面向刑事判决预测的“控辩审”多智能体协作框架研究

韩林睿[1,2,3]，郑日[1,2,4]，丛颖男[5]，李钰鹏[6]

（1. 教育部哲学社会科学实验室——中国政法大学数据法治实验室，北京 100088；
2. 北京高等学校哲学社会科学创新中心数字时代首都新兴领域法治创新中心，北京 100088；
3. 中国政法大学数据法学院，北京 100088；
4. 中国政法大学国家安全学院，北京 100088；
5. 中国政法大学商学院，北京 100088；
6. 香港浸会大学互动媒体系，香港 999077）

摘　要：针对刑事判决预测中现有方法对控辩交互、证据审查和裁判形成过程刻画不足，以及容易出现争点回应不充分、裁判结论与说理不一致等问题，提出一种面向刑事判决预测的“控辩审”多智能体协作框架（PDJ-MACF）。该框架以庭审程序为组织线索，构建开庭、法庭调查、法庭辩论、最后陈述和评议宣判 5 个阶段的协作流程。控方智能体负责指控事实提取与证据组织，辩方智能体围绕案件事实、证据和量刑提出异议与辩护意见，审方智能体归纳争点，完成罪名、法条、刑期和罚金预测，并生成裁判文书。在包含 400 件刑事案件、覆盖 40 种罪名的数据集上开展实验，结果表明：（1）PDJ-MACF 在 12 项自动评价指标中的 9 项取得最高值，在刑罚预测、罪名预测、法条引用和文书语义相似度等任务上表现出较好的综合性能；（2）消融实验中，逐项移除分步事实组织、控方策略路由、辩方参与、检索增强和结构化量刑约束后，相应评价指标出现不同程度下降；（3）人工评测中，PDJ-MACF 在事实还原度、量刑合理性、法律说理质量、文书规范性和整体可用性 5 个维度均取得最高平均分，综合得分为 4.24 分（满分 5 分）。该框架细化了控、辩、审三方在判决预测中的任务边界与信息传递过程，形成的结构化中间产物可用于裁判结果和文书内容的过程核验。

关键词：刑事判决预测；多智能体协作；控辩审关系；裁判文书生成；数字法院
中图分类号：TP183
文献标志码：A
doi: 10.11959/j.issn.2096−6652.ZNKX26135

PDJ-MACF: prosecution-defense-judiciary multi-agent collaborative framework for criminal judgment prediction

Han Linrui[1,2,3], Zheng Ri[1,2,4], Cong Yingnan[5], Li Yupeng[6]

1. Ministry of Education Laboratory of Philosophy and Social Sciences—The CUPL Data Law Lab, China University of Political Science and Law, Beijing 100088, China
2. Beijing Higher Education Innovation Center for Philosophy and Social Sciences, Beijing 100088, China
3. Data Law School, China University of Political Science and Law, Beijing 100088, China
4. National Security School, China University of Political Science and Law, Beijing 100088, China
5. Business School, China University of Political Science and Law, Beijing 100088, China
6. Department of Interactive Media, Hong Kong Baptist University, Hong Kong 999077, China

Abstract: Existing methods for criminal judgment prediction often provide inadequate representations of prosecution-

收稿日期：2026−08−10；**修回日期：**2026−09−02
通信作者：丛颖男，cyn_2010@163.com
基金项目：中国政法大学青年教师学术创新团队支持计划（No.26CXTD02）；教育部人文社会科学研究一般项目（No.22YJC190003）
Foundation Items: The Program for Young Innovative Research Team in China University of Political Science and Law (No.26CXTD02), The General Project of Ministry of Education Foundation for Humanities and Social Sciences (No.22YJC190003)

defense interaction, evidence examination, and judicial decision formation, which can leave disputed issues insufficiently addressed and produce inconsistencies between predicted outcomes and the supporting reasoning. To address these limitations, PDJ-MACF, a prosecution-defense-judiciary multi-agent collaborative framework for criminal judgment prediction, was proposed. Following the sequence of criminal trial proceedings, PDJ-MACF organized collaboration into five stages: opening of the trial, court investigation, courtroom debate, the defendant's final statement, and deliberation and pronouncement of judgment. The prosecution agent extracted facts supporting the charge and organizes the associated evidence; the defense agent raised objections and formulates defense arguments concerning the facts, evidence, and sentencing; and the judiciary agent summarized disputed issues, predicted the charge, applicable statutory provisions, prison term, and fine, and generated the judgment document. Experiments were conducted on a dataset of 400 criminal cases covering 40 charges. The results show that: (1) PDJ-MACF achieves the highest scores on nine of the 12 automatic evaluation metrics and demonstrates strong overall performance in penalty prediction, charge prediction, statute citation, and the semantic similarity of generated judgment documents; (2) individually removing stepwise fact organization, prosecution strategy routing, defense participation, retrieval augmentation, or structured sentencing constraints leads to declines in the corresponding evaluation metrics; and (3) in human evaluation, PDJ-MACF obtains the highest mean score in all five dimensions—factual fidelity, sentencing rationality, legal reasoning quality, normative compliance, and overall usability—with an overall score of 4.24 out of 5. By specifying the responsibilities of the three agents and the flow of information among them, PDJ-MACF produces structured intermediate artifacts that support process-level verification of predicted outcomes and generated judgment documents.

Key words: criminal judgment prediction, multi-agent collaboration, prosecution-defense-judiciary relation, judgment document generation, digital court

0 引言

法律判决预测（legal judgment prediction，LJP）旨在基于案件事实与证据材料，自动得出判决结果，已成为提升司法效率、促进裁判一致性及辅助法官决策的重要研究方向[1-3]。近年来，端到端深度学习模型与预训练语言模型显著提升了罪名、量刑与法条等子任务的预测性能[4-6]。大语言模型（large language model，LLM）通过提示学习与推理设计进一步增强了文本生成与法律分析能力[7-10]。提示工程通过任务指令、示例组织和多步推理调整模型的分析路径，进而影响专业任务的输出质量[11]。然而，在刑事裁判这一高度程序化且对事实认定与法律适用均有严格约束的场景中，现有LJP研究面临诸多挑战，如可给出预测裁判结果却难以给出可核验、可追溯的裁判理由链，可拟合裁判文书的文本表述却难以复现程序化推理过程，制约了该技术在真实司法流程中的可采纳性[12-14]。

究其原因，首先，主流数据驱动方法多将LJP视为静态映射问题，即从案情描述直接输出罪名、法条或刑期，解释通常依赖注意力权重、关键词抽取或事后生成的理由段落。除预测结果外，用户还关心结论如何在证据规则、举证责任与法条适用框架下形成，哪些关键争点被采纳或排除，以及不同观点之间的冲突如何被裁判者化解。缺乏程序化的解释将直接削弱系统输出的可核验性与可质疑性，从而影响其使用价值。

其次，刑事裁判的推理并非单一主体的孤立判断，而是“控方指控—辩方抗辩—审方裁判”三方在法定程序与证据规则约束下的动态互动过程。控方围绕指控构建事实与证据链，辩方针对证据能力、证明力与构成要件提出反驳，审方在争点组织、证据评价与法条适用中形成裁判理由。当前LJP模型主要依赖对历史裁判文书的表层模式学习，根据文本相似性或要件相关性进行预测，其结果更接近对既有裁判样式的拟合，而非对个案事实的推断。由此，易出现结论在形式上与既往案例相近，但缺乏充分理由支撑，或关键争点未被识别与回应、法律适用缺乏必要的中间推理环节等现象，进而导致预测结果难以与真实裁判逻辑对齐。

为此，本文将多智能体协作（multi-agent collaboration，MAC）应用于LJP，提出一种面向刑事判决预测的“控辩审”多智能体协作框架（prosecution-defense-judiciary multi-agent collaborative framework for criminal judgment prediction，PDJ-MACF）。核心思想是以庭审三方角色为结构先验，将判决预测

过程定义为可交互、可记录、可复核的论证与裁断流程，由“控”智能体（prosecution agent，AP）提出指控、证据与法律依据；由“辩”智能体（defense agent，AD）围绕事实争议、证据瑕疵与法律适用，提出反驳或减轻/免责主张；由“审”智能体（judiciary agent，AJ）在程序结构约束下组织争点、评价证据、回应控辩意见，并形成裁判结论与理由。通过多智能体对话与过程信息固化，系统能够输出与庭审结构对齐、可检查的结构化中间产物，为后续人工核验与争点追踪提供依据。

本研究主要贡献如下。

- 面向刑事判决预测，构建以庭审程序为组织线索的信息传递框架。框架围绕控、辩、审三方在5个阶段形成的事实主张、证据意见、异议和争点材料，明确其进入终局裁判的交接关系，使角色互动能够连续服务于罪名、法条、刑罚预测和裁判文书生成。
- 建立与控、辩、审职责相衔接的结构化协作过程：控方以事实命题和证据对齐组织指控，辩方以异议清单保留质证与辩护立场，审方以争点提纲和辩论记录控制交互并形成裁断，为各项判决要素提供可检查的中间依据。
- 在真实刑事案件数据上开展多维实验评估，结合结果指标、消融分析和人工评测，考察该框架在刑罚、罪名、法条、裁判文书生成及过程可用性方面的表现。

1　相关工作

1.1　法律判决预测方法

法律判决预测旨在基于案件事实自动预测判决结果，其核心挑战在于平衡预测精度与决策可解释性[2]。早期研究主要依赖统计模型与规则系统，Kort[15]采用多因素复合分析法预测美国最高法院裁决倾向，Nagel[16]探索相关性分析建模判决结果，而Lawlor[17]和Keown[18]进一步讨论了计算分析与数学模型在法律决策形式化中的可能性。此类方法虽揭示了司法规律，但受限于人工特征工程的复杂性与噪声敏感性——法律文本常含冗余事实与不完整数据，导致模型泛化能力不足[2]。

监督学习技术的发展推动LJP转向文本分类与结构化预测。部分研究者利用支持向量机[19]、随机森林[20]及文本分类模型[21]处理案情描述；Liu等[22]结合PAT树与语义知识库优化中文刑事文书分类，Hu等[23]引入法律属性干预缓解小样本混淆问题。然而，这些方法仍依赖人工设计特征，难以捕捉深层语义关联，而且缺乏对司法逻辑的显式建模。

