AI 大语言模型的数据安全风险与治理措施*

林　杰，王家盛

（同济大学　经济与管理学院，上海　200092）

摘　要：随着 AI 大语言模型的迅猛发展，其在为各领域带来创新机遇的同时，也引发了诸多数据安全风险。深入剖析 AI 大语言模型所面临的数据安全风险，包括数据泄露、数据滥用、数据偏见等方面，并针对性地提出一系列治理措施，旨在为构建安全可靠的 AI 大语言模型应用环境提供全面的理论依据与实践指导，促进该技术在合法合规且安全的轨道上持续发展。

关键词：大语言模型；数据安全风险；治理措施；隐私保护

中图分类号：TP309　　**文献标识码：**A　　**DOI：**10. 19358/j. issn. 2097 - 1788. 2025. 11. 005

引用格式：林杰，王家盛. AI 大语言模型的数据安全风险与治理措施［J］. 网络安全与数据治理，2025，44(11)：24 - 29.

Data security risks and governance measures for AI large language models

Lin Jie，Wang Jiasheng

(School of Economics and Management，Tongji University，Shanghai 200092，China)

Abstract：With the rapid development of AI large language model，it not only brings innovation opportunities for various fields，but also causes many data security risks. This study deeply analyzes the data security risks faced by AI large language model，including data leakage，data abuse，data bias，etc. ，and proposes a series of governance measures，aiming to provide a comprehensive theoretical basis and practical guidance for building a safe and reliable AI large language model application environment，and promote the sustainable development of the technology on the track of legal compliance and safety.

Key words：large language model；data security risks；governance measures；privacy protection

0　引言

AI 大语言模型的快速发展和广泛应用，正在深刻改变各个行业的生产和生活方式。在自然语言处理、图像生成、音视频生成、代码开发等多个领域，大模型都发挥着越来越重要的作用。AI 大模型技术的进步不仅提高了生产效率，节省了人力成本，还为用户提供了更多的创造性工具，丰富了产品的多样性。

随着 AI 大语言模型的不断成熟，其对人们的日常生活的影响也日益显著。然而，伴随而来的数据安全问题也日益突出。大模型在训练过程中通常需要大量的数据，这些数据可能涉及到用户的个人信息和敏感内容。因此，如何确保数据的安全性、合法性和合理使用，成为技术发展过程中的关键议题。

此外，随着 AI 大模型的发展，其生成内容的多样性和复杂性，也增加了对用户隐私和知识产权的威胁。这种复杂性不仅仅源于其技术层面的创新，更涉及伦理、法律和社会等多重维度的思考。因此，在探索 AI 大语言模型潜力的同时，保障数据安全、维护用户隐私和促进合规运营，显得格外重要。

本研究旨在全面识别 AI 大语言模型的数据安全风险，包括数据收集、数据存储、数据使用和数据共享阶段的潜在安全风险。此外，针对这些安全风险提出切实可行的治理策略，包括法规与政策层面、技术层面和管理层面。

借此，在 AI 大模型给生产生活带来帮助的同时，切实守护用户隐私权益，杜绝个人信息泄露风险，维护市场运营秩序平稳有序。此外，填补当前在 AI 大模型数据安全风险深度剖析与综合治理系统性方案设计方面的既有空白，为该项技术沿着合规、稳健的轨道，持续高质

*基金项目：国家社会科学基金重点项目（22AZD136）；上海市 2024 年度科技创新行动计划（24692114200）

量发展注入强大内生动力，助力科技发展与安全保障实现协同共进。

1 文献综述

随着 ChatGPT 在 2022 年底的横空出世，其在生活、学习、工作等多个场景中都展现出惊人的表现，生成式人工智能引起了社会各界的广泛关注。2025 年初，DeepSeek 的爆火又将生成式人工智能和 AI 大模型推至了前所未有的高度。相关研究人员也逐渐意识到其中隐藏的数据安全风险。

