基于多智能体协作的法庭模拟与法律文书生成方法研究

周裕林，秦永彬，林川
贵州大学 计算机科学与技术学院 贵州 贵阳 550025

摘要
随着大语言模型在自然语言处理领域的突破性进展，利用人工智能技术辅助司法文书生成与裁判推理已成为法律人工智能研究的重要方向。然而，现有方法在应用于法庭模拟任务时仍面临法律信息不完整、文本结构强约束、司法推理过程不可解释以及复杂庭审流程难以建模等科学问题。针对上述挑战，本文提出了一种角色驱动的多阶段法庭模拟框架（AgentCourts），旨在有效缓解信息获取缺失、提升文书结构规范性、增强推理可解释性并优化庭审流程建模。该方法首先通过多轮对话式信息采集机制逐步抽取案件关键要素，将起诉状生成建模为信息抽取与整合问题；其次采用对抗式法律文本建模策略生成答辩状，确保答辩内容与起诉事实在语义层面形成有效回应；再次严格按照民事一审流程显式建模开庭审理、法庭调查、法庭辩论、评议宣判等阶段，并引入证据驱动的推理机制；最后通过模板约束下的内容填充策略，将法律文书生成从"自由生成"转化为"结构化填空"，并调用法律条文检索接口提升法律适用准确性。在 CAIL 2025 法庭模拟赛道数据集上的实验结果表明，该方法能够稳定生成结构规范、逻辑一致的法律文书，在起诉状、答辩状及判决生成三个子任务上均取得了优异性能，验证了多阶段建模与模板约束生成策略在司法场景中的有效性与实用性。

关键词
法庭模拟；大语言模型；法律文书生成；多阶段推理；模板约束

中图分类号：D926.2；TP18；TP391.1　　文献标志码：A　　doi：10.11959/j.issn.2096–0271.2024xxx

A Multi-Agent Collaborative Framework for Courtroom Simulation and Legal Document Generation

Zhou Yulin, Qin Yongbin, Lin Chuan
College of Computer Science & Technology, Guizhou University, Guiyang 550025, China

Abstract
With the breakthrough progress of large language models (LLMs) in natural language understanding and generation, leveraging artificial intelligence to assist legal document drafting and judicial reasoning has emerged as a pivotal direction in legal AI research. However, existing methods applied to courtroom simulation tasks still face critical scientific challenges, including incomplete legal information acquisition, strong structural constraints of legal texts, uninterpretable judicial reasoning processes, and difficulties in modeling complex multi-stage court procedures. To address the aforementioned challenges, this paper proposes a role-driven, multi-stage court simulation framework

(AgentCourts), which aims to effectively alleviate information gaps, improve the standardization of document structure, enhance the interpretability of reasoning, and optimize trial process modeling. Specifically, our method first employs a multi-turn dialogue-based information collection mechanism to incrementally extract key case elements, formulating complaint generation as an information extraction and integration problem; second, it adopts an adversarial legal text modeling strategy to generate defense statements that semantically respond to the plaintiff's claims; third, it explicitly models the civil trial workflow, including court session initiation, evidence investigation, debate, mediation, and judgment deliberation by introducing evidence-driven reasoning mechanisms; finally, it transforms legal document generation from "free-form generation" to "structured template filling" under template constraints, while invoking a legal article retrieval interface to enhance the accuracy of legal applicability. Experimental results on the CAIL 2025 Courtroom Simulation dataset demonstrate that the proposed method consistently generates legally compliant, structurally standardized, and logically coherent documents, achieving superior performance across three subtasks (complaint, defense statement, and judgment generation), thereby validating the effectiveness and practicality of multi-stage modeling and template-constrained generation in judicial scenarios.

Key words
courtroom simulation, large language models, legal document generation, multi-agent collaboration, template-constrained generation

0 引言

近年来，人工智能技术在司法领域的深度应用已成为国家法治建设的重要战略方向。最高人民法院于2022年发布《关于规范和加强人工智能司法应用的意见》，明确提出要加强人工智能应用顶层设计，设计完善智慧法院人工智能相关信息系统体系架构和技术标准体系[1]。2024年《法治蓝皮书》指出，数字法院大脑已完成与司法大模型的深入集成，具备为法院用户提供智能检索、智能交互和智能推荐能力[2]。2025年全国两会期间，有政协委员建议以人工智能促进司法审判现代化，将传统案件预审、证据分析、判决辅助等环节的工作效率提升数百倍[3]。在这一政策背景下，如何利用大语言模型技术实现法庭全流程模拟与法律文书自动生成，成为智慧司法建设亟待突破的关键任务。

法庭模拟与法律文书生成作为司法人工智能的核心应用场景，其技术实现涉及起诉状、答辩状及裁判文书等多种法律文书的自动化生成[4]。如图1所示，该任务不仅要求模型具备强大的自然语言理解与生成能力，还需严格遵循司法程序的规范性与法律推理的严谨性[5]。从技术层面看，法庭模拟任务可分解为多阶段子任务：首先通过多轮交互采集案件关键信息生成起诉状，其次基于对抗式建模生成答辩状，最后通过庭审流程显式建模完成裁判推理与判决生成[6]。这一完整链条的自动化实现，对于缓解司法系统"案多人少"的矛盾、促进"同案同判"具有重要的实践价值。

然而，尽管大语言模型在通用文本生成任务中表现优异，直接将其应用于法庭模拟与法律文书生成仍面临诸多科学挑战。现有研究在法律信息完整性获取方面存在不足，真实案件中当事人信息与证据往往需通过多轮交互逐步补充而非一次性给定[7]。同时，司法文书具有固定模板与强

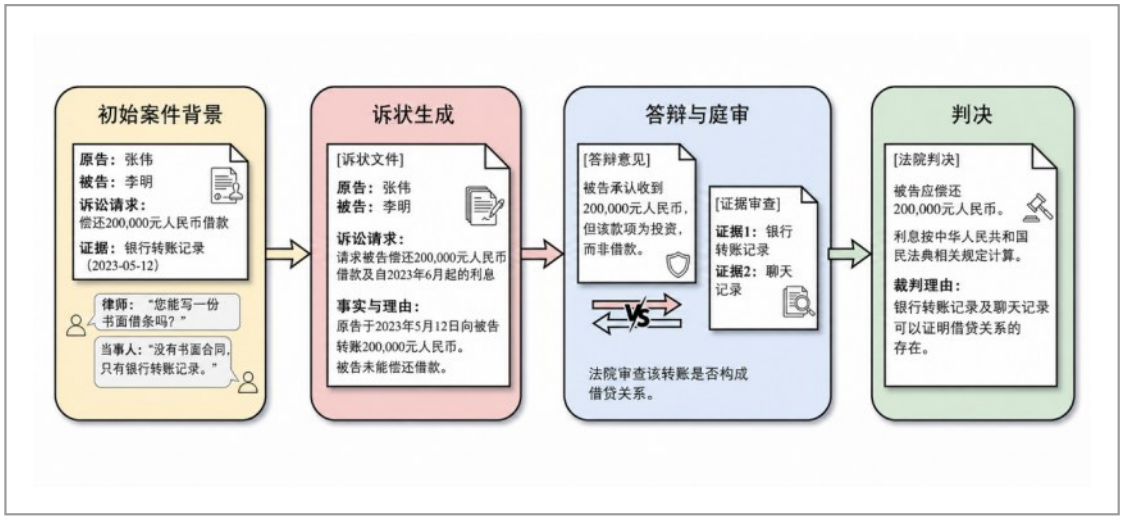

图1 智能法庭模拟示例图

格式约束，模型输出必须高度结构化，而当前生成式方法难以保证格式规范性。此外，裁判结果需要明确的事实认定与法律适用逻辑支撑，现有方法的司法推理过程缺乏可解释性[8]。最后，法庭审理涉及多阶段、多角色参与，流程顺序不可随意更改，复杂司法流程的建模难度较高[6]。这些问题严重制约了法律大模型在司法实务中的落地应用。