深度学习的兴起推动了LJP由特征工程和任务分类转向端到端语义建模。Zeleznikow等[24]融合神经网络与Toulmin论证结构生成解释性判决；Luo等[4]利用注意力机制强化事实–法条关联建模；Zhong等[25]将判决任务间的依赖关系表示为拓扑结构；Yue等[26]从案件情节出发建模司法判断；吕由钢等[3]通过判决链组织罪名、法条和刑罚预测，进一步刻画判决要素之间的依赖关系。法律领域预训练研究首先着眼于专业语义表示和长文档建模。Chalkidis等[27]在法律语料上继续预训练BERT，以适配法律文本分布；Xiao等[28]面向中文法律长文档改进上下文建模；Zheng等[29]进一步以CaseHOLD数据集考察自监督预训练在法律任务中的迁移效果。大语言模型的引入使研究重点扩展到任务组织与推理过程。丛颖男等[30]面向刑事案件智能判决，讨论了大语言模型在案件事实理解、法律适用和裁判生成中的应用；Deng等[7]将案件事实分解、相近罪名辨析与裁判预测串联为判别性推理过程；Liu等[13]提出基于诉讼参与人提交证据预测法律事实的任务，以扩展早期诉讼阶段的判决预测条件。Prompt4LJP[8]将提示学习用于法律判决预测，LegalReasoner[9]把多阶段推理与法律知识整合到预测流程中，思维链提示[10]和法律三段论提示[31]进一步使事实、规范与结论之间的推导步骤显式化。针对通用模型的法律适配，Lawyer LLaMA[32]利用法律指令数据进行训练；LawGPT[33]加入中文法律知识增强，分别从指令对齐与知识注入两条路径推进法律任务建模。

法律大语言模型进入司法场景，还需兼顾专业性、结果可解释性与规范边界[12]。生成式人工智能治理研究进一步表明，技术应用需要在创新发展与风险控制之间保持平衡[34]。现有LJP多将裁判推断处理为单模型的静态预测，结果解释通常来自事后归因或与预测弱对齐的文本生成，控辩互动、证据评价和审方裁断的过程信息因而难以连续进入预测结果，限制了理由链的核验和追溯。

1.2　多智能体协作在法律AI中的应用

多智能体协作通过任务分解、角色分工和信息交互处理单一模型难以完成的复杂任务。相关综述从通信方式、任务分配与协作组织等方面归纳了大

语言模型智能体的实现路径[35]。大语言模型智能体的研究已将规划、记忆、工具调用和协作机制纳入系统框架[36]。多智能体强化学习研究将协作、通信和信用分配视为影响群体协同效能的基础问题[37]。这些研究为法律场景中的角色配置和信息交互提供了方法参照。

大语言模型的发展推动法律智能体由单轮问答向多角色协作拓展，相关综述已从系统架构、推理方式和应用任务等方面归纳这一研究方向[38]。在法律咨询场景中，LawLuo 设置接待员、律师、秘书和主管 4 类角色，通过分工协作完成多轮咨询[39]；ChatLaw 以角色对齐的混合专家架构组织法律助理、研究员等角色协作，用于法律问答、信息检索和咨询意见生成[40]。这类系统侧重法律服务过程，主要评价回答质量、专业性和用户需求匹配程度。

多智能体方法也逐步被用于法律判决预测。AgentsCourt 通过法庭辩论、法律资源检索和判决书生成等模块模拟司法决策[41]。韩林睿等[42]面向刑事案件裁判文书生成，探索以多智能体协作组织法律分析和文书撰写。MAGLJP 面向多被告案件构建法律事件逻辑图，由定罪、法律知识辅助和量刑智能体处理被告关系及量刑信息[43]。AgentsBench 设置法官与陪审员智能体模拟合议讨论[44]；Debate-Feedback 利用多智能体辩论和反馈评估修正预测结果[45]；MALR 则将法律理论分析分解为多个智能体协同完成的推理步骤[46]。相关研究从角色讨论、知识检索和任务分解等角度扩展了法律判决预测的实现路径。

近期研究进一步关注多模态案件信息与法律推理过程的核验。JurisMMA 提出面向多模态案件信息的阶段化法律判决预测框架[47]。在法律推理任务中，Sadowski 等[48]将法定规则转化为可核验的中间结构，并用多智能体分解知识获取、案件要素映射和符号推理。上述研究表明，多智能体协作正由角色分工延伸至证据形态融合和推理过程核验。

SimCourt 按照中国刑事庭审程序设置法官、检察官、辩护律师、被告人和书记员 5 类智能体，并以庭审准备、法庭调查、举证质证、法庭辩论和最后陈述组织交互[49]。系统根据被告人信息、起诉书和证据材料生成庭审记录与判决，分别在 200 件案件上评价刑期、缓刑和罚金预测，在 20 段真实庭审录像上评价不同角色的庭审表现。其研究重点是完整庭审模拟和过程质量评价；庭审信息主要保存在连续的自然语言记录中，公开的判决评测尚未单独考察控辩信息向罪名、法条和裁判文书的结构化传递及其一致性。

现有研究已经形成法庭模拟、合议核验、任务分解和多模态预测等多种协作方式。对于同时预测罪名、法条和刑罚并生成裁判文书的任务，控辩双方形成的事实主张、证据意见和异议如何经过争点组织进入审方裁断，仍需更细致的角色建模。本文据此设置控方、辩方和审方智能体，以庭审程序启发的阶段约束规定协作顺序和信息传递关系，使各阶段形成的结构化中间产物连续进入判决要素预测与裁判文书生成。

2 本研究方法

2.1 问题定义

刑事判决预测同时包含判决要素预测与裁判文书生成两个目标。给定案件输入 $x=\left(x_{\text{base}},x_{\text{indict}},x_{\text{defense}},x_{\text{statement}}\right)$、模型输出判决结果 $\hat{y}$ 与裁判文书 $\hat{j}$：

$$\left(\hat{y},\hat{j}\right)=M\left(x\right),x=\left(x_{\text{base}},x_{\text{indict}},x_{\text{defense}},x_{\text{statement}}\right) \quad (1)$$

其中，x_{base} 为被告人基础信息，x_{indict} 为删除量刑建议、指控罪名及其同义表达、援引法条和裁判结果后的起诉材料及控方证据材料，x_{defense} 为辩方证据材料，$x_{\text{statement}}$ 为量刑态度信息；$\hat{y}$ 包含罪名、刑期、罚金和适用法条，$\hat{j}$ 为依据这些判决要素生成的裁判文书。模型需保证 $\hat{y}$ 与真实判决结果在关键要素上保持一致，同时使 $\hat{j}$ 中的事实认定、法律适用和裁判说理均可追溯至输入材料。后文在语境明确时省略预测量的帽号，真实标签使用下标或“真实”标示。

2.2 模型框架

本文提出的 PDJ-MACF 总体框架如图 1 所示。框架以检察官“控”、律师/当事人“辩”和法官“审”3 个法律角色为执行单元，以刑事庭审程序为组织线索，将事实整理、证据审查、控辩交互和裁判生成连接为一个连续的协作推理过程。各阶段形成的案件概要、事实-证据对齐结果、异议清单、庭审争点提纲、辩论记录和量刑态度要点，均作为后续阶段的结构化输入，并最终用于判决要素预测与裁判文书生成。

为便于描述信息流，本文将审方形成的案件概要记为 P_0，将控方形成的事实-证据对齐结果记为

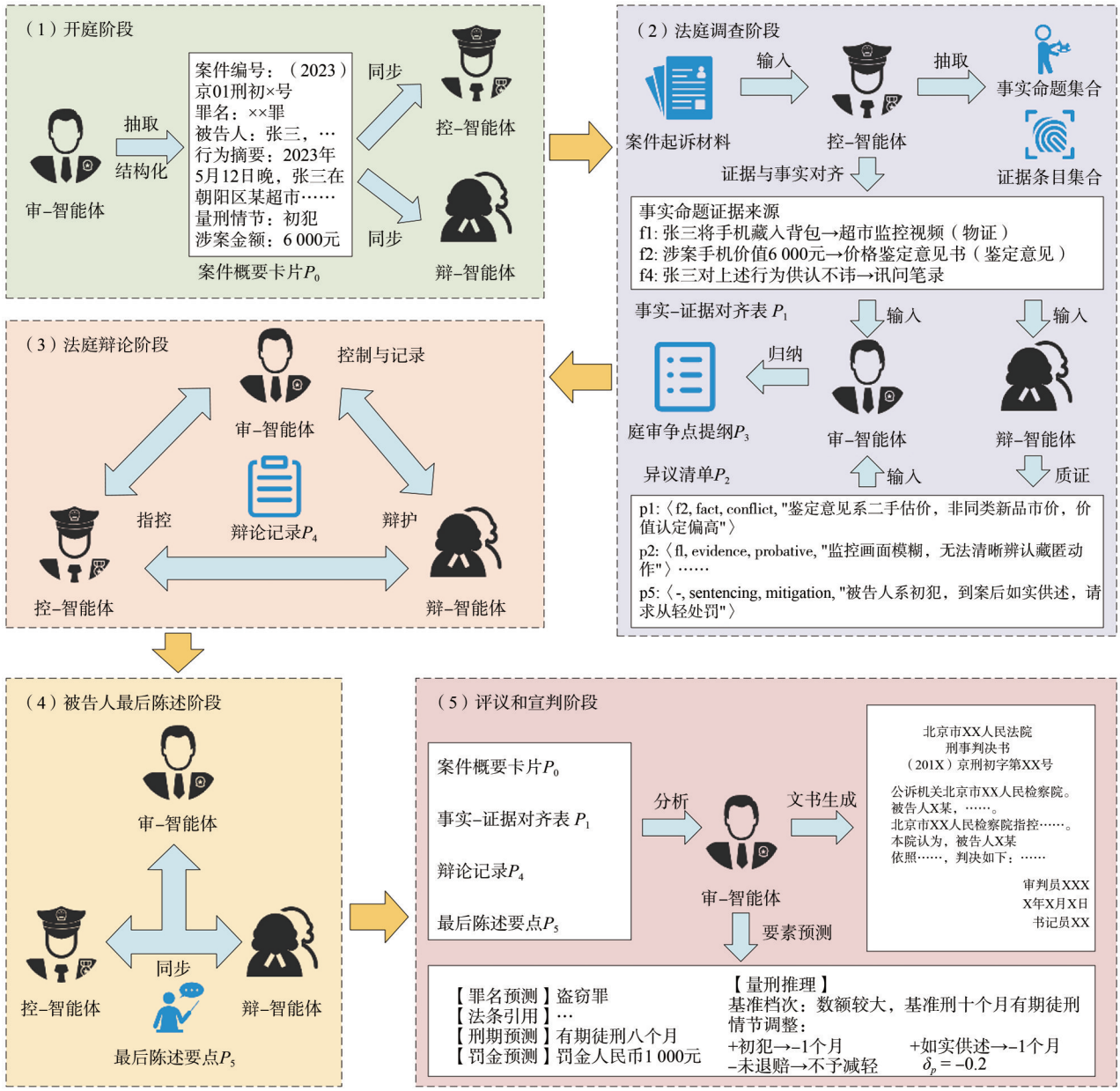


图 1　面向刑事判决预测的“控辩审”多智能体协作框架

P_1，将辩方异议清单记为P_2，将审方归纳的庭审争点提纲记为P_3，将庭审交互的逐轮累积记录记为P_4，将量刑态度信息中明确记录的相关要点记为P_5。各智能体围绕这些中间产物交换信息，但仍分别承担与其法律角色对应的职责。

在开庭阶段，审方从x_{base}与起诉材料中抽取主体、时间、地点、行为、数额和结果等案件要素，形成P_0；控方和辩方读取P_0以统一实体指代和事实基准。

在法庭调查阶段，控方从x_{indict}中整理指控事实，并将每项事实与拟证明的证据片段对齐，形成P_1；辩方结合辩方证据材料进行审阅，针对P_1提出异议并明确辩护方向，形成P_2；审方据此筛选存在实质分歧的事项并形成P_3。

