

法律大模型在裁判文书自动生成中的风险与优化路径

倪慈丹

杭州师范大学沈钧儒法学院，浙江 杭州

收稿日期：2025年6月23日；录用日期：2025年7月4日；发布日期：2025年8月29日

摘要

生成式人工智能大模型的快速发展为法律行业的数字赋能提供了重要契机。作为司法体系的核心主体，人民法院正深入推进大模型技术在智慧司法领域的创新应用。法律大模型具有技术优势，能够基于自然语言处理技术实现法律条文、司法解释及类案的高效检索，从而为法官裁判提供精准的法律依据与决策参考。尤其在司法裁判辅助领域，法律大模型通过自动化生成裁判文书，能够显著提升审判效率与裁判文书质量，对优化司法效能具有积极意义。然而，当前法律大模型在司法裁判文书生成中的应用仍面临多重挑战，制约了法律大模型在司法实践中的功能实现。通过系统性优化与技术迭代，法律大模型有望成为提升司法生产力的关键工具，进而推动更高水平的司法公正实现。

关键词

人工智能，大模型，裁判文书，智慧司法

The Risk and Optimization Pathways of Legal Large Language Models in the Automated Generation of Judicial Documents

Cidan Ni

Shen Junru Law School, Hangzhou Normal University, Hangzhou Zhejiang

Received: Jun. 23rd, 2025; accepted: Jul. 4th, 2025; published: Aug. 29th, 2025

Abstract

The rapid advancement of generative artificial intelligence (AI) large language models (LLMs) presents

文章引用：倪慈丹. 法律大模型在裁判文书自动生成中的风险与优化路径[J]. 法学, 2025, 13(8): 1951-1958.
DOI: 10.12677/ojls.2025.138270

significant opportunities for the digital transformation of the legal industry. As the core entity of the judicial system, People's Courts are actively promoting the innovative application of LLM technology within the domain of smart justice. Leveraging their technical advantages, legal LLMs enable the efficient retrieval of legal provisions, judicial interpretations, and similar cases based on Natural Language Processing (NLP) technology. This capability provides judges with precise legal foundations and decision-making references. Particularly in the field of judicial adjudication assistance, legal LLMs can significantly enhance adjudicative efficiency and the quality of judicial documents through the automated generation of judgments and rulings, demonstrating positive implications for optimizing judicial efficiency. However, the current application of legal LLMs in generating judicial documents faces multiple challenges, which constrain the realization of their full functionality in judicial practice. Through systematic optimization and technological iteration, legal LLMs hold the potential to become critical tools for enhancing judicial productivity, thereby advancing the realization of higher levels of judicial justice.

Keywords

Artificial Intelligence, Large Models, Judicial Documents, Smart Justice

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

当前,人类已逐渐步入智慧时代和数字化社会,以聊天机器人模型(ChatGPT)为代表的大模型技术的迅猛发展,标志着人工智能发展进入了新阶段。大模型是一种基于神经网络架构的深度学习算法,展现出强大的语言理解和内容生成能力[1]。归功于深度学习的发展和硬件计算能力的提升,大模型计算在金融、医疗、广告、营销等领域都有卓越表现。在司法领域,大模型技术亦有广阔的应用前景。《法治蓝皮书·中国法院信息化发展报告 No.8 (2024)》指出,2023年,最高人民法院信息中心搭建起人工智能引擎平台、大模型通用能力和法律法规查询、公文生成及纠错、笔录精简等大模型应用场景,大幅提高了审判工作的效率。值得注意的是,由于裁判文书具有高度结构化和要件化的特征,其自动生成被视为法律大模型最具潜力的应用场景之一。本文将聚焦于辅助司法裁判,探讨法律大模型在裁判文书自动生成中的应用基础、实践困境及优化路径。