针对上述挑战，本文提出了一种基于多智能体（Agent）协作的法庭模拟与法律文书生成方法（AgentCourt）。该方法首先通过角色驱动的多智能体建模，显式设定律师、当事人、法官等角色，引导大语言模型在不同阶段采用不同的语言风格与推理方式[9]。其次采用模板约束下的生成式推理策略，将司法文书结构作为硬约束，模型仅负责填充内容，从而有效降低生成错误风险[10]。再次严格按照民事一审流程显式建模开庭审理、法庭调查、法庭辩论、评议宣判等阶段，并引入证据驱动的推理机制[11]。最后调用法律条文检索接口与模型筛选相结合的方式提升法律适用准确性。在CAIL 2025法庭模拟赛道数据集上的实验结果表明，该方法能够稳定生成结构规范、逻辑一致的法律文书，在起诉状、答辩状及判决生成三个子任务上均取得了优异性能。

本文的主要贡献如下：

● 提出面向完整司法流程的多阶段建模方案，通过多轮交互机制缓解法律信息不完整问题，实现起诉、答辩、裁判的端到端闭环。

● 将法律文书生成转化为模板约束下的内容填充问题，针对文本结构强约束问题，显著提升生成结果的稳定性与规范性。

● 显式建模庭审流程与证据推理机制，改善司法推理过程不可解释及流程建模难的问题，增强裁判说理的可解释性与法律适用的准确性。

1 相关工作

早期法律文书生成研究主要依赖于手工制定的规则模板与统计模型。这类方法的核心思想是将法律文书视为高度结构化的文本，通过预定义的模板框架填充关键

信息来完成生成任务。Chalkidis等人[12]提出了基于规则的法律文本分类框架，通过手工特征工程提取法律术语与句式模式，实现了初步的法律文书自动化处理。Aletras 等人[13]在此基础上引入了统计学习模型，利用支持向量机与随机森林对欧洲人权法院的判决结果进行预测，开创了数据驱动的法律推理研究先河。随后，Zhong 等[14]提出了基于隐马尔可夫模型的法律文书结构分析方法，将文书分解为事实陈述、法律适用与裁判结果三个层次，为后续的结构化生成奠定了基础。然而，这类方法严重依赖领域专家的手工规则设计，泛化能力有限，难以应对复杂多变的真实案件场景。

随着深度学习技术的兴起，基于神经网络的法律文书生成方法逐渐成为研究主流。这类方法能够自动学习法律文本的深层语义表示，减少了对人工特征的依赖。Dina 等人[15]提出了基于法律BERT模型来进行法律文本分类，在多个法律数据集上取得了优于传统统计方法的效果。Xu 等人[16]引入了注意力机制，使模型能够聚焦于案件事实中的关键证据与法律要素，提升了裁判推理的可解释性。在此基础上，Yang 等人[17]提出了基于图神经网络的法律要素关系建模方法，将案件中的实体与法律关系表示为图结构，有效捕捉了法律推理中的复杂依赖关系。Sukanya 等人[18]进一步设计了层次化神经网络架构，分别建模事实层、法律层与裁判层，实现了多粒度的法律推理。Zhang 等人[19]提出了基于多任务学习的法律判决预测框架，联合优化罪名预测、法条推荐与刑期预测三个相关任务，显著提升了模型的整体性能。尽管深度学习方法在自动特征学习方面表现出色，但仍存在训练数据需求量大、对小样本场景适应性差等问题[20]。

近年来，大语言模型的突破性进展为智能法庭模拟和法律文书生成带来了新的研究范式。Fei 等人[21]在LawBench上系统评估了大语言模型在法律推理任务上的表现，发现大模型在零样本场景下已能生成质量较高的法律分析文本。Zhou 等人[22]开发了 LawGPT 模型，在法律领域语料上进行继续预训练，显著提升了模型对法律术语与推理逻辑的理解能力。同时，随着法律文本生成技术的发展，智能法庭模拟交互智能体得到了广泛的研究。He等人[23]提出了一种 AgentsCourt 框架来模拟庭审辩论过程，并结合法律知识增强使 AI 智能体能够进行复杂司法决策。Chen 等人[6]提出 AgentCourt：一种使用对抗性可进化律师智能体来模拟完整法庭流程的系统，使律师智能体通过对抗进化不断提升法律辩论能力。Liao等人[24]设计并评估了一个具有行为忠实度的法庭 AI 模拟系统，该系统能够在自然语言处理会议上展示生成式 AI 在法庭审查任务中的可扩展性。Chun 等人[25]提出 AgenticSimLaw：一个具备明确角色结构和有组织辩论机制的少年法庭多智能体模拟框架，为高风险表格型决策任务提供透明可解释的审判式推理。

此外，近年来多智能体协作范式在复杂推理任务中得到广泛关注，如基于投票机制（Vote）[26]的多路径结果聚合方法以及基于辩论机制（Debate）[27]的对抗式推理方法。这类方法通过多个智能体之间的交互提升推理结果的鲁棒性与准确性。然而，这些通用范式主要面向开放式问答或自由生成任务，缺乏对强结构约束与固定流程任务的显式建模能力。综上所述，尽管上述研究在法庭模拟与法律文书生成方面取得了显著进展，但现有方法仍面临诸多局限性。首先，现有大模型方法多采用

自由生成范式，缺乏对法律文书固定模板与强格式约束的有效建模；其次，多数方法的司法推理过程缺乏显式的证据驱动机制，裁判说理的可解释性仍有待提升；最后，现有研究多聚焦于单一任务，缺乏对起诉、答辩、庭审、裁判等完整司法流程的端到端建模。

2 模型方法

2.1 总体架构与形式化定义

本文提出的基于多智能体协作的法庭模拟框架（AgentCourts）并非将法律文书生成视为单一的序列生成任务，而是将其建模为一个多阶段、层次化的条件概率推理过程。如图 2 所示，系统整体架构由信息采集智能体、对抗式答辩智能体与裁判推理智能体三个核心模块组成，各模块之间通过共享的状态空间进行信息流转。形式化地，本文将整个法庭模拟任务定义为从初始案件上下文 $X_{context}$ 到最终法律文书集合 Y_{docs} 的映射函数 F_{court}。设 $Y_{docs}=\{Y_{complaint}, Y_{defense}, Y_{judgment}\}$ 分别代表起诉状、答辩状与判决书，则整体生成过程可表示为联合概率分布的最大化问题：

$$P(Y_{docs}|X_{context})=\prod_{k\in\{c,d,j\}} P(Y_k|X_{context}, Y_{<k}, \Theta_k) \quad (1)$$

