

基于法条约束与类案融合的可解释司法判决预测方法

摘要 法律判决预测是推动司法智能化的关键研究领域。然而，当前大型语言模型（LLM）在此任务中面临显著挑战：由于缺乏专门的法律知识强化，LLM 主要依赖其固有知识进行推理，这导致它们在处理复杂法律案件时容易产生“幻觉”，即生成不符合法律事实或逻辑的判决。此外，LLM 对专业法律知识的深度理解和精确应用能力不足，严重制约了其在司法实践中的可靠性和可解释性。为克服这些挑战，本文提出了一种基于法条约束与类案融合的可解释司法判决预测方法。该方法构建了一套系统性的外部法律知识引导机制。通过引入外部法律条文数据库作为法律理论知识参考，并借鉴相似案例数据库构成的司法实践经验，本方法显著提升了 LLM 在法律判决预测任务中的性能和透明度。本研究旨在通过为 LLM 提供明确的法律依据和司法实践参考，系统性地提升判决质量。实验结果表明，该方法在罪名预测和刑期预测任务上的 F1 分数分别达到了 0.7743 和 0.5525。相较于同样尝试结合外部知识库的先进基准模型，其 F1 分数分别提升了 2.46% 和 8.34%。这充分验证了本方法通过有效融合多源异构知识，能够显著增强判决预测的准确性。

关键词 大语言模型；检索增强生成；判决预测；智慧法院；

Explainable Judicial Outcome Prediction: A Legal Provision-Constrained and Case-Based Fusion Framework

Abstract Legal judgment prediction is a critical research area driving judicial intelligence. However, current Large Language Models (LLMs) face significant challenges in this task. Without specialized legal knowledge enhancement, LLMs primarily rely on their inherent knowledge for reasoning, which often leads to “hallucinations”—generating judgments inconsistent with legal facts or logic—when handling complex legal cases. Furthermore, LLMs’ insufficient deep understanding and precise application of professional legal knowledge severely limit their reliability and interpretability in judicial practice. Legal judgment prediction is a critical research area driving judicial intelligence. However, current Large Language Models (LLMs) face significant challenges in this task. Without specialized legal knowledge enhancement, LLMs primarily rely on their inherent knowledge for reasoning, which often leads to “hallucinations”—generating judgments inconsistent with legal facts or logic—when handling complex legal cases. Furthermore, LLMs’ insufficient deep understanding and precise application of professional legal knowledge severely limit their reliability and interpretability in judicial practice. This study aims to systematically improve judgment quality by providing LLMs with clear legal grounds and practical judicial references. Experimental results demonstrate that this method achieved F1 scores of 0.7743 and 0.5525 on the charge prediction and sentence prediction tasks, respectively. Compared to advanced baseline models that also attempted to integrate external knowledge bases, its F1 scores improved by 2.46% and 8.34%, respectively. This fully validates that our method, through the effective fusion of multi-source heterogeneous knowledge, can significantly enhance the accuracy of judgment prediction.

Key words Large Language Models; Retrieval-Augmentation Generation; Judgment Prediction; Smart Courts;

1 引言

随着信息技术飞速发展,“数字法院”建设已成为全球司法现代化的重要趋势。这使得对智能化司法审判解决方案的需求日益迫切,尤其是在刑事案件审判领域。该领域的核心目标是提升司法判决的准确性、一致性和效率¹。在此背景下,法律判决预测(LJP)作为法律人工智能领域的基础性、关键性研究任务,受到学术界和实务界的广泛关注。LJP通过分析案件事实描述,自动预测法院判决结果,辅助法官及其他法律从业者提升案件处理效率。

LJP技术的早期研究主要依赖人工构建的规则系统和传统统计机器学习方法²,例如支持向量机(SVM)³。Sulea等人⁴构建了一种结合多个SVM分类器输出平均概率的集成系统。该模型以案情事实描述和时间跨度信息为输入,能够输出判决结果、法律范围和估算判决日期等信息。Katz等人⁵使用随机森林,从案情描述中提取有效特征,预测美国最高法院的判决结果。然而,这些方法未能有效挖掘深层文本特征。由于其人工设计的特性,这些方法需要大量人力成本,难以广泛应用于其他领域。此外,许多早期深度学习LJP模型如同“黑箱”般运作,其决策过程缺乏透明度和可解释性。模型的不可解释性构成了其在司法实践中推广应用的主要障碍^{6,7}。法官难以干预模型的审判逻辑或理解其预测依据,从而削弱了模型的实用价值。缺乏可解释性还可能引发伦理问题,尤其当模型从历史数据中学习潜在偏见时,可能导致不公正或不一致的判决结果⁸。

尽管大型语言模型(LLM)因其卓越的语言理解能力而备受期待⁹,但在法律领域的直接应用暴露出若干固有缺陷。一个核心问题是LLM倾向于产生“幻觉”(Hallucinations),即生成与客观事实或用户输入不符的内容¹⁰。在法律语境下,这可能表现为援引虚假判例、引言或内部引证,从而导致判决不公,甚至判决错误。研究表明,在处理特定法律查询时,LLM的幻觉率可能高达69%至88%¹¹。这种现象通常源于模型在缺乏可验证法律依据的情

况下,尝试进行推理或生成信息¹²。

针对传统方法在可解释性和处理复杂法律逻辑方面的不足,以及通用LLM在直接应用中面临的幻觉、法律知识基础缺乏和专业推理能力弱等问题,本研究提出了一种基于法条约束与类案融合的可解释司法判决预测方法。该方法通过提取判决核心要素(即犯罪核心要素与证据核心要素),旨在减少判决过程中的干扰因素,提供准确且核心的信息。此外,本研究还引入相似案例,旨在为LLM提供司法实践层面的参考,使其理解法律条文在具体情境下的应用方式,并学习裁判经验。最后,LLM作为核心推理引擎,对这些多源异构信息进行综合分析和推理。它综合考量法律原则、司法解释及类案判例的指导作用,输出结构化的判决结果。本研究的方法无需对整个大型语言模型进行重新训练,它将复杂的LJP任务分解为多个子任务,使得整个推理过程更为透明和模块化。与一些端到端的“黑箱”模型相比,这种设计不仅有助于提升整体预测的鲁棒性,还为理解和调试模型行为提供了便利。

2 相关工作

在数据和算力受限的背景下,LJP的早期探索主要依赖统计学方法。例如,Kort¹³通过多因素复合分析,揭示了美国最高法院处理特定案件时的判决规律。该时期的研究还包括应用专家系统将法律知识转化为计算机可处理的规则¹⁴。然而,这些传统方法的核心局限在于它们对噪声数据的高度敏感性以及对人工规则的过度依赖。法律文本的复杂性和模糊性使得规则制定和特征标注异常困难。简单的数学模型难以捕捉司法实践中复杂多样的非线性影响因素,从而限制了其预测性能和泛化能力¹⁵。

为解决此问题,研究转向了机器学习和文本挖掘技术¹⁶。学者们通过将案件事实作为输入、判决结果作为标签,利用支持向量机(SVM)¹⁷或随机森林(Random Forest)¹⁸等模型,从案情描述中自动学习特征以预测判决。例如,Katz等¹⁹应用随机森林模型,有效提取了影响美国最高法院判决的关键特征。这类方法的优势在于增强了模型对非线性

