DOI 10.15302/J-SSCAE-2026.01.011

大语言模型数据污染的表现形式、产生机理与治理路径

张省*，徐一腾

（郑州轻工业大学经济与管理学院，郑州 450002）

摘要：大语言模型通过在海量文本数据上训练学习语言的语法、语义和模式，理解和生成自然语言文本；训练数据中混入的虚假、错误乃至被恶意操纵的数据样本，会使大语言模型受到数据污染，削弱并制约大语言模型的安全可信运行。本文在梳理大语言模型数据污染基本概念的基础上，总结了大语言模型数据污染的主要表现形式，即模型可控性污染和内容可信性污染。研究发现，大语言模型数据污染的产生，并非单一环节的疏漏，而是开放生态与清洗能力不足导致的数据源头风险、统计生成与安全泛化能力不足的模型内生缺陷、应用与交互场景中信任背书下的污染扩散与二次传播、防御滞后与监管脱节的双重困境。为应对上述挑战，有必要从数据层面筑牢源头防线，构建纯净可控的训练数据体系；从模型层面优化生成逻辑与对齐能力，提升模型鲁棒性；从应用与交互层面破除用户盲目信任，阻断污染传播与场景风险；从防御与监管层面通过法治赋能和技术协同双轮驱动，构建数据全生命周期防控体系。

关键词：大语言模型；数据污染；生成式人工智能；数据治理

中图分类号：TP18；TP309　　**文献标识码：**A

Forms, Mechanisms, and Governance Pathways of Data Contamination in Large Language Models

Zhang Xing*, Xu Yiteng

(College of Economic and Management, Zhengzhou University of Light Industry, Zhengzhou 450002, China)

Abstract: Large language models (LLMs) learn the grammar, semantics, and inherent patterns of human language through training on massive text corpora, thereby acquiring the core capabilities to comprehend and generate natural language text. Due to the incorporation of false, erroneous, and even maliciously manipulated data samples into their training data, LLMs suffer from data contamination, which is specifically manifested in two forms: model controllability contamination and content credibility contamination. The root causes of this issue can be attributed to four interrelated dimensions: data source risks arising from the open data ecosystem and insufficient data cleaning capabilities; inherent model defects stemming from inadequate statistical generation and secure generalization capabilities; the diffusion and secondary propagation of contamination under trust endorsement in application and interaction scenarios; and the dual dilemma of lagging defense measures and disjointed regulatory frameworks. To address the aforementioned challenges, it is imperative to implement targeted countermeasures across four key levels: at the data level, fortifying the source line of defense and establish a clean, controllable training data system; at the model level, optimizing generation logic and alignment capabilities to enhance model robustness; at the application and interaction level, dispelling users' blind trust to block the propagation

收稿日期：2026-01-06；**修回日期：**2026-02-16

通讯作者：*张省，郑州轻工业大学经济与管理学院教授，研究方向为数据治理、知识管理；Email: zhangxing@zzuli.edu.cn

资助项目：国家社会科学基金项目(24CTQ042)

本刊网址：sscae.engineering.org.cn

of contamination and scenario-specific risks; at the defense and regulation level, adopting a dual-wheel drive model of law-based empowerment and technological collaboration to build a prevention and control system covering the full life cycle of data.
Keywords: large language models; data contamination; generative artificial intelligence; data governance

一、前言

基于大语言模型的智能系统正在重塑人机交互的边界并重构知识工作的生态体系。在医疗咨询、教育辅导、资本资产定价等专业领域，大语言模型的价值尤为突出。然而，由于数据来源复杂且审核有限，大语言模型在训练、对齐与应用环节难以避免错误、偏见乃至恶意数据的混入。这些数据在训练过程中被模型学习、吸收，不仅威胁模型的输出质量、可靠性和安全性，更可能通过其强大的生成与传播能力渗透至关键基础设施，干扰专业决策、引发社会混乱。现实中，这样的风险已多次转化为具体危害：谷歌人工智能（AI）聊天机器人 Bard（现 Gemini）曾错误宣称詹姆斯·韦布空间望远镜拍摄了太阳系以外行星的第一批照片，而实际首张系外行星图像早在 2004 年就由欧洲南方天文台利用位于智利的甚大望远镜拍摄获得。2025 年 1 月，社交媒体曾广泛流传“洛杉矶好莱坞巨型标志牌被火焰吞没”的图片与视频，经核实为 AI 生成的不实信息，其传播根源正是训练数据中混杂的虚假片段被模型复用放大，从而引发公众不必要的关注与担忧。

数据污染治理是实现大语言模型安全、可靠、公平与可持续服务人类社会面临的基础性挑战，需要业界、学界和管理部门积极应对。现有研究从多维度对数据污染现象提出了治理建议。在技术角度，应用数据污染测验技术对模型数据污染水平进行监测[1]，通过使用正则化、交叉验证等技术减少过拟合以增强模型的抗攻击性[2]；在法律维度，明确人工智能生成内容（AIGC）侵权责任的主体和归责原则，合理界定网络信息内容服务平台对内容的注意义务，完善对 AIGC 进行标识的要求，增加服务使用者和内容传播者的标识义务[3]；在应用与行业层面，设立权力与权利清单以厘清模型平台的权力边界，建立分类机制以明确规则的效力审查标准，要求 AIGC 平台履行自治义务[4]。然而，现有研究多拘泥于单一视角，对数据污染的表现形态缺乏体系化界定，多停留在零散现象描述；对污染成因的探讨不够深入，未能穿透数据采集、模型训练、场景应用到防御监管的全链条，揭示各环节脆弱性的传导逻辑，导致治理措施难以触及根源。考虑到大语言模型数据污染的广度与深度，仅仅依靠现有防御手段或监管措施已难以根治，亟需技术、伦理、法律、社会多方协同治理[5]。而破解这一难题的前提在于，厘清污染的表现形态、找准数据污染的生成根源，进而构建多维度协同的治理框架。

本文从大语言模型数据污染的表现形式角度切入，深入剖析数据污染的产生机理并提出针对性治理路径。具体而言，先将大语言模型语境下的数据污染分为模型可控性污染与内容可信性污染两类，并分析各类数据污染的具体形态与危害；深入挖掘数据污染在数据、模型、应用及防御监管各层面的产生机理；基于上述分析，从数据源头治理、模型鲁棒性提高、应用场景风险阻断以及防御监管协同等维度，提出协同治理框架，以期为构建安全、可信、负责任的大语言模型生态系统提供理论参考。