Shi[1] 认为中国在生成式人工智能治理中应坚持安全与发展并重的平衡方针，推动形成以一般生成型人工智能立法为基础、以具体管理办法为补充的系统性法律监管体系。李森[2] 比较了 GPT 类模型数据的输入、访问以及内容生成等数据运行的风险及其特征，提出从数据源头、内部运行到数据生成的全链条风险防范机制。Dhoni 等[3] 深入研究案例、新兴趋势及潜在对策，揭示生成式人工智能时代网络安全的新维度。高松[4] 聚焦生成式人工智能面临的数据安全挑战，剖析数据处理各环节存在的安全风险，揭示了当前数据安全保护工作的紧迫性和复杂性。Golda[5] 等研究了生成式人工智能固有的隐私和安全挑战，提出了涉及用户、开发人员、机构和决策者的可持续解决方案。Takale 等[6] 剖析了生成式人工智能在文本、图像、视频、音频和代码生成中固有的特定漏洞，阐明每种模式所带来的安全挑战，并探索了必要的保障措施和缓解策略。

此外，一些学者也针对 AI 大模型的数据安全问题，提出了更加具体的数据安全管理和数据治理监管的措施。张欣[7] 提出应以数据解释机制为核心，强化人工智能 2.0 时代个体的信息掌控和自决能力，打造灵活高效的数据治理监管工具体系。斜晓东[8] 提出以风险管控、算法透明原则、多层次的数据管理保障机制以及技术、标准与法律三元架构的生成内容治理机制为应对手段。赵梓羽[9] 提出应使用敏捷治理模式来应对生成式人工智能等新兴技术的数据安全风险。虎嵩林[10] 建议建立国家级大模型安全科技平台，助力人工智能安全产业集群发展。

在治理 AI 大模型的数据安全问题的同时，也有学者发掘了其在网络安全领域的价值，认为它可以大幅提升当前网络安全防御技术的能力。Gupta 等[11] 调查了网络犯罪者如何使用生成式人工智能进行网络攻击，并探讨了使用生成式人工智能的网络安全防御技术，包括网络防御自动化、安全代码生成和检测、网络攻击识别等。

2 AI 大语言模型的数据安全风险分析

2.1 数据收集阶段风险

2.1.1 数据来源合法性问题

在数据收集环节，确保合法性是构建稳固 AI 大语言模型体系的基石，然而现实中违规行径屡见不鲜。部分机构为追求数据的规模与多样性，往往逾越法律红线。就未经用户明确授权收集个人敏感信息而言，诸多智能应用在用户注册或初次使用时，利用冗长晦涩的隐私条款，以默认勾选或模糊表述的方式，悄然收集用户诸如健康状况、财务明细、生物识别特征等高度敏感数据，未给予用户清晰知晓与自主抉择机会，严重违背相关隐私法规精神。

从非法渠道获取数据资源的问题同样存在，有些不良从业者与地下数据交易黑客点勾结，购入通过黑客非法入侵企业系统窃取的用户数据库，或采用违规爬虫技术肆意抓取社交平台、电商网站用户信息，这类行径公然触犯《中华人民共和国网络安全法》《中华人民共和国个人信息保护法》等诸多法规条文。

2.1.2 数据质量与完整性风险

数据质量与完整性对 AI 大模型的训练十分重要，一旦出现问题，模型训练与应用便会陷入困境。在收集方法上，若采用抽样偏差严重的方式，比如仅聚焦特定区域、特定年龄段用户群体收集数据，忽视整体多样性，会使样本缺乏代表性，模型训练后难以泛化至广泛场景。数据源不准确更是严重问题，部分公开数据集标注错误频出，或源于人工标注疏忽、或因标注规则模糊，输入模型后，将严重干扰输出准确性。

数据缺失现象也不容忽视，传感器故障致环境监测数据间断、问卷调研部分题目大量空白未填，都造成信息残缺。基于低质量数据训练的模型，在实际运用时，易引发决策失误、误导用户，大幅降低整个系统的可靠性。

