文章编号:1003 - 2053(2026)04 - 0690 - 11

DeepSeek 类生成式人工智能的新型数据安全风险治理

宗绍昊[1],罗世龙[2]
(1. 中国政法大学数据法治研究院,北京 100088; 2. 武汉大学法学院,湖北武汉 430072)

摘 要:生成式人工智能的数据安全风险治理是当前面临的重要问题。现有生成式人工智能的数据安全治理体系未能针对不同的数据安全风险建立起有效的类型化治理机制,与迭代至 DeepSeek 类生成式人工智能的现实状况相脱节。DeepSeek 系列模型具有多头潜在注意力机制、混合专家模型、纯强化学习等新技术特征,导致原有数据安全风险量变或异化,由此产生运行逻辑各异的新型数据泄露安全风险、数据偏离安全风险和数据删除安全风险。针对数据泄露安全风险,应分别形成技术层面和管理层面的预防机制;针对数据偏离安全风险,应强调不同于人工监督微调模式的数据输入和输出的两端微调并创设数据输出结果的算法解释权;针对数据删除安全风险,应从数据绝对删除转向数据相对删除并实行动态核验机制。

关键词:DeepSeek;ChatGPT;生成式人工智能;数据安全风险;数据治理
中图分类号:G301;G203;D91 **文献标识码**:A

ChatGPT 的问世被誉为生成式人工智能的里程碑事件,此后若干生成式人工智能如雨后春笋般层出不穷,但均未突破 ChatGPT 的框架和能力。由 DeepSeek 研发的 DeepSeek V3 和 DeepSeek R1 模型甫一亮相即引发各方的强烈关注与持续热议,无疑是在催生和引领人工智能领域的新一轮革命浪潮。然而,面对日趋庞大的数据体量,在各大政府网站和软件平台均频繁部署 DeepSeek 的情况下,DeepSeek 的新型技术迭代使得数据安全风险的监管机制面临着严峻考验。在政策层面上,我国业已颁布的数十份政策性文件和技术标准虽极大地改善了生成式人工智能领域的规范匮乏现象,但基本上停留在数据来源安全(内容过滤、合规获取、来源可追溯等)、内容安全(知识产权、个人信息保护等)、模型安全(内容安全、准确与可靠)、数据标注和内容标识等内容。其中,针对数据来源和内容安全,合规获取要求大模型对各类数据进行差异化处理,但现有规定只不过是对法律法规的重述,并未实际解决技术创新与数据保护之间的突出矛盾;实现海量数据的来源可追溯更多的是强调数据本身的合法性问题。并且,数据来源安全风险并非生成式人工智能面临的核心数据安全风险,任何进行大数据分析的软件等均面临着合法合规的要求,因而不具有典型性。针对模型安全,现行规定尤其强调指令微调对于生成式人工智能内容输出的安全性作用,与天然地排斥指令微调的 DeepSeek 系列模型不相适应。针对数据标注和内容标识,对内容准确与可靠的强调只不过是治理目标的重申而无具体方法;数据标注和内容标识的确在预训练数据质量和人工智能生成数据鉴别上发挥功效,但存在数据标注的 DeepSeek 预训练数据只是绝对的少数,且内容标识容易在海量数据混杂之间泯灭价值。通过实证研究可知,数据安全风险治理的倡议性和滞后性是当前各类文件或标准的普遍特征,其未对各类数据安全风险予以类型化并提出针对性的对策,无法形成数据安全风险的完善治理体系,更不可能对不同于 ChatGPT 的 DeepSeek 系列模型所生独特的数据安全风险进行有效治理。在理论层面上,以 ChatGPT 为基础的科学研究颇多,诸如平衡安全与发展并坚持软硬法协同规制[1],治理主体、过程、方法、结果上的包容性[2],伦理先行并提升透明度[3],强化数据监管责任等措施[4],在生成式人工智能治理领域占据主流地位。这些措施抽象性有余而具体化不足,且针对数据安全风险的专题研究数量稀少,针对 DeepSeek 数据安全风险的系统研究更是几乎为零。在 DeepSeek 的新变局背景下,如何妥善应对这些新型数据

收稿日期:2025 - 03 - 11;**修回日期**:2025 - 07 - 22
作者简介:宗绍昊(1998 -),男,博士研究生,E - mail:qmwsz1368@163.com。罗世龙(2001 -),男,博士研究生。