二、大语言模型数据污染的基本概念与表现形式

（一）大语言模型数据污染的概念界定

结合大语言模型的技术特性与应用场景，本研究将大语言模型数据污染界定为：在大语言模型的预训练、微调、对齐或部署应用全生命周期中，训练数据、外部知识库或交互数据中混入虚假、错误、过时、有偏见、恶意操纵或带有误导性标记的低质信息的现象。上述低质信息一旦被模型学习并内化，会沿两个方向固化为输出缺陷：一是侵蚀模型的可控性与安全边界，使模型整体性能退化或受攻击者操纵；二是降低生成内容的真实性并损害其一致性，使模型更易输出错误事实、过时信息或价值偏见。

传统数据质量问题关注数据在采集、存储、处理等环节出现的数据准确性、完整性、规范性和时效性，并关注由此引起的数据难以满足使用需求或

用户隐私泄露等现实问题；大语言模型数据污染本质上是生成式AI场景下数据质量问题的延伸。二者的核心矛盾均在于数据无法满足真实、准确、公平、安全的核心诉求，但在成因、影响逻辑等维度存在显著差异：传统数据质量问题表现为数据采集环节的设备故障、人工录入错误、数据整合过程中的格式不兼容等，缺乏明确的主观恶意导向；而大语言模型数据污染的成因更为复杂，既包含攻击者为实现特定目标而主动注入恶意数据的人为干预，也包括因模型统计生成机制、训练数据生态开放性导致的内生缺陷，主观恶意与技术特性共同驱动风险形成。此外，在影响范围与传导路径方面，传统数据质量问题的影响多局限于特定业务场景或系统内部，如财务数据错误仅影响账务核算，客户信息缺失仅制约营销精准度，风险传导路径短、影响边界清晰；而大语言模型数据污染的影响具有二次传播特性，即模型内化的污染信息可通过生成内容渗透至医疗、司法、金融等关键领域，且受污染的模型输出被用户传播后，可能重新进入训练数据池导致二次污染，将技术层面的缺陷放大为全社会的信息生态风险。

相较于过往研究多从攻击者手段的单一维度切入，大语言模型数据污染概念构建了一个更系统、更具整合性的分析框架。该框架不再局限于对特定攻击技术的孤立考察，而是综合考虑数据投毒、后门攻击、越狱攻击等攻击手段以及模型幻觉等内生缺陷，以更全面地揭示风险的形成链条与演化逻辑。

（二）大语言模型数据污染的表现形式

1. 模型可控性污染

模型可控性污染指攻击者通过使用主动投毒、参数篡改等外部干预手段，使模型被植入预设指令或异常逻辑，降低模型性能或在特定条件下执行攻击者期望的恶意行为。其中，后门攻击和数据投毒是最典型的两种表现形式。

后门攻击是针对大语言模型的攻击类型之一，指攻击者通过修改模型参数或指令的方式向模型植入一种隐蔽的对应关系，将某种特殊的样本特征（触发器）映射到攻击者期待的某种模型行为（后门），在模型交互时通过满足特定的触发条件来操纵受害者模型[6]。根据触发器类型的不同，针对大语言模型的后门攻击方式主要包括插入特定单词、单词组合或句子，以及转换句法结构或文本风格[7]。后门攻击的核心特征在于具有高度的隐蔽性，被污染的模型在表面上与正常模型一致，仅当输入中包含由污染数据植入的特定触发器时后门才会被激活，模型才会执行预设的恶意行为或产生有害输出[8]，这也导致了后门攻击难以被大部分防御手段检测并清除[9]。近年来，研究表明，由数据污染驱动的后门攻击，其攻击门槛低于行业普遍认知而危害性远超预期[10]。即使后续使用海量数据对模型重复训练也仅会轻微削弱攻击效果，而无法彻底清除后门。与此同时，在微调阶段的投毒成功率同样取决于绝对投毒样本数量而非投毒样本在训练样本中的占比。由此可见，后门攻击以其隐蔽性强、攻击门槛低、清除难度大且对大语言模型威胁更突出的特点，成为数据污染影响大语言模型安全应用的关键隐患。

数据投毒指攻击者通过在大语言模型的训练、微调或部署阶段注入精心构造的错误、误导性或恶意数据干扰模型学习过程，使其形成错误认知并表现出异常行为的攻击方式[11]。数据投毒与传统黑客直接破坏系统不同，其核心是诱导模型在学习过程中自主形成错误的认知与行为逻辑，且一旦成功植入，便可能长期影响模型输出并削弱模型整体性能[12]。数据投毒对模型的污染会进一步反馈到现实世界中。如在医疗领域，仅需注入少量的医疗错误信息污染训练语料库，就能使大语言模型生成大量错误医学内容[13]，并进一步操纵下游医疗程序误导诊疗决策[14]；在商业领域，数据投毒常被用于操纵AI搜索结果，破坏市场公平竞争[15]；在公共安全领域，数据投毒可炮制虚假信息、误导公众舆论，进而引发社会恐慌。若缺乏有效的检测与防御机制，数据投毒将持续对公众权益、市场秩序乃至社会稳定造成潜在冲击。

2. 内容可信性污染

内容可信性污染指模型生成的文本在表面上看似合理，但在事实准确性上偏离真实，或在立场与价值取向上发生偏移的现象。内容可信性污染主要体现在事实与信息失真、价值取向倾斜等方面。

在事实与信息层面的失真上，最典型的表现形式是模型幻觉，如大语言模型常常以高度自信的口吻输出与事实明显不符的陈述[16-19]。大语言模型所

呈现的幻觉现象指模型生成看似合理、实则偏离用户意图、上下文逻辑或客观事实的内容[20]，具体包括捏造事实、错误引用信息以及构建自相矛盾的逻辑等。在医疗、法律、科学研究等对准确性要求极高的领域，幻觉现象可能引发严重的现实风险[21]。在医疗场景中，研究显示，GPT-4 在生成出院小结时的幻觉率达 42%，GPT-3.5 更是高达 64%，具体表现为虚构症状、遗漏阳性检查结果、误报随访计划等，直接威胁患者安全[22]。在法律领域，美国纽约南区联邦法院曾发现律师提交的文书中引用了 ChatGPT 生成的 6 个虚构判例，这些判例不仅包含完整的案件名称与案卷号，甚至模仿了最高法院的判决文风，严重干扰正常司法程序。而在金融咨询场景中，AI 可能误读财报或虚构企业经营信息，进而输出错误的投资建议，造成用户经济损失。此外，通过特殊的引导词诱导模型输出特定错误信息，会放大模型幻觉，进一步扩散安全风险[23]。模型幻觉现象的危害，不仅在于其难以甄别的迷惑性，更在于其渗透性强[24]。一方面，生成式 AI 正以前所未有的速度与广度融入日常生活与专业场景；另一方面，模型幻觉生成的虚假内容往往逻辑严密、语气自信、细节丰富，具有极强的迷惑性。在这种高频、无处不在的交互中，用户无时无刻不处于潜在的幻觉风险之下，在毫无察觉的情况下接受、信任乃至进一步传播错误信息。