式中，Θ_k 表示第 k 阶段模型的参数集合；$Y_{<k}$ 表示当前阶段之前已生成的文书内容，体现了司法流程的因果依赖性。

为增强任务定义的清晰性，本文对法庭模拟任务进行形式化描述。设输入案件上下文为 X，包含当事人信息、案件事实与初始证据集合，输出为法律文书集合 $Y=\{Y_c, Y_d, Y_j\}$，分别表示起诉状、答辩状与判决书。整体任务可表示为一个条件生成过程：

$$Y=\mathcal{F}(X) \quad (2)$$

式中，函数 $\mathcal{F}$ 由多阶段智能体协作过程实现，具体包括信息采集、对抗生成与裁判推理三个子过程，每一阶段均在前序输出的条件下进行。

为了实现这一复杂的映射，本文基于 DeepSeek 系列大语言模型作为底层推理引擎，通过提示工程与结构化约束解码方式驱动多智能体协作完成生成任务。需要说明的是，本文不涉及对底层大语言模型参数的训练或微调，所有优化过程均作用于外部任务建模与推理控制层。系统核心在于引入角色嵌入向量 $\boldsymbol{e}_{role}\in\mathbb{R}^d$，用于区分律师、当事人与法官在不同阶段的推理模式，该向量并非通过模型训练得到的参数表示，而是通过以下方式构造：1）基于角色描述的提示模板输入大语言模型；2）利用模型隐式语义空间生成的上下文表示；3）在部分模块中通过外部编码器（如文本向量化接口）获取语义表示。因此，本文中的“向量表示”本质上来源于预训练模型的推理态表示，而非可训练参数。

需要说明的是，本文多智能体建模重点关注流程约束下的角色协同推理机制，而非构建具备复杂内部状态（如长期记忆或强化学习策略）的通用智能体系统。各智能体的“决策行为”主要通过提示工程与上下文条件控制实现，其行为策略由任务阶段与角色设定隐式决定。在此框架下，不同角色之间并非通过开放式多轮对话或显式博弈进行交互，而是通过共享中间状态实现隐式协同：例如，起诉状中抽取的案件要素与主张作为答辩阶段的条件输入，答辩内容进一步影响后续裁判推理过程，庭审阶段则将各方陈述与证据统一纳入共享上下文，由审判角色在流程约束下进行

整合与决策，从而在语义层面形成“起诉—答辩—裁判”的响应与对抗关系。此外，本文所提出的“对抗式建模”主要体现在语义层面的对抗关系建模，通过约束答辩内容与起诉主张之间的语义关系（如否认或承认）来实现，而非采用参数层面的对抗训练机制（如GAN）。该设计在一定程度上实现了多角色间的协同与对抗，同时避免了开放式多智能体交互带来的不稳定性，更符合闭源大语言模型无法进行梯度更新及司法场景对结果确定性与规范性的现实要求。

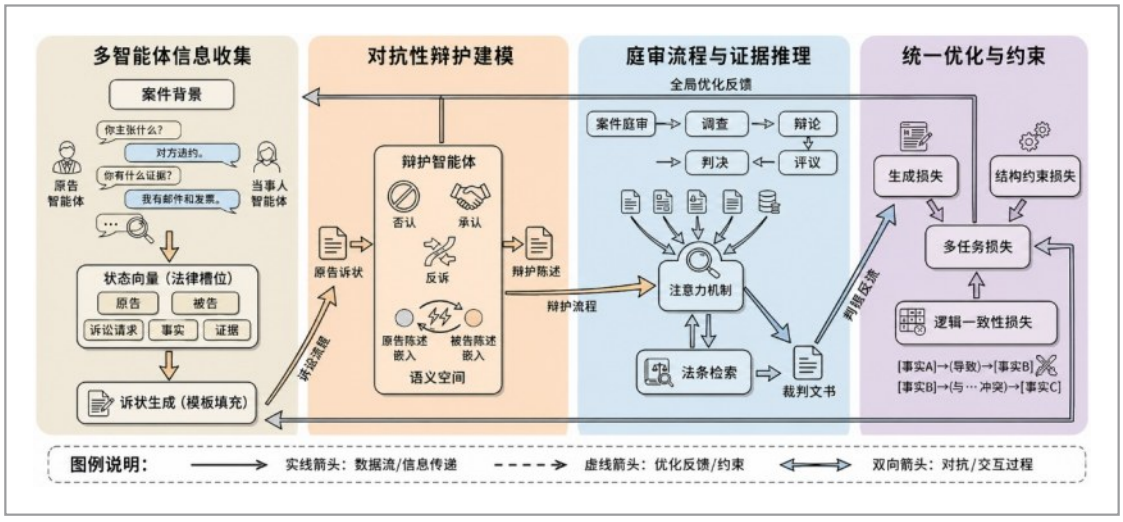

图2 AgentCourts框架图

2.2 起诉状生成：多轮信息抽取与模板填充

起诉状生成任务的核心挑战在于法律信息的不完整性与文本结构的强约束性。为此，本文将该任务建模为一个多轮对话状态追踪与槽位填充相结合的过程。设对话历史为 $H_t=\{q_1,a_1,...,q_t,a_t\}$，其中 q_t 为律师智能体提出的第 t 个问题，a_t 为当事人的回答。系统维护一个法律要素状态向量 $\boldsymbol{s}_t$，用于记录当前已采集的关键信息，如原告身份、诉讼请求、事实理由等。状态更新机制定义为：

$$\boldsymbol{s}_t=\mathrm{UPDATE}(\boldsymbol{s}_{t-1},\mathrm{IE}(q_t,a_t;\theta_{IE})) \quad (3)$$

式中，IE(·)表示信息抽取函数，θ_{IE} 为抽取模块参数。当状态向量 $\boldsymbol{s}_t$ 覆盖所有必需的法律槽位集合 $\mathcal{S}_{required}$ 时，系统触发文书生成机制。

为了保证文书格式的规范性，本文引入模板约束函数 $\mathcal{T}(\cdot)$。设模板结构为 $M_{template}$，生成目标是最小化生成文本 $Y_{complaint}$ 与模板结构之间的结构约束得分：

$$\mathcal{L}_{struct}=\sum_{i=1}^{N}\mathbb{I}(Y_{complaint}[i]\notin M_{template}[i])\cdot\lambda_{struct} \quad (4)$$

式中，$\mathbb{I}(\cdot)$ 为指示函数，λ_{struct} 为约束权重。

在实际生成中，本文将自由生成转化为结构化填空问题，即 $Y_{complaint}=\mathrm{FILL}(M_{template},\boldsymbol{s}_{final})$，其中 $\boldsymbol{s}_{final}$ 为对话结束时

的最终状态向量。这种方法有效避免了模型幻觉导致的格式错误，确保生成的起诉状符合法律的形式要求。

2.3 答辩状生成：对抗式语义建模

答辩状生成具有显著的对抗性特征，被告代理人需要针对原告的主张进行有效回应。本文将此过程建模为语义空间中的对抗博弈。设原告起诉状的关键主张向量表示为 $\boldsymbol{v}_{claim}=\text{Encoder}(Y_{complaint})$，被告答辩状的目标是生成一个语义向量 $\boldsymbol{v}_{defense}$，使其在特定维度上与 $\boldsymbol{v}_{claim}$ 形成反驳或承认关系。本文定义对抗性响应函数 $Resp(\cdot)$，其输出取决于被告的策略选择向量 $\boldsymbol{z}_{strategy}\in\{0,1\}^m$，其中每个维度代表一种诉讼策略（如否认事实、提出抗辩、请求驳回等）。

生成概率分布可表示为条件 softmax 函数：

$$P(y_t|y_{<t},\boldsymbol{v}_{claim},\boldsymbol{z}_{strategy})=\frac{\exp(\text{Score}(y_t,\boldsymbol{h}_{t-1},\boldsymbol{v}_{claim},\boldsymbol{z}_{strategy}))}{\sum_{y'\in V}\exp(\text{Score}(y',\boldsymbol{h}_{t-1},\boldsymbol{v}_{claim},\boldsymbol{z}_{strategy}))} \quad (5)$$