2. 法律大模型辅助司法裁判的应用基础:必要性与可行性

(一) 必要性:法院智慧司法对人工智能技术的渴望

1) 运用法律大模型是实现“司法现代化”的现实需要。“司法现代化”是法治建设的重要组成部分,也是国家治理现代化的关键内容。它不是单纯的司法体系的自身建设,而是要从经济、社会发展的时代背景下、司法体制的变革来实现的[2]。随着社会经济的快速发展和民众法律意识的不断提升,司法领域面临的诉讼需求日益增长,“司法生产力”的相对滞后已成为制约司法效率的重要因素。传统的司法模式难以应对案件数量激增和复杂性提升的双重挑战,导致案件积压、审理周期延长等问题日益凸显。在此背景下,法院对智慧司法的需求愈发迫切,亟需通过技术创新提升司法效能。人工智能技术,尤其是法律大模型的引入,为司法领域提供了新的解决方案。法律大模型凭借其强大的数据处理能力、法律文本理解能力以及智能化辅助决策功能,能够在法律法规查询、文书生成、案件分析等方面显著提升司法

效率,助力实现“智慧法院”和“数字法院”的建设目标。通过科学构建和合理应用法律大模型,司法系统有望更好地满足民众的诉讼需求,推动司法生产力的现代化转型,为社会治理提供更加高效、公正的司法保障。

2) 运用法律大模型建设人民法院信息化 4.0 版的必然选择。智慧法院就是新一代信息技术同司法深度融合的产物[3]。自 2010 年最高人民法院成功申报人民法院电子政务项目国家司法审判信息系统工程并启动我国的人民法院信息化 1.0 版的建设依赖,经过十余年的迭代升级,人民法院建成了以互联互通为主要特征的人民法院信息化 2.0 版,并加快推进全业务网上办理网络化、全流程依法公开阳光化、全方位服务智能化的人民法院信息化 3.0 版[4]。2021 年,《人民法院信息化建设五年发展规划(2021~2025)》明确了“十四五”时期人民法院信息化建设目标、思路、任务和路线,提出打造人民法院信息化 4.0 版,即全方位智能化、全系统一体化、全业务协同化、全时空泛在化、全体系自主化。当前已进入“十四五”的收尾阶段,最高人民法院信息中心已搭建人工智能引擎平台,在线立案、在线调解、在线庭审等服务已经在全国范围内推广,部分法院已经运用大数据和区块链技术应用,人民法院在信息化 4.0 版的推进中已初步实现部分目标。在这一背景下,大模型技术将成为全面实现信息化 4.0 版的关键技术支撑,为司法现代化注入新的动力。

3) 大模型是实现“数字正义”的实践载体。“在数字社会中,自由、平等、民主以及法律、秩序和正义都将被重新定义,数字正义将是更高的正义”[5]。在数字技术深刻影响司法系统的背景下,社会对公平正义有了更高水平的需求,正义的理念被赋予新的内涵。将数字正义定义为数字技术应用,尤其是算法应用满足人权、正义、法治价值的一种理想状态[6]。数字正义的实现要求发挥技术力量在正义生产和正义实现中的潜力[7],强调通过技术手段实现司法效率的提升、司法公正的保障以及司法服务的普惠性,大模型作为最新的数字技术,应用于立法、执法、司法、守法等司法系统的各个领域,通过技术手段弥补传统司法模式的不足,必将提升司法效率,为实现“数字正义”提供新的实践路径。

(二) 可行性:法律大模型自动生成裁判文书的应用潜力

1) 大模型自动生成裁判文书的优势。训练完成的大模型有参数规模大、可以生成新内容、涌现新能力、呈现通用性的特征[8],基于大模型的上述技术特征,相较于传统人工撰写模式,法律大模型自动生成裁判文书具有多重优势。首先,在裁判一致性方面,一眼的人工智能根据过往类案发现裁判规律,从而生成裁判结果,能够有效避免入额里法官可能受到的案外因素的影响,这种智能系统自动运行的司法决策条件下,同案同标、同案同判将有机会走进现实生活。其次,在事实发现层面,大模型的涌现能力可能使其发现法官所忽视的案件细节与线索,这种能力源于模型对海量司法数据的深度挖掘和分析,有助于更全面地还原案件事实,提高裁判的准确性[9]。此外,在法律适用方面,法律大模型能够为法官处理疑难法律问题提供智能辅助,如快速检索相关法律条文和司法解释、生成规范化的法律论证框架等。这些功能能够显著减轻法官在信息处理和文书撰写方面的工作负担,有助于提高司法效率。