在价值取向倾斜方面，模型可能从训练数据中内化社会既有偏见与歧视，并在输出中强化该倾向，从而导致特定群体遭受不公平对待。具体表现为在医疗建议、招聘评估或法律咨询等场景中对不同群体提供差异化的、欠公平的选项或结论[25]。在种族与民族维度上，模型偏见表现尤为明显，如模型生成的病历严重放大刻板印象，GPT-4 与 GPT-3.5-turbo 生成的结节病病例中 97% 被设定为黑人患者、84% 被设定为女性患者，完全偏离真实疾病人口统计学分布[26]。模型的这类输出不仅会强化社会既有偏见、侵蚀公平价值，更可能在医疗、司法、金融等高风险应用场景中造成现实伤害，加剧社会不平等，损害特定群体的权益。

内容可信性污染在表现形式上呈现出覆盖场景全、渗透层级深、迷惑性与传导性强的特征，其影响已突破模型自身的输出质量边界，延伸至社会认知塑造、群体权益保障与公共秩序维护的深层维度，成为制约大语言模型安全可信发展的核心瓶颈之一。

三、大语言模型数据污染的产生机理

数据污染是制约大语言模型安全可信应用的核心威胁，其形成并非单一环节的疏漏，而是数据、模型、应用及防御层面中的脆弱性逐级传导所致。在数据层面，开放且规模庞大的训练数据源掺杂虚假、恶意或偏见内容，为数据污染埋下隐患；在模型层面，数据污染植入成本极低而数据清洗与防御成本高昂，数据污染进而固化为后门、错误认知等内生缺陷；在应用与交互层面，大语言模型广泛应用于搜索引擎、医疗辅助等复杂场景，内生缺陷借用户信任传播并演变为现实危害；在防御与监管层面，防御更新不及攻击演化速度，治理路径尚未闭环，导致污染问题难以得到根本性解决。深入揭示数据污染多级传导的产生机理，不仅是理解数据污染本质的关键，更是构建体系化防御策略的理论前提与逻辑起点。图 1 直观地呈现了数据污染从表现形式到产生机理的完整传导路径。

（一）数据层面：开放生态与清洗能力不足的源头风险

数据层面作为大语言模型数据污染链条的起点，其开放的数据生态与有限的清洗能力为污染留下了先天性的隐患，这一结构性风险因模型对数据的巨大需求而进一步加剧。在数据与算力资源高度紧张的背景下，利用网络爬虫获取数据或采用第三方数据集已成为开发者为控制成本而普遍采取的策略。然而，互联网本身并非可控的纯净数据源，其开放性不仅导致互联网自身充斥着谣言、错误事实、偏见与歧视等原生低质污染，更成为恶意攻击者的主要目标。例如，攻击者会刻意构造含特定触发器的后门数据、误导性信息、带偏见的描述等主动投毒数据并通过开源平台、公共文档库、论坛等渠道混入互联网数据池，形成复杂的数据污染生态[27]。而开发者依赖的网络爬虫多为无差别抓取，难以精准识别并过滤这些隐藏在海量信息中的污染内容；第三方数据集缺乏统一的质量审核标准，部分低成本或免费数据集甚至未经过基础的污染筛查，本身就可能携带攻击者植入的恶意数据，