在法庭辩论阶段，审方依据P_3调度发言轮次，控方围绕争点回应辩方异议，辩方进行质证、反驳或提出从宽主张。每轮结束后，审方将双方发言、引用的事实和证据以及争点状态增量写入P_4；庭审结束后形成完整的辩论记录。

在最后陈述阶段的计算抽象中，审方从量刑态度信息中提取认罪认罚及其他明确记录的量刑态度，形成P_5，并将其作为终局裁判的阶段性输入。

评议和宣判阶段由审方综合P_0、P_1、P_4和P_5，确定判决要素并生成裁判文书。下面分别说明3个智能体的设计。

2.3 检察官“控”智能体设计

检察官“控”智能体（AP）承担指控事实整理、举证组织和控方意见生成职责。AP以x_{indict}和P_0为主要输入，先把起诉叙事转化为可核验的事实命题，再将命题与证据片段绑定，最后根据P_3和辩方异议生成面向争点的控方回应。其输出P_1供辩方质证和审方归纳争点，庭审中的控方回应则进入P_4。

（1）指控事实结构化

AP首先对起诉书中的事实叙事进行分段识别和实体归一化，每条事实命题用五元组$\langle a_i, v_i, o_i, \ell_i, t_i \rangle$表示，依次对应行为人、行为动作、行为对象或被害对象、时间地点，以及结果、数额或后果。缺失字段记为∅，用于区分信息缺失、否定陈述和数值为零3种情况。

事实抽取采用分步提示约束：先定位案件事实段落并识别事件陈述，再将候选陈述转写为五元组，最后统一实体指代、合并重复命题并保留原文定位。这一分步过程与链式思维提示的组织方式一致[50]，使事实命题的字段含义和来源都可被后续模块检查。以“张三在超市内带离价值6 000元的手机，随后被保安拦截”为例，AP分别记录带离行为及其对象、时间地点和数额，并将拦截行为单独记录为事实命题。

（2）事实-证据对齐

AP从起诉书的证据目录和附件中抽取可引用的证据片段，并为每条事实命题记录其证据集合，对齐时同时比较证据的拟证明内容与命题的行为人、行为、对象、时间地点和结果等核心要素；语义匹配借鉴句向量方法[51]，候选对齐关系还需依据证据类型、文本来源和拟证明事项进行规则校验。

P_1以“事实命题-证据片段集合”的形式保存。一个命题可以对应多个相互印证的片段；无匹配证据的命题标记为证据缺口，并在P_2和P_3中列为优先审查对象。该对齐关系为辩方逐项质证提供定位，也使审方能够区分事实争议与证据争议。

（3）争点回应生成

控方意见生成围绕P_3中的庭审争点展开。AP先使用控方辩论策略图谱（prosecution argumentation strategy graph，PASG）确定回应路径，用于限定回应目标、可引用材料与发言结构；随后从P_1中检索与争点相关的事实命题及证据片段，按策略要求组织控方发言。

PASG依据相关法律规范人工构建，图谱分为争点域、异议类型和策略叶子3层，如图2所示。争点域包括事实、证据、法律适用和量刑4类问题。事实类节点关注关键事实是否得到证明以及事实命题之间是否一致，证据类节点对应证据合法性、真实性、关联性和证明力的审查，两类节点的划分参照《中华人民共和国刑事诉讼法》关于证据审查、证明标准和非法证据排除的规定。法律适用类节点依据《中华人民共和国刑法》总则、分则及相关司法解释处理罪名定性与构成要件争议；量刑类节点结合刑法规定和最高人民法院、最高人民检察院的《关于常见犯罪的量刑指导意见（试行）》设置从重、从轻和减轻处罚情节。每个异议类型均配置反驳式和调整式策略叶子，用于限定回应目标、可引用材料和发言结构。图谱架构经5名法律专家审阅，并根据专家意见修订争点分类、类型边界和策略适用条件。该分层设计与争点驱动的法律论证逻辑一致[52]。

对于庭审争点提纲P_3中的争点及其对应异议，AP先将其映射到争点域和异议类型，再依据现有证据与论证对该异议的可反驳度$\rho(z)$选择反驳式或调整式策略：

$$(d,o)=\text{Map}(q,\text{z}),s=\begin{cases}S^{R}_{d,o},\ \rho(z)\geqslant\theta\\ S^{A}_{d,o},\ \rho(z)<\theta\end{cases}\tag{2}$$

其中，映射函数将争点及其对应的异议归入PASG中相应的争点域和异议类型；S_R表示在控方证明目标不变的情况下，以已有材料直接回应并补强证据链；S_A表示在异议成立或证据不足时，收缩事实主张、调整法律适用或修正量刑建议。阈值θ用于区分两种策略。$\rho(z)$取值范围为[0,1]，由AP依据与异议相关的事实命题、证据覆盖情况及反驳依据的充分程度进行判断，数值越高，表示现有材料越足以直接回应该异议。完成路由后，AP只检索策略叶子允许引用的材料，并按“复述异议—给出判断—逐项列示依据”的顺序生成控方意见。

例如，辩方主张受贿案件中财物交付的时间和方式不清，AP将其归入关键事实未明，并检索银行流水、电子数据和相关供述中的交付信息。若这些材料能够补足交付要素，则采用反驳式策略逐项

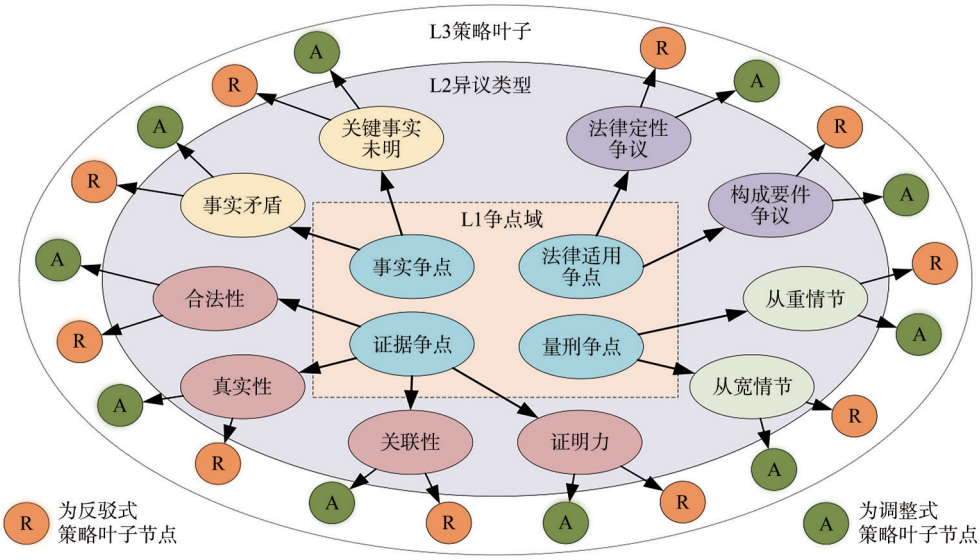


图2　控方辩论策略图谱

回应；若仍存在缺口，则采用调整式策略明确尚不能支持的证明范围。

2.4　律师/当事人“辩”智能体设计

律师/当事人“辩”智能体（AD）以P_0、P_1和辩方证据材料为输入，承担控方材料审查、异议提出和辩护意见生成职责。AD首先输出结构化异议P_2，随后结合P_3和上一轮记录在P_4中的控方回应形成当前轮次的辩护发言。辩方侧的核心任务是识别证据链缺口、程序瑕疵和推理不一致；证据足以支持指控事实时，论证资源转向量刑争点。

（1）质证意见生成

AD按照“证据覆盖—证据可靠性—命题一致性”的顺序审查P_1。每条异议p_k记录为$\langle q_k, d_k, o_k, r_k \rangle$，分别表示被质疑的事实命题、争点域、异议类型和质疑理由。首先，检查每项命题是否有证据覆盖，以及是否仅由单一片段支撑；其次，依据证据来源和形式审查取证程序、真实性、关联性和证明力；最后，比较不同命题之间的主体、时间、行为和结果是否相互一致，并核对控方结论是否超出证据能够支持的范围。

审查结果被归入事实不足、证据瑕疵、证明力不足、命题矛盾或法律适用不当等类型，并保留对应的原文片段和证据位置。对没有形成实质分歧的问题，AD不重复生成泛化异议；会影响罪名、法条或量刑的异议，则优先提交给审方形成P_3。

（2）辩护意见生成

AD在辩论开始前，依据P_3为每个争点建立初始辩护记录，记录争点、关联异议、辩方立场和主张文本。立场包括3种基本状态：deny表示否认控方对该争点的证明；mitigate表示承认部分事实但主张从宽处理；concede表示在控方证据充分时停止争执该事实，并将论证资源转向其他争点。立场选择综合P_2中的异议理由、P_1中的证据覆盖和既定辩护方向，生成环节据此组织表述。

在每轮交互中，AD判断控方回应是否带来新的事实或证据。新信息影响既有争点时，仅更新该争点的立场和引用；未受影响的争点维持原有记录。当前发言及其引用在本轮结束后写入P_4。

2.5　法官“审”智能体设计

法官“审”智能体（AJ）负责建立统一事实基准、归纳庭审争点、控制辩论程序并做出终局裁判。AJ先基于x_{base}和起诉材料建立P_0；在法庭调查结束后，依据P_1和P_2归纳P_3。辩论阶段，AJ依据P_3维护发言顺序并形成P_4；最后陈述对应的计算阶段从量刑态度信息中提取P_5，随后综合各阶段材料确定判决要素并生成裁判文书。

（1）案件要素抽取

AJ将x_{base}和起诉材料中的案情叙述整理为案件概要P_0。每个案件要素记录为“类别-规范化值-原文位置”三元组，类别覆盖主体、时间、地点、行为、对象、数额、结果和量刑情节等。抽取采用分类引导流程，按类别扫描文本并同步完成实体规范化，将“其”“该被告人”等指代表达统一为规范实体名；对自首、坦白、从犯、未遂等量刑情节，则

保留触发该情节的原文依据，供后续裁判说理核验。

（2）庭审争点归纳

AJ综合P_1和P_2，将每条异议映射到事实、证据、法律适用或量刑争点，并定位其关联事实命题及证据集合。若异议没有改变控方事实或证据结论，则降低其庭审优先级；若涉及证据缺口、程序合法性、构成要件或量刑情节，则纳入P_3。AJ再按争点域和优先级排序，明确每个争点的控方初始主张、辩方初始立场以及需要通过辩论解决的问题。

（3）庭审控制与记录

AJ依据P_3采用争点驱动的轮次调度。每轮优先处理尚未解决且影响判决要素的争点，由控方先回应，辩方随后质证或让步。双方形成一致或辩方撤回异议时，AJ将该争点标记为已处理；达到程序允许的轮次上限仍有分歧的，保留为待裁断状态，再转入下一争点。