2.2 数据存储阶段风险

2.2.1 数据泄露风险

在 AI 大模型的架构中，数据存储基础设施十分关键，却深陷风险。AI 大模型面临的外部网络攻击猖獗，黑客将恶意软件藏于常规更新或插件，诱使下载窃取数据；或用翻新的网络钓鱼手段，仿冒信息诱骗用户输入关键数据以破防护；以及利用系统漏洞潜入窃取。内部威胁也不容忽视，部分员工受利益诱惑，违规私自拷贝、售卖高敏感数据，致企业数据外流；也有部分员工因意识淡薄、误操作等使数据暴露。系统故障方面，硬件老化、软件崩溃、电力中断等，会导致数据存储无序脆弱，

给不法分子可乘之机。

2.2.2 数据存储管理不善风险

数据存储管理对 AI 大模型至关重要。缺乏有效分类管理，海量多元数据在存储介质杂乱堆砌，医疗、企业等各类数据混淆，调取关键信息耗时费力，影响系统功能发挥。访问控制机制若有漏洞、权限随意，无关人员易触敏感数据，开发人员可越界浏览，不法者借机篡改窃取，导致数据不准、保密失效，干扰模型决策。

此外，数据备份与恢复策略若不周全，遇存储介质故障或天灾，若无合理异地备份，数据易永久丢失；恢复流程拖沓，系统将会瘫痪，导致服务中断、企业运营受阻、客户流失。数据存储管理不善将削弱数据可用性，阻碍系统正常运转。

2.3 数据使用阶段风险

2.3.1 数据滥用风险

在 AI 大模型蓬勃发展并广泛渗透至各应用场景之际，数据滥用问题如影随形，成为不容忽视的风险隐患。诸多企业在运营过程中，为追求更高效精准的营销成效，常罔顾用户权益，擅自将收集而来的数据超出既定授权范畴加以运用。以精准广告推送与用户画像分析为例，部分互联网平台在获取用户基础信息时，仅以宽泛模糊的条款征得初始同意，后续却深挖用户浏览历史、搜索偏好、地理位置等多维度数据，整合勾勒出详尽用户画像，用以投放高度靶向性广告，全然不顾及用户对隐私边界被肆意突破的抵触情绪，导致用户隐私权被严重侵犯。

此外，有些不法分子将黑手伸向数据资源，利用获取的数据实施诈骗、恶意营销等违法犯罪活动。他们借助掌握的用户身份、联系方式及消费习惯等信息，精心编织骗局，伪装成正规金融机构或知名电商平台，以虚假优惠、高额回报为诱饵，诱骗用户转账汇款，导致大量用户财产受损，社会信任根基遭受重创，公众安全感急剧下降，扰乱正常市场经济秩序，严重损害用户权益与社会公共利益。

2.3.2 数据偏见与歧视风险

AI 大语言模型的输出公正性深受训练数据特质左右，一旦训练数据裹挟偏见，其衍生的不良影响将渗透至诸多关键决策领域。从根源追溯，训练数据收集环节易受社会既有刻板印象浸染，在性别维度，传统行业分工认知偏差致使数据集中男性主导行业样本丰富，女性相关样本失衡，模型训练后在职业推荐场景倾向男性职业，边缘化女性职业发展可能；种族层面，若数据反映种族聚居区资源分布不均，会导致生成的回答出现偏差。例如：谷歌的 AI 大模型 Gemini 回答“美国的国父是谁?”这一问题时，生成了黑人男性的图片。这与事实严重不符。

2.4 数据共享阶段风险

2.4.1 数据共享协议不规范风险

在 AI 大模型的生态下，数据共享愈发频繁，然而数据共享协议的不规范却成为潜在风险，随时可能引爆合作中的矛盾冲突。部分企业拟定协议时，条款措辞含糊不清，对于数据使用范围仅是宽泛概述，未精准界定应用场景与边界，例如仅提及“用于产品优化与研发”，却不细化涵盖的具体产品、可开展的研发操作，易使使用方肆意扩大利用范畴。