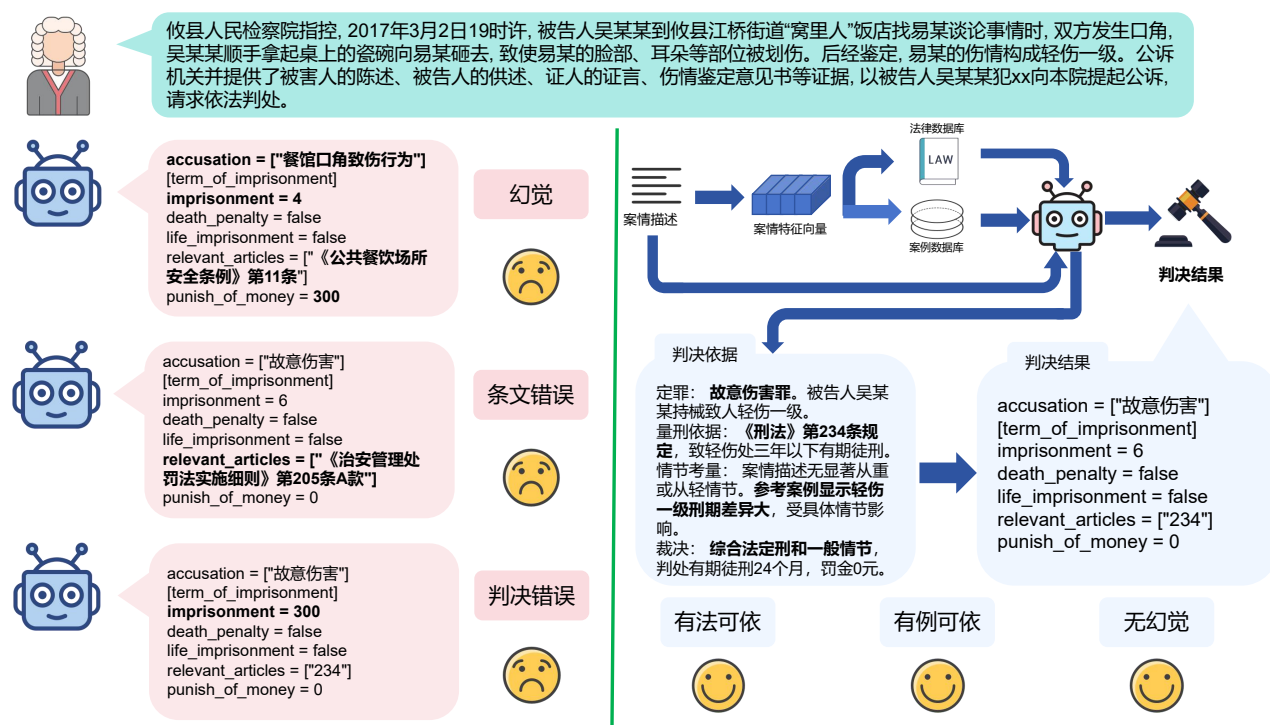


Fig. 1 基于法条约束与类案融合的可解释司法判决预测动机

关系的建模能力, 并初步实现了特征提取的自动化。然而, 其短板也十分明显: 模型性能高度依赖于人工设计的特征工程, 这不仅耗费大量人力, 而且难以挖掘文本背后深层次的语义信息, 导致其难以迁移至更广泛的法律场景。

随着深度学习的兴起, LJP 研究迎来了新的突破。基于深度神经网络的模型能够自动捕捉文本中更复杂、更抽象的特征 ????. Luo 等 ? 提出的模型利用注意力机制, 动态识别并聚焦于事实描述中最关键的部分, 有效关联了案件事实与罪名认定, 显著提升了预测准确性。为了更全面地模拟司法过程, Yue 等 ? 构建了一个情境感知的多任务学习框架 (NeurJudge), 它通过协同法条预测、罪名预测等多个子任务, 使模型学习到任务间的共享信息, 从而提升了主任务的性能。受 BERT 等模型成功的启发, 法律 AI 领域涌现出 Legal-BERT ?????、Lawformer ???? 等在海量法律语料上预训练的模型。它们通过迁移学习, 将从大规模无标注文本中学到的丰富语言知识应用于下游任务, 相较于传统机器学习模型, 获

得了显著的性能提升 ?????。尽管深度学习极大地推动了 LJP 的发展, 但深度学习模型如同“黑箱”般运作, 其决策过程缺乏透明度和可解释性。尤其当数据存在潜在偏见时, 深度学习模型可能直接导致不公正或不一致的判决结果。

LLM 因其强大的复杂上下文推理能力、庞大的预训练知识库和良好的可解释性, 被期望成为下一代 LJP 的新范式 ???。当前, 以 LLM 为核心的技术范式已成为研究热点。其强大的常识推理和零样本/少样本学习能力为 LJP 带来了新的可能性 ?????。将 LLM 直接应用于严肃的法律领域仍面临严峻挑战。未经充分优化的通用 LLM 在处理法律问题时, 容易出现“幻觉” (Hallucination) ???。这可能表现为捏造不存在的法律条文、引用错误的案例, 或给出超出法定范围的量刑建议 ?。为解决 LLM 的幻觉问题, 主要优化思路分为两条路径。一是通过提示工程 (Prompt Engineering) 引导模型, 例如利用“思维链” (Chain of Thought) ????? 或司法三段论 ???? 来规范其推理逻辑。然而, 为特定任务精心