P_4按争点保存发言角色、发言内容、引用的事实和证据、争点状态以及控辩双方的核心观点摘要。AJ将原始发言与归纳摘要分别保存；终局推理时，AJ通过摘要定位原始发言及其引用，摘要不替代原始交互内容。

（4）裁判文书生成

终局阶段，AJ先根据P_0、P_1、P_4和P_5对罪名、适用法条、刑期和罚金逐项作出判断，再以这些判决要素为约束生成裁判文书。罪名判断核验犯罪构成要件与争点处理结果，法条选择与罪名和争点结论保持一致，量刑判断则区分基准刑与情节调整。刑期与罚金计算分别写为：

$$\begin{aligned} p &= \mathrm{clip}\left(p_{\mathrm{base}}\left(1+\delta_p\right), p_{\min}, p_{\max}\right) \\ f &= \mathrm{clip}\left(f_{\mathrm{base}}\left(1+\delta_f\right), f_{\min}, f_{\max}\right) \end{aligned} \tag{3}$$

其中，p_{base}为所适用量刑档次的基准刑期，δ_p为已认定量刑情节形成的综合调整项，$p_{\min}$和$p_{\max}$为相应的法定刑期边界。f_{base}根据所适用罪名、涉案数额或违法所得、危害后果等已认定事实确定基准罚金，δ_f为与罚金相关的从重、从轻情节形成的调整项；$f_{\min}$和$f_{\max}$为法律或司法解释规定的数额边界。$\mathrm{clip}(\cdot)$表示区间截断，初始值低于下界时取下界，高于上界时取上界，处于区间内时保持不变。法律或司法解释未规定可用于截断的罚金数额边界时，式（3）中的罚金项不执行$\mathrm{clip}(\cdot)$操作，可视为恒等映射，不另设经验上限。该表达用于组织数值型刑罚要素，具体数额仍由AJ依据P_0、P_1、P_4和P_5中可核验的材料，结合相应法律依据和裁判理由作出判断。

AJ在生成阶段引入检索增强生成（retrieval-augmented generation，RAG）[53]，用于规范裁判文书的表述和法条引用。RAG案例库由150个训练案件构建，并依据案号、case_id与事实文本相似度进行去重。案例库中的每条记录包含案件事实、判决要素和对应裁判文书。文本相似度沿用句向量检索思路[51]。BAAI/bge-m3[54]编码当前案件事实与候选案例事实，综合排序进一步加入罪名一致性和刑期接近度。案例库返回综合得分最高的K个案例，K固定为3。

候选案例b_n的综合得分同时考虑事实文本相似度、罪名一致性和刑期接近度：

$$\mathrm{sim}_n = \lambda_1 s_{\mathrm{text}}\left(q_{\mathrm{ret}}, b_n\right) + \lambda_2 I\left(c_n = c\right) + \lambda_3 s_{\mathrm{term}}\left(p_n, p\right) \tag{4}$$

其中，q_{ret}为当前案件事实的编码表示，b_n为第n个候选案例的编码表示，s_{text}为q_{ret}与b_n的归一化余弦相似度，$I\left(c_n = c\right)$为罪名一致性的指示函数，$s_{\mathrm{term}} = \exp\left(-\left|p - p_n\right| / p_{\max}\right)$为刑期接近度，其中分母取当前案件所适用量刑档次的法定最高刑期；c和p为当前案件的预测罪名和刑期，c_n和p_n为候选案例的对应标签。3项权重满足$\lambda_1+\lambda_2+\lambda_3=1$。检索案例用于组织裁判说理和校正规范表述，当前案件的事实认定与判决要素仍由P_0、P_1、P_4和P_5确定。最终，AJ依据判决要素设置一致性约束，参考类案文本生成$\hat{j}$，并检查文书中的罪名、刑期和罚金是否与预测结果一致，以及核心适用法条是否正确写入文书。为统一3项得分的取值范围，文本余弦相似度经$(1+\cos)/2$归一化为s_{text}，其取值范围为[0,1]；罪名一致性取0或1，刑期接近度保持(0,1]范围。在模型主干、解码参数和检索数保持不变的条件下，本文在50个验证案件上比较候选权重，以法条引用F1占0.5、裁判理由BERTScore与判决结果BERTScore各占0.25的综合值为选择依据。最终取$\lambda_1=0.55$、$\lambda_2=0.30$、$\lambda_3=0.15$，测试集仅用于最终评测。

3 实验与分析

3.1 实验设置

（1）实验数据

本文所用数据集以SimCourt[49]开源的刑事判

决预测数据为基础。SimCourt 数据来源于真实刑事判决书，经匿名化处理（人名替换为“某”，日期替换为“20XX 年 X 月 X 日”），其量刑预测子集包含 200 个案件，每个案件均提供起诉书文本、结构化证据材料、量刑态度信息及对应的刑期、罚金、缓刑、罪名与判决结果等标注字段。

为扩充数据规模并提升罪名分布的均衡性，本研究首先从中国裁判文书网检索 2024 年至 2025 年的刑事判决书，根据案号、案件内容和罪名范围筛选出 200 件符合研究要求的案件。依托实验室与最高人民法院司法大数据研究院签订的科研合作协议，在核验案号对应关系后，获取入选案件对应的真实起诉书、控辩双方证据材料和庭审记录。起诉书与证据材料按照 SimCourt 的数据字段进行匿名化和规范化整理，量刑态度信息从真实庭审记录中提取。最终形成包含 400 个案件的实验数据集，覆盖危险驾驶罪、盗窃罪、故意伤害罪、交通肇事罪、诈骗罪、走私、贩卖、运输、制造毒品罪等 40 种主要罪名，每种罪名均包含 10 件案件。数据字段结构见表 1。

其中，被告人信息（defendant_information）用于构成 x_{base}；删除量刑建议、指控罪名及其同义表达、援引法条和裁判结果后的起诉材料（prosecution_statement_sub）与 evidence 字段中的控方证据组共同构成 x_{indict}；evidence 字段中的辩方证据组构成 x_{defense}；量刑态度信息（defendant_attitude）构成 $x_{\text{statement}}$。实验不读取包含上述标签信息的原始 prosecution_statement 字段。刑期、罚金与罪名字段构成真实判决标注，用于与预测值 y 对照；适用法条从裁判文书全文（result）中抽取，并与生成文书中的法条引用进行集合匹配。裁判文书全文作为参考文书，用于评估生成质量。400 个案件按罪名分层划分为 150 件训练集、50 件验证集和 200 件测试集。测试集中每种罪名均为 5 件，训练集和验证集中每种罪名分别为 3~4 件和 1~2 件；划分清单以 case_id 形式为补充材料提供。LoRA 训练使用 150 个训练案件，模型选择与阈值调参仅使用训练集和验证集，最终结果在 200 个测试案件上报告。数据集统计特征见表 2。

（2）对比方法

为评估 PDJ-MACF 的有效性，本研究从抽取式方法、预训练生成模型、大语言模型方法与多智能体方法 4 个层次设置对比基线。

①统计与图模型方法方面，选取 TextRank[55] 与基于 TF-IDF 词项加权[56]的抽取式方法作为对比基线，二者均从案件材料中抽取关键句组合输出，不具备生成与推理能力，用于衡量深度模型相对于抽取式方法的性能增益。

②生成式预训练模型方面，选取 BART[57]、BERTSumAbs[58] 与 CPT[59]3 种序列到序列生成模型。其中 BART 与 BERTSumAbs 为通用生成式摘要模型，CPT 为面向中文的预训练生成模型，三者均以案件材料为输入直接生成裁判文书，不引入任何法律角色结构或程序结构约束。

③基于大语言模型方法方面，以 Qwen3-8B[60] 为统一主干，设置 Few-Shot、SFT、CoT 和 ReAct 4 种对比方法。Direct Prompting 作为跨模型实验的统一参照，用于比较不同模型引入 PDJ-MACF 前后的性能变化。其中，Direct Prompting 直接将案件材料输入模型并要求输出裁判文书，不附加任何示例或推理引导；Few-Shot[61]，从训练集中选取 3 个人工标注的案件示例并附加至输入，引导模型参照示例格式生成；SFT[62]，在 150 个训练案件

表 1　数据字段结构

标签	字段	内容
被告人信息	defendant_information	被告人黄某，男，1995 年……
量刑态度信息	defendant_attitude	认罪认罚
真实起诉书	prosecution_statement	河北省承德市……起诉书全文……
证据	evidence	[[控方证据组],[辩方证据组]]
刑期（月）	term	4.0
罚金（元）	fine	20 000.0
缓刑（月）	reprieve	6.0
罪名	charge	["危险驾驶罪"]
裁判文书全文	result	"裁判文书全文"
标签遮蔽后的起诉材料	prosecution_statement_sub	起诉材料

表 2 数据集统计特征

统计项	数值	统计项	数值
案件总数/个	400	平均刑期/月	67.81
其中 SimCourt 原始数据/个	200	含缓刑案件/个	103（25.75%）
其中本研究补充数据/个	200	含罚金案件/个	270（67.50%）
覆盖罪名种类/种	40	各类罪名样本数	10

上对 Qwen3-8B 进行监督微调，采用 LoRA[63]方式（rank=16，alpha=32，目标模块为 q/k/v/o 投影层），学习率 5×10^{-5}，训练 3 个 epoch；CoT[50]，通过思维链提示引导模型逐步推理后输出裁判结论；ReAct[64]，采用“推理–行动”交替框架，引导模型在生成过程中动态调用法条检索工具，将检索结果融入推理链后输出裁判文书。

④多智能体方法方面，MAGLJP[43]面向多被告案件建模被告人之间的责任关系与量刑关联，本文使用单被告庭审案件；AgentsBench[44]在已给定案件事实、罪名和适用法条的条件下预测刑期，与本文联合预测罪名、法条、刑罚并生成裁判文书的设置不同；Debate-Feedback[45]未提供与原文实验设置对应的完整实现和配置，因而无法按原文条件复现；MALR[46]主要考察候选罪名构成要件的理论推理，输出不包含刑罚、法条或裁判文书。因此，本文选取 AgentsCourt[41]和 SimCourt[49]作为对比基线。AgentsCourt 通过法庭辩论模拟、法律资源检索与判决书生成模块的协同，模拟法官决策过程并输出判决结果。SimCourt 配置法官、检察官、辩护律师、被告人和书记员等角色，围绕庭审准备、法庭调查、举证质证、法庭辩论和最后陈述组织交互。此外，为了控制模型主干的影响，本文统一以 Qwen3-8B 运行两种多智能体方法。

此外，为验证不同大语言模型主干对 PDJ-MACF 框架性能的影响，本文进一步在 Hunyuan-7B-Instruct[65]、Baichuan2-7B-Chat[66]、Llama-3-8B-Instruct[67]、Qwen3-8B[60]、DeepSeek-R1-0528-Qwen3-8B[68]与 GLM-4-9B-chat[69]6 种模型上分别以 Direct Prompting 与 PDJ-MACF 两种方式运行，对比框架在不同模型上的一致性增益。