在数据保护责任方面，协议常未明确规定安全防护等级、漏洞修复时限等关键细则，一旦出现数据泄露，双方推诿扯皮，提供方认为使用方运维不力，使用方指责数据初始状态不安全。至于数据所有权归属，协议文本表述不清，当数据经加工、整合产生价值增值时，双方对新成果权属各执一词，或争抢权益，或都不愿担责。这类争议不仅迟滞数据流动节奏、打乱业务协同计划，更会破坏合作双方的信任根基，阻碍行业数据高效整合利用。

2.4.2 数据跨境传输风险

伴随全球化进程，AI 大模型的跨国数据共享司空见惯，可背后风险重重。不同国家和地区数据保护法规大相径庭，欧盟秉持严格的《通用数据保护条例》，强调用户权利至上、数据出境严苛审批；而部分新兴市场国家法规尚处完善阶段，管控宽松。跨国企业常陷入两难：遵循高标准则成本剧增，迁就低标准则易造成违规。

数据主权问题凸显，各国视数据为关键资产，跨境传输中，政府担忧本国数据失控或被他国不当利用，加强审查干预。数据传输过程中，网络攻击、信道劫持等安全威胁不断，技术短板与管理漏洞使加密、监控手段失效，隐私数据易泄露。一旦出现问题，将引发国际法律冲突，削弱国际协作动力，制约技术跨境拓展。

3 AI 大语言模型的数据安全治理措施

3.1 法律法规与政策层面

3.1.1 完善数据安全相关法律法规

AI 大模型发展迅猛，现有法律难以适配其复杂的数据处理。在数据收集上，部分应用为获取海量、多元数据，未经充分明示授权收集敏感信息，如智能创作助手违规采集用户隐私细节。修订法规时，要明确合法授权流程，强制清晰告知收集目的、范围、期限与处置方式，保障用户知情权与选择权。

使用阶段，模型训练与成果运用常现数据超范围调

用、滥用，比如精准营销中将数据挪作他用、交易隐私数据等。需精准划分使用边界，设可追溯机制，明晰越界责任与惩处，以便于监管。存储环节，面对海量数据，应规范存储设施安全、备份冗余及留存时长标准，以防泄露与隐私风险。共享时，因产业链长、数据流动频繁，要厘清各主体权责、审核义务，依危害提升违法成本，筑牢法律根基。

3.1.2 制定行业规范与标准

在数据安全管理体系构建层面，应出台一套涵盖组织架构搭建、人员职责明晰、流程制度落地的完整规范集。例如，指导企业内部设立独立数据安全管理部门，赋予其全程监督、审核数据业务权限，依据规范严格把控风险。此外，制定详尽数据分类分级标准，按敏感程度、重要价值将数据标签化管理，对涉及个人身份、财务等核心敏感数据，施以高级别加密、严格访问限制防护。

3.2 技术层面

3.2.1 数据加密与脱敏技术

在 AI 大模型复杂的数据生态中，数据加密与脱敏技术构成首道坚固防线。先进的数据加密算法，如对称加密的 AES（高级加密标准），凭借其高效加密和解密速度，对大规模敏感数据在存储与传输环节进行“密锁封装”。以医疗领域生成式 AI 应用为例，海量患者病历、基因检测等敏感数据，运用 AES 算法加密后存储于本地或云端，密钥由严格权限管控体系保管，确保数据流转全程保密性，即便数据意外泄露，也呈密文形态，无法被轻易利用。

非对称加密的 RSA 算法，则依托公钥、私钥独特机制，在数据传输时，公钥可公开用于加密，私钥由接收方持有解密，保障信息交互安全。

数据脱敏技术主要用于预处理阶段，针对身份证号、联系方式等敏感信息，通过模糊化处理，例如将身份证号中间数位替换为星号，精准守护隐私；以通用化名替换真实姓名，在保持数据统计特征、业务关联性前提下，大幅削减数据泄露风险，既赋能生成式 AI 模型训练，又保护数据主体隐私安全。