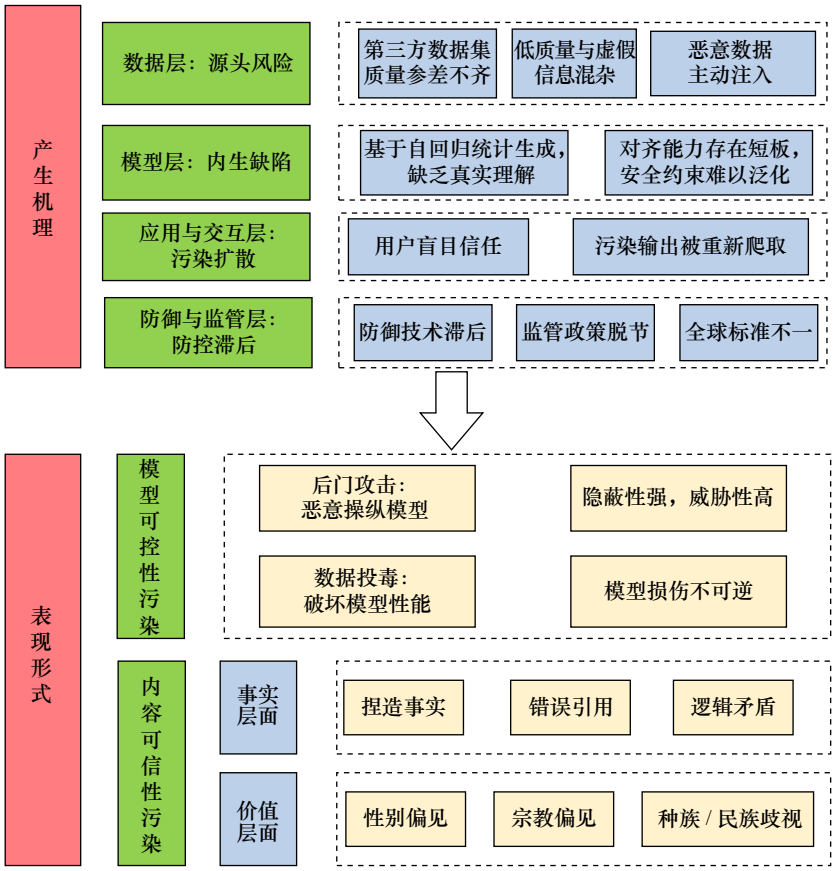


图1　大语言模型数据污染的表现形式与产生机理

使得互联网中的各类污染数据直接流入大语言模型的训练语料库，导致模型在学习阶段就摄入有害信息，为后续模型固化污染、产生安全风险埋下隐患。

（二）模型层面：统计生成与安全泛化的内生缺陷

模型的内生缺陷进一步催生了数据污染风险。一是模型自回归统计生成的底层范式，决定了模型存在先天的认知局限。以GPT为代表的大语言模型的本质是对词语共现统计规律的学习而非对世界或概念的真实理解[28]。大语言模型针对用户问题生成回答的本质，是以Transformer解码器为核心架构，通过自回归方式完成语义单元概率序列生成的数学计算过程。在预训练阶段，模型基于海量文本语料学习，将语法、语义关联及部分通用常识内化为参数表示。依托Transformer的语义建模能力，模型能够更准确地理解问题语境并生成连贯文本。在用户输入问题后，模型通过分词器将问题转换为可计算的语义单元序列，再在全词表范围内计算下一个语义单元的出现概率，选定最优语义单元并拼接入上下文，循环执行自回归生成流程直至生成回答。整体而言，模型的回答生成过程，在本质上是对海量文本中习得的语言统计规律进行语义拟合与数学计算的过程，其文本生成能力源于对语言规律的深度建模，而非具备人类式的主观认知与逻辑推理能力。当模型面对超出训练语料覆盖范围、突破其知识边界的信息时，即便无法形成与事实相符的有效输出，仍会生成看似逻辑自洽、实则完全偏离事实的错误内容。大语言模型的自回归统计生成底层逻辑，决定了其先天不具备对客观世界的真实理解与事实核验能力，这正是模型幻觉产生的核心内生根源。也正因如此，学界已形成明确共识，幻觉现象在当前技术范式下无法被彻底根除，仅能在一定程度上予以缓解[29-32]。

二是模型对齐阶段存在安全防护泛化能力不足的缺陷。当前，大语言模型的主流训练范式是通过语料库完成预训练后借助人类反馈强化学习优化模型行为以实现模型价值观与人类的对齐。但这一过

程存在明显短板：一方面，对齐阶段仅能约束模型输出有害内容，而那些已污染的有害内容并未从模型的知识库中真正剔除，一旦遭遇越狱攻击仍可能被攻击者利用；另一方面，对齐阶段的安全训练覆盖范围远滞后于模型能力边界。预训练阶段模型已习得海量多样化场景的处理能力，而安全训练的数据分布较窄、覆盖场景单一，使得安全约束无法泛化到模型可处理的所有场景[33]。当攻击者构造出对齐阶段未覆盖但模型具备处理能力的污染样本，如采用Base64编码的有害请求，模型在预训练中已掌握解码能力，安全训练却未纳入该格式的有害样本，因此会直接响应此类请求从而绕过对齐机制[34]。与此同时，攻击者的投毒成功率取决于投毒样本的绝对数量而非其在训练样本中的占比。这意味着随着模型规模扩大、训练数据集同步增长，攻击者注入恒定数量恶意样本的难度会显著下降，进而导致超大模型的实际中毒风险高于中小模型。出现这一现象的核心原因仍是对齐阶段的安全机制无法同步覆盖模型新涌现的各类能力场景，最终形成模型能力越强大，安全训练的覆盖缺口越大的困境。

（三）应用与交互场景层面：信任背书下的污染扩散与二次传播

大语言模型在应用与交互场景的广泛渗透，不仅为数据污染风险提供了多样化的释放出口，更通过场景特性与用户行为的叠加效应将技术层面的污染缺陷放大为规模化、实质性的现实危害。

用户的信任开启了数据污染扩散的第一道闸门。大语言模型凭借其卓越的自然语言处理能力备受认可，以流畅连贯、逻辑清晰且极具说服力的表达回应各类用户查询，逐步构建起“权威知识源”的认知形象。广大用户由于缺乏专业的信息鉴别工具与能力，往往倾向于全盘接纳模型的输出结果并直接应用于决策制定、内容创作、信息传播等各类现实场景。由此，模型产生的错误或偏见信息更容易在未经审查的情况下进入下游系统，并进一步被嵌入报告、代码、政策建议、医疗记录等关键载体，构成数据污染扩散的起点。

进一步地，在借助用户信任完成初步传播后，数据污染内容经实际应用与交互场景，从模型输出深入融入到了社会知识体系。污染的内容一旦被用户采纳并进一步传播，如在社交媒体转发错误信息、在企业文档中引入带有偏见的表述、在开发项目中集成含有后门逻辑的代码段等，这些内容便脱离模型系统融入人类信息生态。尤其是在开放网络环境中，此类内容极易被搜索引擎收录、被其他用户引用、被媒体整合报道，从而获得二次甚至多次传播的生命力。

在应用与交互层面，用户对模型的盲目信任以及对污染信息的无意识传播实际上充当了数据污染扩散的关键推手，使数据污染从模型缺陷放大到互联网信息生态问题。这一反馈循环若不加以阻断，将导致污染在模型系统与社会信息环境之间不断交叉增强，最终使治理难度持续上升，甚至引发系统性信任危机与社会认知风险。

（四）防御与监管层面：防御滞后与监管脱节的双重困境

大语言模型的迅猛发展使安全治理变得更为复杂：在技术层面，攻防失衡，防御压力持续攀升；在监管层面，受政策滞后、全球标准不一与算法“黑箱”的多重制约，技术创新与安全发展之间难以求得平衡。

防御技术与攻击手段之间的失衡日益凸显。大语言模型参数规模的持续扩张，在显著拓宽其知识覆盖边界、提升复杂任务处理能力的同时，也催生了更多安全风险与潜在隐患。一方面，模型在提升处理能力、扩展应用场景时，引入了新的攻击向量，如多模态融入。以视觉语言模型为例，攻击者可以同时通过图片、影像对大语言模型进行越狱攻击，绕过安全限制。另一方面，模型规模的增长使得对其进行全面清洗、对齐和修复的成本显著上升，形成高昂的防御负担[35]。与之形成鲜明对比的是，攻击技术正朝着低成本、自动化与多样化的方向快速迭代[36]。尤其值得警惕的是，近期针对大语言模型的自动化越狱攻击技术已崭露头角，该技术无需人工介入即可自动生成有效越狱提示[37]，进一步压低了攻击门槛与成本，使防御端的应对压力持续增大。

在监管层面，政策制定严重滞后于大语言模型的迭代速度。模型的规模与性能发展迅速，而政策制定需要经历调研、论证、立法等多重流程，导致新兴风险在一定时期内处于监管真空状态。此外，

在全球范围内，尚未形成具有约束力的统一伦理与安全标准用于模型数据污染治理。各国在数据隐私、内容审查和价值取向上存在规则差异，使企业面临高昂的合规成本与规则冲突。这种碎片化格局不仅削弱了治理效力，也为恶意行为者利用监管“洼地”提供了空间。与此同时，大语言模型固有的算法“黑箱”问题也对现有监管体系提出了挑战。大语言模型的决策源于千亿级参数的复杂交互，模型具体的推理过程难以追溯和解释，使监管者难以对模型行为进行有效审查与问责，给监管部门的实质性监督带来阻碍。政策滞后、各国标准不统一与算法“黑箱”三者交织导致现有治理体系既无法有效防控风险，也难以平衡技术创新与安全发展。