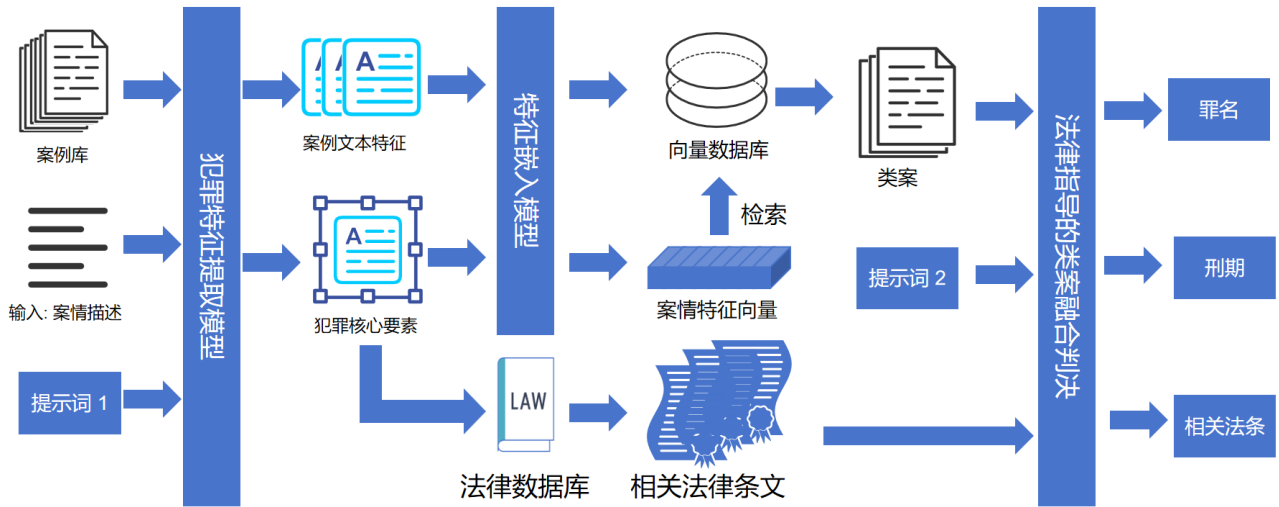


Fig. 2 基于法条约束与类案融合的可解释司法判决预测方法

设计的提示词（Prompt）往往缺乏通用性，难以直接迁移至其他法律任务，这限制了其可扩展性；二是以法律专业数据对 LLM 进行微调，例如 Lawyer LLaMA ??，使其具备更强的“法律素养”。尽管微调是提升专业性的有效途径 ???，但 LLM 微调存在以下问题：首先，微调的成功依赖于高质量标注数据和大量计算资源。其次，微调后的 LLM 知识体系是静态的，其知识停留在训练数据集的时间点，无法实时跟进法律法规的更新和司法解释的演进 ??。最后，微调过程还可能导致模型对其通用知识的“灾难性遗忘” ?，损害其基础推理能力。因此，本研究提出一种基于 LLM 并由法律引导的案例融合方法。该方法将提示词工程与检索增强技术相结合，以实现司法智能审判的可信判决。该方法通过特定的提示词提取案例描述的关键判决核心要素，从而保留司法审判理论的核心要素并去除冗余噪声。此外，通过引入案例数据库和法律条文数据库，为 LLM 提供精确的罪名定义和司法实践案例，以综合理论知识与实践经验。最后，利用多源异构信息进行综合分析 & 推理，做出最终的司法判决。该方法无需微调 LLM，并且可以通过引入新的法律条文和案例，提供最新的法理知识，从而避免 LLM 因缺少相关知识而产生幻觉。

3 方法

3.1 总体流程

本研究提出的司法判决预测方法，其核心在于构建一个能有效整合案件事实、法律法规及类案判例的智能推理框架。整体流程如图??所示。首先，系统接收待判决案件的详细事实描述（ C ），并利用预训练的 LLM 进行初步语义分析和关键信息抽取。LLM 从复杂的案情描述中识别并初步推断出罪名类别、犯罪构成要件（包括主体、主观、客体、客观）及证据特征等核心法律要素，形成结构化的犯罪核心要素表示（ F ）。其次，系统基于抽取的犯罪核心要素 F ，并行地从法律条文数据库和案例数据库中检索相关信息。一方面，从权威的法律条文数据库中检索出与案件特征 F 高度相关的法律法规条款集合（ L ）。另一方面，从海量的历史案例数据库中智能检索出与当前案件在罪名构成、事实情节和证据方面最为相似的判例集合（ S ）。最后，将原始案件事实描述 C 、初步提取的犯罪核心要素 F 、检索到的相关法律条文 L 以及筛选出的相似判例 S 作为共同的上下文信息，输入至 LLM。LLM 将进行综合分析 & 推理，输出最终的判决结果（ J ）。

该模型作为核心推理引擎，通过整合这些多源

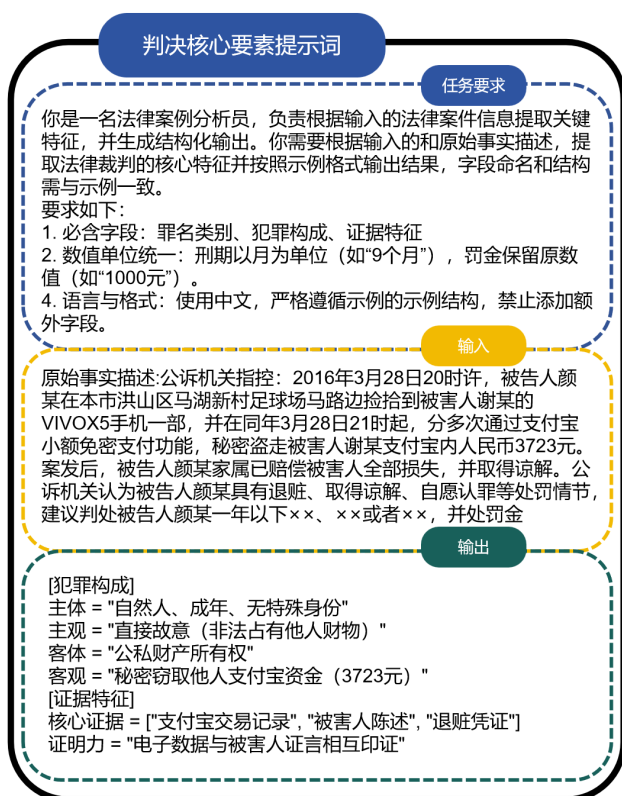


Fig. 3 判决核心要素提示词

异构信息，综合考量法律原则、司法解释及类案判例的指导作用。这不仅有效弥补了 LLM 在法律专业知识和复杂逻辑推理方面的固有局限性，而且显著增强了判决预测的专业性、准确性和可解释性。

3.2 判决核心要素提取

在大陆法系的刑法理论中，一项行为若要被判定为犯罪，其客观事实必须符合刑法分则具体条文所规定的全部构成要件。此外，法官在进行司法判决时，还需要综合考量检察机关提供的证据、犯罪嫌疑人的行为动机以及造成的事实性后果等因素。然而，传统方法未能识别法律要素间的逻辑依赖，导致判决预测缺乏法律推理的前提，最终可能造成判决结果不可信甚至错误^{???}。为解决此问题，本研究提出了判决核心要素提取方法。判决核心要素 F 由犯罪核心要素和证据核心要素构成。其中，犯罪核心要素包括犯罪主体、犯罪客体、犯罪主观和犯罪客观四个方面。证据核心要素由证据和证明力组

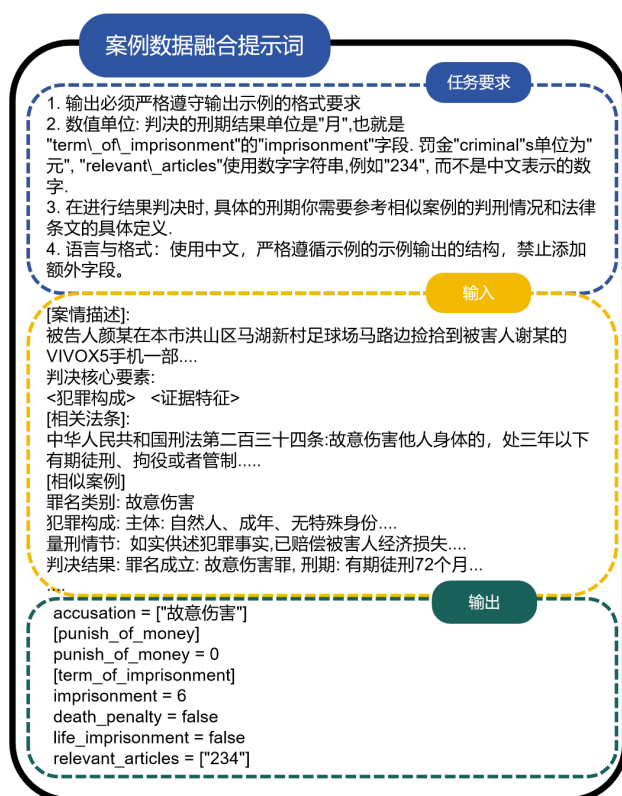


Fig. 4 案例数据融合提示词

成。证据指检察机关提供的直接证据，而证明力是证据的可信程度，表明证据之间相互印证、逻辑连贯，能够充分证明犯罪行为。该证据特征通过 LLM 定制提取。明确识别判决核心要素有助于去除冗余信息和噪声，为后续的法律条文检索与相似案例检索提供精确、简要的数据，并为逻辑关联分析和推理奠定基础。判决核心要素提示词从事实描述 C 中提取，其示例详见图??所示。明确提取犯罪核心要素和证据核心要素，不仅优化了案件结构化理解的基础，也为后续的法律条文检索和判决文本生成提供了更为精准和逻辑严谨的输入。这从而提升了整个智能判决系统的透明度、准确性以及与人类司法实践的一致性。