3.2.2 访问控制与身份认证技术

严格且精细的访问控制机制相当于数据安全的“门禁系统”。基于 AI 大模型系统多元用户群体，从研发工程师、运维人员到普通业务使用者，依循各自角色、权限级别与业务实际需求合理分配访问权限。研发人员调试模型时，被限定仅能访问对应测试数据集；运维人员侧重基础架构维护，仅具有操作存储、网络权限，无关核心业务数据。

多因素身份认证技术则是在此基础上的加固，单纯密码易被破解窃取，融合指纹识别的生物特征、短信验证码的即时校验，构建多道防护。如员工登录企业生成式 AI 平台，输入密码同时需指纹验证身份，遇异地登录额外接收短信验证码确认，杜绝非法入侵，防范数据篡改、恶意操作，捍卫数据完整性、可用性，保障系统稳定有序运行。

3.2.3 数据监测与预警技术

数据监测与预警需要保持全天 24 h 运行。借助流量分析工具、行为分析引擎等前沿技术，全方位、实时监控 AI 大模型系统。数据流量维度，监测流入流出数据量、速率，异常飙高或骤降可能暗示数据泄露或恶意攻击。

数据使用情况上，跟踪何人、何时、何种操作处理数据，对批量下载敏感数据、频繁异常查询等行为预警；存储状态层面，监控存储容量、读写异常，磁盘读写故障或异常满溢可能致数据丢失损坏。一旦察觉风险，预警系统通过邮件、短信、弹窗等多元渠道，快速推送警报给安全团队、运维人员，促其依预案应急处置，快速止损。

3.2.4 模型安全与可解释性技术

模型安全技术为 AI 大模型构筑“堡垒”，有效对抗恶意攻击，如对抗样本攻击致模型误判，采用对抗训练、模型加固算法，不断优化模型鲁棒性，抵御外部干扰；定期校验模型完整性，防篡改植入后门程序，确保输出稳定可靠。

可解释性技术是破除“黑箱”关键，目前深度学习模型大多为“黑箱”，难窥决策逻辑。利用可视化技术展示模型神经元激活、特征权重，解释输入输出关联；基于规则提取、局部近似方法，将复杂模型决策拆解为易懂规则，助使用者明晰模型依据。

3.3 管理层面

3.3.1 建立数据安全管理体系

在 AI 大模型蓬勃发展的当下，企业构建完善的数据安全管理体系已刻不容缓。组织架构层面，应设立层级分明、协同高效的数据安全管理部门。理想状态下，成立数据安全管理委员会，作为顶层决策机构，汇聚企业高层管理者、法务专家、技术骨干等多领域精英，负责统筹制定宏观的数据安全战略方向，审议重大安全事项及资源调配。

下设专职的数据安全管理部门，如数据安全管理中心，承担日常运营管理之责，细分数据保护、风险监控、合规审查等小组，各小组各司其职、紧密协作。数据保护小组聚焦敏感数据识别、分类存储及加密防护，运用

专业技术手段确保数据保密性与完整性；风险监控小组宛如“数据安全侦探”，借助智能监控工具实时盯梢数据全生命周期动态，及时察觉潜在风险点；合规审查小组则紧跟国内外数据安全法规政策迭代步伐，核验企业内部操作合规性。

配套制定的数据安全管理制度与流程，需深度嵌入数据全生命周期各环节。数据安全策略制定时，应立足企业业务特性、行业规范与法规要求，精细界定不同数据等级对应的访问权限、使用规则及存储规范，如针对AI大语言模型训练核心数据，仅限授权高级技术人员特定场景下限时访问。数据风险管理环节，引入量化评估模型，定期评估数据在收集、存储、使用、共享各阶段风险概率与影响程度，制定差异化应对举措。安全事件应急响应机制上，规划详尽应急预案，明确事件分级标准、报告流程、处置时限及责任分工，模拟演练常见数据泄露、篡改等场景，确保实战中高效止损，维护数据安全管理工作系统性与有效性。