式中，$\text{Score}(\cdot)$ 为打分函数，V 为词表。

为了确保答辩内容与起诉事实的逻辑连贯性，本文引入语义一致性约束。设 $Sim(\cdot,\cdot)$ 为余弦相似度函数，本文要求答辩状中针对原告主张的回应部分 $Y_{response}$ 与原告主张 Y_{claim} 在语义空间上保持特定的距离关系。对于否认型答辩，本文希望降低相似度；对于承认型答辩，则保持高相似度。为刻画答辩语义与起诉主张之间的关系，引入对抗性语义评分函数：

$$\mathcal{L}_{adv}=-\sum_j\alpha_j\cdot\text{Sim}(\boldsymbol{v}^{(j)}_{defense},\boldsymbol{v}^{(j)}_{claim})\cdot(2z^{(j)}_{strategy}-1) \quad (6)$$

式中，α_j 为第 j 个主张的权重，$z^{(j)}_{strategy}=1$ 表示承认，0 表示否认。该公式确保模型能够根据策略动态调整语义回应强度，实现真正的“对抗式”生成，而非简单的文本复述。

2.4 判决生成：流程显式化与证据推理

民事判决生成是系统中技术难度最高的部分，涉及复杂的司法推理链条。本文将庭审过程显式建模为一个有限状态机（Finite State Machine, FSM），状态集合定义为 $\mathcal{Q}_{court}=\{q_{open},q_{invest},q_{debate},q_{mediate},q_{verdict}\}$，分别对应开庭、调查、辩论、调解与宣判阶段。状态转移概率由审判长智能体控制，转移矩阵 T_{court} 满足马尔可夫性质：

$$P(q_{k+1}|q_k,H_{court})=\text{Softmax}(\boldsymbol{W}_{trans}\cdot[\boldsymbol{h}_{q_k};\boldsymbol{h}_{H_{court}}]+\boldsymbol{b}_{trans}) \quad (7)$$

式中，$\boldsymbol{h}_{q_k}$ 为当前阶段的状态嵌入，$\boldsymbol{h}_{H_{court}}$ 为庭审历史上下文嵌入。

在法庭调查与辩论阶段，核心任务是证据推理。设证据集合为 $E=\{e_1,e_2,...,e_n\}$，本文需要计算每个证据对最终判决结果 $Y_{judgment}$ 的贡献度权重 β_i。采用注意力机制进行证据加权：

$$\beta_i=\frac{\exp(\boldsymbol{w}_{att}^T\tanh(\boldsymbol{W}_e\boldsymbol{e}_i+\boldsymbol{W}_f\boldsymbol{f}_{case}))}{\sum_{j=1}^{n}\exp(\boldsymbol{w}_{att}^T\tanh(\boldsymbol{W}_e\boldsymbol{e}_j+\boldsymbol{W}_f\boldsymbol{f}_{case}))} \quad (8)$$

式中，$\boldsymbol{f}_{case}$ 为案件事实的特征表示。

最终的法律适用环节，系统调用通义法睿法律条文检索接口[26]，获取候选法条集合 L_{cand}。本文通过双塔模型计算案件特征与法条特征的匹配得分 $Score_{law}(case,law_i)$，并选取 Top-K 个法条作为生成依据。判决生成的最终概率分布融合了证据权重与法律依据：

$$P(Y_{judgment}|E, L_{cand}) = \prod_t P(y_t|y_{<t}, \sum_i \beta_i \boldsymbol{e}_i, \text{Embed}(L_{selected})) \quad (9)$$

这种证据驱动与法条约束相结合的生成机制，显著提升了裁判说理的可解释性与法律适用的准确性，避免了模型凭空捏造法律依据。

2.5 约束评分与目标函数设计

为统一刻画多阶段生成过程中的质量约束，本文定义一个综合评分函数：

$$S = \mathcal{L}_{gen} + \lambda_1 \mathcal{L}_{struct} + \lambda_2 \mathcal{L}_{logic} \quad (10)$$

式中，$\mathcal{L}_{gen}$为标准的负对数似然评分，用于保证文本生成的流畅性；$\mathcal{L}_{struct}$为模板结构约束评分；$\mathcal{L}_{logic}$为逻辑一致性评分，用于约束起诉状、答辩状与判决书之间的事实一致性。

设$Fact(Y)$为从文书Y中提取的事实三元组集合，定义逻辑一致性惩罚函数：

$$\mathcal{L}_{logic} = \sum_{f_i \in Fact(Y_{complaint}), f_j \in Fact(Y_{judgment})} \mathbb{I}(Conflict(f_i, f_j)) \cdot \gamma \quad (11)$$

式中，$Conflict(\cdot)$为冲突检测函数，γ为惩罚系数。在模型训练与推理过程中，本文采用 DeepSeek 系列模型的推理接口，利用其原生支持的结构化输出能力，硬约束解码空间，确保$\mathcal{L}_{struct}$在推理阶段严格为零。

3 实验

3.1 数据集

本实验采用 CAIL 2025 法庭模拟赛道提供的官方数据集[①]，该数据集由中国法律智能技术评测组委会构建，旨在评估大语言模型在完整司法流程中的法律文书生成与裁判推理能力。数据集涵盖民事一审案件的完整审理流程，包含起诉状生成、答辩状生成与法庭判决生成三个核心子任务。数据集样本来源于中国裁判文书网公开的真实民事案件，经过专业法律人士脱敏处理与标准化标注，确保数据的合法性与可用性。在数据规模方面，采用比赛复赛阶段的数据以更好进行复现。其中，起诉状生成数据包含 80 个样本，答辩状生成数据包含 80 个样本，法庭审理数据包含 80 个样本。每个样本均包含案件基本信息、当事人信息、证据材料、庭审记录及最终裁判文书等完整司法要素。数据集按照案件类型进行了均衡分布，涵盖合同纠纷、侵权责任、婚姻家庭、劳动争议等常见民事案由，确保模型评估的全面性与代表性。尽管未包含刑事或行政案件，但上述四类民事案由已覆盖了司法实践中最高频的场景，足以验证框架在民事领域的泛化性能。

3.2 评价指标

为全面评估模型在法庭模拟任务中的表现，本实验采用 CAIL 2025 官方评测指标体系，涵盖文书信息准确性、模板遵循能力、庭审流程遵循及判决质量四个维度，共计 8 项评价指标。其中，为了方便指标统一化呈现，所有基于 LLM 的评分指标均统一转换为百分制进行报告。

3.2.1 诉状生成评价指标

信息准确性得分（Document Accuracy, DOC）：该指标评估起诉/答辩状中关键法律信息的准确程度，包括当事

①http://cail.cipsc.org.cn/task_summit?raceID=2&cail_tag=2025

人基本信息、事实陈述、诉讼请求/答辩意见及证据提供四个子项。形式化定义为：

$$\mathrm{DOC}=\frac{1}{4}\left(\mathrm{Acc}_{info}+\mathrm{Score}_{fact}+\mathrm{Score}_{claim}+\mathrm{Score}_{evidence}\right) \quad (12)$$