3.3 相关法律条文检索

为解决 LLM 在法律知识方面的不足和潜在的“幻觉”问题，相关法律条文检索通过引入真实的法律条文，为 LLM 提供准确、权威的法律规范依据。

这确保了模型在进行判决预测时，能够明确犯罪的构成要件、法律定义以及量刑的法定边界。首先，本研究构建了一个包含各类法律法规的法律条文数据库 (DB_{law})。数据库中的每条法律条文均通过文本嵌入模型转换为高维向量表示，并存储于向量数据库中，以支持高效检索。当处理新案件时，从案件事实描述 C 中提取的犯罪核心要素 F 同样被转换为查询向量 F_e 。随后，系统采用近似最近邻搜索算法，通过计算查询向量与数据库中法律条文向量之间的相似度，检索出与案件最为相关的 k 条法律条文，形成集合 L 。

$$L = (l_1, l_2, \dots, l_k) = \text{Topk}(\text{Sim}(F_e, DB_{law})), \quad (1)$$

其中， Topk 表示取相似度最高的 k 个结果， embed 表示文本嵌入函数， Sim 是相似度计算函数。

数据库中的每条法律条文均通过文本嵌入模型转换为高维向量表示，并存入向量数据库以支持高效检索。当处理新的案件时，从案件事实描述 C 中提取的犯罪核心要素 F 同样被转换为查询向量 F_e 。随后，系统采用近似最近邻搜索算法，通过计算查询向量与数据库中法律条文向量之间的相似度，检索出与案件最为相关的 k 条法律条文，形成集合 L 。该过程可表示为：

3.4 相似案例检索

为解决仅凭法条和案例描述难以覆盖所有复杂情况，并促进“同案同判”的问题，本研究提出了相似案例检索方法。该方法旨在从历史判例中寻找与当前待审案件在核心特征上相似的案例，为 LLM 提供司法实践层面的参考。相似案例检索能使 LLM 理解法律条文在具体情境下的应用方式，学习既往判决中蕴含的裁判经验和量刑酌情考量。

首先，本研究构建了一个结构化的案例数据库。该数据库中的每个案例都包含详细字段，例如“罪名类别”、“犯罪构成”、“量刑情节”、“证据特征”、“法律适用”、“裁判逻辑”和“判决结果”。对于案例中的关键文本字段，特别是“罪名类别”、“犯罪构成”和“证据特征”，采用文本嵌入模型将其转换为向量表示，并构建相应的向量索引。当需要检索相似案例时，针对当前案件提取的犯罪核心要素 F 中的对应

字段（令检索字段集合为 $R = \{\text{罪名, 构成, 证据}\}$ ），即 F_i （其中 $i \in R$ ），分别将其通过相同的文本嵌入模型转换为查询向量 F_{i_e} 。接着，对每个查询向量，在案例数据库对应字段的向量索引中，利用近似最近邻搜索算法和相似度计算，各自独立检索出 m 个最相似的案例。为获得与当前案件整体最为匹配的案例，需要对上述各字段检索出的候选案例进行综合评估。对于每个候选案例 s_j ，计算其与当前案件在 R 中所有字段上的平均相似度

$$\text{AvgSim}(s_j) = \frac{1}{|R|} \sum_{i \in R} \text{Sim}(F_{i_e}, \text{embed}(s_j, i)), \quad (2)$$

其中， $\text{embed}(s_j, i)$ 表示候选案例 s_j 在特定字段 i 上的嵌入向量。最后，选取平均相似度最高的 n 个案例 ($n < m$) 作为最终的相似案例集合 S

$$S = (s_1, s_2, \dots, s_n) = \text{Topn}(\text{AvgSim}(s_j)) \quad (3)$$

通过引入与当前案件高度相似的历史判例，为 LLM 提供了宝贵的经验性知识。这不仅有助于模型更准确地把握特定罪名的构成要件和量刑尺度，还能使判决建议更符合司法实践，从而增强判决结果的合理性和可接受性。

3.5 法律约束的类案融合判决

为整合前述模块获取的全部信息，有效融合多源异构信息并进行复杂法律推理，法律约束的类案融合判决将原始案件事实描述 C 、LLM 从 C 中提取的结构化犯罪核心要素 F 、从法律条文数据库中检索到的相关法律条文集合 L ，以及从案例数据库中检索到的相似案例集合 S ，共同组织成一个全面的上下文——案例数据融合提示词，其示例详见图??所示。此提示被输入到预训练的 LLM 中。LLM 利用其强大的自然语言理解、知识整合和逻辑推理能力，对这些输入信息进行深度分析和融合。模型在推理过程中，会考量法律条文 L 的规定（作为法律依据），并参考相似案例 S 中的裁判思路和判决结果（作为实践经验）。最终，LLM 生成结构化的判决结果 J ，其内容通常包括建议的罪名、刑期、是否适用死刑或无期徒刑、相关的法律条文编号以及可能的罚金等。

$$J = \text{LLM}(C|F, S, L), \quad (4)$$

罪名						相关法条					
CAIL Small			CAIL			CAIL Small			CAIL		
类别	数量	占比	类别	数量	占比	类别	数量	占比	类别	数量	占比
盗窃	5330	19.59%	盗窃	48438	20.36%	133	6122	22.00%	133	59638	24.10%
危险驾驶	4166	15.32%	危险驾驶	40309	16.94%	264	5227	18.78%	264	48023	19.41%
故意伤害	2712	9.97%	故意伤害	24625	10.35%	234	2556	9.18%	234	23901	9.66%
贩毒	2104	7.73%	交通事故	19457	8.18%	347	2074	7.45%	347	18694	7.56%
交通事故	1980	7.28%	贩毒	18889	7.94%	354	1001	3.60%	266	7834	3.17%
容留吸毒	1024	3.76%	诈骗	7273	3.06%	266	887	3.19%	354	6969	2.82%
其他	9886	36.34%	其他	78926	33.17%	其他	9965	35.80%	其他	82365	33.29%

Table 1 CAIL 和 CAIL Small 的实验数据分布

其中, $C|F, S, L$ 表示以 F, S, L 为条件上下文信息, 结合原始描述 C 进行判决。

法律约束的类案融合判决通过整合多源数据和 LLM 的综合推理, 实现了对案件事实、法律规范和司法判例的有效融合。它不仅提升了判决预测的准确性和专业性, 而且通过结合明确的法律条文和相似案例, 增强了判决结果的可解释性和说服力。

4 实验

4.1 实验数据

本实验采用开源的 CAIL 2018 数据集[?]。该数据集包含 1,927,685 个案例, 涵盖 202 种刑事罪名和 183 条刑法法规。其中, 训练集有 1,927,685 条数据, 测试集有 216,829 条数据。为验证本文提出方法的有效性, 本研究从 CAIL 2018[?] 中按罪名分层采样 10% 的数据, 构建了 CAIL Small 数据集, 用于消融实验和分析实验。CAIL Small 和 CAIL 的数据分布如表 ?? 所示。数据集中, 多被告案件的法条分布常呈现长尾现象。如图 ?? 所示, 各法条在判决结果中的出现频率及占比存在较大差异。在测试数据的法条分布中, 占比最高的五种罪行是盗窃、危险驾驶、故意伤害和交通肇事, 其占比分别为 20.63%、17.17%、10.49%、8.29% 和 7.33%。占比最高的十种罪行占总测试数据的 73.58%, 而占比最少的 100 种罪行仅占总数据的 1.83%。如图 ?? 所示, 法条数据

