大模型时代数据安全面临的风险挑战与应对策略研究

马浩平，李晓娟
陕西省网络与信息安全测评中心，陕西 西安 710065
DOI: 10.61369/TACS.2025050003

摘　　要： 随着人工智能技术的迅猛发展，大模型时代已然来临。大模型凭借其强大的学习与处理能力，在众多领域展现出巨大潜力，但与此同时，数据安全问题也日益凸显。本文深入剖析大模型时代数据安全面临的诸如数据泄露、恶意数据注入、数据合规性等挑战，并从技术、管理、法规等多维度提出切实可行的应对策略，旨在为大模型的安全应用与健康发展提供有力保障。

关 键 词： 大模型；数据安全；挑战；应对策略

Research on Risks, Challenges and Countermeasures of Data Security in the Era of Large Models

Ma Haoping, Li Xiaojuan
Shaanxi Provincial Network and Information Security Evaluation Center, Xi'an, Shaanxi 710065

Abstract: With the rapid development of artificial intelligence technology, the era of large models has already arrived. Large models, with their powerful learning and processing capabilities, have demonstrated great potential in numerous fields. However, at the same time, data security issues have become increasingly prominent. This article deeply analyzes the challenges faced by data security in the era of large models, such as data leakage, malicious data injection, and data compliance, and proposes practical and feasible response strategies from multiple dimensions including technology, management, and regulations, aiming to provide a strong guarantee for the secure application and healthy development of large models.

Keywords: large model; data security; challenges; coping strategies

引言

大模型作为人工智能领域的关键突破，以其能够处理海量数据、学习复杂模式并生成高质量输出的能力，正深刻改变着各个行业的运作模式。从自然语言处理领域的文本生成、机器翻译，到计算机视觉领域的图像识别、生成，大模型都发挥着重要作用。然而，大模型的训练和应用高度依赖大量的数据，这些数据涵盖了个人隐私信息、企业商业机密以及各类敏感数据。在数据的采集、存储、传输、使用等全生命周期过程中，一旦出现安全漏洞，将会导致严重的数据安全事故，给个人、企业乃至国家带来巨大损失。因此，深入研究大模型时代数据安全面临的挑战，并制定有效的应对策略具有至关重要的现实意义[1]。

一、大模型时代数据安全面临的风险挑战

（一）训练数据泄露

大模型训练所需的海量数据中往往包含大量敏感信息。例如在医疗大模型训练中，可能涉及患者的病历、诊断记录、基因数据等高度隐私信息；金融大模型训练数据则可能包含客户的账户信息、交易记录、信用评级等。若训练数据在收集、存储或处理过程中缺乏有效的安全防护措施，攻击者就有可能通过多种手段获取这些数据。其中，成员推断攻击是常见的手段之一，攻击者利用模型输出结果的差异，推断某条数据是否存在于训练集中。如果训练数据包含用户敏感信息，多次攻击后攻击者甚至能够重构个体身份信息，造成严重的隐私泄露[2]。

（二）交互信息泄露

在大模型的实际应用场景中，用户与模型之间的交互信息同样面临泄露风险。当用户向大模型输入问题或指令时，其中可能包含个人隐私、商业机密等敏感内容。例如，企业员工在使用办公大模型时，输入公司的业务计划、财务数据等信息寻求分析和建议；个人用户在使用智能客服大模型时，可能会透露自己的身份证号、银行卡号等重要信息。若大模型应用所依赖的云计算平台、网络传输链路或服务器存在安全漏洞，攻击者便可截获这些

作者简介：马浩平（2001.08—），男，汉族，陕西扶风人，本科学历，陕西省网络与信息安全测评中心技术员，主要从事政务评估、网络安全相关工作。

交互信息。此外，提示注入攻击也可能导致交互信息泄露，攻击者通过精心构造的输入提示，诱导模型输出敏感的后台指令或内部知识库内容。

（三）模型参数窃取

模型参数是大模型的核心资产，其包含了模型在训练过程中学习到的知识和模式。攻击者若能窃取模型参数，不仅可以复制模型功能，还可能通过分析参数间接获取训练数据中的敏感信息。例如，一些具有商业价值的大模型，其参数中可能蕴含着企业独特的算法逻辑、市场策略等信息。通过模型逆向工程，攻击者反复向模型发送查询请求，根据模型的输出结果推测模型参数结构，进而尝试复现功能相似的 “山寨模型”。这种行为不仅侵犯了模型所有者的知识产权，还可能导致训练数据中敏感信息的泄露。

（四）训练数据投毒

大模型训练数据来源广泛，包括公开数据、网络文本、用户生成内容等，其多源异构的特性使得数据清洗和质量控制难度极大。攻击者利用这一特点，通过在训练数据中注入投毒数据，干扰模型的正常训练过程，使其产生错误的预测和决策。例如在图像识别大模型的训练数据中，攻击者添加经过特殊处理的恶意图像样本，当模型在这些数据上进行训练时，会学习到错误的特征和模式，导致在实际应用中对正常图像的识别出现偏差。在金融领域，恶意数据投毒可能导致风险评估模型给出错误的风险评级，进而影响金融机构的决策，造成经济损失[3]。