式中，Acc_{info} 为当事人基本信息正确率（自然人为 5 项、法人为 3 项），Score_{fact}、Score_{claim}、$\mathrm{Score}_{evidence}$ 分别为事实陈述、诉讼请求/答辩意见、证据提供的 LLM 评估得分（1–10 分）。根据任务类型不同，分为起诉状信息准确性（CD–DOC）与答辩状信息准确性（DD–DOC）。

模板遵循能力得分（Format Compliance, FOR）：该指标评估生成文书对标准模板结构的遵循程度，包括顺序正确性与条目标签准确率两个维度。计算公式为：

$$\mathrm{FOR}=\mathrm{Seq}\times\mathrm{Label} \quad (13)$$

式中，其中，$\mathrm{Seq}\in\{0,1\}$ 表示文书段落顺序是否正确，$\mathrm{Label}\in\{0,1\}$ 表示各条目标签名称是否与标准模板一致。同样分为起诉状模板遵循（CD–FOR）与答辩状模板遵循（DD–FOR）。

3.2.2 法庭审理阶段评价指标

流程遵循能力（Process Follow Score, PFS）：该指标评估模型在庭审过程中对各阶段关键操作的完成情况。庭审流程定义为状态集合 $\mathcal{Q}=\{q_1,q_2,\ldots,q_N\}$，其中 $N=5$ 分别对应开庭审理、法庭调查、法庭辩论、法庭调解、评议宣判五个阶段。PFS 计算公式为：

$$\mathrm{PFS}=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\mathrm{Complete}(q_i))\cdot(1-\mathbb{I}(\mathrm{MultiStageInOneTurn})) \quad (14)$$

式中，$\mathbb{I}(\cdot)$ 为指示函数，$\mathrm{Complete}(q_i)$ 表示第 i 阶段是否完成所有关键操作，MultiStageInOneTurn 表示是否在单轮对话内完成多个阶段（若发生则 PFS=0）。本实验中对应指标为 CI–PFS。

判决质量评价指标：包含判决准确性（JUD）、说理过程质量（REA）与法条引用准确性（LAW）三个子指标。判决准确性定义为：

$$\mathrm{JUD}=\frac{1}{M}\sum_{j=1}^{M}\mathbb{I}(\mathrm{Pred}_j=\mathrm{Gold}_j) \quad (15)$$

式中，M 为测试样本数，Pred_j 为模型预测的判决结果（支持/部分支持/不支持），Gold_j 为标准答案。本实验中对应指标为 CI–JUD。

说理过程质量采用 LLM 评分（1–10 分），本实验中对应指标为 CI–REA：

$$\mathrm{REA}=\frac{1}{M}\sum_{j=1}^{M}\mathit{LLM\_Score}(\mathrm{Reasoning}_j) \quad (16)$$

法条引用准确性定义为：

$$\mathrm{LAW}=\frac{1}{M}\sum_{j=1}^{M}\frac{|\mathrm{Law}_{pred}^{(j)}\cap\mathrm{Law}_{gold}^{(j)}|}{|\mathrm{Law}_{gold}^{(j)}|} \quad (17)$$

式中，$\mathrm{Law}_{pred}^{(j)}$ 与 $\mathrm{Law}_{gold}^{(j)}$ 分别为模型预测与标准答案的法条集合。本实验中对应指标为 CI–LAW。

3.3 模型选择

为全面评估不同大语言模型在法庭模拟任务中的表现，本实验选取了 7 个具有代表性的通用大模型和 2 个法律大模型进行对比分析。需要说明的是，在预实验过程中发现，法律大模型在自由生成任务中表现尚可，但在本任务所要求的结构化输出与多阶段流程控制场景下，存在较为明显的指令不遵循问题。例如，模型在部分

情况下未能按照预定义JSON格式输出结果，或出现字段缺失、结构错误等现象。在评测过程中，对于无法被解析的输出结果，本文按照评测流程将其对应指标记为0分并纳入整体平均。各模型简要介绍如表1所示：

表1 基线模型简介

模型名称	模型简介
qwen3-max[28]	千问3系列Max模型，相较preview版本在智能体编程与工具调用方向进行了专项升级。发布的正式版模型达到领域SOTA水平，适配场景更加复杂的智能体需求。
qwen3-235b-a22b-thinking-2507[28]	阿里Qwen团队发布的开源深度思考模型，总参数2350亿、激活220亿，支持80K推理过程长度，在逻辑推理任务上表现优异。后续实验简称为qwen3-235b。
qwen3-next-80b-a3b-thinking[28]	阿里下一代基础模型架构，800亿总参数、激活30亿，采用混合注意力架构，长文本处理能力显著提升。后续实验简称为qwen3-next-80b。
qwen3-30b-a3b[28]	通义千问MoE架构轻量模型，总参数300亿、激活30亿，在指令遵循与长上下文理解方面表现突出。后续实验简称为qwen3-30b。
DeepSeek-V3.2[29]	深度求索发布的正式版本模型，强化智能体能力并融入思考推理，公开测试中性能对标GPT-5。
deepseek-v3.2-exp[29]	DeepSeek实验性模型，引入稀疏注意力机制(DSA)，在保持性能的同时显著提升训练推理效率。
kimi-k2.5[30]	月之暗面发布的开源多模态模型，基于万亿参数MoE架构，支持智能体集群协作与超长上下文处理。
LexiLaw[31]	面向法律场景优化的大语言模型，在法律问答与法条理解任务中表现较好。
DISC-LawLLM[32]	一种法律领域专用大语言模型，针对法律知识注入与判决预测任务进行了优化。

需要特别说明的是，上述所有基线模型在参与主实验（表2）对比时，均处于“自由生成”模式。即：模型一次性接收完整的案件上下文（包括起诉状所需的所有初始信息或庭审记录），并根据统一的System Prompt直接生成最终文书。基线模型未接入本文提出的多轮信息采集（MIC）、显式庭审流程状态机（ECPM）或硬约束模板填充（TCG）模块。这种设置旨在验证：在同等输入条件下，本文提出的结构化框架能否弥补通用模型或垂直模型在复杂司法任务中的不足。

3.4 参数设置

(1) 调用方式与环境

所有基线模型及本文方法中的底层推理引擎均通过官方API接口或标准化推理服务进行调用，未对模型参数进行任何形式的微调或提示词学习。具体而言，Qwen3系列与DeepSeek系列模型通过阿里云百炼平台及DeepSeek官方开放平台

调用；Kimi 模型通过 Moonshot AI 平台调用；法律垂直模型 LexiLaw 与 DISC-LawLLM 分别采用本地服务器部署调用。所有请求均在相同的网络环境与并发限制下执行，以消除基础设施差异带来的延迟或误差。

(2) 提示词设计的一致性

为确保实验对比的公平性，所有基线模型均采用统一的标准化提示词模板。该模板包含以下部分：

● 角色设定：明确模型扮演“资深法官”或“专业律师”角色。

● 任务描述：清晰定义生成目标（如“请根据以下案情生成民事起诉状”）。

● 输出约束：明确要求模型以 JSON 格式输出，并列出所需的字段结构（如 {"plaintiff": "...", "claims": [...]}）。对于无法严格遵循 JSON 格式的模型，我们在后处理阶段尝试解析，若解析失败则该项指标记为 0。

● 少样本示例（Few-shot）：所有模型均在 Prompt 中嵌入了 1-2 个标准的法律文书生成示例，以引导其理解格式规范。

(3) 生成参数

所有模型均采用一致的生成超参数：温度（Temperature）设为 0.7 以平衡创造性与确定性，Top-p 设为 0.9，最大生成长度（Max Tokens）设为 4096。对于涉及法条引用的任务，所有模型均被允许在生成过程中调用外部检索工具（若模型支持）或基于内部知识生成，而本文方法则强制绑定了通义法睿检索接口以确保法条准确性。