（3）参数设置

所有实验均在 CY5468M 服务器上本地部署运行，服务器配置双路 Intel Xeon 5320 处理器，256 GB DDR4-3200 ECC 内存，8 块 NVIDIA A800 GPU。推理阶段统一采用 8 卡并行，temperature 设为 0.7，top-p 设为 0.8，top-k 设为 20，最大生成长度为 2 048 个 token。SFT 训练采用 LoRA 微调，最大输入长度 1 536 个 token，批大小为 2，其余超参数与上述一致。控方辩论策略图谱中的异议可反驳度阈值 θ 设为 0.6，即当现有证据与论证对辩方异议的直接反驳充分程度低于 0.6 时，控方采用调整式策略；反之，采用反驳式策略。RAG 检索固定返回 3 个案例，文本向量编码器采用 BAAI/bge-m3。案例排序中，事实文本语义相似度、罪名一致性和刑期接近度的权重分别设为 0.55、0.30 和 0.15；该组合由验证集选择，测试集不参与参数确定。

（4）评价指标

为评估判决要素 y 的预测准确性及生成裁判文书 j 的质量，本文参考 JuDGE 基准[70]构建评价体系，从刑罚预测准确度、罪名预测准确度、法条引用准确度与文书语义相似度 4 个维度共计 12 项指标进行评估。具体指标定义如下。

①刑罚预测准确度（penalty accuracy）：分别计算刑期（prison）与罚金（fine）两项指标。对任一数值型刑罚要素，设预测值为 L_{pred}、真实值为 L_{true}，采用归一化相似度得分：

$$S=1-\frac{\left|L_{\text{pred}}-L_{\text{true}}\right|}{\max\left(L_{\text{pred}},L_{\text{true}}\right)} \tag{5}$$

当预测值和真实值均为 0 时，定义 S=1；其他情况按式（5）计算。S 的取值范围为[0, 1]，得分越高，表示预测越接近真实裁判结果。各实验表格所列刑期与罚金准确度均为有效数值型刑罚案例上 S 的算术平均值。

②罪名预测准确度（charge accuracy）：设测试集包含 N 个案件，第 i 个案件的真实罪名集合与预测罪名集合分别为 C_i 和 $\hat{C}_i$。案件级精确率、召回率及测试集总体指标计算如下：

$$p_i^C=\frac{\left|\hat{C}_i\cap C_i\right|}{\left|\hat{C}_i\right|},\ r_i^C=\frac{\left|\hat{C}_i\cap C_i\right|}{\left|C_i\right|}$$
$$P_C=\frac{1}{N}\sum_{i=1}^{N}p_i^C,\ R_C=\frac{1}{N}\sum_{i=1}^{N}r_i^C,\ F1_C=\frac{2P_CR_C}{P_C+R_C} \tag{6}$$

其中，P_C 和 R_C 为各案件精确率与召回率的算术平均值，采用案件级宏平均统计口径。若相应集合为空导致分母为 0，则该项记为 0；当 $P_C+R_C=0$ 时，$F1_C=0$。

③法条引用准确度（statute reference accuracy）：

从生成文书和真实文书中抽取《中华人民共和国刑法》条文编号。设第 i 个案件的真实法条集合与预测法条集合分别为 A_i 和 $\hat{A}_i$，则：

$$p_i^A=\frac{\left|\hat{A}_i\cap A_i\right|}{\left|\hat{A}_i\right|},\ r_i^A=\frac{\left|\hat{A}_i\cap A_i\right|}{\left|A_i\right|}$$
$$P_A=\frac{1}{N}\sum_{i=1}^{N}p_i^A,\ R_A=\frac{1}{N}\sum_{i=1}^{N}r_i^A,\ \mathrm{F1}_A=\frac{2P_AR_A}{P_A+R_A} \tag{7}$$

其中，精确率衡量所引法条的准确程度，召回率反映真实文书所引法条的覆盖程度。零分母情况按式（6）的规则处理。

④文书语义相似度（document semantic similarity）：在“裁判理由”与“判决结果”两个文本片段上分别计算 METEOR[71] 和 BERTScore[72]。设 d_i^q 和 $\hat{d}_i^q$ 分别为第 i 个案件的真实文本与生成文本，其中 $q\in\{r,j\}$，r 表示“裁判理由”，j 表示“判决结果”。两个片段的语义相似度计算如下：

$$M_q=\frac{1}{N}\sum_{i=1}^{N}\mathrm{METEOR}\left(d_i^q,\hat{d}_i^q\right)$$
$$B_q=\frac{1}{N}\sum_{i=1}^{N}\mathrm{BERTScore}_{\mathrm{F1}}\left(\hat{d}_i^q,d_i^q\right),\ q\in\{r,j\} \tag{8}$$

其中，M_q 和 B_q 分别表示 METEOR 与 BERTScore 的样本均值，BERTScore 取 $F1$ 分量。由此得到裁判理由 METEOR、裁判理由 BERTScore、判决结果 METEOR 和判决结果 BERTScore 共 4 项指标，数值越高，表示生成文书与真实文书越接近。

3.2　实验结果及分析

（1）不同裁判文书生成方法性能对比

表 3 汇总了不同方法在刑事判决预测任务上的结果。以 Qwen3-8B 为统一主干时，PDJ-MACF 在 12 项具体指标中有 9 项达到最高值；SimCourt 在法条引用 $F1$ 和判决结果 BERTScore 上领先，ReAct 的法条引用精确率最高。各项结果呈现出不同方法在预测环节上的性能差异。

在刑罚预测中，PDJ-MACF 的刑期和罚金准确度分别为 0.686 1 和 0.571 6，均为表中最高值。相对 SimCourt，两个指标分别提高 0.012 5 和 0.031 3；相对 AgentsCourt，增幅为 0.034 9 和 0.053 5。刑期与罚金均以预测值和真实值的接近程度计分，两个数值同时提高，表明其量刑结果更接近测试集标注。

在罪名预测的案件级宏平均口径下，PDJ-MACF 的精确率、召回率和 $F1$ 值分别为 0.970 6、0.941 0 和 0.955 6。SimCourt 的对应结果为 0.948 1、0.914 4 和 0.930 9，PDJ-MACF 在 3 项指标上分别提高 0.022 5、0.026 6 和 0.024 7。与 AgentsCourt 相比，PDJ-MACF 的罪名 $F1$ 提高 0.035 3。

法条引用指标给出了更有区分度的结果。SimCourt 的精确率、召回率和 $F1$ 值分别为 0.829 5、0.794 8 和 0.811 8，法条引用 $F1$ 为表中最高。PDJ-MACF 的召回率达到 0.801 7，高于 SimCourt 的 0.794 8，但精确率为 0.761 5，$F1$ 值为 0.781 1。ReAct 的精确率最高，为 0.877 8，其召回率仅为 0.582 9，$F1$ 值为 0.700 6。3 种方法在法条集合匹配上体现出不同的精确率和召回率取向。

文书语义相似度方面，PDJ-MACF 在裁判理由 METEOR、裁判理由 BERTScore 和判决结果 METEOR 上分别取得 0.668 9、0.912 7 和 0.648 3，

表 3　不同裁判文书生成方法在刑事判决预测任务上的性能对比

模型	刑罚预测准确度		罪名预测准确度			法条引用准确度			文书语义相似度			
	刑期	罚金	P	R	$F1$	P	R	$F1$	裁判理由		判决结果	
									METEOR	BERTScore	METEOR	BERTScore
TextRank	0.184 3	0.093 4	0.581 2	0.574 3	0.577 7	0.112 4	0.048 7	0.068 0	0.623 4	0.718 7	0.498 7	0.231 2
TF-IDF	0.201 2	0.098 7	0.603 0	0.677 3	0.638 0	0.218 7	0.193 4	0.205 3	0.512 7	0.753 9	0.351 2	0.284 3
BART	0.141 2	0.127 4	0.661 3	0.646 7	0.653 9	0.294 4	0.114 0	0.164 4	0.158 7	0.631 8	0.122 4	0.271 9
BERTSumAbs	0.253 4	0.189 8	0.546 1	0.617 9	0.579 8	0.519 8	0.417 9	0.463 3	0.304 9	0.793 1	0.426 7	0.757 8
CPT	0.231 4	0.180 1	0.570 8	0.650 3	0.608 0	0.275 5	0.182 9	0.219 8	0.235 1	0.719 8	0.226 2	0.441 1
Qwen3-8B@Few-Shot	0.539 9	0.371 8	0.694 8	0.639 4	0.665 9	0.748 1	0.766 4	0.757 1	0.458 4	0.772 2	0.533 1	0.493 2
Qwen3-8B@SFT	0.586 4	0.469 6	0.837 5	0.841 6	0.839 5	0.664 2	0.665 7	0.664 9	0.448 9	0.823 7	0.537 2	0.814 7
Qwen3-8B@CoT	0.606 1	0.449 3	0.820 7	0.751 7	0.784 7	0.630 6	0.608 1	0.619 1	0.402 1	0.797 4	0.545 0	0.781 6
Qwen3-8B@ReAct	0.624 5	0.499 5	0.893 4	0.901 2	0.897 3	**0.877 8**	0.582 9	0.700 6	0.481 2	0.842 8	0.582 2	0.884 3
Qwen3-8B@AgentsCourt	0.651 2	0.518 1	0.947 4	0.894 7	0.920 3	0.718 7	0.723 4	0.721 0	0.583 4	0.908 1	0.621 4	0.870 2
Qwen3-8B@SimCourt	0.673 6	0.540 3	0.948 1	0.914 4	0.930 9	0.829 5	0.794 8	**0.811 8**	0.612 7	0.893 6	0.627 1	**0.917 5**
Qwen3-8B@PDJ-MACF	**0.686 1**	**0.571 6**	**0.970 6**	**0.941 0**	**0.955 6**	0.761 5	**0.801 7**	0.781 1	**0.668 9**	**0.912 7**	**0.648 3**	0.886 1

均为表中最高值。判决结果BERTScore由SimCourt取得最高值0.917 5，PDJ-MACF为0.886 1。METEOR与BERTScore分别主要反映词元匹配和上下文语义相似性，判决结果需要结合两项得分考察。

综合表3各项结果，PDJ-MACF在数值型刑罚、罪名判别及裁判理由生成方面取得稳定优势，SimCourt则在法条引用*F*1与判决结果BERTScore上更高。两种方法的性能差异主要集中在法条集合选择和判决结果文本表达。

（2）不同大语言模型主流的性能对比

6种模型上的性能见表4。

从Direct Prompting的基础性能来看，不同模型在各评价维度上存在明显差异。Hunyuan-7B-Instruct与GLM-4-9B-chat的罪名预测*F*1分别为0.838 7和0.896 2，两者的法条引用*F*1分别为0.444 2和0.609 4。Llama-3-8B-Instruct的罪名预测*F*1为0.453 2；DeepSeek-R1-0528-Qwen3-8B的罪名预测和法条引用*F*1分别为0.741 6和0.480 0，刑期预测准确度为0.347 9。不同主干在罪名、法条与数值型刑罚预测上的表现并不一致。

