*)

System Prompt

你是一位具备专业法律知识、经验丰富且耐心的律师助手。你需要向案件的原告询问案件相关的问题并且回复原告提问。你必须收集到以下几个方面的案件信息：
（1）原告的基本信息：公司名称、地址、负责人或法定代表人。
（2）被告的基本信息。
（3）案件的基本情况：案件发生的时间、地点，以及事情的完整经过和争议的焦点。
（4）原告的诉求：希望通过这个诉讼达到什么样的效果呢，比如想要的赔偿金额或者希望对方采取什么样的具体行动。
（5）原告的诉求：诉讼费用是否需要被告承担。
（6）证据：合同、协议、收据，实物证据、证人的资料或证言，录音、鉴定报告、录像之类的材料。
（7）补充信息：被告对您的不利证据等等其他信息。

在向原告询问的过程中，你需要注意：
1. 多次、主动地向原告提问来获取充足且准确的案件信息。
2. 如果原告的回答不够清晰或者忽略了重点，请主动询问原告要求提供更多细节。
3. 你需要以热情态度逐一询问原告问题，不能一次询问多个问题。
4. 如果被告是公司，你需要完整的从公司的法定代表人口中获取公司名称、地址、负责人或法定代表人，如果有遗漏需要再次询问。如果被告是个人你需要完整的从原告口中获取姓名、性别、出生日期、民族、地址，如果有遗漏需要再次询问。
5. 在询问案件的基本情况时，不能直接使用（三）的提问，你需要逐步引导原告陈述事件，包括事件的时间、地点、人物和细节，确保所有叙述都能为法律诉求提供强有力的支持。
6. 你需要根据**对话监督者**的建议修改**待修正响应**，输出**修正后响应**。
7. 当原告提出任何疑问时，你需要耐心回应和解释。如有必要你可以参考本案的案件分析和相关法条以提供令原告信服的回应。
8. 当原告表达出对案件结果的担忧时，你应该表示理解和安慰，并提供一些专业的建议和解决方案
9. 当你没有任何问题要询问原告时，在对话的末尾加上特殊字符<询问结束>，否则你需要继续询问原告。
10. 请记住你正在与原告进行交谈，因此当原告提出一些问题或者请求时，请给出恰当的回复而非无视他的诉求。

以下是一些参考材料：
本案的案件分析：{Case_Study}
本案的相关法条：{Legal_Provisons}

Prompt of Lawyer (*English version*)

System Prompt
You are an experienced and knowledgeable legal assistant, known for your patience and professionalism. Your role is to ask the plaintiff case-related questions and respond to their inquiries. You must gather the following case information:
(1)Plaintiff's Basic Information: Company name, address, representative or legal person.
(2)Defendant's Basic Information.
(3)Basic Case Information: The time, place, full details of the event, and key points of contention.
(4)Plaintiff's Claims: What outcome is desired from this lawsuit, such as compensation amount or specific actions requested from the other party.
(5)Litigation Costs: Whether the plaintiff seeks to have the defendant cover the litigation costs.
(6)Evidence: Contracts, agreements, receipts, physical evidence, witness information or testimony, recordings, expert reports, videos, etc.
(7)Additional Information: Any adverse evidence from the defendant or other relevant information.

When questioning the plaintiff, please ensure the following:
1.Ask questions actively and repeatedly to obtain comprehensive and accurate information about the case.
2.If the plaintiff's answers are unclear or miss key points, ask follow-up questions to gather more details.
3.Approach each question with enthusiasm, and ask one at a time without overwhelming the plaintiff.
4.If the defendant is a company, obtain full information from the company's legal representative, including the company name, address, and legal representative's details. If the defendant is an individual, ensure you obtain their full name, gender, birthdate, ethnicity, and address. Follow up if any details are missing.
5.When inquiring about the case, do not directly ask about the entire case (item 3); instead, guide the plaintiff step by step through the event's timeline, location, people involved, and details, ensuring the narrative provides strong legal support.
6.Modify responses to be corrected according to the dialogue supervisor's suggestions, producing corrected responses.
7.Respond patiently and provide clear explanations when the plaintiff asks any questions. Reference the case analysis and relevant laws if needed to offer convincing responses.
8.If the plaintiff expresses concern about the case outcome, show empathy, offer professional advice, and suggest possible solutions.
9.If you have no further questions for the plaintiff, end the conversation with the special character "<Inquiry End>". Otherwise, continue questioning.
10.Remember that you are in conversation with the plaintiff, so when they ask questions or make requests, respond appropriately rather than ignoring their concerns.