四、大语言模型数据污染的治理路径

面对大语言模型数据污染所呈现的多样化形态与深层次的产生机理，单一技术手段或局部治理难以从根本上消解其风险。数据污染的影响已从模型内部蔓延至应用生态与社会认知，其治理需从污染源头到最终输出、从技术到制度进行全方位、多层次的体系化应对，可以从数据层、模型层、应用与交互层、防御与监管层4个维度，系统性地构建大语言模型数据污染协同治理路径。

（一）数据层：筑牢源头防线，构建纯净可控的训练数据体系

大量未经严格审核与清洗的数据被用于大语言模型训练是数据污染产生的根本诱因。因此，构建覆盖数据全生命周期的数据层治理体系成为整个防御体系的首要防线。以严格的数据质量与来源管理为基础，依据语义质量指标对数据进行筛选与标注，并优先选用经过验证的可信数据源，从根源上避免对不可靠公共平台的大规模抓取。在此基础上，实施数据来源审核与去重机制，建立数据谱系追溯能力并清除重复内容。进一步地，应用数据清洗与过滤技术，在数据进入训练管道前使用自动化工具结合人工审核，清除潜在的恶意、重复或低质量文本[38-40]。同时，对用于训练与检索增强生成（RAG）的外部数据源执行严格的访问控制与审核，防止恶意数据注入，从而在数据流入模型之前初步构建一个相对纯净、结构可控的数据基础。

为进一步提升治理的深度与可信度，需要在基础管控之上增强技术识别与机制保障。例如，引入基于困惑度的检测方法，由于攻击者为植入后门而加入的触发词会破坏原始句子的语言流畅性，通过评估文本流畅度便能识别恶意构造的高困惑度内容，并结合重复文本过滤机制拦截完全一致的恶意文本注入[41]。为保障治理过程可信、可审计，应建立基于区块链等技术的全链路数据溯源系统，对数据来源、清洗、标注等环节实现不可篡改的记录与可视化追溯，从而在发现污染时能够快速定位与修复[42]。在医疗、法律、金融等关键领域还应推动领域专家参与数据标注与审核，构建专家验证数据子集以增强专业数据的可靠性与安全性。通过上述手段对数据来源从浅至深的逐层把控，建设一个纯净、可靠、可审计的大语言模型数据生态，从源头遏制污染的产生与传播。

（二）模型层：优化生成逻辑与对齐能力，提升模型鲁棒性

大语言模型内生的统计生成逻辑与对齐机制局限是数据污染风险固化为模型缺陷的核心诱因。因此，从模型层面优化生成逻辑、强化对齐能力与抗干扰性是提升模型鲁棒性、破解数据污染难题的关键环节。

针对大语言模型因固有文本生成策略而产生的幻觉现象，可在生成阶段引入RAG技术。RAG可以通过实时检索外部权威知识库为模型提供最新、最准确的信息参考，减少模型对参数记忆中可能存在的错误或过时信息的依赖，从而提升输出的事实准确性。此外，可应用语义熵评估生成内容的可靠性。该方法的核心是对同一提示进行多次采样生成，并利用自然语言推理技术对生成答案进行语义聚类，进而计算语义分布的熵值。若生成答案在语义上高度一致、聚类集中则熵值较低，表明模型输出确定性高；反之，若答案语义分散、聚类分散则熵值较高，反映出模型在该问题上的不确定性较大，可能伴随幻觉现象[43]。

针对对齐机制泛化能力不足的问题，可以引入对抗性训练作为防御手段。对抗性训练通过引入对抗性样本，使模型在学习过程中识别并抵抗诱导性、攻击性输入，从而提升模型面对对抗提示时的

鲁棒性[44]。此外，可定期组织专家开展红队测试，模拟越狱攻击、后门攻击等恶意行为，主动发现并修复模型安全漏洞[45]。针对攻击者通过后门攻击或越狱攻击导致的用户隐私泄露，可以结合差分隐私、联邦学习和区块链技术对模型进行隐私保护。其中，差分隐私通过在原始数据中添加噪声，降低了攻击者从模型中提取敏感信息的可能性[46]；联邦学习基于不同设备上的数据集构建机器学习模型，保证数据不出本地，减少数据暴露风险[47]；区块链通过记录每一次模型更新，防止攻击者篡改或注入有毒的模型参数，并在模型版本更新之前对本地模型进行性能检测，拒绝不满足性能阈值的恶意更新[48]。考虑到区块链吞吐效率低下的现实情况[49]，可结合区块链与星际文件系统（IPFS），将大语言模型更新产生的版本日志等大容量原始数据存储于IPFS网络，通过其内容寻址机制生成唯一不可篡改的内容标识符，解决大体积模型数据直接上链的存储成本与性能瓶颈问题[50]。

综上，应对模型层的内生缺陷，需要协同应用RAG、语义熵检测等多重技术手段，从生成逻辑与对齐能力两个核心维度破解模型层安全难题，削弱模型内生缺陷对最终输出与模型应用的影响。

（三）应用与交互层：破除用户盲目信任，阻断污染传播与场景风险

针对应用与交互层面数据污染的治理，应从源头破除用户对模型的盲目信任入手。通过在输出中嵌入强制性风险提示、面向不同用户群体开展分层教育，并对生成信息进行可信度分级标注，以主动管理用户预期并提升其信息鉴别能力。同时，在使用环节部署幻觉检测与偏见筛查等自动化核验工具，对输出内容进行实时预警与过滤，从技术上阻断污染信息的进一步传播。

为切断污染扩散与自我强化的循环，必须协同干预信息传播与数据采集两端。这要求内容平台建立针对AIGC的识别与审核机制，主动拦截虚假、偏见及恶意代码的传播并规范模型训练中的数据抓取行为，避免已被标记的污染数据重新进入训练语料库[51]。与此同时，建立高效、透明的用户反馈与模型迭代通道，通过简化上报流程、公示处理结果并将已验证的反馈用于模型微调与优化，使用户转化为积极的治理参与者。特别是在专业领域，应激励专家参与反馈和标准制定以提升模型在高风险场景下的可靠性。对于医疗、司法等高风险应用场景，需实施严格的场景化管控，明确划定模型不可替代人类决策的核心禁区，如最终临床诊断或司法判决。对已进入医疗、司法等领域的模型应用，实施准入备案与安全评估，并要求模型提供者定期接受输出质量、核验率及污染处理情况的审计，确保治理措施在高风险场景中切实落地，最大限度减少数据污染对高风险领域的影响。