引入PDJ-MACF框架后，6种主流模型在全部指标上均出现不同程度的提升。其中，Qwen3-8B在刑期、罪名*F*1、法条*F*1和裁判理由METEOR 4项指标上的增幅较为突出：刑期预测准确度由0.351 1提升至0.686 1，提升0.335 0；罪名预测*F*1由0.599 6提升至0.955 6，提升0.356 0；法条引用*F*1由0.311 6提升至0.781 1，提升0.469 5；裁判理由METEOR由0.312 9提升至0.668 9，提升0.356 0。在Direct Prompting条件下，Qwen3-8B的基础表现与引入任务结构、角色分工和程序结构约束后的表现存在明显差异。案件事实抽取、证据组织、异议生成、争点归纳和裁判生成等步骤的结构化安排，与该模型各项指标的增益同时出现。

对于基础性能较强的模型，PDJ-MACF同样带来增益，但提升幅度相对收敛。例如，Hunyuan-7B-Instruct在Direct Prompting下已具备较好的罪名识别和案件理解能力，引入PDJ-MACF后，其刑期预测准确度和罪名预测*F*1分别提升约6.6和6.0个百分点。GLM-4-9B-chat在直接提示下整体表现较为均衡，引入PDJ-MACF后，罪名预测和法条引用F1进一步提升至0.931 8和0.787 5。Baichuan2-7B-Chat与Llama-3-8B-Instruct经PDJ-MACF处理后亦实现提升，但最终性能与Qwen3-8B、GLM-4-9B-chat仍存在差距。结果显示，框架增益会随主干模型的中文法律语义理解、指令遵循和长文本整合能力而变化。

跨模型实验中6种主流模型均有指标提升，表明PDJ-MACF在本实验范围内具有一定适用性，但其向其他模型和任务的迁移能力仍待检验。

（3）消融实验

为考察各组成部分对模型性能的影响，本文设置5组消融实验。除被移除的组件外，各变体均采用相同的模型主干、数据划分和解码参数。具体设置如下。

①w/o Stepwise Fact Organization：取消控方智能体在事实命题抽取中的分步提示，改为端到端直

表4 不同大语言模型对比分析

模型	方法	刑罚预测准确度		罪名预测准确度			法条引用准确度			文书语义相似度			
		刑期	罚金	*P*	*R*	*F*1	*P*	*R*	*F*1	裁判理由		判决结果	
										METEOR	BERTScore	METEOR	BERTScore
Hunyuan-7B-Instruct	Direct Prompting	0.544 4	0.273 7	0.832 3	0.845 3	0.838 7	0.707 5	0.323 7	0.444 2	0.352 1	0.801 2	0.248 7	0.683 7
	PDJ-MACF	0.610 3	0.348 6	0.895 6	0.902 4	0.899 0	0.731 8	0.451 2	0.558 2	0.413 2	0.873 8	0.368 3	0.813 3
Baichuan2-7B-Chat	Direct Prompting	0.312 4	0.154 3	0.583 2	0.591 4	0.587 3	0.321 5	0.218 7	0.260 3	0.284 0	0.762 4	0.213 9	0.531 2
	PDJ-MACF	0.428 7	0.261 4	0.712 4	0.725 3	0.718 8	0.563 4	0.482 8	0.520 0	0.351 2	0.804 3	0.328 6	0.775 7
Llama-3-8B-Instruct	Direct Prompting	0.357 1	0.108 6	0.449 4	0.457 1	0.453 2	0.249 8	0.112 3	0.154 9	0.261 9	0.774 3	0.312 4	0.281 8
	PDJ-MACF	0.443 8	0.218 7	0.682 3	0.691 2	0.686 7	0.452 1	0.368 7	0.406 2	0.328 7	0.812 9	0.394 7	0.630 7
DeepSeek-R1-0528-Qwen3-8B	Direct Prompting	0.347 9	0.208 3	0.728 2	0.755 5	0.741 6	0.582 1	0.408 4	0.480 0	0.360 7	0.794 3	0.268 7	0.681 2
	PDJ-MACF	0.512 4	0.428 7	0.893 4	0.901 2	0.897 3	**0.763 4**	0.551 2	0.640 2	0.451 2	0.848 1	0.512 9	0.857 1
GLM-4-9B-chat	Direct Prompting	0.449 1	0.192 0	0.896 2	0.896 2	0.896 2	0.610 7	0.608 1	0.609 4	0.464 7	0.864 3	0.331 2	0.751 4
	PDJ-MACF	0.583 4	0.418 7	0.931 9	0.931 7	0.931 8	0.757 1	**0.820 5**	**0.787 5**	0.542 1	0.901 9	0.618 7	0.858 7
Qwen3-8B	Direct Prompting	0.351 1	0.197 6	0.595 5	0.603 8	0.599 6	0.339 8	0.287 8	0.311 6	0.312 9	0.798 7	0.284 3	0.712 8
	PDJ-MACF	**0.686 1**	**0.571 6**	**0.970 6**	**0.941 0**	**0.955 6**	0.761 5	0.801 7	0.781 1	**0.668 9**	**0.912 7**	**0.648 3**	**0.886 1**

接生成，其余流程保持不变，用于分析分步推理对事实组织及下游任务的影响。

②w/o PASG：移除控方辩论策略图谱，控方不再根据异议类型选择回应策略，而是依据争点提纲和辩方异议直接生成回应，用于分析策略路由对控方论证及后续裁判的影响。

③w/o AD：移除辩方智能体，不生成质证意见和辩护意见，审方直接根据控方形成的事实命题、证据对齐结果和指控意见归纳争点并作出裁决，用于分析控辩交互的作用。

④w/o RAG：固定完整模型生成的罪名、法条、刑期和罚金，仅在裁判文书生成阶段移除Top-3类案上下文。该变体与完整模型共享前序判决要素，用于分析类案检索对法条引用和文书语义质量的影响。

⑤w/o Structured Sentencing：固定案件要素、罪名和法条，取消刑期与罚金生成中的结构化量刑约束，改由LLM直接给出数值，其余流程按新的刑罚结果继续执行，用于分析该约束对刑罚预测和文书生成的影响。

消融实验结果见表5。

消融结果反映了各模块在协作链条中的不同作用。取消分步事实组织后，罪名$F1$下降9.67个百分点，其他指标也都不同程度下降。这表明缺少分步提示时，事实命题与证据之间的对应关系被弱化，后续争点识别、罪名判断和裁判说理均缺少稳定的事实基础。PASG消融后，法条$F1$下降9.42个百分点，刑罚指标变化较小。PASG依据异议所属争点和类型选择回应策略；取消策略路由后，控方仍能提出基本指控，但对法律适用争议的回应不够集中，法条引用受到的影响更明显。

移除AD后，裁判理由与判决结果METEOR分别下降8.16和7.55个百分点。辩方提出的证据异议、构成要件质疑和从宽情节是审方归纳争点并校正控方主张的重要来源。缺少这些信息，文书对争议事实和量刑理由的覆盖随之减弱。RAG消融使法条引用$F1$下降12.89个百分点，文书指标也有较大回落。类案文本所提供的法条引用方式、说理结构和司法表达，有助于审方将既定判决要素组织为完整文书。

取消结构化量刑约束后，刑期和罚金准确度分别下降14.76和17.44个百分点，裁判理由的METEOR与BERTScore下降1.49和1.94个百分点，判决结果相应下降2.94和1.69个百分点。结构化量刑将基准刑期、基准罚金、量刑情节的调整方向与法定边界组织为明确的计算关系，使数值结论在文书生成前得到约束。改由LLM直接给出数值后，量刑判断更依赖自由文本生成，情节权衡与最终数值之间的对应关系变弱，误差进一步反映在量刑说理和判决主文中。

3.3　人工评测

尽管自动评价指标能够从刑罚预测准确度、罪名识别效果、法条引用精度和文书语义相似度等方面量化模型性能，但对于刑事裁判文书生成任务而言，仅依赖自动指标仍存在一定局限。裁判文书要求结果要素准确，同时还要求事实认定清楚、证据采信合理、法律适用恰当、量刑说理充分，并符合司法文书的规范表达要求。这些细粒度特征涉及法律推理质量、裁判逻辑严密性和司法规范性，难以由上述自动指标充分衡量。为此，本文进一步开展人工评测，以补充评价不同方法生成裁判文书的综合质量。

评测样本与流程：本文从测试集中随机抽取50个案件样本作为人工评测对象，样本覆盖盗窃、

表5　消融实验结果

模型	刑罚预测准确度		罪名预测准确度			法条引用准确度			文书语义相似度			
	刑期	罚金	P	R	$F1$	P	R	$F1$	裁判理由		判决结果	
									METEOR	BERTScore	METEOR	BERTScore
Qwen3-8B@PDJ-MACF	**0.686 1**	**0.571 6**	**0.970 6**	**0.941 0**	**0.955 6**	**0.761 5**	**0.801 7**	**0.781 1**	**0.668 9**	**0.912 7**	**0.648 3**	**0.886 1**
w/o Stepwise Fact Organization	0.628 7	0.513 8	0.881 9	0.837 1	0.858 9	0.693 0	0.731 8	0.711 9	0.571 3	0.883 9	0.569 6	0.851 9
w/o PASG	0.663 4	0.548 1	0.951 2	0.924 7	0.937 8	0.657 1	0.719 6	0.686 9	0.618 4	0.901 1	0.597 1	0.877 2
w/o AD	0.641 2	0.528 7	0.938 1	0.911 7	0.924 7	0.691 2	0.748 8	0.718 8	0.587 3	0.892 4	0.572 8	0.853 0
w/o RAG	0.686 1	0.571 6	0.970 6	0.941 0	0.955 6	0.624 8	0.682 2	0.652 2	0.506 7	0.839 2	0.528 0	0.817 4
w/o Structured Sentencing	0.538 5	0.397 2	0.970 6	0.941 0	0.955 6	0.761 5	0.801 7	0.781 1	0.654 0	0.893 3	0.618 9	0.869 2

故意伤害、诈骗、危险驾驶和交通肇事等常见罪名。为避免案件来源、审理法院及当事人身份等信息影响评估判断，所有样本在评测前均进行匿名化处理，删除或替换可能暴露案件背景的敏感内容。评测采用双盲方式进行，评估人员不知晓各生成结果对应的模型来源，不同模型的输出结果经随机排序后统一呈现，以减少顺序效应、模型先验认知和主观偏好对评分结果的影响。

评测团队：人工评测团队由具有不同法律背景的人员组成，包括 2 名刑事法官、2 名刑事律师和 6 名法学硕士研究生。该设置兼顾司法裁判经验、辩护实务经验和法学理论分析能力，有助于从多元视角评价生成文书质量。正式评分前，本文组织所有评估人员进行统一培训，详细说明评测目的、评分维度、评分尺度及刑事裁判文书的基本规范要求，并提供若干标准裁判文书用于评分尺度校准，以提高不同评估人员之间评分标准的一致性。