2) 人工智能在司法领域的应用已积累了丰富的实践经验。最高人民法院关于规范和加强人工智能司法应用的意见提出:“加快推进人工智能技术与审判执行、诉讼服务、司法管理和服社会治等工作的深度融合,规范司法人工智能技术应用,提升人工智能司法应用实效,促进审判体系和审判能力现代化,为全面建设社会主义现代化国家、全面推进中华民族伟大复兴提供有力司法服务。”在人工智能时代浪潮的席卷下,我国各级法院积极探寻并实施了“人工智能+”的发展模式,围绕立案、审判、送达、执行等诉讼流程,各地都在积极开发网上立案系统、电子送达系统、语音识别系统、在线诉讼平台、智慧辅助办案系统、网络查控系统等,切实解决立案难、送达难、执行难等问题。这些实践不仅展示了人工智能技术在司法领域的广泛应用,也为大模型在法院系统中的进一步应用奠定了坚实的基础。

3) 人民法院为法律大模型自动生成裁判文书提供充足数据支持。司法智能化的两个重要条件是大数据与算法[10]。其中,算法的发展依托于数字化技术的进步,而法律大模型所需的大数据则来源于各类法律规范、法律注释、裁判案例以及案件事实等多维数据源。在这一维度上,人民法院相较于其他法律实践主体具有三重独特优势:其一,数据规模优势,人民法院因其庞大的案件审理量以及全国四级法院之间的数据共享机制,积累了更为丰富的司法数据资源为法律大模型训练提供了充足的语料资源。其二,数据质量优势。依托“中国司法大数据服务平台”的协同机制,四级法院系统实现了裁判数据的标准化采集与结构化处理,还可联通人民法院案例库、法答网、法信等最高人民法院权威知识服务体系。相较于分散的法律服务数据,法院系统数据具有更高的完整性、规范性和时效性特征。其三,应用场景优势。法官等司法实践主体对法律大模型的接受度和应用能力也相对较高,进一步凸显了人民法院在法律大模型应用中的主体优势。

3. 法律大模型自动生成裁判文书的实践风险

(一) 技术性约束:技术外包造成权力“外溢”,司法问责流于形式

由于兼具法律专业知识与计算机技术能力的复合型人才稀缺,人民法院在独立开发法律大模型方面面临显著的技术瓶颈,因而不得不依赖外部科技公司研究核心算法技术。一般而言,人民法院通过购买或者外包的形式向科技公司寻求帮助,如法信法律基座大模型就是由最高人民法院与清华大学及面壁智能合作研发的成果。此外,人民法院不仅在法律大模型的开发阶段需要科技公司的技术支持,在后续的模型部署、运维、优化及升级过程中,仍需持续依赖科技公司的专业服务。然而,科技公司出于技术垄断与商业利益保护的考虑,往往以涉及商业秘密或知识产权为由,拒绝将核心技术开源或公开,进一步加剧了人民法院的技术依赖[11]。