分布同样呈现长尾趋势。占比最高的十种相关法条占测试数据的 0.73%, 而占比最低的 100 种法条仅占 2.44%。

4.2 实验设置

本研究利用 CAIL 2018[?] 的训练集构建案例数据库。为在控制数据库规模的同时兼顾检索效率和案例代表性, 本研究从训练集中每种罪名最多抽取 100 条数据, 共构建 1,676 个案例。法条数据库使用中国刑法典作为文本数据。案例数据库和法条数据库的文本嵌入模型使用 BAAI/BGE-m3 模型[?]。向量数据库采用 Milvus^{??}, 相似度函数为内积 (Inner Product, IP), 向量索引方法为倒排索引 (Inverted File, IVF), 聚类中心设为 200 个。搜索算法采用暴力搜索 (Flat Search), 即在每个聚类内部使用相似度函数进行比较。要素提取模型采用 Qwen-Turbo 模型, 判决模型使用 Qwen-Plus 模型[?]。CAIL 2018 法律判决预测中的罪名和刑期预测两项子任务, 均属于多标签多分类问题[?]。

4.3 评价指标

本研究采用精确率 (Precision, P)、召回率 (Recall, R) 和 F1 分数三项指标来衡量模型的预测效果。

$$P = \frac{\sum_{i=1}^n TP_i}{\sum_{i=1}^n TP_i + \sum_{i=1}^n FP_i} \quad (5)$$

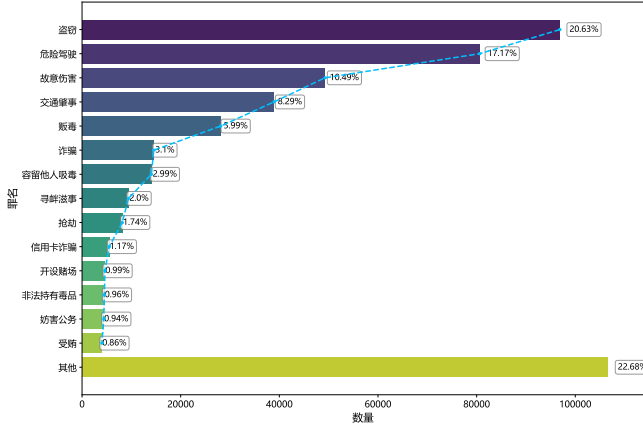


Fig. 5 测试数据的罪名分布

$$R = \frac{\sum_{i=1}^n TP_i}{\sum_{i=1}^n TP_i + \sum_{i=1}^n FN_i} \quad (6)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (7)$$

其中, i 为分类任务中类别的种类; TP_i 为 True Positive, 指被正确地划分为类别 i 的样本个数; FP_i 为 False Positive, 指实际为其他类但被分类器划分为 i 类的样本数; FN_i 为 False Negative, 指实际为 i 类, 但是被分类器划分错误的样本数。

4.4 基准模型

为检验基座 LLM 的性能、本研究提出的“基于法条约束与类案融合的可解释司法判决预测方法”的有效性, 以及其与基于域外语料训练的法律大模型相比的优劣, 本研究在 CAIL 2018 数据集上进行了罪名和刑期两个子任务的对比实验, 旨在比较本文模型与四种基线模型的性能差异。基线模型包括基于词嵌入的深度学习模型 MTL-Fusion[?]、中国司法长文本文书预训练模型 Lawformer[?]、预训练模型 BERT[?], 以及结合司法数据微调和 RAG 技术的 LawChatGLM 模型[?]。

4.5 实验结果与分析

表??的实验结果清晰揭示了不同技术路径在司法判决预测任务上的性能差异。从传统模型 MTL-Fusion 和 Lawformer, 到基于预训练的 BERT, 再到结合法律知识增强的 LawChatGLM, 模型在罪

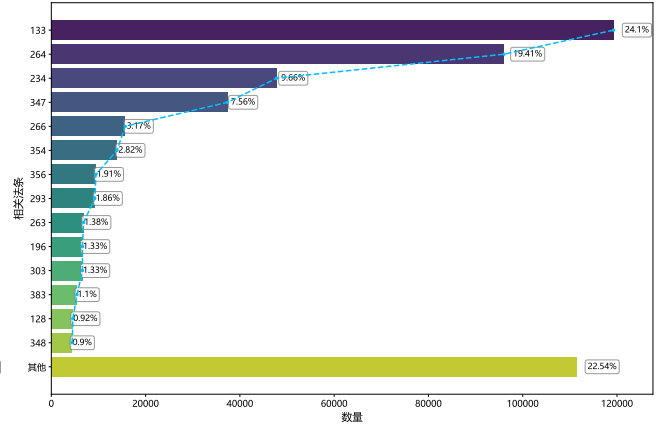


Fig. 6 测试数据的法条数据分布

名和刑期预测上的性能呈现稳步提升趋势。这证明了更强的语义理解能力和外部知识引入是提升 LJP 性能的关键。然而, 即使是表现最强的基准模型 LawChatGLM, 尽管通过检索法律条文提升了罪名预测的准确性, 但在更需酌情裁量的刑期预测上仍有较大提升空间 (F1 分数为 0.4691)。这表明仅依赖成文法条作为外部知识源尚不足以完全捕捉司法实践的复杂性。

本文提出的方法在所有评价指标上均取得了最优表现, 尤其在刑期预测任务上实现了关键突破, F1 分数达到 0.5525, 相较于 LawChatGLM 提升了 8.34%。这一显著优势的核心原因在于我们创新的“法条约束与类案融合的可解释司法判决预测方法”。该框架不仅通过检索法律条文为罪名判定提供了权威的法律依据, 更关键的是, 通过引入“相似案例检索”模块, 为模型提供了来自真实司法实践的量刑参考。这些相似判例中蕴含的裁判经验和量刑逻辑, 有效弥补了成文法条在具体刑期裁量上的模糊性, 使得模型的预测更贴近真实的司法裁判思维。实验结果有力证明, 将 LLM 的推理能力与法律条文的规范性指引、相似案例的实践性参考进行深度融合, 是构建高精度、高可靠性智能司法判决系统的有效路径。

模型	罪名			刑期		
	精确率	召回率	F1 分数	精确率	召回率	F1 分数
MTL-Fusion	0.6861	0.6911	0.6886	0.3512	0.3567	0.3539
Lawformer	0.6927	0.7082	0.7004	0.3581	0.3629	0.3605
BERT	0.7011	0.7178	0.7094	0.4311	0.4308	0.4309
LawChatGLM	0.7517	0.7478	0.7497	0.4712	0.4671	0.4691
Ours	0.7797	0.7689	0.7743	0.5578	0.566	0.5525

Table 2 法律判决预测结果的对比

4.6 消融实验

为验证本研究方法的有效性，我们对 CAIL Small 数据集的三个子任务进行了消融实验，结果如表 ?? 所示。消融实验的重点在于评估法律条文检索模块和相似案例检索模块的作用，因此分别对这两个模块进行了实验。