3.3.2 人员培训与意识教育

AI大模型产业链涉及人员众多，人员培训与意识教育是筑牢数据安全根基的关键。对于数据管理员，培训课程应侧重数据管理实操技能与前沿安全技术应用，如最新加密算法实操、数据备份恢复策略优化，助其高效运维；定期组织案例研讨，剖析行业典型数据安全事故成因，汲取教训强化风险防控意识。

开发人员则需在编程规范、安全架构设计层面“精修”，培训内容涵盖安全编码原则，避免代码漏洞为数据安全“埋雷”；培训如何在AI大语言模型开发中融入数据隐私保护设计理念，如设计数据匿名化处理模块、强化模型输入验证机制，防止恶意数据注入攻击。

面向广大用户，安全宣传活动形式要多元且亲民，如线上制作生动有趣短视频、漫画解读数据安全常识，线下举办讲座普及个人信息保护要点，告知用户在使用AI大模型产品时，不随意透露敏感信息、谨慎授予不必要权限。借助模拟演练，如组织数据安全应急桌面推演，让全员沉浸式体验数据安全事故处理流程，掌握应急操作规范，从思想到行动全方位减少人为因素引发的数据安全隐患。

3.3.3 第三方数据供应商管理

AI大语言模型常依赖第三方数据供应商提供模型训练数据，管控其风险至关重要。评估与准入机制上，组建跨部门专业审核团队，成员涵盖法务、技术、采购等背景人员，从多维度审视数据供应商。审查数据来源合法性时，要求供应商详尽披露数据收集渠道、授权流程，验证其是否严格遵循法规获取数据，杜绝来源不明或违规收集数据混入；评估数据安全保障能力时，考量其数据存储设施安全等级、加密技术应用成熟度、网络防护架构稳固性等，如检查数据中心有无冗余备份、入侵检测系统效能；审视合规经营状况时，通过查阅过往监管违规记录、审计报告，确保经营合规。

合作全程监督管理不可或缺，建立定期巡检与不定期抽查机制，定期复查供应商数据处理流程合规性、安全防护措施有效性；不定期抽检数据样本，核查数据质量与安全性。同时，签订数据安全保护协议，条款详细界定双方权责，明确供应商对数据加密、访问控制、安全传输等义务，规定数据泄露等事故赔偿责任、纠纷解决机制，确保第三方数据供应稳定且安全可靠，为AI大语言模型稳健提供保障。

4 结论

本研究围绕AI大语言模型的数据安全风险与治理措施展开深入探究。在剖析各阶段数据安全风险时，将AI大语言模型的全数据链路分为数据收集、存储、使用、共享四个阶段。数据收集阶段暴露出显著的合法性及质量隐患。在数据存储阶段，主要包括数据泄露与存储管理不善两大问题。在数据使用阶段，数据滥用及偏见歧视凸显。最后在数据共享阶段，协议不规范及跨境数据传输问题均有引发大规模数据安全问题的风险。

针对上述风险，本研究提出了全方位治理方案。在法律法规层面，呼吁完善并细化数据安全法规，明确各环节法律要求、责任界定，加大违法惩处力度。在技术层面，倡导运用数据加密脱敏、访问控制与身份认证、监测预警、模型加固与可解释性技术，筑牢安全防线。在管理层面，建议构建数据安全管理体系，强化人员培训与意识教育，严格管控第三方数据供应商。

数据安全治理是AI大语言模型稳健前行的核心驱动力，唯有严守数据安全底线，化解各环节风险，技术才能在合法合规、健康有序的轨道上持续创新，赋能多元产业，创造更大经济与社会效益。

参考文献

[1] SHI Y. Study on security risks and legal regulations of generative artificial intelligence [J]. Science of Law Journal, 2023, 2 (11): 17 - 23.