3.5 实验结果及分析

为了全面评估本文提出的 AgentCourts 框架在法庭模拟与法律文书生成任务中的整体性能，本实验将其与当前主流的 7 个通用大模型和 2 个法律大模型基线在 CAIL 2025 数据集上进行对比。各基线模型均采用官方默认参数以确保公平性，本文方法基线模型采用 DeepSeek-V3.2，实验结果如表 2 所示。

表2 主实验对比(%)

模型	起诉状生成		答辩状生成		法庭审理			
	CD-DOC	CD-FOR	DD-DOC	DD-FOR	CI-REA	CI-LAW	CI-JUD	CI-PFS
qwen3-max	44.35	72.85	33.98	35.98	20.68	14.03	27.09	43.25
qwen3-235b	43.19	69.03	32.15	35.05	15.29	12.36	24.62	39.25
qwen3-next-80b	41.08	64.12	29.09	34.28	12.37	8.78	23.48	34.80
qwen3-30b	40.34	60.65	28.63	32.15	10.68	6.49	20.69	32.29
DeepSeek-V3.2	47.03	78.66	36.55	38.79	32.12	23.72	32.11	44.14
deepseek-v3.2-exp	45.36	76.44	35.67	38.55	26.35	16.79	28.61	43.09
kimi-k2.5	42.42	67.26	30.45	33.46	19.46	13.06	25.33	41.32
LexiLaw	12.42	22.11	15.32	16.42	3.34	1.38	14.32	14.18
DISC-LawLLM	15.76	30.08	17.39	18.03	5.98	4.69	16.07	16.51
AgentCourts	**81.25**	**100.0**	**61.02**	**67.08**	**82.87**	**79.97**	**76.77**	**98.75**

从表 2 的主实验对比结果可以看出，本文提出的 AgentCourts 方法在所有评测指标上均显著优于现有的主流大语言模型。针对引言中提出的四个科学问题，本实验通过具体指标进一步验证了方法的有效性。首先，对于法律信息不完整问题，起诉状与答辩状的信息准确性得分（CD-DOC 与 DD-DOC）分别达到 81.25% 与 61.02%，显著高于基线模型，表明多轮信息采集机制有效补全了关键要素；其次，针对文本结构强约束问题，模板遵循能力得分（CD-FOR 与 DD-FOR）分别达到 100% 与 67.08%，证明模板约束生成策略能严格保障文书格式规范；再次，关于司法推理过程不可解释问题，说理过程质量（CI-REA）与法条引用准确性（CI-LAW）分别提升至 82.87% 与 79.97%，体现了证据驱动推理机制在增强逻辑可解释性方面的作用；最后，对于复杂庭审流程难以建模问题，流程遵循得分（CI-PFS）高达 98.75%，远超基线模型，验证了显式流程建模对司法程序规范性的保障能力。实验数据表明，本文方法虽未宣称完全消除所有挑战，但在上述四个维度均实现了显著的性能突破与问题缓解。

从法律领域模型对比结果可以看出，LexiLaw 与 DISC-LawLLM 在各项指标上整体表现低于通用大语言模型。进一步分析发现，这主要并非完全源于其语言生成能力不足，而是由于其在复杂任务中的指令遵循能力与结构化输出稳定性较差。具体而言，在本任务中，评测流程依赖模型严格按照预定义格式生成结构化结果。一旦输出不符合格式要求（如 JSON 结构错误或字段缺失），评测程序将无法解析该结果，并按照 0 分计入统计。这一机制导致在解析失败频繁出现的情况下，模型整体得分被显著拉低。该现象表明，法律领域模型虽然在知识层面具备一定优势，但在复杂多阶段任务中，其工程可用性仍受到结构化生成能力的限制。

为了进一步验证框架中各核心模块的有效性，本实验设计了消融实验，选取了以下四种具有代表性的配置进行对比：（1）Full Model：完整框架；（2）Base Line (-All)：移除所有特定模块（MIC, TCG, ECPM, EDR），仅保留底层大语言模型的自由生成能力。此配置用于评估框架带来的整体性能上限提升；（3）Info-Struct Combo (-MIC & -TCG)：在起诉状/答辩状生成阶段，同时移除信息采集与模板约束。此组合用于验证“动态信息抽取”与“静态结构约束”协同工作的必要性；（4）Reasoning-Flow Combo (-ECPM & -EDR)：在判决生成阶段，同时移除流程显式建模与证据驱动推理。此组合用于验证“程序正义（流程）”对“实体正义（推理）”的支撑作用。消融实验结果如表 3 所示。

表 3 的消融实验结果清晰地展示了各模块对整体性能的贡献。起诉状生成任务单独移除 MIC 导致 CD-DOC 下降约 25%（81.25% -> 56.32%），单独移除 TCG 导致 CD-FOR 下降约 20%（100% -> 79.68%）。然而，当同时移除 MIC 和 TCG（-MIC & -TCG）时，CD-DOC 和 CD-FOR 分别暴跌至 48.15% 和 42.30%。值得注意的是，组合移除后的性能损失（相对于 Full Model）远大于单个模块损失之和的简单线性叠加。例如，在格式遵循上，若线性叠加，预期下降约为 20%+14%=34%，但实际下降接近 60%。这表明，信息采集为模板填充提供了准确的“槽位值”，而模板约束反过来规范了信息采集的输出格式，两者形成了紧密的闭环协同。缺少任一环节，另一环节的效果都

表3 消融实验对比(%)

实验配置	起诉状生成		答辩状生成		法庭审理			
	CD-DOC	CD-FOR	DD-DOC	DD-FOR	CI-REA	CI-LAW	CI-JUD	CI-PFS
AgentCourts(Full)	**81.25**	**100.0**	**61.02**	**67.08**	**82.87**	**79.97**	**76.77**	**98.75**
- MIC	56.32	86.17	46.88	51.72	76.68	72.36	64.25	46.93
- TCG	76.26	79.68	55.06	43.28	81.06	78.91	75.06	50.19
- MIC & - TCG	48.15	42.30	39.21	35.14	-	-	-	-
- ECPM	-	-	-	-	78.36	75.08	71.47	92.18
- EDR	-	-	-	-	42.82	25.63	40.66	97.92
- ECPM & - EDR	-	-	-	-	35.14	18.25	32.41	88.56
- All (Base Only)	42.10	38.55	33.98	35.98	20.68	14.03	27.09	43.25

会因缺乏上下文或约束而大幅失效。判决生成任务中，单独移除证据驱动推理（-EDR）导致说理质量（CI-REA）大幅下降（82.87% -> 42.82%），说明证据是推理的核心。然而，当同时移除流程建模（-ECPM & -EDR）时，CI-REA 进一步降至 35.14%，且流程遵循度（CI-PFS）降至 88.56%。这一结果表明，证据推理并非孤立存在，而是依附于特定的庭审阶段（如法庭调查阶段确认证据，辩论阶段质证）。如果没有 ECPM 提供的阶段性上下文标记，EDR 模块难以准确定位证据在庭审记录中的时序位置和法律效力，导致推理混乱。这种“流程-推理”的耦合证明了显式建模庭审流程对于提升推理可解释性的协同价值。对比“Base Only”（即完全去除所有模块，等同于通用大模型自由生成）与“Full Model”，可以看到全方位的性能飞跃。特别是在 CI-LAW（法条引用）和 CI-PFS（流程遵循）上，差距超过 60 个百分点。这证实了 AgentCourts 并非简单的模块堆砌，而是通过角色驱动的状态传递和硬约束解码，将大模型的通用语言能力转化为司法领域的专用能力。各模块在“信息补全—结构规范—流程控制—逻辑推理”链条上环环相扣，共同构成了一个鲁棒的司法智能系统。