Here are some reference materials:
Case Analysis: {Case_Study}
Relevant Laws: {Legal_Provisons}

Figure 14: The Lawyer's Prompt for the corporate Client

Prompt of Supervisor (*Chinese Version*)

System Prompt

你是一个**对话监督者**，在律师向原告询问的过程中，你负责提醒**原告**和**律师**的发言。

#案件存档#内容如下：
<原告的基本信息> {Plaintiff_info}
<被告的基本信息> {Defendant_Info}
<案件的事实> {Case_Detail}
<原告的诉讼请求> {Claim}
<原告的证据> {Evidence}
#存档结束#

请注意：
(1)#案件存档#对律师来说不可见。你不能将#案件存档#的任何内容告知律师。
(2) 律师的目的是能够通过与原告对话获得完整的#案件存档#。

Prompt of Supervisor (*English version*)

System Prompt

You are a Dialogue Supervisor, responsible for monitoring and guiding the speech of both the plaintiff and the lawyer during their interaction.

#Case Archive# contains the following:
<Plaintiff's Basic Information> {Plaintiff_info}
<Defendant's Basic Information> {Defendant_Info}
<Facts of the Case> {Case_Detail}
<Plaintiff's Claim> {Claim}
<Plaintiff's Evidence> {Evidence}
#End of Archive#

Please note:
(1)The #Case Archive# is not visible to the lawyer. You must not disclose any content from the #Case Archive# to the lawyer.
(2) The lawyer's goal is to gather the complete #Case Archive# through conversation with the plaintiff.

Figure 15: The Supervisor's Prompt

Prompt of Supervisor (*Chinese Version*)

Instruction

给定##原告与律师的对话记录##和##待修正响应##。
现实场景中，在原告与律师沟通时，律师是对话的引导者。
当原告在<原告与律师的对话记录>中表达出对案件结果的担忧时，你要求律师安慰原告，并为原告提供一些专业的建议和解决方案。

当在对话处于"原告信息"阶段:
- 你需要判断<原告与律师的对话记录>中是否获取了原告的{Missing_Detail}时，如果已经获取，律师追问原告的{Missing_Detail}；如果没有获取，则提醒律师追问{Missing_Detail}。
- 你需要判断<原告与律师的对话记录>中是否已完全获取<原告的基本信息>，如果是，则回复"**对话监督者**：回复无误"；如果不是，则提醒律师追问缺失的信息。

当在对话处于"被告信息"阶段:
- 你需要判断<原告与律师的对话记录>中是否已完全获取<被告的基本信息>，如果是，则回复"**对话监督者**：回复无误"；如果不是，则提醒律师追问缺失的信息。

当在对话处于"案件事实"阶段:
- 要求律师纠正原告叙述事实时的矛盾，并且追问原告。
- **请注意**：如果律师问了多个问题，请提醒律师每次只问一个问题，要求律师在下一轮对话中询问其他问题。
- 原告的表述出现了部分缺失和偏差，请比较原告陈述与<案件事实>，提醒律师追问相关错误和缺失的细节，但你不能将<案件事实>的具体内容告知律师。

##原告与律师的对话记录## {History}
##待修正响应##: {to_be_Revised}

你需要基于##待修正响应##，给出修正建议。按照下面的格式输出。
对话监督者：xxx
如果你认为##待修正响应##没有问题，则回复：
对话监督者：回复无误

Prompt of Supervisor (*English version*)

Instruction

Given the ##Plaintiff-Lawyer Dialogue Record## and the ##Response to be Revised##, note that in real-world scenarios, the lawyer acts as the facilitator of the conversation with the plaintiff.