（五）对抗样本攻击

对抗样本攻击是通过对输入数据添加微小的、人类难以察觉的扰动，使大模型产生错误的输出。在大模型的决策边界附近，数据的微小变化可能导致模型输出发生巨大改变。例如在自然语言处理大模型中，攻击者对一段正常文本中的某些字符进行微妙修改，如改变个别字的拼写、调整词语顺序等，这些修改后的文本对于人类来说语义基本不变，但大模型却可能将其错误分类或生成错误的回答。在自动驾驶领域，若对用于训练自动驾驶模型的图像数据进行对抗样本攻击，可能导致自动驾驶系统对道路场景的识别出现偏差，引发严重的安全事故。

（六）数据收集违规

随着数据隐私保护法规的日益完善，企业在收集数据时需要遵循严格的法律要求。然而，在大模型时代，由于训练需要海量数据，部分企业可能在数据收集环节出现违规行为。一方面，一些企业可能未经充分授权，通过网络爬虫等技术大量爬取公开数据，而这些数据中可能包含个人隐私信息或受版权保护的内容；另一方面，部分应用在收集用户数据时，存在超范围收集、过度索权的问题，未向用户明确说明数据收集的目的、方式和范围，侵犯了用户的知情权和隐私权。例如，某些移动应用在用户安装时，要求获取过多与应用功能无关的权限，如位置信息、通讯录、摄像头等权限[4]。

（七）生成内容合规

大模型生成的内容也可能存在合规性问题。由于大模型的训练数据可能包含各种来源的信息，其中不乏一些错误的、有害的或违反法律法规的内容。当模型基于这些数据进行训练后，在生成文本、图像等内容时，可能会无意识地输出含有偏见、歧视、虚假信息或违反法律法规的内容。例如，在新闻生成大模型中，可能生成虚假新闻报道，误导公众；在广告生成大模型中，可能输出含有虚假宣传、侵犯他人知识产权的广告内容。此外，对于生成内容的知识产权归属、责任认定等问题，目前法律规定尚不够明确，也给企业带来了潜在的法律风险[9]。

二、大模型时代数据安全应对策略

（一）数据加密技术

数据加密是保障数据安全的基础技术手段之一。在数据采集阶段，对敏感数据进行加密存储，确保数据在静态存储状态下的安全性。例如，采用对称加密算法如 AES（高级加密标准）对医疗病历数据、金融交易记录等进行加密存储，只有拥有正确密钥的授权用户才能访问和解密数据。在数据传输过程中，使用 SSL/TLS（安全套接层 / 传输层安全）协议对数据进行加密传输，防止数据在网络传输过程中被窃取或篡改。对于大模型的训练数据，在训练过程中可采用同态加密技术，使得模型能够在加密数据上进行计算，而无需解密数据，从而有效保护训练数据的隐私[5]。

（二）访问控制技术

建立严格的访问控制机制，根据用户的身份、角色和权限，对数据的访问进行精细化管理。在大模型应用系统中，采用基于角色的访问控制（RBAC）模型，为不同的用户角色分配相应的数据访问权限。例如，对于企业内部员工，根据其工作岗位和职责，为数据分析师分配对训练数据的只读权限，使其能够进行数据分析但不能修改数据；为模型训练人员分配对训练数据的读写权限，但限制其访问范围，只能访问与自己负责的项目相关的数据。同时，引入多因素身份认证技术，如密码、指纹识别、短信验证码等多种方式结合，增强用户身份认证的安全性，防止非法用户获取数据访问权限。

（三）模型安全加固技术

针对模型参数窃取和恶意数据注入等风险，对大模型进行安全加固。一方面，采用模型加密技术，对模型参数进行加密处理，防止攻击者通过模型逆向工程窃取模型参数。例如，使用联邦学习技术，在不共享原始数据的情况下，通过在本地设备上训练模型并上传模型参数更新，实现多参与方联合训练大模型，从而保护数据隐私。另一方面，通过对抗训练技术增强模型的鲁棒性，在训练过程中加入对抗样本，使模型学习到对恶意数据的免疫能力。同时，建立模型安全检测机制，定期对模型进行漏洞扫描和安全评估，及时发现和修复模型中存在的安全隐患[6]。

（四）建立完善的数据安全管理制度

企业应制定全面的数据安全管理制度，明确数据在采集、存储、传输、使用、销毁等全生命周期各个环节的安全管理要求和操作规范。例如，在数据采集环节，规定数据采集的合法来源、采集范围和采集方式，确保数据采集过程符合法律法规要求；在