（四）防御与监管层：法治赋能与技术协同双轮驱动，构建数据全生命周期防控体系

为解决防御与监管层面的现有痛点，应构建以法律规制为核心、技术防御为支撑、主体责任明确的系统性治理框架。

在防御层面，构建以预防为主，覆盖数据采集、标注、训练、输出全生命周期的防控体系：在数据采集阶段，强化数据源准入过滤与合法性核验，优先接入权威可信数据源并建立交叉验证机制；在标注阶段，建立多方交叉验证与偏差审查机制，防范主观偏见与恶意注入；在训练阶段，推行对抗训练、模型健康度认证与安全蒸馏技术，阻断缺陷传导；在输出阶段，构建场景化、差异化的人机协同验证网络，高风险场景强制引入跨领域专家联合审议，低风险场景保留人工干预通道，形成全链条闭环防控。

在监管层面，构建坚实的法治基础与清晰的责任框架是首要任务。近年来，我国围绕生成式AI的合规要求、风险处置与安全治理已形成一系列制度供给与政策指引。例如，2023年7月，我国发布《生成式人工智能服务管理暂行办法》，在兼顾产业发展与治理平衡的基础上，明确了AIGC标注义务、违法内容与虚假信息处置机制，并设立用户投诉和通报制度；2024年7月，我国提出的《加强人工智能能力建设国际合作决议》在联合国大会通过，将技术普及和互惠共享作为核心重点，为全球AI治理提供了聚焦弥合“数字鸿沟”、实现各国平等受益的务实合作框架；2024年9月，我国发布《人工智能安全治理框架》1.0版，围绕风险识别、技术应对、综合治理、安全指引四大核心维度，将以技术手段防控AI安全风险的方案进一步明确化、体系化。在此基础上，仍需推动专项立法或出台司法

解释以填补法律空白，将数据投毒行为明确纳入现有法律调整范围，明确其可适用故意毁坏财物罪、破坏生产经营罪、破坏计算机信息系统罪等罪名[52]；同时，将倡导性文件中的核心要求转化为具有法律约束力的强制性义务，确保制度规范向刚性约束转变。此外，相关机构可以建立投诉举报渠道，接收社会发现的幻觉输出、错误信息、AI传播虚假内容线索，反向定位数据污染问题。

在责任界定方面，针对责任界定模糊的痛点，需对恶意投毒者秉持从严追责原则，践行主观归责理念，只要证明投毒行为与损害结果存在条件关系即可确认归责；对数据提供者、开发者、服务提供者、数据管理者按数据生命周期明确各自义务：数据提供者需履行来源认证与安全筛查义务，开发者需构建抗投毒模型架构并开展安全性测试，服务提供者需落实数据来源审核与运行监控责任，数据管理者需部署动态监测与异常报告机制[53]。由此，在数据流转的全过程中确保权责清晰、环环相扣，将抽象的安全责任转化为各参与方可具体执行与核查的操作规范，确保安全要求落到实处。

此外，应对数据投毒的跨境传播风险与国际管控标准碎片化格局，深度参与跨境数据投毒治理的国际标准制定，建立跨国监管信息共享与执法协作机制，构建数据安全跨境联防联控体系。推动国内行业自律规则与国际通行治理标准的衔接适配，引导头部企业建立覆盖模型全生命周期的跨境数据投毒风险防控合规体系，明确跨境训练数据采集、开源数据集跨境流转、模型跨境部署环节的投毒风险防控主体义务。最终在维护国家数据安全、社会公共利益与促进AI技术全球创新协同之间实现动态平衡，筑牢大语言模型跨境应用与发展的安全可控底线。

五、结语

数据污染因其表现形态多样、社会影响广泛，正在持续削弱并制约大语言模型的安全可信运行。从潜伏性极强的后门攻击、全周期渗透的数据投毒，到难以根除的模型幻觉与价值倾斜，污染风险不仅内生于模型的技术架构与训练范式，在开放的数据生态、薄弱的防御体系以及滞后的监管框架中被不断放大与扩散，威胁着模型自身的安全可信运行，更可能穿透数字边界，对公众权益、市场秩序乃至社会稳定造成实质性冲击。

面对这一复杂且动态演变的安全风险，本文从数据、模型、应用与交互、防御与监管4个层面系统剖析了污染的产生逻辑，并据此提出协同治理路径。在数据源头上，构建纯净、可控、可溯源的训练数据体系；在模型内部优化生成逻辑、增强对齐能力与抗干扰性；在应用场景中，破除用户对模型的无条件信任，阻断受污染数据的二次乃至多次传播；在防御与监管层面，推动法治完善、责任明晰与国际协同。应对数据污染的安全挑战，是一场无法速胜的攻坚战与持久战，必须依靠多维度、多主体的协同治理与严格把控。

大语言模型的技术迭代仍在加速，数据污染的攻击手段也朝着低成本、自动化、多样化方向演进，治理工作是长期且动态的博弈过程。这既需要技术层面的持续创新，如更高效的污染检测算法、更鲁棒的模型架构、更智能的实时干预工具，也离不开制度与伦理层面的同步建设。最终，大语言模型的健康发展必须在能力突破与安全可控之间寻求平衡，需要开发者、使用者、监管者及社会各界共同努力维护与承担。未来的研究需要在持续优化现有防御策略的基础上，针对可能的新型攻击手段开发更具适应性的防御技术与监管模式，协同构建覆盖模型全生命周期的安全体系，以应对日益复杂的数据污染威胁。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: January 6, 2026; **Revised date:** February 16, 2026

Corresponding author: Zhang Xing is a professor from College of Economic and Management, Zhengzhou University of Light Industry. His major research fields include data governance and knowledge management. E-mail: zhangxing@zzuli.edu.cn

Funding project: The National Social Science Fund of China Project (24CTQ042)

参考文献

[1] Golchin S，Surdeanu M. Data contamination quiz：A tool to detect and estimate contamination in large language models[J]. Transactions of the Association for Computational Linguistics，2025，13：809-830.

[2] 陈俊秀，徐玉琴. 人工智能大语言模型引发的数据污染风险及其规制路径[J]. 大连理工大学学报（社会科学版），2025，46（4）：55-62.

Chen J X，Xu Y Q. The risk of data pollution caused by large language model and its regulatory paths[J]. Journal of Dalian University of Technology（Social Sciences），2025，46(4)：55-62.

[3] 李铭轩，文继荣. AIGC时代网络信息内容的法律治理——以大语言模型为例[J]. 北京理工大学学报（社会科学版），2023，25(6)：83-92.
Li M X，Wen J R. Legal governance of Internet information content in AIGC era—Taking large language model as an example[J]. Journal of Beijing Institute of Technology（Social Sciences Edition），2023，25(6)：83-92.