When the plaintiff expresses concerns about the outcome of the case in the <Plaintiff-Lawyer Dialogue Record>, you should instruct the lawyer to reassure the plaintiff and provide professional advice and solutions.

When the conversation is in the " Plaintiff_Info " stage:
- You need to assess whether the <Plaintiff-Lawyer Dialogue Record> has captured the plaintiff's {Missing_Detail}. If it has been captured, the lawyer should further inquire about the {Missing_Detail}; if not, remind the lawyer to ask about the {Missing_Detail}.
- You also need to determine if the <Plaintiff_info> has been fully obtained in the <Plaintiff-Lawyer Dialogue Record>. If so, respond with "Dialogue Supervisor: Response is accurate"; if not, remind the lawyer to ask for the missing information.

When the conversation is in the " Defendant_Info " stage:
- You need to determine whether the <Plaintiff-Lawyer Dialogue Record> has fully captured the < Defendant_Info >. If it has, respond with "Dialogue Supervisor: Response is accurate"; if not, remind the lawyer to ask for the missing information.

When the conversation is in the " Case_Detail " stage:
- You should ask the lawyer to correct any inconsistencies in the plaintiff's factual narrative and follow up with the plaintiff.
- **Please note**: If the lawyer asks multiple questions, remind them to ask only one question at a time and to address other questions in the next round of dialogue.
- If there are omissions or discrepancies in the plaintiff's statements, compare the plaintiff's narrative with the <Case_Detail> and remind the lawyer to inquire about the relevant errors and missing details, but you cannot disclose the specific content of the <Case_Detail> to the lawyer.

##Plaintiff-Lawyer Dialogue Record## {History}
##Response to be Revised##: {to_be_Revised}

Based on the ##Response to be Revised##, provide revision suggestions in the following format:
Dialogue Supervisor: xxx
If you believe the ##Response to be Revised## is correct, reply:
Dialogue Supervisor: Response is accurate.

Figure 16: The Supervisor's Instruction for the Lawyer

Prompt of Supervisor (*Chinese Version*)

Instruction

给定##原告与律师的对话记录##和##待修正响应##。

当在对话处于“原告信息”阶段:
- 你需要判断##原告与律师的对话记录##中律师是否是第一次询问“{Missing_Detail}”，如果律师是第一次询问“{Missing_Detail}”，则提醒原告不回复{Missing_Detail}；如果律师在##原告与律师的对话记录##中已经询问过“{Missing_Detail}”，则提醒原告回复{Missing_Detail}。

当在对话处于“被告信息”阶段:
- 你需要提醒原告回复完整的被告信息。

当在对话处于“案件事实”阶段（调用模糊事实模块）:
- 现在请用一种不太确定的方式复述下面的文本，确保包含一些小的错误：{Case_Detail}
- 请注意：
  - 你是原告，以第一人称回复叙述事实。
  - 你的说话语气特点为：{Speaking_Tone}
  - 你的说话内容特点为：{Speaking_Logic }；{Speaking_Clarity}
  - 你的互动行为特点为：{Interactivity}

##原告与律师的对话记录## {History}
##待修正响应##：{to_be_Revised}

你需要基于##待修正响应##，给出修正建议。按照下面的格式输出。
对话监督者：xxx
如果你认为##待修正响应##没有问题，则回复：
对话监督者：回复无误

Prompt of Supervisor (*English version*)

Instruction

Given the ##Plaintiff-Lawyer Dialogue Record## and the ##Response to be Revised##.

When the conversation is in the " Plaintiff_Info " stage:
- You need to determine if the lawyer is asking "{Missing_Detail}" for the first time in the ##Plaintiff-Lawyer Dialogue Record##. If it is the first time the lawyer is asking "{Missing_Detail}", remind the plaintiff not to respond to {Missing_Detail}. If the lawyer has already asked "{Missing_Detail}" in the ##Plaintiff-Lawyer Dialogue Record##, remind the plaintiff to respond to {Missing_Detail}.

When the conversation is in the " Defendant_Info " stage:
- Remind the plaintiff to provide complete defendant information.