法律条文检索消融实验。为验证法律条文检索模块的有效性，我们在 CAIL Small 数据集中进行了消融实验。具体而言，在 LLM 进行多源信息融合推理时，我们移除了相关法律条文信息作为输入，但保留了三个相似案例作为参考。实验结果显示，在不使用法律条文信息后，相关法条预测的精确率下降了 8.1%，召回率下降了 5.8%，F1 分数下降了 6.9%。此外，其他任务的性能也有一定程度的降低，这表明法律条文检索模块为 LLM 提供了重要的法律理论支持。

相似案例检索消融实验。为验证相似案例检索模块的有效性，我们同样在 CAIL Small 数据集中进行了消融实验。具体而言，在 LLM 进行多源信息融

合推理判决时，我们移除了相似案例库作为信息源，但保留了三个相关法律条文作为提示。根据表 ?? 的结果，在不使用相似案例库后，刑期预测的 F1 分数下降了 0.7%，其余任务的性能也有所下降。这说明相似案例检索模块为 LLM 提供了关键的司法实践经验参考。

4.7 分析实验

为确定最佳的相似案例检索和相关法律条检索数量，我们在 CAIL Small 数据集上进行了分析实验。具体而言，我们对比了不同数量的相似案例和相关法律条的 Top-K 参数对模型性能（F1 分数）的影响。实验结果如图 ?? 所示。

对于不同检索数量的相关法律条，模型的整体 F1 分数在不同的 Top-K 值下保持稳定。这表明在检索相关法律条文时，模型的性能在此范围内对 Top-K 值不敏感。其原因在于法律条文通常明确且每个案件通常仅涉及一条或少数几条相关法条，因此即使 Top-K 值较小，相关法条检索也能提供足够的信息以进行准确判决。

方法		罪名			刑期			相关法条		
法条	类案	精确率	召回率	F1 分数	精确率	召回率	F1 分数	精确率	召回率	F1 分数
✓	✓	0.698	0.669	0.683	0.424	0.419	0.421	0.715	0.673	0.693
	✓	0.697	0.669	0.683	0.424	0.414	0.418	0.634	0.615	0.624
✓		0.695	0.671	0.683	0.424	0.405	0.411	0.655	0.635	0.645
		0.684	0.670	0.677	0.364	0.387	0.349	0.634	0.615	0.624

Table 3 消融实验

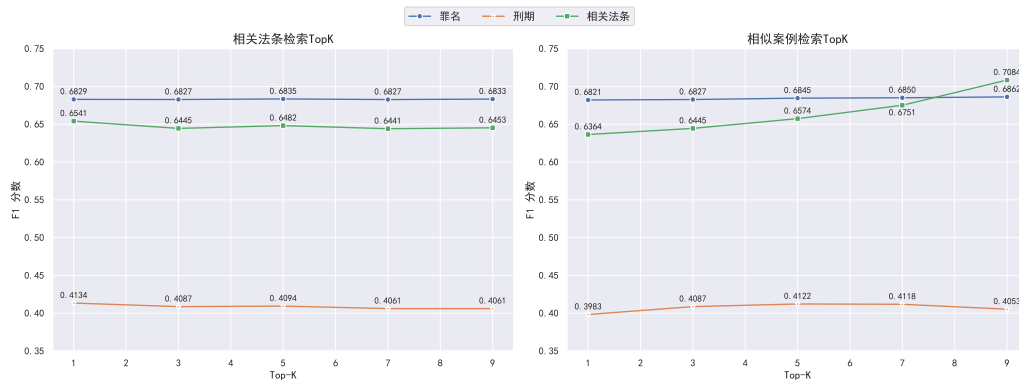


Fig. 7 超参数 TopK 对模型性能的影响

对于不同数量的相似案例参考，模型在罪名预测和刑期预测上的 F1 分数相对稳定，但在相关法条预测上，F1 分数随 Top-K 值的增加而逐渐提升。这表明在相似案例检索中，更多案例可提供更丰富的司法实践经验，从而提高模型的判决准确性。特别是对于法律定义边界模糊的罪名，更多的司法实践参考能帮助 LLM 更好地理解不同法律条文在实际应用中的区别，并更合理地应用法律条文。

4.8 案例研究

本研究提出的方法将以一个具体的盗窃案件为例进行说明。首先，系统接收一段原始的案件事实描述，例如：“公诉机关指控，2016年3月28日20时许，被告人颜某在…盗走被害人谢某支付宝内人民币3723元”。接着，系统利用 LLM 对该文本进行“判决核心要素提取”，生成结构化的犯罪特征 F ：

主体 = “自然人、成年、无特殊身份”
 主观 = “直接故意，非法占有他人财物”
 客体 = “公私财产所有权”
 客观 = “秘密窃取他人支付宝资金 3723 元”
 核心证据 = [“支付宝交易记录”，“被害人陈述”，“退赃凭证”]
 证明力 = “电子数据与被害人证言相互印证”

随后，进入本方法的核心——双重知识检索阶段。系统会将提取出的核心要素 F 通过文本嵌入模

型 BAAI/BGE-m3 转换为高维查询向量。该查询向量被用于在两条并行路径上执行检索：第一条路径中，系统在预先向量化的法律数据库中进行近似最近邻（ANN）搜索，通过计算内积相似度，找出与案件特征最匹配的法律条文 L ：

中华人民共和国刑法第二百三十四条：故意伤害他人身体的，处三年以下有期徒刑、拘役或者管制……

第二条路径中，系统同样利用该查询向量，在向量化的案例数据库中检索相似判例 S ：

罪名类别：故意伤害；
 犯罪构成：主体：自然人、成年…；
 客体：公私财产所有权；
 量刑情节：如实供述犯罪事实，已赔偿被害人经济损失…；
 判决结果：
 罪名成立：故意伤害罪；
 刑期：有期徒刑 72 个月…；

此处的案例检索更为精细，它会综合考量罪名、犯罪构成、证据特征等多个维度的相似度，以确保筛选出的历史判例与当前案件具有高度可比性。

最后，将原始案情 (C)、提取的核心要素 F 、检索到的法律条文 L 及相似案例 S 共同作为输入，送入核心 LLM 推理引擎，由其进行综合分析推理，

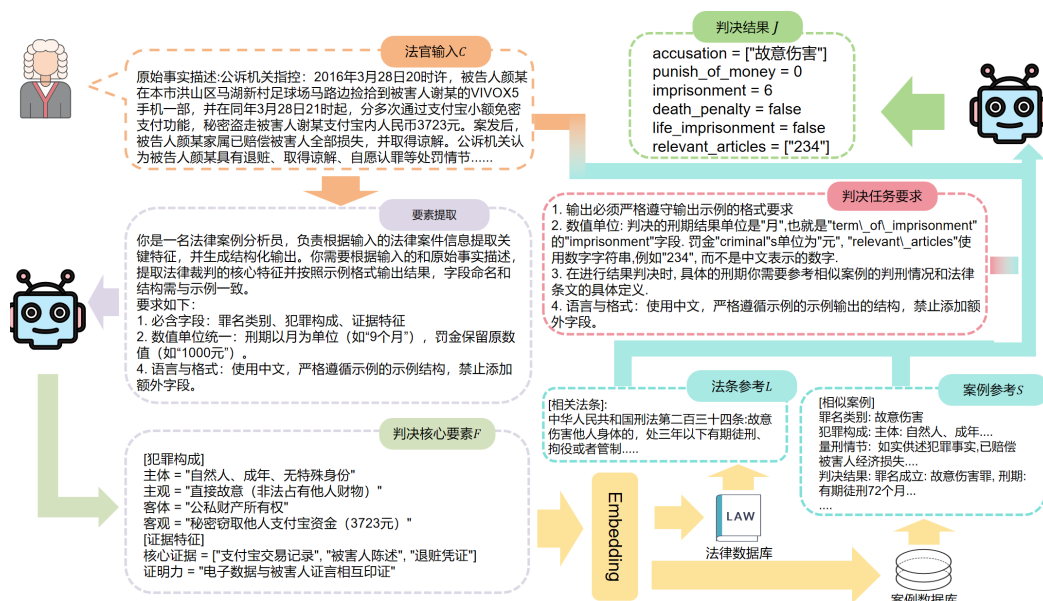


Fig. 8 案例说明

最终生成一个的结构化判决结果 (J):

```
accusation = [“盗窃罪”];
relevant_articles = [“264”];
punish_of_money = 3723;
imprisonment = 7;
death_penalty = false;
life_imprisonment = false;
```