评测维度：结合刑事裁判文书生成任务的特点，本文从以下 5 个维度开展人工评测。

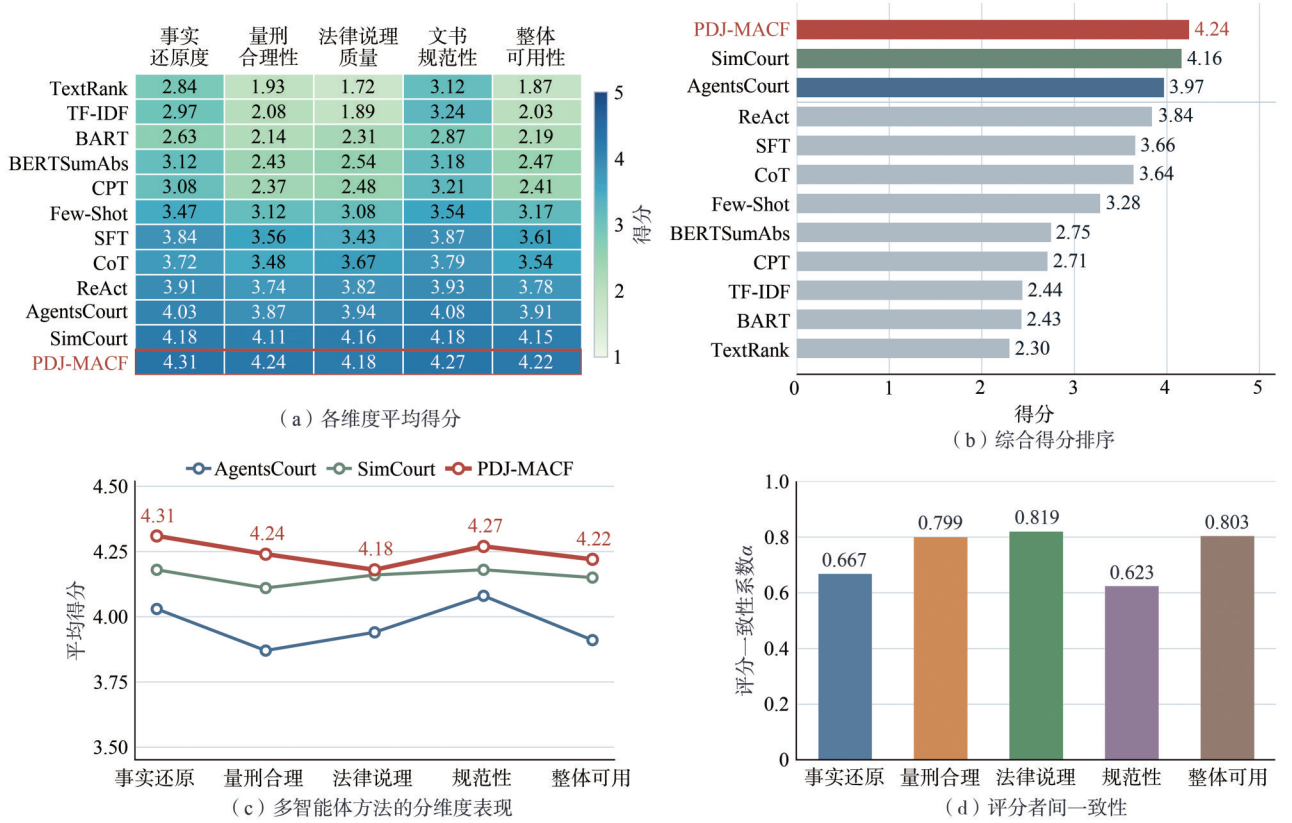
- 事实还原度（factual fidelity）：衡量生成文书对案件基本事实、关键情节和结构化要素的还原程度，重点考察是否存在事实遗漏、关键情节误写、主体关系混淆或前后表述矛盾等问题。
- 量刑合理性（sentencing rationality）：衡量文书中的刑期、罚金及量刑说理是否与案件事实和情节相匹配，重点考察模型能否结合从重、从宽、认罪认罚、退赔谅解、共同犯罪等量刑因素，形成合理的裁判结果表达。该维度用于衡量判决要素预测与文书生成的衔接质量。
- 法律说理质量（legal reasoning quality）：衡量文书法律论证的充分性、严密性与连贯性，重点考察事实认定、证据采信、法条适用和量刑分析之间的逻辑衔接是否清晰，是否能够形成完整的裁判论证链条，并避免法条堆砌、结论先行或说理跳跃等问题。
- 规范性（normative compliance）：衡量文书整体结构、段落组织、语言表达和法律术语使用是否符合刑事裁判文书写作规范，重点考察是否存在结构失衡、格式失范、术语使用不当或司法表达不严谨等问题。
- 整体可用性（overall usability）：从司法辅助应用角度综合评价生成文书的实际参考价值，即模型输出能否作为具有较高初稿质量的裁判文书，为后续法官审核、修改和完善提供有效基础。

每个维度采用五分制有序量表：5 分表示表现优秀，基本符合司法实践要求；4 分表示整体可接受，存在少量局部瑕疵；3 分表示基本合格，仍具有参考价值；2 分表示问题较明显，仅具有限参考价值；1 分表示质量较差，难以作为司法辅助文本使用。10 名评测人员对每个案例–方法输出独立评分。各方法在各维度上的得分均为 50 个案例、10 名评测人员评分的算术平均值，综合得分为 5 个维度得分的平均值。以同一案例和同一方法的输出为评分单元，采用 Krippendorff's alpha 检验五级有序评分的评分者间一致性。事实还原度、量刑合理性、法律说理质量、规范性和整体可用性的 alpha 分别为 0.667、0.799、0.819、0.623 和 0.803。图 3 展示了各方法的逐维得分、综合得分及评分者间一致性。

图 3（a）和图 3（b）呈现了不同方法在文书形式与裁判理由上的差异。TextRank 和 TF-IDF 的规范性得分分别为 3.12 和 3.24，直接复用案件材料中的表述，能够保留常用术语和基本结构；量刑合理性均不超过 2.08，法律说理质量仅为 1.72 和 1.89，抽取片段难以支撑事实认定、法律适用和量刑情节之间的连贯推演，综合得分因此仅为 2.30 和 2.44。BART、BERTSumAbs 和 CPT 的综合得分为 2.43~2.75，但量刑合理性和法律说理质量仍低于 2.55，通用生成模型对刑事裁判论证的把握仍有限。

大语言模型方法将整体得分带入 3 分区间。Few-Shot 的综合得分为 3.28；引入监督微调、链式推理或行动式推理后，SFT、CoT 和 ReAct 分别达到 3.66、3.64 和 3.84。SFT 的事实还原度与规范性分别为 3.84 和 3.87，均略高于 CoT，说明监督微调有助于模型学习较为稳定的文书结构。CoT 的法律说理质量为 3.67，比 SFT 高 0.24 分，显式展开推理过程有助于保持论证的连贯性。ReAct 的 5 项得分均不低于 3.74，表现较为均衡。

多智能体方法的综合得分进一步升至 3.97 以上。AgentsCourt、SimCourt 和 PDJ-MACF 的综合得分依次为 3.97、4.16 和 4.24。PDJ-MACF 在 5 个维度均居首位，但相对 SimCourt 的提升并不均匀：事实还原度和量刑合理性均高 0.13 分，规范性和整体可用性分别高 0.09 分和 0.07 分，法律说理质量仅高



（a）各维度平均得分　（b）综合得分排序

（c）多智能体方法的分维度表现　（d）评分者间一致性

图3　人工评测结果

0.02 分。法律说理质量上的接近结果说明，两种方法均能形成较完整的法律论证，差距主要集中在事实材料与量刑情节的组织，并反映在整体可用性上。图 3（d）的 Krippendorff's α 为 0.623~0.819。法律说理质量和整体可用性的 α 分别为 0.819 和 0.803，评估人员的判断较为一致；规范性为 0.623，对文书格式、语体和细部表述的判断仍存在一定差异。

4　结束语

本文提出面向刑事判决预测的“控辩审”多智能体协作框架 PDJ-MACF，以控方、辩方和审方为角色先验，在开庭、法庭调查、法庭辩论、最后陈述和评议宣判 5 个阶段组织协作推理。控方负责事实与证据组织，辩方负责异议和辩护意见生成，审方负责争点归纳、判决要素预测和裁判文书生成。

实验结果表明，PDJ-MACF 在 12 项自动评价指标中的 9 项取得最高值，6 种主干模型均出现指标增益，人工评测的 5 个维度也取得最高评分。框架同时保留事实命题、证据对齐、异议清单、争点提纲和辩论记录等可检查的结构化中间产物，为人工核验和过程追溯提供支持。

但本研究仍存在一定局限。面对多被告、多罪名和证据关系复杂的案件，PDJ-MACF 的争点组织和信息传递还需要更多检验。真实司法流程中的长期使用，也应交由后续研究评估。下一步笔者将扩展案件结构、补充法律知识约束，并探索由专业人员确认和修正中间产物的人机协同流程。

参考文献：

[1] Feng Y, Li C Y, Ng V. Legal judgment prediction: a survey of the state of the art[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2022: 5461-5469.

[2] Cui J Y, Shen X Y, Wen S C. A survey on legal judgment prediction: datasets, metrics, models and challenges[J]. IEEE Access, 2023, 11: 102050-102071.

[3] 吕由钢, 郝继泰, 王梓涵, 等. 基于判决链的法律判决预测[J]. 计算机研究与发展, 2025, 62(8): 1918-1930.
Lyu Y G, Hao J T, Wang Z H, et al. Legal judgment prediction based on chain of judgment[J]. Journal of Computer Research and Development, 2025, 62(8): 1918-1930.

[4] Luo B F, Feng Y S, Xu J B, et al. Learning to predict charges for criminal cases with legal basis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2017: 2727-2736.

[5] Zhao J, Guan Z Y, Xu C, et al. Charge prediction by constitutive ele-

ments matching of crimes[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2022: 4517-4523.

[6] Zhang H, Dou Z C, Zhu Y T, et al. Contrastive learning for legal judgment prediction[J]. ACM Transactions on Information Systems, 2023, 41(4): 1-25.

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

[8] Huang Q Y, Xia Y H, Long Y F, et al. Prompt4LJP: prompt learning for legal judgment prediction[J]. The Journal of Supercomputing, 2025, 81(2): 420.

[9] Wang X R, Zhang X G, Hoo V, et al. LegalReasoner: a multi-stage framework for legal judgment prediction via large language models and knowledge integration[J]. IEEE Access, 2024, 12: 166843-166854.

[10] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners[C]//Proceedings of the Advances in Neural Information Processing Systems 35. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 22199-22213.

[11] 黄峻, 林飞, 杨静, 等. 生成式AI的大模型提示工程: 方法、现状与展望[J]. 智能科学与技术学报, 2024, 6(2): 115-133.
Huang J, Lin F, Yang J, et al. From prompt engineering to generative artificial intelligence for large models: the state of the art and perspective[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(2): 115-133.

[12] 魏斌. 法律大语言模型的司法应用及其规范[J]. 东方法学, 2024(5): 57-73.
Wei B. Judicial applications and regulation of large legal language models[J]. Oriental Law, 2024(5): 57-73.