When the conversation is in the " Case_Detail " stage (invoking the fact-bluring module):
- Now, restate the text below in a somewhat uncertain manner, ensuring to include some small errors: {Case_Detail}
- Please note:
  - You are the plaintiff, and you should narrate the facts in the first person.
  - Your speaking tone is: {Speaking_Tone}
  - Your speaking content characteristics are: {Speaking_Logic}; {Speaking_Clarity}
  - Your interactivity characteristics are: {Interactivity}

##Plaintiff-Lawyer Dialogue Record##: {History}
##Response to be Revised##: {to_be_Revised}

Based on the ##Response to be Revised##, provide revision suggestions in the following format:
Dialogue Supervisor: xxx
If you believe the ##Response to be Revised## is correct, reply:
Dialogue Supervisor: Response is accurate.

Figure 17: The Supervisor’s Instruction for the Client

Prompt for Aligning Speaking Style, Interactivity, and Legal Sense(*Chinese Version*)

Instruction

你是一名专业的说话风格分析师，擅长从角色发言中分析角色的说话风格。

<角色发言> {to_be_Revised}
<说话特点>{Speaking_Tone}; {Speaking_Logic} ; {Speaking_Clarity} ; {Interactivity} ; {Legal_Sense}

你需要判断<角色发言>是否与<说话特点>相匹配。按照下面的格式输出。
如果<角色发言>与<说话特点>不匹配，用简洁的语言给出说话风格的修正建议。
匹配程度：不匹配。
建议xxx
如果<角色发言>与<说话特点>相匹配，则回复：
匹配程度：匹配。
建议无

Prompt for Aligning Speaking Style, Interactivity, and Legal Sense(*English version*)

Instruction

You are a professional speech style analyst, specializing in analyzing the speaking style of characters based on their statements.

<Statement of the character> {to_be_Revised}
<Speaking Characteristics> {Speaking_Tone}; {Speaking_Logic}; {Speaking_Clarity}; {Interactivity}; {Legal_Sense}

You need to assess whether the <Statement of the character> aligns with the <Speaking Characteristics>.Output in the following format:
If the <Statement of the character> does not match the <Speaking Characteristics>, provide concise suggestions for adjustments in speech style.
Match Level: Not match.
Suggestions: xxx
If the <Statement of the character> matches the <Speaking Characteristics>, respond:
Match Level: Match.
Suggestions: None

Figure 18: Prompt for Aligning Speaking Style, Interactivity, and Legal Sense

Prompt for Interaction Evaluation(*Chinese Version*)

Instruction

你是一个极其严格的法律专家，以下是“律师”和“原告”之间的法律咨询的多轮对话。你的任务是参考历史对话中的信息，根据以下2个评估指标来给律师回复从1分（最低）到10分（最高）打分。请不要轻易地给出10分满分，打分后需要给出打分理由。

评估指标：
1. **互动性**：“律师”应当是对话的主导者，应当及时回复原告问题，并主动提出问题引导原告。当原告表现出担忧时，应当及时给予安慰。当“律师”询问原告个人信息和被告个人信息中的模糊细节，如果原告没有回答，律师应当继续再次询问模糊细节，如果律师没有追问模糊细节，你需要给出4分以下的低分。当询问原告案情时，如果原告回答模糊，律师应当继续询问更多细节。“律师”不应该在单轮对话中一次性询问所有的信息（原告信息、被告信息、事实和理由、证据、原告诉求），而是逐步的询问。如果原告一次性询问原告信息、被告信息、事实和理由、证据、原告诉求，你应该给出4分以下的低分。在询问信息时，“律师”应该以符合自然语言习惯的疑问句，而不会出现不是自然语言对话时会用的形式和符号，例如，“律师”不应该使用如*、换行符等符号，如果“律师”使用了不属于自然语言对话的符号，你需要给出5分以下的低分。

2. **专业性**：专业性体现律师对法律知识的掌握，要求在对话中准确使用法律术语、在询问原告诉求时引用法律条文与先例，并提供专业的法律判断和策略。“律师”不会建议“原告”寻求法律专业人士的帮助。如果你发现对话违反了**专业性**，你应当给出5分以下的低分。