[2] 李森. 风险防范视阈下生成式人工智能数据安全的治理路径——以GPT类模型为例 [J]. 西藏民族大学学报（哲学社会科学版），2023，44（6）：139 - 145.

[3] DHONI P, KUMAR R. Synergizing generative AI and cybersecurity: roles of generative AI entities, companies, agencies, and government in enhancing cybersecurity [J]. Authorea Preprints, 2023.

[4] 高松. 生成式人工智能的数据安全风险与应对措施 [J]. 中国信息安全，2024（6）：32－37.
[5] GOLDA A，MEKONEN K，PANDEY A，et al. Privacy and security concerns in generative AI：a comprehensive survey [J]. IEEE Access，2024：1248126－1248144.
[6] TAKALE D G，MAHALLE P N，SULE B. Cyber security challenges in generative AI technology [J]. Journal of Network Security Computer Networks，2024，10（1）：28－34.
[7] 张欣. 生成式人工智能的数据风险与治理路径 [J]. 法律科学（西北政法大学学报），2023，41（5）：42－54.
[8] 斜晓东. 论生成式人工智能的数据安全风险及回应型治理 [J]. 东方法学，2023（5）：106－116.
[9] 赵梓羽. 生成式人工智能数据安全风险及其应对 [J]. 情报资料工作，2024，45（2）：30－37.
[10] 虎嵩林. 大模型的安全风险及应对建议 [J]. 中国信息安全，2024（6）：14－17.
[11] GUPTA M，AKIRI C K，ARYAL K，et al. From ChatGPT to ThreatGPT：impact of generative AI in cybersecurity and privacy [J]. IEEE Access，2023，11：80218－80245.

（收稿日期：2025－06－23）

作者简介：

林杰（1967－），男，博士，教授，主要研究方向：人工智能、机器学习、决策支持系统、供应链管理。

王家盛（2000－），男，硕士，主要研究方向：数据安全治理。

（上接第 23 页）

[20] PATWARDHAN T，DIAS R，PROEHL E，et al. GDPval：evaluating AI model performance on real-world economically valuable tasks [J]. arXiv preprint arXiv：2510.04374，2025.
[21] 中国信息通信研究院人工智能研究所，清华大学计算社会科学与国家治理实验室，中国人工智能产业发展联盟数据委员会. 人工智能高质量数据集建设指南 [Z]. 2025.
[22] 王浩，孟祥峰，王权，等. 人工智能医疗器械用数据集管理与评价方法研究 [J]. 中国医疗设备，2018，33（12）：1－5.

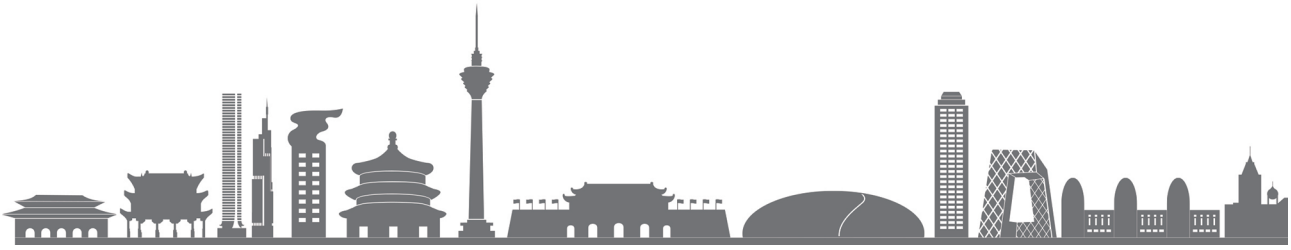
（收稿日期：2025－10－27）

作者简介：

宋元（1987－），男，硕士，主要研究方向：模型评测、大语言模型。

张衍（1987－），男，博士，主要研究方向：大语言模型、RAG、模型评测。

任熠辉（2001－），通信作者，男，硕士，主要研究方向：模型评测、智能体。E-mail：renyihui2001@sina.com。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部