[13] Liu J K, Tong Y J, Huang H, et al. Legal fact prediction: the missing piece in legal judgment prediction[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2025: 6345-6360.

[14] Virkar S, Alexopoulos C, Tsekeridou S, et al. A user-centred analysis of decision support requirements in legal informatics[J]. Government Information Quarterly, 2022, 39(3): 101713.

[15] Kort F. Predicting supreme court decisions mathematically: a quantitative analysis of the “right to counsel” cases[J]. American Political Science Review, 1957, 51(1): 1-12.

[16] Nagel S. Applying correlation analysis to case prediction[J]. Texas Law Review, 1963, 42: 1006.

[17] Lawlor R C. What computers can do: analysis and prediction of judicial decisions[J]. American Bar Association Journal, 1963, 49(4): 337-344.

[18] Keown R. Mathematical models for legal prediction[J]. Computer L.J., 1980, 2: 829.

[19] 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.

[20] Katz D M, Bommarito M J, Blackman J. A general approach for predicting the behavior of the Supreme Court of the United States[J]. PLoS One, 2017, 12(4): e0174698.

[21] Sulea O M, Zampieri M, Malmasi S, et al. Exploring the use of text classification in the legal domain[PP]. V1. (2017-10-25)[2026-08-10]. arXiv: arXiv.1710.09306.

[22] Liu C L, Hsieh C D. Exploring phrase-based classification of judicial documents for criminal charges in Chinese[M]//Foundations of Intelligent Systems. Berlin: Springer, 2006: 681-690.

[23] Hu Z K, Li X, Tu C C, et al. Few-shot charge prediction with discriminative legal attributes[C]//Proceedings of the 27th International Conference on Computational Linguistics. Kerrville: Association for Computational Linguistics, 2018: 487-498.

[24] Zeleznikow J, Stranieri A. The split-up system: integrating neural networks and rule-based reasoning in the legal domain[C]//Proceedings of the 5th International Conference on Artificial Intelligence and Law. New York: ACM, 1995: 185-194.

[25] Zhong H X, Guo Z P, Tu C C, et al. Legal judgment prediction via topological learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 3540-3549.

[26] Yue L N, Liu Q, Jin B B, et al. NeurJudge: a circumstance-aware neural framework for legal judgment prediction[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 973-982.

[27] Chalkidis I, Fergadiotis M, Malakasiotis P, et al. LEGAL-BERT: the muppets straight out of law school[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: ACL, 2020: 2898-2904.

[28] Xiao C J, Hu X Y, Liu Z Y, et al. Lawformer: a pre-trained language model for Chinese legal long documents[J]. AI Open, 2021, 2: 79-84.

[29] Zheng L, Guha N, Anderson B R, et al. When does pretraining help?: assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings[C]//Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. New York: ACM, 2021: 159-168.

[30] 丛颖男, 韩林睿, 马佳羽, 等. 基于大语言模型的刑事案件智能判决研究[J]. 计算机科学, 2025, 52(5): 248-259.
Cong Y N, Han L R, Ma J Y, et al. Research on intelligent judgment of criminal cases based on large language models[J]. Computer Science, 2025, 52(5): 248-259.

[31] Jiang C, Yang X L. Legal syllogism prompting: teaching large language models for legal judgment prediction[C]//Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. New York: ACM, 2023: 417-421.

[32] Huang Q Z, Tao M X, Zhang C, et al. Lawyer LLaMA technical report[PP]. V2. (2023-10-14)[2026-08-10]. arXiv: arXiv.2305.15062.

[33] Zhou Z, Shi J X, Song P X, et al. LawGPT: a Chinese legal knowledge-enhanced large language model[PP]. V1. (2024-06-07)[2026-08-10]. arXiv: arXiv.2406.04614.

[34] 迟舜雨, 杜安琪. 生成式人工智能敏捷治理的国际趋势与建构模式[J]. 智能科学与技术学报, 2024, 6(3): 402-412.
Chi S Y, Du A Q. International trends and constructive models of agile governance for generative artificial intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(3): 402-412.

[35] Tran K T, Dao D, Nguyen M D, et al. Multi-agent collaboration mechanisms: a survey of LLMs[PP]. V1. (2025-01-10)[2026-08-10]. arXiv: arXiv.2501.06322.

[36] 奚志恒, 陈文翔, 郭昕, 等. 基于大语言模型的智能体: 发展与未来展望[J]. 中国科学: 信息科学, 2026, 56(2): 485-486.
Xi Z H, Chen W X, Guo X, et al. The rise and potential of large language model based agents: a survey[J]. Scientia Sinica (Informationis), 2026, 56(2): 485-486.

[37] 项凤涛, 罗俊仁, 谷学强, 等. 群视角下的多智能体强化学习方法综述[J]. 智能科学与技术学报, 2023, 5(3): 313-329.
Xiang F T, Luo J R, Gu X Q, et al. Survey on multi-agent reinforce-

ment learning methods from the perspective of population[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(3): 313-329.

[38] Yang S, Yang Z, Liu Y T, et al. From single-agent to multi-agent: a comprehensive review of LLM-based legal agents[J]. AI Agent, 2025, 1(1). DOI: 10.20517/aiagent.2025.06.

[39] Sun J Y, Dai C X, Luo Z Z, et al. LawLuo: a multi-agent collaborative framework for multi-round Chinese legal consultation[PP]. V3. (2024-12-16)[2026-08-10]. arXiv: arXiv.2407.16252.

[40] Cui J X, Ning M N, Li Z J, et al. Chatlaw: a multi-agent legal assistant based on a role-aligned mixture-of-experts architecture[PP]. V3. (2026-06-08)[2026-08-10]. arXiv: arXiv.2306.16092.

[41] He Z T, Cao P F, Wang C H, et al. AgentsCourt: building judicial decision-making agents with court debate simulation and legal knowledge augmentation[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg: ACL, 2024: 9399-9416.

[42] 韩林睿, 宋高捷, 郑日, 等. 基于多智能体协作的刑事案件裁判文书智能生成方法[J]. 计算机科学, 2026, 53(8): 336-356.
Han L R, Song G J, Zheng R, et al. Automated judicial document generation for criminal cases based on multi-agent collaboration[J]. Computer Science, 2026, 53(8): 336-356.

[43] Yuan W K, Song K S, Jiang Z R, et al. A multi-agent framework with legal event logic graph for multi-defendant legal judgment prediction[J]. Information Processing & Management, 2026, 63(1): 104319.

[44] Jiang C, Yang X L. AgentsBench: a multi-agent LLM simulation framework for legal judgment prediction[J]. Systems, 2025, 13(8): 641.

[45] Chen X, Mao M, Li S, et al. Debate-feedback: a multi-agent framework for efficient legal judgment prediction[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). Stroudsburg: ACL, 2025: 462-470.

[46] Yuan W K, Cao J J, Jiang Z R, et al. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg: ACL, 2024: 7577-7597.

[47] Kang Z L, Gong J H, Chen Q X, et al. Multimodal multi-agent empowered legal judgment prediction[C]//Proceedings of the ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2026: 12202-12206.

[48] Sadowski A, Chudziak J A. On verifiable legal reasoning: a multi-agent framework with formalized knowledge representations[C]//Proceedings of the 34th ACM International Conference on Information and Knowledge Management. New York: ACM, 2025: 2535-2545.

[49] Zhang K Y, Li J Q, Wu Y Y, et al. Chinese court simulation with LLM-based agent system[PP]. V1. (2025-08-24)[2026-08-10]. arXiv: arXiv.2508.17322.

[50] Wei J, Wang X Z, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the Advances in Neural Information Processing Systems 35. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 24824-24837.

[51] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: ACL, 2019: 3980-3990.

[52] Prakken H, Sartor G. Law and logic: a review from an argumentation perspective[J]. Artificial Intelligence, 2015, 227: 214-245.

[53] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[54] Chen J, Xiao S, Zhang P, et al. M3-Embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation[PP]. V4. (2024-06-28)[2026-08-10]. arXiv: arXiv.2402.03216.

[55] Mihalcea R, Tarau P. TextRank: bringing order into text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. [S.l.:s.n.], 2004: 404-411.

[56] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.

[57] Lewis M, Liu Y H, Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 7871-7880.

[58] Liu Y, Lapata M. Text summarization with pretrained encoders[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: ACL, 2019: 3728-3738.

[59] Shao Y F, Geng Z C, Liu Y T, et al. CPT: a pre-trained unbalanced transformer for both Chinese language understanding and generation[J]. Science China Information Sciences, 2024, 67(5): 152102.

[60] Yang A, Li A, Yang B, et al. Qwen3 technical report[PP]. V1. (2025-05-14)[2026-08-10]. arXiv: arXiv.2505.09388.

[61] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[62] Wei J, Bosma M, Zhao V, et al. Finetuned language models are zero-shot learners[C]//International Conference on Learning Representations (ICLR). [S.l.:s.n.], 2022.

[63] Hu E J, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models[C]//International Conference on Learning Representations. [S.l.:s.n.], 2022.

[64] Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models[C]//International Conference on Learning Representations (ICLR). [S.l.:s.n.], 2023.

[65] TENCENT-HUNYUAN. Hunyuan-7B[EB]. (2025)[2026-08-10].

[66] Yang A Y, Xiao B, Wang B N, et al. Baichuan 2: open large-scale language models[PP]. V4. (2025-04-17)[2026-08-10]. arXiv: arXiv.2309.10305.

[67] AI@META. Llama 3 model card[EB]. (2024)[2026-08-10].

[68] DeepSeek-AI, Guo D Y, Yang D J, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning[PP]. V2. (2026-01-04)[2026-08-10]. arXiv: arXiv.2501.12948.

[69] Team GLM, Zeng A H, Xu B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[PP]. V2. (2024-07-30)[2026-08-10]. arXiv: arXiv.2406.12793.

[70] Su W H, Yue B Q, Ai Q Y, et al. JuDGE: benchmarking judgment document generation for Chinese legal system[C]//Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2025: 3573-3583.

[71] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Kerrville: Association for Computational Linguistics, 2005: 65-72.

[72] Zhang T, Kishore V, Wu F, et al. BERTScore: evaluating text generation with BERT[C]//International Conference on Learning Representations. [S.l.:s.n.], 2020.

[作者简介]



韩林睿（2000− ），男，教育部哲学社会科学实验室——中国政法大学数据法治实验室助理研究员，中国政法大学数据法学院数据法学博士生，主要研究方向为数据法学、法律人工智能、司法大模型。



郑日（2000− ），女，教育部哲学社会科学实验室——中国政法大学数据法治实验室助理研究员，中国政法大学国家安全学院博士生，主要研究方向为国家安全学、数据法学、法律人工智能。



丛颖男（1985− ），男，中国政法大学商学院副教授，产业经济系副主任，主要研究方向为法商大数据、法律人工智能、区块链、金融科技、监管科技、复杂系统。



李钰鹏（1990− ），男，香港浸会大学互动媒体系副教授，人工智能与社会向善实验室主任，主要研究方向为面向社会的人工智能、网络科学及可信大模型。