根据我国宪法与相关法律规定,审判权专属于人民法院行使,这是司法独立性与权威性的重要体现。然而,在法律大模型的开发与建设过程中,科技公司的深度技术介入对审判权的专属性原则构成了潜在挑战。尽管在开发过程中有司法机关与法律专业人员提供专业知识指导,但核心算法技术的研发仍主要由科技公司的技术人员主导。这种技术主导模式可能导致设计者的个人偏见或商业利益考量被无意识地嵌入算法之中,进而影响模型的输出结果与决策逻辑,造成审判权力的“外溢”现象。具体而言,技术公司可能通过算法设计、数据选择或模型优化等环节,间接影响司法审判的公正性与独立性,从而对司法权的专属性与权威性形成侵蚀[12]。早在人工智能技术初步应用于司法领域时就有学者担忧,让人工智能技术参与司法裁判过程将导致司法问责制度流于形式,“让人工智能自动生成判决、根据大数据矫正法律决定的偏差等做法势必形成审判主体的双重结构、甚至导致决定者的复数化,事实上将出现程序员、软件工程师、数据处理商、信息技术公司与法官共同作出决定的局面。一旦审判主体和决定者难以特定,那么权力边界也就变得模糊不清,司法问责制就很容易流于形式,至少推卸责任的可能性被大幅度扩充了”[13]。当法官过度依赖于法律大模型自动生成裁判文书时,实际的决策权可能向技术开发者转移,导致技术标准无形中取代了司法判断标准。

(二) 数据库桎梏:数据库搭建难度高,易造成信息泄漏

数据是人工智能技术的基石,是人工智能技术运行的原料性基础[14]。在法律大模型的构建过程中,法律数据库的搭建是其核心基础,数据质量直接决定了法律大模型的可靠性。有学者曾总结人工智能运用于我国司法裁判在数据方面存在的四大障碍:(1) 法律数据不充分,例如大量裁判文书不上网、裁判者心证过程等决策信息未体现于裁判文书中;(2) 法律数据不真实,例如公开的裁判理由与实际的裁判理由并不一致;(3) 法律数据不客观,不同法官对同一问题可能有不同看法,导致裁判数据的主体性、主观性;(4) 法律数据结构化不足,例如法律语词的不统一导致人工智能识别困难[15]。在应用法律大模型自动生

成裁判文书的过程中,这些数据难题必然存在,并因数据的地域性和不确定性加剧了数据库搭建的难度。首先,法院数据具有显著的地域性特征,不同地区的法院在裁判尺度、法律适用以及司法实践方面存在差异,这种地域性差异导致数据分布不均且标准不统一。例如,经济发达地区与欠发达地区在案件类型、裁判倾向以及法律解释上可能存在显著差异,这种地域性特征使得数据难以在全国范围内形成统一的训练样本,从而影响法律大模型的泛化能力与适用性。其次,法院数据的不确定性进一步加剧了数据库搭建的复杂性。由于司法裁判过程中法官的主观判断、心证过程以及自由裁量权的存在,同一类案件在不同法院甚至同一法院的不同法官手中可能产生截然不同的裁判结果。这种不确定性不仅体现在裁判结果的多样性上,还反映在裁判文书的表述方式、法律理由的阐释以及证据采纳的标准等方面,导致数据的噪声较多且难以标准化处理。此外,法院数据的公开程度与质量也存在较大差异,部分地区的数据公开不充分或存在滞后性,进一步增加了数据采集与清洗的难度[4]。

数据安全问题在法律大模型自动生成裁判文书过程中同样不容忽视。人民法院的裁判文书中不可避免地包含当事人的个人信息与隐私数据,例如姓名、身份证号、住址等敏感信息,还可能涉及敏感的商业秘密、国家秘密等信息。如果在法律大模型的开发过程中未能建立完善的数据保护机制,例如数据脱敏、加密存储、访问控制等技术措施,极有可能导致个人隐私的泄露。这种泄露不仅可能侵犯当事人的合法权益,还可能引发公众对司法系统数据安全性的信任危机。特别是在法律大模型的训练过程中,海量裁判文书数据的集中使用与多次迭代,进一步放大了隐私泄露的风险。

(三) 算法性局限: 技术固有缺陷影响法律大模型效能发挥

由于大模型的数据采集、训练、精调、部署等流程相比于通常模型更加复杂,其中的不可控因素影响了其效能发挥。首先,大模型幻觉问题尤为突出,即大模型生成无意义或与用户提示词不对应的内容,通常包含与用户输入冲突幻觉、上下文冲突幻觉和事实冲突幻觉[16]。法律大模型在自动生成法律文书时可能产生与事实不符或缺乏法律依据的内容,导致生成的内容可能是虚假的或者完全与事实相反的,这种幻觉现象在复杂案件或边缘案例中尤为明显,可能导致误导性结论,影响司法决策的准确性与公正性。