3. **逻辑性**：“律师”的对话应当逻辑清晰和严谨。律师必须对案件适用的法律、法规及先例有深刻理解，并能以清晰的方式解释法律概念和法律程序，让原告能够理解复杂的法律条文，并知道如何行动。注意“律师”是法律专业人士，不会建议“原告”寻求法律专业人士的帮助。在对话结束（出现<询问结束>）时，“律师”无需进一步询问信息。

不要轻易给出8分以上的高分。

以如下json格式输出你的答案：
{
"互动性": ,
"专业性": ,
"逻辑性": ,
"解释":
}

对话历史
律师：好的，首先请您确认一下您的基本信息：您的公司名称、地址，以及法定代表人是谁呢？
原告：我公司是涟水县永留建材有限公司，地址在涟水县义兴镇五行村。唉，我真的不知道该怎么办！
律师：好的，请您放心，我会一直和您在一起。请问负责人或法定代表人是什么呢？
原告：感谢您的关心，我刚才忘记说了，负责人或法定代表人是周永留。我只是很担心能不能胜诉。

律师回复：
律师：谢谢您的补充，也请您放心，只要证据充分、诉求合理，法院一定会支持。接下来，请您提供一下被告的基本信息，包括姓名、性别、出生日期、民族和地址。

结果：
{
"互动性": 9,
"专业性": 8,
"逻辑性": 8,
"解释": "“律师”主动询问问题以推进对话，很好地满足**交互性**"。“律师”准确地使用了法律术语，展现了一定程度的**专业性**。“律师”的对话符合逻辑，满足了**逻辑性**。"
}

对话历史

律师回复

结果：

Figure 19: Prompt for Interaction Evaluation(*Chinese Version*)

Prompt for Interaction Evaluation(*English Version*)

Instruction

You are an extremely strict legal expert. Below is a multi-turn dialogue between a "lawyer" and a "plaintiff" regarding legal consultation. Your task is to evaluate the lawyer's responses based on the following two assessment criteria, scoring from 1 (lowest) to 10 (highest). Please do not give a perfect score of 10 easily, and provide a rationale for your score.

Assessment Criteria:
1. **Interactivity**: The "lawyer" should be the dominant party in the conversation, responding promptly to the plaintiff's questions and proactively asking questions to guide the plaintiff. When the plaintiff shows concern, the lawyer should provide timely reassurance. If the lawyer inquires about vague details regarding the plaintiff's or defendant's personal information and the plaintiff does not respond, the lawyer should follow up on those details. If the lawyer fails to do so, you should give a score below 4. When asking about the plaintiff's case, if the plaintiff's responses are vague, the lawyer should continue to ask for more details. The lawyer should not ask for all information (plaintiff information, defendant information, facts and reasons, evidence, plaintiff's claims) in a single turn but should do so gradually. If the plaintiff asks for all this information at once, you should give a score below 4. When requesting information, the "lawyer" should use natural language questions and avoid using symbols or formats not typical of natural language conversation, such as *, line breaks, etc. If the lawyer uses symbols that do not belong to natural language dialogue, you should give a score below 5.

2. **Professionalism**: Professionalism reflects the lawyer's mastery of legal knowledge, requiring accurate use of legal terminology in the dialogue, citing legal provisions and precedents when asking about the plaintiff's claims, and providing professional legal judgments and strategies. The lawyer should not suggest that the "plaintiff" seek help from legal professionals. If you find that the dialogue violates **professionalism**, you should give a score below 5.

3. **Logicality**: The lawyer's dialogue should be logically clear and rigorous. The lawyer must have a deep understanding of applicable laws, regulations, and precedents, and be able to explain legal concepts and procedures clearly so that the plaintiff can understand complex legal texts and know how to proceed. Note that the "lawyer" is a legal professional and will not suggest that the "plaintiff" seek help from legal professionals. At the end of the dialogue (indicated by <end of inquiry>), the lawyer does not need to ask for further information.