5 结论

这项研究旨在解决当前法律判决预测领域的两大难题: 传统模型因为“黑箱”操作缺乏解释性, 而大型语言模型 (LLM) 则存在“幻觉”和专业知识不足的问题, 导致难以直接应用。为此, 我们提出了一种创新的、可解释的司法判决预测方法, 该方法融合了法条约束和类案参考。我们的方法将复杂的判决任务拆解成几个清晰的步骤: 首先, 利用 LLM 精确提取判决的核心要素。接着, 我们同步从外部知识库中检索权威的法律条文和高度相似的司法判例。最后, LLM 会将这些来自不同来源的信息整合起来, 生成最终的判决结果。实验结果充分证明了我们方法的有效性。在罪名和刑期预测任务上, 我们

的方法表现出色, F1 分数分别高达 0.7743 和 0.5525, 超越了所有对比模型。与那些同样整合了外部知识库的先进模型相比, 我们的方法在罪名预测的 F1 分数上提升了约 3.3%, 而在高度依赖司法实践经验的刑期预测上, 更是实现了 17.8% 的显著飞跃。这项性能的巨大提升, 主要归功于我们框架的两大核心创新: 一是引入了法律条文数据库, 为模型的推理提供了坚实的法律依据, 有效缓解了“幻觉”问题; 二是引入了相似案例数据库, 让模型能够学习和借鉴海量的司法实践经验, 特别是在需要酌情裁量的刑期预测任务上展现出卓越的性能。此外, 我们的消融实验也清楚地表明, 本研究提出的各个模块都为方法的有效性贡献良多, 验证了其设计的合理性。这些成果共同强调了将大型语言模型的强大语言能力、法律条文的规范性以及司法判例的实践性深度融合, 是构建精准可靠的智能司法判决系统的关键所在。

参考文献

- ALETRAS N, TSARAPATSANIS D, PREOTIUC-PIETRO D, et al. Predicting judicial decisions of the european court of human rights: A natural language processing perspective [J/OL]. PeerJ computer science, 2016, 2: 93. DOI: [10.7717/](https://doi.org/10.7717/peerj.computer-science.2016.2.93)

- peerj-cs.93.
- 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): 0174698.
- KEOWN R. Mathematical models for legal prediction, 2 computer l. j. 829 (1980)[J]. UIC John Marshall Journal of Information Technology & Privacy Law, 1980, 2(1): 29.
- BOELLA G, DI CARO L, HUMPHREYS L. Using classification to support legal knowledge engineers in the eunomos legal document management system[C]//Fifth International Workshop on Juris-informatics. 2011.
- KIM M, XU Y, GOEBEL R. Legal question answering using ranking svm and syntactic/semantic similarity[C]//New Frontiers in Artificial Intelligence. 2015: 244-258.
- SULEA O, ZAMPIERI M, MALMASI S, et al. Exploring the use of text classification in the legal domain[C/OL]//Proc of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts Co-located with the 16th Int Conf on Artificial Intelligence and Law. 2017. <https://ceur-ws.org/Vol-2143/paper5.pdf>.
- LING W, YOGATAMA D, DYER C, et al. Program induction by rationale generation: Learning to solve and explain algebraic word problems[C]//Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017: 158-167.
- MA C S. Law towards a digital society[M]. Beijing: Law Press China, 2021.
- NYE M, ANDREASSEN A, ARI G, et al. Show your work: Scratchpads for intermediate computation with language models[A]. 2021.
- LUO B, FENG Y, XU J, et al. Learning to predict charges for criminal cases with legal basis[C]//Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 2727-2736.
- LV Y, WANG Z, REN Z, et al. Improving legal judgment prediction through reinforced criminal element extraction[J]. Information Processing & Management, 2022, 59(1): 102780.
- JIANG C, YANG X. Legal syllogism prompting: Teaching large language models for legal judgment prediction[C]//Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. 2023: 417-421.
- 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.
- ZHENG L, GUHA N, ANDERSON B, et al. When does pre-training 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. 2021: 159-168.
- DAHL M, MAGESH V, SUZGUN M, et al. Large legal fictions: Profiling legal hallucinations in large language models[J/OL]. Journal of Legal Analysis, 2024, 16(1): 64-93. <http://dx.doi.org/10.1093/jla/laae003>.
- ZHONG H, WANG Y, TU C, et al. Iteratively questioning and answering for interpretable legal judgment prediction[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 1250-1257.
- ZHONG H, XIAO C, TU C, et al. Jec-qa: a legal-domain question answering dataset[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 9701-9708.
- KORT F. Predicting supreme court decisions mathematically: A quantitative analysis of the “right to counsel” cases[J/OL]. The American Political Science Review, 1957, 51(1): 1-12. DOI: [10.2307/1951767](https://doi.org/10.2307/1951767).
- SUSSKIND R. Expert systems in law: A jurisprudential approach to artificial intelligence and legal reasoning[J]. The Modern Law Review, 1986, 49(2): 168-194.
- DENG W, PEI J, KONG K, et al. Syllogistic reasoning for legal judgment analysis[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 13997-14009.
- DONG Q, NIU S. Legal judgment prediction via relational learning[C]//Proc of the 44th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2021: 983-992.