其次,算法的“黑箱”特性使得其决策过程缺乏透明性与可解释性。所谓“黑箱”,是指这样一种系统或机制,它的输入和输出是明确的,但是输入到输出是如何转化却是部分或全部隐藏的[17]。这种“黑箱”并不意味着不能观察其内部逻辑,还意味着即使计算机试图向我们解释,我们也无法理解[18]。法在法律大模型的开发与应用中,设计人员可能通过“黑箱”机制生成具有偏向性的辅助审判结果,误导法官作出有偏见的司法判决,同时当事人难以理解算法的决策过程和裁决理由,这不仅可能影响法官的中立性,还可能损害公众对司法公正的信任。

此外,算法歧视问题同样不容忽视,由于训练数据中可能隐含的历史偏见或地域性差异,模型可能在裁判建议中无意识地放大这些偏见,导致对特定群体或案件类型的不公平对待。以美国为例,犯罪风险评估软件系统性地歧视黑人,白人更多被错误地评估为具有低犯罪风险,而黑人则被错误地评估为具有高犯罪风险的几率两倍于白人。这就使得算法裁判的准确性大打折扣,影响到其在司法实践中的更广泛运用[19]。

(四) 价值观冲突: 司法裁决的价值判断难以被纳入算法决策

《最高人民法院关于加强和规范裁判文书释法说理的指导意见》中提出,裁判文书说理要做到四个层面,即事理、法理、情理和文理的统一。深刻揭示了司法裁判活动的复合性本质。这种复合性体现在三个相互关联的维度:其一,司法裁判绝非简单的事实认定与法律适用的机械过程,而是要求裁判者在具体情境中综合运用法律专业知识、价值判断能力和社会经验智慧的创造性活动;其二,作为权利义务分配的特殊社会机制,司法裁判必然承载着优先保护特定利益、弘扬主流价值的规范功能;其三,裁判活动需要实现法律效果与社会效果、政治效果的有机统一。这种复合性特质决定了优质的司法裁判必须

同时具备法律专业性、价值正当性和社会适应性三个基本要素。

在法律专业性维度,“法律知识表达的核心问题是法律解释。”[20]法律知识以其高度的专业性与复杂性著称,其不仅包含成文法、判例法、司法解释等多元化的法律渊源,还涉及法律原则、法律逻辑以及法律适用中的具体情境分析。因此,法律大模型的构建不仅需要具备通用文本的理解能力,还需在此基础上实现对法律文本的深度解析与精准把握。具体而言,法律文本的理解不仅要求模型能够识别法律术语、条文结构以及语法规则,还需进一步捕捉法律条文背后的立法意图、司法逻辑以及价值取向。例如,法律文本中常见的模糊性表述、例外条款以及隐含的法律原则,均需要模型具备高度的语义理解能力。这也是法律大模型构建过程中普遍无法解决的问题。另外,基于法律体系的开放性,法律知识不断发展和变化,法律大模型的开发构建不可避免存在滞后性,也可能导致对法条的理解和应用存在偏颇。同时,法律领域对逻辑性的要求极高,法律推理不仅需要对法律条文进行精准解读,还需结合案件事实、司法先例以及法律原则进行复杂的逻辑推演。然而,当前的大模型尽管在自然语言处理与模式识别方面取得了显著进展,但其在法律推理中的表现仍难以满足高标准要求。机器学习算法的目标并非复现法律推理,而是寻找判决中各个参数间的相关性。在组成司法判决的不同词汇组之间构建分类链接:输入阶段的特定词汇组(表征案件事实)对应于输出阶段的特定词汇组(表征裁判结论)。它的基本原理近似于“讯飞”这样的机器翻译系统,只能在一组词汇和已经完成的译文之间对最佳匹配作可能的估计值,而无法真正“理解”所处理的句子的意思[21]。法律大模型仅能识别现有数据中的统计模型,而不能复制或模拟人类的推理过程,其根本无法理解法官作出如此判决的原因,因而其作出的数据推断也完全不同于法官作出的法律逻辑推理。