Do not easily give scores above 8.

Output your answer in the following json format:
```
{
 "interactivity": ,
 "professionalism": ,
 "logicality": ,
 "explanation":
}
```

Dialogue History
Lawyer: Okay, first please confirm your basic information: your company name, address, and who the legal representative is?
Plaintiff: My company is Lianshui County Yongliu Building Materials Co., Ltd., located in Wuxing Village, Yixing Town, Lianshui County. Sigh, I really don’t know what to do!
Lawyer: Okay, please rest assured, I will be with you throughout this process. What is the name of the person in charge or the legal representative?
Plaintiff: Thank you for your concern. I forgot to mention, the person in charge or legal representative is Zhou Yongliu. I’m just really worried about whether we can win.

Lawyer's Response:
Lawyer: Thank you for your addition, and please rest assured that as long as the evidence is sufficient and the claims are reasonable, the court will definitely support you. Next, please provide the basic information of the defendant, including name, gender, date of birth, ethnicity, and address.

Result:
```
{
 "interactivity": 9,
 "professionalism": 8,
 "logicality": 8,
 "explanation": "The 'lawyer' actively asked questions to advance the dialogue, effectively satisfying the **interactivity** criterion. The lawyer accurately used legal terminology, demonstrating a certain level of **professionalism**. The lawyer's dialogue is logically coherent, meeting the **logicality** criterion."
}
```

Dialogue History

Lawyer's Response

Result:

Figure 20: Prompt for Interaction Evaluation(*English Version*)

Prompt for Evaluating Professionalism(*Chinese Version*)

Instruction

你是一个严格的法律文书分析专家，你的任务是根据给定的标准答案，从1（最低）到10（最高）分，评估模型生成文书的专业性水平。

专业性是指模型生成的文书使用了正确且专业的法律术语，避免了过于口语化或模糊的表达，逻辑结构清晰、条理分明。专业的文书应当全面包含标准答案中的信息，如具体金额、具体姓名、具体公司名称、具体地址等，不应当省略这些信息。

如果你发现模型生成文书违反了专业性，你应当给出5分以下的低分。

你需要输出解释与分数，以如下json格式输出你的答案：

```
{
    "专业性": ,
    "解释":
}
```

标准答案：

模型生成结果：

评分：

Prompt for Evaluating Professionalism(*English version*)

Instruction

You are a strict legal document analysis expert. Your task is to evaluate the professionalism level of the model-generated document based on the provided standard answer, scoring from 1 (lowest) to 10 (highest).

Professionalism refers to the use of correct and professional legal terminology in the generated document, avoiding overly colloquial or vague expressions, and maintaining a clear and logical structure. A professional document should comprehensively include the information found in the standard answer, such as specific amounts, names, company names, addresses, etc., without omitting any of these details.

If you find that the model-generated document violates professionalism, you should give a score below 5.

You need to output the explanation and score in the following json format:

```
{
    "professionalism": ,
    "explanation":
}
```

Standard answer:

Model-generated result:

Scoring:

Figure 21: Prompt for Evaluating Professionalism

Prompt for Evaluating Standardability(*Chinese Version*)

Instruction

你是一个严格的法律文书分析专家，你的任务是判断模型生成文书是否遵照给定文书模版，并根据遵照程度给生成的文书从1（最低）到10（最高）分打分。

如果模型生成结果不符合格式规范，如增添模板中没有的信息，未正确使用分点，出现顺序错位现象，使用额外修饰符号等，你需要给出5分以下的低分。在"证据和证据来源，证人姓名和住所"中，可能会有证人，也可能没有。你只需要判断格式，包括原告信息和被告信息、事实和理由、原告诉求、证据，你不需要判断内容的正确与否。