- CHEN Y, LIU Y, HO W. A text mining approach to assist the general public in the retrieval of legal documents[J]. Journal of the American Society for Information Science and Technology, 2013, 64(2): 280-290.
- GONÇALVES T, QUARESMA P. Evaluating preprocessing techniques in a text classification problem[C]//Proceedings of the Conference of the Brazilian Computer Society. 2005.
- KIANMEHR K, ALHAJJ R. Crime hot-spots prediction using support vector machine[C]//IEEE International Conference on Computer Systems and Applications. 2006: 952-959.
- CHENG F D. Legal consultation data and corpus from china law network: Replication data for design and research of text classification system[EB/OL]. 2025. <https://doi.org/10.18170/DVN/OLO4G8>.
- FENG Y, LI C, VINCENT N. Legal judgment prediction via event extraction with constraints[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 648-664.
- JIANG X, YE H, LUO Z, et al. Interpretable rationale augmented charge prediction system[C]//Proc of the 27th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2018: 146-151.
- WANG P, FAN Y, NIU S, et al. Hierarchical matching network for crime classification[C]//Proc of the 42nd int ACM SIGIR Conf Research and Development in Information Retrieval. New York: ACM, 2019: 325-334.
- HUANG N, HE J, SUN J, et al. Improved lawformer-based approach for forecasting crimes[J]. Journal of Beijing University of Technology, 2019, 45(8): 742-748.
- YUE L, LIU Q, JIN B, et al. Neurjudge: A circumstance-aware neural framework for legal judgment prediction[C]//Proc of the 44th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2021: 973-982.
- LIU Y, OTT M, GOYAL N, et al. A robustly optimized bert pre-training approach with post-training[C]//China National Conference on Chinese Computational Linguistics. Cham: Springer International Publishing, 2021: 471-484.
- CHALKIDIS I, FERGADIOTIS M, MALAKASIOTIS P, et al. Legal-bert: Preparing the muppets for court[C]//Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 2898-2904.
- DEEPA M. Bidirectional encoder representations from transformers (bert) language model for sentiment analysis task [J]. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021, 12(7): 1708-1721.
- DEVLIN J, CHANG M, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 4171-4186.
- FAN A M, WANG Y C. Multi task intelligent legal judgment method based on bert model[J]. Microelectronics & Computer, 2022, 39(9): 107-114.
- XIAO C, HU X, LIU Z, et al. Lawformer: A pre-trained language model for chinese legal long documents[J/OL]. AI Open, 2021, 2: 79-84. DOI: [10.1016/j.aiopen.2021.06.003](https://doi.org/10.1016/j.aiopen.2021.06.003).
- DU Z, QIAN Y, LIU X, et al. Glm: General language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 320-335.
- FEI Z, SHEN X, ZHU D, et al. Lawbench: Benchmarking legal knowledge of large language models[A]. 2023.
- OANA-MARIA C, TIM R, THOMAS L, et al. e-snli: Natural language inference with natural language explanations[C]//Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018: 9560-9572.
- CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese bert[J/OL]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514. DOI: [10.1109/TASLP.2021.3124365](https://doi.org/10.1109/TASLP.2021.3124365).
- HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for nlp[C]//Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019: 2790-2799.
- HU Z, LI X, TU C, et al. Few-shot charge prediction with discriminative legal attributes[C]//Proc of the 27th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2018: 487-498.

- ZHANG H, DOU Z, ZHU Y, et al. Contrastive learning for legal judgment prediction[J]. *ACM Transactions on Information Systems*, 2023, 41(4): 1-25.
- WU Y, KUANG K, ZHANG Y, et al. De-biased court's view generation with causality[C]//*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2020: 763-780.
- YANG A, XIAO B, WANG B, et al. Baichuan 2: Open large-scale language models[A]. 2023.
- YAO H, CHEN Y, YE Q, et al. Refining language models with compositional explanations[C]//*Proc of the 34th Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2020: 8954-8967.
- 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.
- HUANG J, CHANG K. Towards reasoning in large language models: A survey[A]. 2022.
- SHU W, LI R, SUN T, et al. Large-scale language modeling: Principles, implementation and development [J/OL]. *Computer Research and Development*, 2024, 61(2): 351-361. DOI: [10.7544/issn1000-1239.202330303](https://doi.org/10.7544/issn1000-1239.202330303).
- WEN S, QIAN L, HU H, et al. Review of research progress on question-answering techniques based on large language models[J]. *Data Analysis and Knowledge Discovery*, 2024, 8(6): 16-29.
- CUI J, SHEN X, WEN S. A survey on legal judgment prediction: Datasets, metrics, models and challenges[J]. *IEEE Access*, 2023, 11: 102050-102071.
- RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. 2025. <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf>.
- RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. *Journal of Machine Learning Research*, 2020, 21(140): 1-67.
- KOJIMA T, GU S, REID M, et al. Large language models are zero-shot reasoners[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 22199-22213.
- IZACARD G, GRAVE E. Leveraging passage retrieval with generative models for open domain question answering[C]//*Proc of the 16th Conf of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: ACL, 2021: 874-880.
- RAJANI N, MCCANN B, XIONG C, et al. Explain yourself! leveraging language models for commonsense reasoning[C]//*Proc of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: ACL, 2019: 4932-4942.
- TALMOR A, TAFJORD O, CLARK P, et al. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge[C/OL]//*Proc of the 33rd Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2019. https://proceedings.neurips.cc/paper_files/paper/2020/hash/e992111e4ab9985366e806733383bd8c-Abs-tract.html.
- WANG X, WEI J, SCHUURMANS D, et al. Self-consistency improves chain of thought reasoning in language models[C/OL]//*Proc of the 11th Int Conf on Learning Representations*. 2023. <https://openreview.net/forum?id=1P L1NIMMrw>.
- WEI J, WANG X, SCHUURMANS D, et al. Chain of thought prompting elicits reasoning in large language models[A]. 2022.
- HUANG Q, TAO M, ZHANG C, et al. Lawyer llama technical report[A]. 2023.
- TRAUTMANN D, PETROVA A, SCHILDER F. Legal prompt engineering for multilingual legal judgement prediction[A]. 2022.
- XU L, LI A, ZHU L, et al. Superclue: A comprehensive chinese large language model benchmark[A]. 2023.
- YU F, QUARTEY L, SCHILDER F. Legal prompting: Teaching a language model to think like a lawyer[A]. 2022.
- CHEN S, HOU Y, CUI Y, et al. Recall and learn: Fine-tuning deep pretrained language models with less forgetting[C]//

- Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 7870-7881.
- YUE L, LIU Q, WU H, et al. Circumstances enhanced criminal court view generation[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 1855-1859.
- HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models[A]. 2021.
- HU E, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models[C/OL]//Proc of the 10th Intl Conf on Learning Representations. 2022. <https://openreview.net/forum?id=nZeVKeeFYf9>.
- ZELIKMAN E, WU Y, MU J, et al. Star: self-taught reasoner bootstrapping reasoning with reasoning[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. 2024: 15476-15488.
- LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[A]. 2021.
- ZHANG Q T, WANG Y C, WANG H X, et al. A comprehensive review of large language model fine-tuning[J]. Computer Engineering and Applications, 2024, 60(17): 17-33.
- 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(05): 248-259.
- ZHAO J, GUAN Z, XU C, et al. Charge prediction by constitutive elements matching of crimes[C]//International Joint Conferences on Artificial Intelligence Organization. 2022: 4517-4523.
- XIAO C, ZHONG H, GUO Z, et al. Cail2018: A large-scale legal dataset for judgment prediction[A/OL]. 2018. arXiv: 1807.02478. <https://arxiv.org/abs/1807.02478>.
- CHEN J, XIAO S, ZHANG P, et al. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation: arXiv:2402.03216[M/OL]. arXiv, 2024. DOI: 10.48550/arXiv.2402.03216.
- GUO R, LUAN X, XIANG L, et al. Manu: a cloud native vector database management system[J]. Proceedings of the VLDB Endowment, 2022, 15(12): 3548-3561.
- WANG J, YI X, GUO R, et al. Milvus: A purpose-built vector data management system[C]//Proceedings of the 2021 International Conference on Management of Data. 2021: 2614-2627.
- QWEN, :, YANG A, et al. Qwen2.5 Technical Report: arXiv:2412.15115[M/OL]. arXiv, 2025. DOI: 10.48550/arXiv.2412.15115.
- XIAO C, ZHONG H, GUO Z, et al. Cail2018: A large-scale legal dataset for judgment prediction[A]. 2018.
- ZHUOPENG X, XIA L, YINLIN L, et al. Multi-task legal judgement prediction combining a subtask of seriousness of charge[C/OL]//SUN M, LI S, ZHANG Y, et al. Proceedings of the 19th Chinese National Conference on Computational Linguistics. Haikou, China: Chinese Information Processing Society of China, 2020: 1132-1142. <https://aclanthology.org/2020.ccl-1.105/>.