[4] 邓栩健. 生成式人工智能平台自治规则的体系化构建[J]. 东方法学，2025(3)：76-92.
Deng X J. Systematic construction of autonomous rules for generative artificial intelligence platforms[J]. Oriental Law，2025(3)：76-92.

[5] 刘洵，张馨，张琪琪，等. 医疗领域大模型伦理风险识别、治理及前瞻研究[J]. 中国工程科学，2025，27(6)：54-67.
Liu L，Zhang X，Zhang Q Q，et al. Ethical risk identification，governance，and foresight of medical foundation models[J]. Strategic Study of CAE，2025，27(6)：54-67.

[6] Yang H M，Xiang K L，Ge M Y，et al. A comprehensive overview of backdoor attacks in large language models within communication networks[J]. IEEE Network，2024，38(6)：211-218.

[7] Sainz O，Campos J，García-Ferrero I，et al. NLP evaluation in trouble：On the need to measure LLM data contamination for each benchmark[C]//The 2023 Conference on Empirical Methods in Natural Language Processing，2023：10776-10787.

[8] 何静，沈阳，谢润锋. 大语言模型幻觉现象的分类识别与优化研究[J]. 计算机科学与探索，2025，19(5)：1295-1301.
He J，Shen Y，Xie R F. Research on categorical recognition and optimization of hallucination phenomenon in large language models[J]. Journal of Frontiers of Computer Science and Technology，2025，19(5)：1295-1301.

[9] Souly A，Rando J，Chapman E，et al. Poisoning attacks on LLMs require a near-constant number of poison samples[PP/OL]. arXiv(2025-10-08)[2026-01-06]. https://doi.org/10.48550/arXiv.2510.07192.

[10] Improta C. Poisoning programs by un-repairing code：Security concerns of AI-generated code[C]//2023 IEEE 34th International Symposium on Software Reliability Engineering Workshops（ISSREW），2023：128-131.

[11] 陈晋音，席昌坤，郑海斌，等. 多模态大语言模型的安全性研究综述[J]. 计算机科学，2025，52(7)：315-341.
Chen J Y，Xi C K，Zheng H B，et al. Survey of security research on multimodal large language models[J]. Computer Science，2025，52(7)：315-341.

[12] Shi X Y，Chen S F，Zhang G，et al. Jailbreak attack with multimodal virtual scenario hypnosis for vision-language models[J]. Pattern Recognition，2026，172：112391.

[13] Alber D A，Yang Z H，Alyakin A，et al. Medical large language models are vulnerable to data-poisoning attacks[J]. Nature Medicine，2025，31(2)：618-626.

[14] Akinci D'Antonoli T，Tejani A S，Khosravi B，et al. Cybersecurity threats and mitigation strategies for large language models in health care[J]. Radiology：Artificial Intelligence，2025，7(4)：e240739.

[15] Yang Y F，Jin Q，Huang F R，et al. Adversarial prompt and fine-tuning attacks threaten medical large language models[J]. Nature Communications，2025，16：9011.

[16] 贾铮，王东波. 大语言模型技术悬浮现象：特征表现与影响因素研究[J]. 情报理论与实践，2025，48(11)：94-100.
Jia Z，Wang D B. Research on the suspension phenomenon of large language modeling technology：Characteristics and influencing factors[J]. Information Studies：Theory & Application，2025，48(11)：94-100.

[17] 林珊，邓小昭. AIGC情境下用户信息验证行为过程及影响因素研究[J]. 图书情报工作，2025，69(22)：136-148.
Lin S，Deng X Z. Research on the process and influencing factors of user information verification behavior in the AIGC context[J]. Library and Information Service，2025，69(22)：136-148.

[18] 周建新，王福如. 人工智能算法嵌入文化产业的技术异化及其消解路径[J]. 南京社会科学，2025(6)：138-148.
Zhou J X，Wang F R. The alienation of technology and resolution paths for artificial intelligence algorithms embedded in the cultural industry[J]. Nanjing Journal of Social Sciences，2025(6)：138-148.

[19] 张海，段荟. 大语言模型用户中辍行为的现象学研究[J]. 情报杂志，2025，44(10)：181-186.
Zhang H，Duan H. A phenomenography study of large language model users' dropout behavior[J]. Journal of Intelligence，2025，44(10)：181-186.

[20] 曾建光，尤澜涛，严江南，等. 思政创新与大语言模型"幻觉"——基于大语言模型中文评价平台的证据[J]. 产业经济评论，2025(4)：180-200.
Zeng J G，You L T，Yan J N，et al. Innovation-driven development of ideological work and hallucination in large language models[J]. Review of Industrial Economics，2025(4)：180-200.

[21] Hiriyanna S，Zhao W B. Multi-layered framework for LLM hallucination mitigation in high-stakes applications：A tutorial[J]. Computers，2025，14(8)：332.

[22] Williams C Y K，Bains J，Tang T Y，et al. Evaluating large language models for drafting emergency department encounter summaries[J]. PLoS Digital Health，2025，4(6)：e0000899.

[23] 刘泽垣，王鹏江，宋晓斌，等. 大语言模型的幻觉问题研究综述[J]. 软件学报，2025，36(3)：1152-1185.
Liu Z Y，Wang P J，Song X B，et al. Survey on hallucinations in large language models[J]. Journal of Software，2025，36(3)：1152-1185.

[24] 姜毅，杨勇，印佳丽，等. 大语言模型安全与隐私风险综述[J]. 计算机研究与发展，2025，62(8)：1979-2018.
Jiang Y，Yang Y，Yin J L，et al. Survey on security and privacy risks in large language models[J]. Journal of Computer Research and Development，2025，62(8)：1979-2018.

[25] 郭园方，余梓彤，刘艾杉，等. 多模态大模型安全研究进展[J]. 中国图象图形学报，2025，30(6)：2051-2081.
Guo Y F，Yu Z T，Liu A S，et al. Recent progress of the security research for multimodal large models[J]. Journal of Image and Graphics，2025，30(6)：2051-2081.

[26] Zack T，Lehman E，Suzgun M，et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care：A model evaluation study[J]. The Lancet Digital Health，2024，6(1)：e12-e22.

[27] Zhang Q J，Wang D，Qian H T，et al. Speculating LLMs' Chinese

training data pollution from their tokens[C]//The 2025 Conference on Empirical Methods in Natural Language Processing，2025：26124-26144.

[28] Marcus G. The next decade in AI：Four steps towards robust artificial intelligence[PP/OL]. V3. arXiv(2020-02-19)[2026-01-06]. https://doi.org/10.48550/arXiv.2002.06177.