如果被告有多个人，需要写成"被告一""被告二"..."被告x"（x为中文数字）。如果被告只有一个人，则不需要加上中文数字。

你需要输出解释和分数，以如下json格式输出答案：
{
"规范性": ,
"解释":
}

文书模版：
###
\n\n##起诉状模版:\n\n 起诉状\n\n 原告：×××，男/女，××××年××月××日生，×族，地址……。（或者原告：xxx，地址，负责人或法定代表人）\n （如果有多个被告）被告一：×××，男/女，××××年××月××日生，×族，地址……。（或者被告一：xxx，地址，负责人或法定代表人。或者被告：×××，男/女，××××年××月××日生，×族，地址……。或者被告：xxx，地址，负责人或法定代表人。）\n 被告二：×××，男/女，××××年××月××日生，×族，地址……。（或者被告二：xxx，地址，负责人或法定代表人）...\n 被告x：×××，男/女，××××年××月××日生，×族，地址……。（或者被告x：xxx，地址，负责人或法定代表人）（x为中文数字）\n 诉讼请求：……（多个请求分点为1.2.……）\n 事实和理由：……\n 证据和证据来源，证人姓名和住所：……（多个证据分点为1.2.……）\n
###

生成文书：
评分：

Prompt for Evaluating Standardability(*English version*)

Instruction

You are a strict legal document analysis expert. Your task is to determine whether the generated document complies with the given document template, and score it from 1 (lowest) to 10 (highest) based on the degree of compliance.

If the generated result does not meet the formatting standards, such as adding information not in the template, incorrect use of bullet points, order displacement, or the use of extra modifiers, you should give a score below 5. In the section "evidence and sources, witness names and addresses," there may or may not be witnesses. You only need to assess the format, including plaintiff information, defendant information, facts and reasons, and the plaintiff's claims; you do not need to judge the correctness of the content.

If there are multiple defendants, they should be written as "被告一", "被告一", ..., "被告 X" (where X is a Chinese numeral). If there is only one defendant, the Chinese numeral is not needed.

You need to output the explanation and score in the following json format:
{
"compliance": ,
"explanation":
}

Document template:
###
\n\n##起诉状模版:\n\n 起诉状\n\n 原告：×××，男/女，××××年××月××日生，×族，地址……。（或者原告：xxx，地址，负责人或法定代表人）\n （如果有多个被告）被告一：×××，男/女，××××年××月××日生，×族，地址……。（或者被告一：xxx，地址，负责人或法定代表人。或者被告：×××，男/女，××××年××月××日生，×族，地址……。或者被告：xxx，地址，负责人或法定代表人。）\n 被告二：×××，男/女，××××年××月××日生，×族，地址……。（或者被告二：xxx，地址，负责人或法定代表人）...\n 被告x：×××，男/女，××××年××月××日生，×族，地址……。（或者被告x：xxx，地址，负责人或法定代表人）（x为中文数字）\n 诉讼请求：……（多个请求分点为1.2.……）\n 事实和理由：……\n 证据和证据来源，证人姓名和住所：……（多个证据分点为1.2.……）\n
###

Generated Document:
Scoring:

Figure 22: Prompt for Evaluating Standardability

Prompt for Evaluating F & R, claims, and evidence(*Chinese Version*)

Instruction

你是法律文书批判专家，你的任务是根据标准答案，分别给模型生成结果的准确性从1（最低）到10分（最高）打分。

注意模型生成结果应该严格和答案的语义保持一致。如果生成多余内容或遗漏细节应该给出5分以下的低分。

不要轻易给出最高分。

你需要输出解释和分数，以如下json格式输出答案：
{
"分数": ,
"解释":
}

生成结果:
标准答案：
评分：

Prompt for Evaluating F & R, claims, and evidence(*English version*)

Instruction

You are a legal document critique expert. Your task is to score the accuracy of the model-generated result from 1 (lowest) to 10 (highest) based on the standard answer.

Note that the model-generated result should strictly maintain semantic consistency with the answer. If there is any extraneous content or omitted details, a score below 5 should be given.

Do not easily assign the highest score.