在价值正当性和社会适应性维度,算法系统的技术逻辑与司法裁判的实践逻辑存在做出差异。前者追求形式化、确定性和标准化,后者则强调情境化、创造性和价值判断。司法裁判过程中的价值判断是必不可少的,因为司法裁判本身就是通过权利和义务的分配,来优先保护某种利益,弘扬某些价值。通常来说,法官所运用的价值判断应当是特定社会的主流价值观[22]。然而,算法裁判依赖于预设的数据与规则,难以充分理解与体现法律条文背后的价值取向与社会背景,可能导致裁判结果缺乏人文关怀与社会适应性。例如,算法可能无法准确识别某些案件中的特殊情境或社会文化背景,从而在价值判断上出现偏差,影响裁判的公正性与社会接受度。

我国宪法虽然赋予了人民法院独立审判的地位,然而,受到国家政治体制和历史传统的影响,政治因素在司法裁判中的渗透仍然较为显著,“司法裁决仍不是一种单纯的司法行为,而是带有明显的政治考量。”[23]比如,“民愤极大”就曾是一种从政治上、道义上来考虑的从重处罚情节,“领导批示”也会对案件的定性和裁判具有重要“指导”作用。因政治性因素往往是无形的,不确定性的,“领导批示”也仅能体现在裁判者的裁判思路中,而不能直接反映在卷宗里,这些政治性因素很难进入算法决策的考量范围[24]。这些非正式政治性因素因其无形性和不确定性难以被算法系统有效识别和量化,导致算法生成的裁判建议与实际情况存在显著偏差。

4. 法律大模型自动生成裁判文书的优化路径

(一) 完善人机协同,提升法官掌控力

作为司法裁判的唯一主体,法官既是法律大模型的使用者,也是其实际效能的最终评判者。技术优化司法效能的上限,始终以不突破审判权专属性和法官心证自由为边界。因此,应当构建以法官为核心、技术为辅助的智能化裁判体系。在技术开发阶段,应充分吸纳一线法官干警的意见和建议,通过问卷调查、座谈会、试点应用等多种形式,深入了解其在办案过程中遇到的实际问题和技术需求。此外,应建立法官干警与技术开发团队之间的常态化协作机制,确保技术开发始终围绕司法实践的核心需求展开。

同时,应注重培养既精通法律业务又熟悉技术应用的复合型人才,通过在岗培训、业务研修、跨部门交流等方式,系统提升法官运用智能技术的能力,将数字司法能力作为新时代法官专业素养的重要组成部分,使其既能有效利用技术工具提升办案效率,又能保持独立的司法判断力,避免形成技术依赖。

在具体应用层面,必须通过制度设计筑牢审判权保障防线。为避免技术工具不当介入司法裁判核心环节,要明确大模型始终处于辅助地位,严禁将自动生成的文书直接作为裁判依据,严格要求法官对模型输出内容进行实质性审查和必要修改,并在文书中完整记载技术工具的使用范围和人工修正情况,确保全程留痕、可追溯。在程序上,要建立“人类法官最后签字”制度,规定未经法官亲笔签名的大模型生成文书不具有法律效力,确保审判权不被技术取代。同时,要将法官对智能文书的审查修改情况纳入案件质量评查体系,建立相应的考核激励机制。针对不同类型案件的特点,还应当建立差异化的应用规则,例如在事实清楚、争议较小的简单民商事案件中,可允许模型协助完成事实梳理和法条检索等基础性工作,但必须严格禁止其介入裁判结论形成等核心司法判断环节,在复杂、高风险刑事案件中,大模型仅能用于辅助法条检索,而不能介入事实认定和量刑裁判,始终确保法官对裁判结果的绝对主导权。通过这种人机协同的精细化制度安排,既能充分发挥技术增效作用,又能从根本上保障审判权的依法独立行使。