[29] Dumit J，Roepstorff A. AI hallucinations are a feature of LLM design，not a bug[J]. Nature，2025，639(8053)：38.

[30] Artsi Y，Sorin V，Glicksberg B S，et al. Challenges of implementing LLMs in clinical practice：Perspectives[J]. Journal of Clinical Medicine，2025，14(17)：6169.

[31] Das B C，Amini M H，Wu Y Z. Security and privacy challenges of large language models：A survey[J]. ACM Computing Surveys，2025，57(6)：1-39.

[32] 梁昭. AI“幻觉”：认知困境、术语反思与范式嬗变[J]. 民族学刊，2025，16(8)：82-87，161.
Liang Z. AI “hallucination”：Cognitive dilemma，terminological reflection，and paradigm shift[J]. Journal of Ethnology，2025，16(8)：82-87，161.

[33] 李希陶，吴江，郑庆华，等. 大语言模型越狱攻击：模型、根因及其攻防演化[J]. 中国科学：信息科学，2025，55(6)：1372-1405.
Li X T，Wu J，Zheng Q H，et al. Jailbreaking large language models：Models，origins，and evolution of attacks and defenses[J]. Scientia Sinica Informationis，2025，55(6)：1372-1405.

[34] Wei A，Haghtalab N，Steinhardt J. Jailbroken：how does llm safety training fail[J]. Advances in Neural Information Processing Systems，2023，36：80079-80110.

[35] 李南，丁益东，江浩宇，等. 面向大语言模型的越狱攻击综述[J]. 计算机研究与发展，2024，61(5)：1156-1181.
Li N，Ding Y D，Jiang H Y，et al. Jailbreak attack for large language models：A survey[J]. Journal of Computer Research and Development，2024，61(5)：1156-1181.

[36] 梁思源，何英哲，刘艾杉，等. 面向大语言模型的越狱攻击与防御综述[J]. 信息安全学报，2024，9(5)：56-86.
Liang S Y，He Y Z，Liu A S，et al. A review of jailbreak attacks and defenses for large language models[J]. Journal of Cyber Security，2024，9(5)：56-86.

[37] 付志远，陈思宇，陈骏帆，等. 大语言模型安全的挑战与机遇[J]. 信息安全学报，2024，9(5)：26-55.
Fu Z Y，Chen S Y，Chen J F，et al. Challenges and opportunities of large language model security[J]. Journal of Cyber Security，2024，9(5)：26-55.

[38] 顾欢欢，李千目，刘臻，等. 基于虚假演示的隐藏后门提示攻击方法研究[J]. 信息网络安全，2025，25(4)：619-629.
Gu H H，Li Q M，Liu Z，et al. Research on hidden backdoor prompt attack methods based on false demonstrations[J]. Netinfo Security，2025，25(4)：619-629.

[39] Hu M，Yang Y T，Pan D，et al. Syntactic paraphrase-based synthetic data generation for backdoor attacks against Chinese language models[J]. Information Fusion，2025，124：103376.

[40] Peng Z Q，Fu J M，Zou L X，et al. Backdoor samples detection based on perturbation discrepancy consistency in pre-trained language models[J]. Neural Networks，2026，193：108025.

[41] Li J Z，Wu Z F，Ping W，et al. Defending against insertion-based textual backdoor attacks *via* attribution[C]//The Association for Computational Linguistics：ACL 2023，2023：8818-8833.

[42] 孙毅，范灵俊，洪学海. 区块链技术发展及应用：现状与挑战[J]. 中国工程科学，2018，20(2)：27-32.
Sun Y，Fan L J，Hong X H. Technology development and application of blockchain：Current status and challenges[J]. Strategic Study of CAE，2018，20(2)：27-32.

[43] Farquhar S，Kossen J，Kuhn L，et al. Detecting hallucinations in large language models using semantic entropy[J]. Nature，2024，630(8017)：625-630.

[44] Kumar A，Salim M M，Camacho D，et al. A comprehensive survey on large language models for multimedia data security：Challenges and solutions[J]. Computer Networks，2025，267：111379.

[45] 李子浩，李天云. 技术向善：人工智能大模型伦理风险识别及治理路径[J]. 学术交流，2024 (8)：30-42.
Li Z H，Li T Y. Technology for good：Ethical risk identification of artificial intelligence model and its governance paths[J]. Academic Exchange，2024 (8)：30-42.

[46] 张熙，李朝卓，许诺，等. 面向可信大语言模型智能体的安全挑战与应对机制[J]. 信息通信技术与政策，2025，51(1)：33-39.
Zhang X，Li C Z，Xu N，et al. Security challenges and response mechanisms for trustworthy large language model agents[J]. Information and Communications Technology and Policy，2025，51(1)：33-39.

[47] Weeks C，Cheruvu A，Abdullah S M，et al. A first look at toxicity injection attacks on open-domain chatbots[C]//The 39th Annual Computer Security Applications Conference，2023：521-534.

[48] Salim M M，Deng X J，Park J H. A privacy-preserving local differential privacy-based federated learning model to secure LLM from adversarial attacks[J]. Human-Centric Computing and Information Sciences，2024，14：57.

[49] 支凤稳，云仲伦，张闪闪. 基于区块链的个人科学数据共享模式研究[J]. 现代情报，2021，41(12)：69-78.
Zhi F W，Yun Z L，Zhang S S. Research on personal scientific data sharing mode based on blockchain[J]. Modern Information，2021，41(12)：69-78.

[50] 陆丽娜，尹丽红，于啸，等. 基于区块链的农业科学数据溯源模型构建研究[J]. 情报科学，2024，42(11)：54-60.
Lu L N，Yin L H，Yu X，et al. Construction of agricultural science data provenance model based on blockchain[J]. Information Science，2024，42(11)：54-60.

[51] 翟羽佳. 大语言模型应用于图书馆建设未来学习中心的适应性、风险与策略[J]. 图书馆学研究，2024(7)：77-85.
Zhai Y J. Adaptability，risk and strategies of large language models in the construction of future learning centers in libraries[J]. Research on Library Science，2024(7)：77-85.

[52] 徐蕴杰. 数据投毒的罪名适用与刑事归责[J]. 天府新论，2025 (4)：38-48，152.
Xu Y J. The application of the crime of data poisoning and its criminal liability[J]. New Horizons from Tianfu，2025 (4)：38-48，152.

[53] 翟岩，李小波. 人工智能安全视域下“数据投毒”的内涵特征、层级风险与治理路径[J]. 河南社会科学，2025，33(11)：113-124.
Zhai Y，Li X B. The connotative characteristics，tiered risks and governance pathways of “data poisoning” in the context of AI security[J]. Henan Social Sciences，2025，33(11)：113-124.