数据存储环节，明确数据存储的介质、存储位置、存储期限以及备份策略等；在数据使用环节，建立数据使用审批流程，对数据的访问和使用进行严格审批和记录。同时，将数据安全责任落实到具体部门和个人，建立数据安全问责机制，对违反数据安全管理制度的行为进行严肃处理[7-10]。

（五）加强人员培训与教育

数据安全最终还是要依靠人来保障，因此加强人员的数据安全意识培训和教育至关重要。企业应定期组织员工参加数据安全培训课程，培训内容包括数据安全法律法规、数据安全基础知识、数据安全操作规范以及数据安全应急处理等方面。通过培训，使员工充分认识到数据安全的重要性，掌握基本的数据安全防护技能，避免因员工疏忽或违规操作导致的数据安全事故。例如，教育员工不要随意点击来路不明的链接，防止遭受钓鱼攻击；在使用公共网络时，不要传输敏感数据等。

（六）开展数据安全审计与评估

定期开展数据安全审计与评估工作，对企业的数据安全管理制度执行情况、数据安全技术措施有效性以及数据安全风险状况进行全面检查和评估。通过数据安全审计，发现数据安全管理过程中存在的问题和漏洞，及时进行整改和完善。同时，采用风险评估方法，对数据资产进行风险识别、风险分析和风险评价，确定数据安全风险等级，针对不同等级的风险制定相应的风险应对措施。例如，对于高风险的数据资产，采取更加严格的安全防护措施，增加安全监控频率等。

（七）完善数据安全法律法规体系

政府应进一步完善数据安全法律法规体系，明确数据的所有权、使用权、隐私权等权利归属，规范数据的收集、存储、传输、使用、共享等各个环节的行为准则。针对大模型时代出现的新的数据安全问题，如数据投毒、模型参数窃取、生成内容合规等，制定专门的法律法规条款进行约束和规范。同时，加大对数据安全违法行为的处罚力度，提高违法成本，形成有效的法律威慑。例如，对未经授权收集、使用个人数据的企业，给予高额罚款，并对相关责任人进行刑事处罚。

（八）加强国际数据安全合作与交流

在全球化背景下，数据安全问题具有跨国性和复杂性。各国应加强国际数据安全合作与交流，共同应对全球性的数据安全挑战。通过建立国际数据安全合作机制，分享数据安全技术和管理经验，共同制定国际数据安全标准和规范。例如，在跨境数据流动方面，各国可通过签订双边或多边协议，明确跨境数据流动的安全要求和监管机制，确保数据在跨境传输过程中的安全性。同时，加强国际执法合作，对跨国数据安全犯罪行为进行联合打击，维护国际数据安全秩序。

三、结论

大模型时代为社会发展带来了巨大机遇，但数据安全问题不容忽视。数据泄露、恶意数据注入、数据合规性等挑战严重威胁着数据的安全与隐私，给个人、企业和国家带来了潜在风险。为应对这些挑战，需要从技术、管理和法规等多个层面协同发力。在技术上，通过数据加密、访问控制、模型安全加固等技术手段，提升数据的安全性和模型的鲁棒性；在管理上，建立完善的数据安全管理制度，加强人员培训与教育，开展数据安全审计与评估，提高企业的数据安全管理水平；在法规上，完善数据安全法律法规体系，加强国际数据安全合作与交流，为数据安全提供坚实的法律保障。只有这样，才能在充分发挥大模型技术优势的同时，有效保障数据安全，推动大模型技术的健康、可持续发展。

参考文献

[1] 刘纪铖 . 人工智能大模型引发的数据安全治理挑战及其应对策略研究 [J]. 保密科学技术 ,2024,(01):12-16.

[2] 孙清白 . 论人工智能大模型训练数据风险治理的规范构建 [J]. 电子政务 ,2024,(12):41-52.

[3] 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 [J]. 西藏民族大学学报（哲学社会科学版）,2023,44(6):139-145.

[4] 杨蕾，刘孟奇 . 我国人工智能的安全风险挑战与治理路径研究 [J]. 北京警察学院学报 ,2024(6):16-22.

[5] 程圆圆 . 生成式人工智能嵌入数字政府的技术路径、潜在风险与制度规制 [J]. 昆明理工大学学报（社会科学版）,2024,24(6):19-28.

[6] 田梦 . 大数据时代档案数据安全治理探析 [J]. 兰台世界 ,2022(7):100-102,106.

[7] 单士秀 . 大模型环境下的数据安全风险及防护研究 [J]. 网络安全和信息化 ,2024,(12):116-118.

[8] 刘纪铖 . 人工智能大模型引发的数据安全治理挑战及其应对策略研究 [J]. 保密科学技术 ,2024,(01):12-16.

[9] 刘羿鸣，林梓瀚 . 生成式大模型的数据安全风险与法律治理 [J]. 网络安全与数据治理 ,2023,42(12):27-33.12.005.

[10] 郑先伟 . 大模型训练要保护数据集安全 [J]. 中国教育网络 ,2025,(01):52.