(二) 加强审慎评估,完善技术安全保护机制

如前所述,法律大数据中可能包含大量个人隐私信息、商业秘密、国家机密信息等,在构建法律大模型过程中,首先要对法律大数据进行安全审查和评估。根据数据敏感性、法律重要性以及用途,对数据进行分类和分级,并根据数据的不同级别制定相应的保护措施,采用数据加密和匿名化等措施,增强数据的安全性,防止数据滥用。司法机关还可以引入法律和数据安全专家,对数据处理方案进行审核,评估数据处理可能带来的风险,特别是对数据主体权利和数据安全的影响[25]。

为此,为减轻大模型自身技术存在的“黑箱”问题,在技术设计中应注重破解技术黑箱的透明化改造,研发司法专用的可解释性算法框架。例如通过可视化工具或逻辑推理链条,向用户展示模型的决策依据和推理过程,使大模型推理过程能够还原为可理解的司法逻辑链条。为降低大模型“幻觉”生成错误事实的概率,除加强人工复核外,还可训练专门的事实核查大模型,重点检测实践、金额、主体资格等关键事实要素的准确性。

总之,作为司法权的承载者和司法程序的主导者,人民法院在正式投入使用法律大模型前,应加强审查评估义务,必须经过专业技术人员和法律专业人员的双重评估,确保算法、程序设置的无误性,在产品使用过程中也要进行周期性的检查、维护和更新,以应对突发情况。为增强司法体制内部的责任规制,可以借鉴近年来数字检察改革的成果,充分利用数字检察系统和大数据监督的核心优势,及时发现异常数据,实现权力内部的有效监督制约[26]。如果存在虚假评估或者技术操作问题,需要追究相应人员的责任。

(三) 规范价值对齐,训练模型“公正”裁判

大模型价值对齐指使具有大量参数和复杂结构的机器学习模型的输出结果与人类的意图和价值观相一致[27]。价值对齐贯穿于法律大模型系统的全生命周期,从数据采集、模型训练、性能优化到实际应用,均需确保模型的输出符合人类社会的法律原则和道德规范。在法律大模型的场景中,价值对齐的核心挑战在于如何使模型生成的判决“以事实为依据,以法律为准绳”,同时与人类伦理价值观保持一致。首先,应将习近平总书记提出的“以人为本、智能向善”人工智能治理理念[28]。等核心价值观融入价值对齐的场景中,确保法律大模型的输出内容不仅符合法律规范,还应体现正向的价值观和高度的社会责任感。例如,在涉及道德判断的案件中,模型不仅需要依据法律条文作出裁决,还需考虑社会公众的伦理期待和情感认同,避免因技术局限性导致的伦理冲突或社会争议。其次,在数据采集和模型训练过程中,必须坚持“公平公正”的法治原则,避免大模型在收集和处理法律数据时产生针对不同国别、民族、信

仰、性别、身份等的偏见和歧视。法律大模型的训练数据通常来源于历史裁判文书、法律法规和司法案例，这些数据中可能隐含历史偏见或结构性不公。因此，在数据预处理阶段，需通过科学的清洗和标注方法，剔除带有偏见或歧视性的内容，确保数据的多样性和代表性。同时，在模型训练过程中，应引入公平性评估机制，通过多维度指标(如群体公平性、个体公平性等)对模型的输出进行监测和优化，确保其不同社会群体中均能实现公正地裁决。