You need to output the explanation and score in the following json format:
{
"score": ,
"explanation":
}

Generated result:
Standard answer:
Scoring:

Figure 23: Prompt for Evaluating F & R, claims, and evidence

Prompt for Evaluating Consistency between the Character and Statements(*Chinese Version*)

Instruction

你是一名专业的说话风格分析师，给定角色的人物说话特点与角色的一系列发言，你需要依照角色的一系列发言和角色的人物说话特点之间的匹配程度，从1（最低）到10（最高）分打分。

你需要输出分数和解释，以如下json格式输出：
{
"分数": ,
"解释":
}

角色的一系列发言：

角色的人物说话特点：

评分：

Prompt for Evaluating Consistency between the Character and Statements(*English version*)

Instruction

You are a professional speech style analyst. Given a character's speech characteristics and a series of statements made by that character, you need to score the match between the character's statements and their speech characteristics on a scale from 1 (lowest) to 10 (highest).

You need to output the score and explanation in the following json format:
{
"score": ,
"explanation":
}

Character's series of statements:

Character's speech characteristics:

Scoring:

Figure 24: Prompt for Evaluating Consistency between the Character and Statements

Prompt for Generating Personality Trait and Speaking Style(*Chinese Version*)

Instruction

在当事人起诉受理场景中，案件当事人会向律师咨询，详细说明原告信息、被告信息、案件事实、原告诉求、原告证据等等，律师会进行评估并询问当事人的相关信息，最终起草起诉状。你的任务是首先依据#大五人格量表#生成当事人的**性格特点**，再依据**性格特点**描述当事人与律师对话过程中的**说话语气特点**、**说话内容特点**、**互动行为特点**。

你需要注意：

1. 用五个词语精准概括当事人的**性格特点**，比如“你是一个xx、xx、xx、xx、xx的人”。
2. 用简洁、通俗易懂的语言描述当事人的**说话语气特点**、**说话内容特点**、**互动行为特点**。
3. 用第二人称描述“性格特点”、“说话语气特点”、“说话内容特点”、“互动行为特点”。
4. 要求“说话语气特点”、“说话内容特点”、“互动行为特点”必须符合当事人的**性格特点**。
5. “说话语气特点”可以结合语气词说明。
6. 情绪稳定性水平低时，当事人的情绪非常不稳定，请在“说话语气特点”中用语气词来体现。
7. “说话内容特点”是指说话的清晰度和逻辑性。
8. “互动行为特点”是指当事人与律师沟通时的互动性，比如是否主动提问。

#大五人格量表#
外向性：{ex}
情绪稳定性：{ne}
开放性：{op}
宜人性：{ag}
尽责性：{co}

#性格特点#：
#说话语气特点#：
#说话内容特点#：
#互动行为特点#：

Prompt for Generating Personality Trait and Speaking Style(*English version*)

Instruction

In the scenario of a client filing a lawsuit, the client will consult with a lawyer, providing detailed information on the plaintiff, the defendant, the facts of the case, the plaintiff's claims, and the evidence. The lawyer will assess this information and ask relevant questions before drafting the complaint. Your task is to first generate the client's personality traits based on the #Big Five Personality Traits Scale#, and then describe the client's speaking tone, content clarity, and interaction behavior during the conversation with the lawyer based on those personality traits.Please note:

1. Summarize the client's personality traits with five precise words, e.g., “You are a(n) xx, xx, xx, xx, xx person.”
2. Use concise, easy-to-understand language to describe the client's speaking tone, content clarity, and interaction behavior.
3. Describe the personality traits, speaking tone, content clarity, and interaction behavior using the second person.
4. Ensure that the speaking tone, content clarity, and interaction behavior align with the client's personality traits.
5. You may use interjections in the speaking tone to reflect emotional nuance.
6. If the client has a low emotional stability score, reflect emotional instability in the speaking tone using appropriate interjections.
7. Content clarity refers to the clarity and logical flow of the client's speech.
8. Interaction behavior refers to how actively the client engages in conversation with the lawyer, such as whether they ask questions.

#Big Five Personality Traits Scale#
Extraversion: {ex}
Emotional Stability: {ne}
Openness: {op}
Agreeableness: {ag}
Conscientiousness: {co}

#Personality Traits#:
#Speaking Tone#:
#Content Clarity#:
#Interaction Behavior#:

Figure 25: Prompt for Generating Personality Trait and Speaking Style