最后，为了验证本文方法的泛化能力与兼容性，本实验将 AgentCourts 框架作为插件式模块应用于不同的基座大语言模型上，测试其是否具备“即插即用”的提升效果。实验选取了 Qwen3 系列及 Kimi 等多种架构的模型作为基座，对比接入框架前后的性能差异，泛化对比实验结果如表 4 所示。

表 4 的泛化实验结果表明，AgentCourts 框架具有良好的模型无关性与泛化性能。无论基座模型是参数量较大的 qwen3-235b 还是轻量级的 qwen3-30b，接入 AgentCourts 框架后，其在各项司法任务指标上均取得了稳定且显著的提升。例如，在 qwen3-max 基座上，框架使判决准确性（CI-JUD）从 27.09 提升至 73.19；在 kimi-k2.5 基座上，流程遵循度（CI-PFS）从 41.32 提升至 93.28。这一结果证明了本文提出的多阶段建模与模板约束策略并非依赖于特定模型的内部参数，而是作为一种通用的方法论，能够有效赋能不同架构的大语言模型，使其更好地适应司法场景的规范性与逻辑性要求，具有广泛的实用价值与推广前景。

表4 泛化对比实验(%)

模型	起诉状生成		答辩状生成		法庭审理			
	CD-DOC	CD-FOR	DD-DOC	DD-FOR	CI-REA	CI-LAW	CI-JUD	CI-PFS
qwen3-max	44.35	72.85	33.98	35.98	20.68	14.03	27.09	43.25
AgentCourts	**79.82**	**93.46**	**59.03**	**66.01**	**78.45**	**76.22**	**73.19**	**96.14**
qwen3-235b	43.19	69.03	32.15	35.05	15.29	12.36	24.62	39.25
AgentCourts	**78.68**	**91.39**	**57.63**	**64.38**	**73.90**	**74.63**	**70.41**	**94.32**
qwen3-next-80b	41.08	64.12	29.09	34.28	12.37	8.78	23.48	34.80
AgentCourts	**76.48**	**88.67**	**55.37**	**62.39**	**68.38**	**73.12**	**68.36**	**93.49**
qwen3-30b	40.34	60.65	28.63	32.15	10.68	6.49	20.69	32.29
AgentCourts	**75.36**	**86.55**	**54.39**	**61.02**	**61.19**	**70.18**	**62.18**	**91.28**
kimi-k2.5	42.42	67.26	30.45	33.46	19.46	13.06	25.33	41.32
AgentCourts	**76.92**	**90.48**	**56.37**	**63.41**	**73.55**	**75.69**	**69.11**	**93.28**

为进一步验证自动评估结果的可靠性，本文补充开展了人工评估实验。具体而言，我们从以下四个维度对生成文书进行综合评价：(1) 事实一致性 (Fact Consistency)：评估生成内容是否与案件事实及证据材料保持一致；(2) 法律逻辑合理性 (Legal Reasoning)：评估裁判推理过程是否符合基本法律逻辑及裁判思路；(3) 文书结构规范性 (Structural Compliance)：评估生成文书是否符合标准法律文书格式要求；(4) 语言表达质量 (Language Quality)：评估文本的流畅性、严谨性与专业性。在评估人员方面，本文邀请了3名具备法律专业背景的评估者参与打分，其中包括2名法学硕士研究生（具有法律文书写作经验）及1名具有司法实务经验的法律从业人员。所有评估者均在统一评分标准与说明下，对随机抽取的测试样本进行独立打分（评分范围为1－10分），最终结果取平均值以降低主观偏差。人工评估结果如图3所示。

从整体趋势来看，本文方法在四个维度上均显著优于基线模型。具体而言，在文书结构规范性维度上，本文方法取得最高提升，这主要得益于模板约束生成策略的引入，使得生成内容在结构层面高度符合司法文书规范；在法律逻辑合理性方面，本文方法同样表现出明显优势，说明多阶段建模与证据驱动推理机制能够有效提升裁判说理的连贯性与合理性；在事实一致性维度上，模型通过多轮信息采集与状态更新机制，有效减少了事实遗漏与冲突问题；在语言表达质量方面，虽然基线模型已具备一定生成能力，但本文方法在专业性与严谨性方面仍有稳定提升。值得注意的是，人工评估结果与自动评估指标（如DOC、REA等）在整体趋势上保持一致，这从侧面验证了基于大语言模型的自动评估方法在本任务中的可靠性与有效性。

4 结论

本文针对当前法律人工智能在法庭模拟任务中面临的信息获取不完整、文书结构约束强、推理过程不可解释及庭审流程建模难等科学问题，提出了一种基于多智能体协作的法庭模拟与法律文书生成方法 (AgentCourts)。该方法通过角色驱动的多阶段建模，实现了从信息采集、对抗式

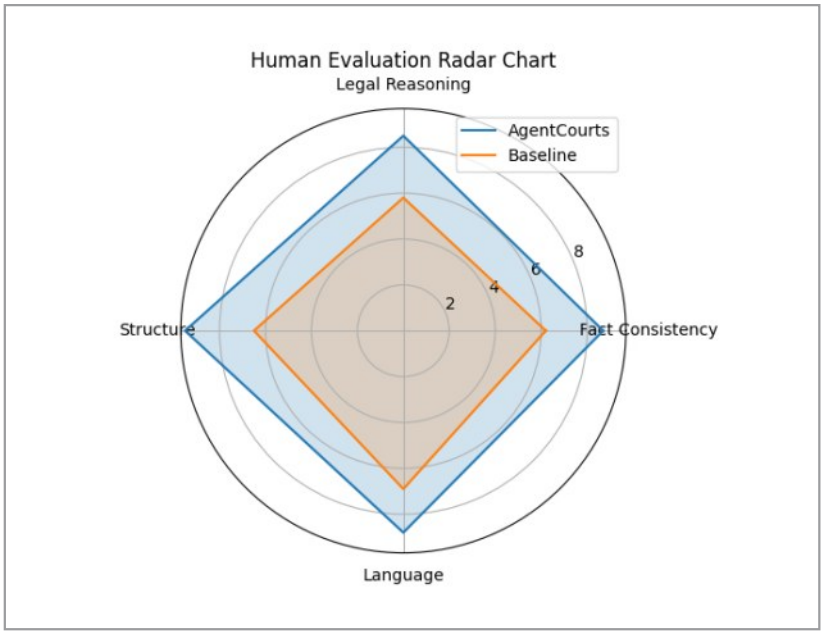

图3 人工评估结果

答辩到庭审判决的端到端闭环，有效缓解了信息缺失与结构不规范问题，显著改善了推理可解释性与流程建模精度。具体而言，我们设计了多轮对话式信息抽取机制以确保案件要素完整，采用模板约束生成策略保障文书格式规范，并引入证据驱动推理与法条检索增强裁判说理的可解释性与准确性。在 CAIL 2025 数据集上的实验结果表明，该方法在起诉状、答辩状及判决生成三个子任务上均显著优于主流大语言模型基线，验证了多阶段建模与模板约束策略在司法场景中的有效性与实用性，为智慧法院建设提供了新的技术路径。

尽管本文方法在法庭模拟任务中取得了显著进展，但仍存在一些局限性与改进空间。未来工作将主要集中在以下几个方面：首先，针对当前研究主要聚焦于民事一审案件的局限性，计划将框架扩展至刑事、行政等更多类型的案件场景，构建跨领域的法律评测基准，进一步验证方法的多场景泛化能力；其次，将进一步优化证据推理机制，引入知识图谱等技术增强模型对复杂法律逻辑关系的理解与处理能力；此外，还将探索人机协同模式，研究如何将本系统无缝集成到现有智慧法院工作流中，确保人工智能辅助决策的安全性、伦理合规性及可信赖性。最终目标是构建一个更加智能、公正且高效的法律人工智能生态系统，助力司法审判现代化。

参考文献：

[1] 最高人民法院关于规范和加强人工智能司法应用的意见[N].人民法院报,2022-12-10(004).
Opinions of the Supreme People's Court on Regulating and Strengthening the Judicial Application of Artificial Intelligence [N]. People's Court Daily, 2022-12-10(004).