参考文献

- [1] 余圣琪. 公共大模型决策的法治化约束[J]. 国家检察官学院学报, 2025, 33(1): 160-176.
- [2] 江国华. 数据共享与中国司法现代化[J]. 中国高校社会科学, 2017(1): 80-88.
- [3] 胡铭, 周翔, 等. 数字法治实践与变革[M]. 杭州: 浙江大学出版社, 2022.
- [4] 李琴, 姚依哲, 刘翔. 大型语言模型引入智能化审判的论证规则与实现路径[C]//上海市奉贤区人民法院, 上海市奉贤区人民法院审判监督庭审判团. 上海市奉贤区人民法院商事审判庭《法学前沿》集刊 2024 年第 1 卷——第十二届京津沪渝法治论坛文集. 上海: 上海市法学会, 2023: 107-117.
- [5] 周强. 深入学习贯彻党的十九届四中全会精神不断推进审判体系和审判能力现代化[J]. 人民司法 2020(1): 4-8.
- [6] 周尚君, 罗有成. 数字正义论: 理论内涵与实践机制[J]. 社会科学, 2022(6): 166-177.
- [7] 帅奕男. 数字时代的司法范式转型[J]. 求是学刊, 2021, 48(6): 121-133.
- [8] 刘金瑞. 生成式人工智能大模型的新型风险与规制框架[J]. 行政法学研究, 2024(2): 17-32.
- [9] 邱遥堃. 机遇与风险: 文生视频大模型将如何影响智慧法院建设升级迭代[J]. 中国应用法学, 2024(2): 89-98.
- [10] 张卫平, 冉博. 对司法智能化应用局限性的观察与思考[J]. 财经法学, 2024(5): 94-108.
- [11] 张凌寒. 智慧司法中技术依赖的隐忧及应对[J]. 法制与社会发展, 2022, 28(4): 180-200.
- [12] 吴惠. 司法裁判人工智能化的困境及未来走向[J]. 中国人民警察大学学报, 2024, 40(6): 29-35.
- [13] 季卫东. 人工智能时代的司法权之变[J]. 东方法学, 2018(1): 125-133.
- [14] 郑曦. 生成式人工智能在司法中的运用: 前景、风险与规制[J]. 中国应用法学, 2023(4): 81-93.
- [15] 左卫民. 关于法律人工智能在中国运用前景的若干思考[J]. 清华法学, 2018, 12(2): 108-124.
- [16] 腾讯研究院. 大模型安全与伦理研究报告 2024: 以负责任 AI 引领大模型创新[EB/OL]. <https://www.tisi.org/27403>, 2024-01-24.
- [17] Pasquale, F. (2015) The Blackbox Society: The Secret AI Algorithms that Control Money and Information. Harvard University Press, 3-8. <https://doi.org/10.4159/harvard.9780674736061>
- [18] 许可. 人工智能的算法黑箱与数据正义[N]. 社会科学报, 2018-03-29(6).
- [19] 高学强. 人工智能时代的算法裁判及其规制[J]. 陕西师范大学学报, 2019, 48(3): 161-168.
- [20] 张保生. 人工智能法律系统的法理学思考[J]. 法学评论, 2001, 19(5): 11-21.
- [21] 郭锐著. 人工智能的伦理和治理[M]. 北京: 法律出版社, 2020.
- [22] 冯洁. 大数据时代的裁判思维[J]. 现代法学, 2021, 43(3): 41-55.
- [23] 韩显阳. 美国最高法院大法官席位争夺战开打[N]. 光明日报, 2016-2-17(12).
- [24] 马长山. 司法人工智能的重塑效应及其限度[J]. 法学研究, 2020, 42(4): 23-40.
- [25] 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): 43-73.
- [26] 吕子逸. 数字司法的制度检视与样态建构[J]. 中国海商法研究, 2024, 35(3): 96-112.
- [27] 吉嘉铭, 邱天异, 陈博远, 等. 人工智能对齐: 全面性综述[Z]. 北京大学人工智能研究院 AI 安全与治理中心译. 人工智能对齐网(AIAlignment). 2024-01.
- [28] 新华社. 习近平向联合国贸易和发展会议成立 60 周年庆祝活动开幕式发表视频致辞[EB/OL]. https://www.gov.cn/yaowen/liebiao/202406/content_6956945.htm, 2024-06-12.