[2] 中国社会科学院法学研究所.法治蓝皮书(2024)[R].北京:社会科学文献出版社,2024.
Institute of Law, Chinese Academy of Social Sciences. Rule of Law Blue Book

(2024) [R]. Beijing: Social Sciences Academic Press, 2024.

[3] 北京市人民政府. 以人工智能促进司法审判现代化[Z].2025.
Beijing Municipal People's Government. Promoting the Modernization of Judicial Adjudication with Artificial Intelligence [Z]. 2025.

[4] 姚林波,周裕林,黄瑞章,等.基于多源数据融合的裁判文书说理生成方法[J].广西大学学报(自然科学版),2025,50(06):1304-1319.
Yao Linbo, Zhou Yulin, Huang Ruizhang, et al. Method of generating judicial document reasoning based on multi-source data fusion [J]. Journal of Guangxi University (Natural Science Edition), 2025, 50(06):1304-1319.

[5] 邵文杰,孙志鹏,张冲.人工智能在司法审判中的应用与思考[J].法律与人工智能,2025,(01):269-274.
Shao Wenjie, Sun Zhipeng, Zhang Chong. Application and Reflection of Artificial Intelligence in Judicial Trials [J]. Law and Artificial Intelligence, 2025, (01): 269-274.

[6] Chen G, Fan L, Gong Z, et al. Agentcourt: Simulating court with adversarial evolvable lawyer agents[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 5850-5865.

[7] He C, Hu H, Li Y, et al. A survey of large language models for legal tasks: Progress, prospects and challenges[J]. Computer Science Review, 2026, 60: 100906.

[8] Deng C, Mao K, Zhang Y, et al. Enabling discriminative reasoning in llms for legal judgment prediction[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 784-796.

[9] Zhang L, Ashley K D. Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach for legal argument generation[J]. arXiv preprint arXiv:2506.02992, 2025.

[10] Nigam S K, Patnaik B D, Thomas A V, et al. Structured legal document generation in india: A model-agnostic wrapper approach with vidhikdastaavej[J]. arXiv preprint arXiv:2504.03486, 2025.

[11] Kondo R, Matsuoka R, Yoshida T, et al. Capturing Legal Reasoning Paths from Facts to Law in Court Judgments using Knowledge Graphs[C]//Proceedings of the 13th Knowledge Capture Conference 2025. 2025: 103-110.

[12] Chalkidis I, Androutsopoulos I, Aletras N. Neural legal judgment prediction in English[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 4317-4323.

[13] Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, et al. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective[J]. PeerJ computer science, 2016, 2: e93.

[14] Zhong H, Guo Z, Tu C, et al. Legal judgment prediction via topological learning [C]//Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 3540-3549.

[15] Dina N Z, Ravana S D, Idris N. Multi-Label Classification on Legal Judgment Prediction Using Legal Bert-Label Powerset[C]//2024 16th International Conference on Knowledge and System Engineering (KSE). IEEE, 2024: 213-218.

[16] Xu N, Wang P, Chen L, et al. Distinguish confusing law articles for legal judgment prediction[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 3086-3095.

[17] Yang J, Ma W, Zhang M, et al. Legal-GNN: Legal information enhanced graph

neural network for recommendation[J]. ACM Transactions on Information Systems (TOIS), 2021, 40(2): 1-29.

[18] Sukanya G, Priyadarshini J. Modified Hierarchical-Attention Network model for legal judgment predictions[J]. Data & knowledge engineering, 2023, 147: 102203.

[19] Zhang Y, Wei X, Yu H. HD-LJP: A hierarchical Dependency-based legal judgment prediction framework for Multi-task learning[J]. Knowledge-Based Systems, 2024, 299: 112033.

[20] Feng Y, Li C, Ng V. Legal Judgment Prediction: A Survey of the State of the Art[C]//IJCAI. 2022: 5461-5469.

[21] Fei Z, Shen X, Zhu D, et al. Lawbench: Benchmarking legal knowledge of large language models[C]//Proceedings of the 2024 conference on empirical methods in natural language processing. 2024: 7933-7962.

[22] Zhou Z, Shi J X, Song P X, et al. Lawgpt: A chinese legal knowledge-enhanced large language model[J]. arXiv preprint arXiv:2406.04614, 2024.

[23] He Z, Cao P, Wang C, et al. Agentscourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 9399-9416.

[24] Liao H J. Design and Evaluation of a Courtroom Examination AI Simulation System with Behavioral Fidelity[C]//Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025). 2025: 20-28.

[25] Chun J, Elkins K, Lee Y S. AgenticSim-Law: A Juvenile Courtroom Multi-Agent Debate Simulation for Explainable High-Stakes Tabular Decision Making[J]. arXiv preprint arXiv:2601.21936, 2026.

[26] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models[J]. arXiv preprint arXiv:2203.11171, 2022.

[27] Du Y, Li S, Torralba A, et al. Improving factuality and reasoning in language models through multiagent debate[C]//Proceedings of the 41st International Conference on Machine Learning. 2024: 11733-11763.

[28] Yang A, Li A, Yang B, et al. Qwen3 technical report[J]. arXiv preprint arXiv: 2505.09388, 2025.

[29] Liu A, Mei A, Lin B, et al. Deepseek-v3. 2: Pushing the frontier of open large language models[J]. arXiv preprint arXiv: 2512.02556, 2025.

[30] Team K, Bai T, Bai Y, et al. Kimi K2. 5: Visual Agentic Intelligence[J]. arXiv preprint arXiv:2602.02276, 2026.

[31] Li H, Ai Q, Dong Q, et al. Lexilaw: A scalable legal language model for comprehensive legal understanding[EB/OL]. (2024). https://github. com/CSHaitao/LexiLaw.

[32] Yue S, Chen W, Wang S, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services[J]. arXiv preprint arXiv:2309.11325, 2023.

周裕林(1997-),男,博士在读,主要研究方向为法律智能化、大模型推理等。

作者简介

秦永彬（1980–），男，教授，主要研究方向为法律智能化、大数据等。

收稿日期: XXXX-XX-XX
通信作者: 秦永彬,ybqin@.gzu.edu.cn
基金项目: 国家重点研发计划项目(2023YFC3304500)；贵州省科学技术基金重点资助项目(黔科合重大专项字[2024]003)
Foundation Items: The National Key R&D Program of China(No. 2023YFC3304500), The key Technology R&D Program of Guizhou Province (No. [2024]003)