安全风险以使数据在合规的轨道上流通和利用亟待深入研究。

1　DeepSeek 技术迭代征表多样化数据安全风险

在 DeepSeek 系列模型横空出世之前,规模定律(Scaling Laws)始终主导着机器学习领域中大型语言模型(Large Language Models,LLMs)的开发与应用活动。依该定律可知,模型的性能往往会随着模型的参数量、训练量和训练数据量等计算资源的增多而日趋强化[5]。在这种认识的指引下,OpenAI 的 GPT 系列、Anthropic 的 Claude 以及谷歌的 Gemini 等,都纷纷致力于不断扩张模型的规模以求进一步增强产品的综合性能。构成模型规模基础的 GPU 等硬件设施的部署和使用成本随之急剧升高,成为影响基座模型厂商核心竞争力的关键因素和制约大模型普及化应用的一大瓶颈。为实现弯道超车,我国初创企业 DeepSeek 通过优化算法的方式带动大模型成本的显著下降及其性能的极大改善,从而成功开辟出推动人工智能技术进步的新模式。

DeepSeek 系列产品的出现不仅意味着全球人工智能产品竞争格局的悄然改变,更昭示着新一轮算法技术迭代周期的陡然来临。与以 ChatGPT 系列为代表的生成式人工智能相比,DeepSeek 的技术革新主要体现在以下 6 个方面:一是开创性地提出并运用多头潜在注意力机制(Multi - head Latent Attention, MLA),大幅降低显存占用,使模型在相同硬件条件下的计算效率和推理速度得到质的提升。具言之,DeepSeek 的 MLA 架构通过降低模型维度来高效地存储 KV 缓存,对传统的多头注意力计算过程中的高维矩阵运算予以优化,以引入低维的潜在空间为抓手来简化计算。其实质上是在以重塑查询(Query)、键(Key)和值(Value)的方式来构建出一种更为便捷的注意力计算机制。二是采用全新的混合专家模型(Mixed Expert Models, MoE)结构,借助负载均衡的算法设计,有效缓解传统 MoE 架构因负载不均衡而存在的训练效率低下难题。这种模型之"新"主要在于,一方面,通过精心设置的门控机制,其能够更准确地将输入端的多样数据分配到对应的专家模型中去,如将通用知识压缩至共享专家处以减轻路由专家的参数冗余,甚或基于深度神经网络学习,根据输入端数据的多方面特征进行智能判断,将数学问题和自然语言生成任务分别配送给擅长数学推理和在语言生成上表现出色的专家模型处理。另一方面,依托 MLA 架构对不同专家模型之间的协作和融合方式予以创新,使得彼此的输出在互相融汇的过程中避免信息冲突或丢失现象的发生,明显提升整个模型的泛化能力。三是采行纯强化学习(Reinforcement Learning, RL)方法,不依赖监督微调(Supervised Fine - Tuning, SFT),让模型自行进化推理能力。这意味着,只要配备合适的激励机制并辅之以冷启动数据和多阶段训练管道,大模型就可以纯粹依靠强化学习技术来实现自主进化,同时兼具令人满意的推理水平与高质量的表达效果[6]。四是系统运用知识蒸馏(Knowledge Distillation, KD)技术,将预先训练完成的大模型学习成果整体迁移至参数更低的轻量模型中去,充分发挥"低成本、高性能"的优势,在应用推广领域形成规模化效应。五是采取大量精细化的系统工程优化举措,推进系统软件层次的创新。例如,其使用 FP8 混合精度训练框架加速训练过程、设计 Dualpipe 算法以改善训练过程中的流水线并行效率、采纳高效的跨节点通信以节省通信开销并提升大规模分布式训练的效益,等等。六是舍弃闭源系统,实施开源创新策略,营建开放生态,以促进行业内的知识共享与技术交流。

基于此,此前针对以 ChatGPT 为代表的生成式人工智能数据安全风险治理的诸多讨论恐已落后于时代发展,无法针对性地回应大模型的最新发展态势,实有更新之必要。例如,开源大模型会给数据泄露安全风险带来革命性变化。在开源创新模式下,对多个开源库的依赖会造成极其庞大的攻击面,其中一个组件中的漏洞就可能危及整个系统的安全性能,这会致使数据泄露安全风险呈指数倍对外扩散而出现深刻的质变。同时,在强推理外表的包装下,即便使用老旧落伍的错误数据亦使得 DeepSeek 具有相当的可信服力,再加之监督微调的缺失,如何避免此种因机器幻觉而产生的数据偏离安全风险所带来的威胁,尚需更新的制度安排。此外,DeepSeek 的激增参数为其数据删除义务的履行带来诸多挑战。综合来看,数据泄露安全风险、数据偏离安全风险和数据删除安全风险,是生成式人工智能迭代至 DeepSeek 时代所面临的三大数据安全风险,这些风险或许在外观上与 ChatGPT 所对应的数据安全风险具有某种程度的相似性,但内在运作逻辑截然不同,

其中既有量变也有质变,需要探索与之相匹配的类型化治理机制。

2 DeepSeek 数据泄露安全风险的预防机制

在 DeepSeek 场景下,诸多内外部数据泄露环节使得数据泄露安全风险在异化的同时趋于严重化。针对此种特殊的数据泄露安全风险,应在技管融合治理的基本模式下建构起健全的预防机制并予以时刻贯彻,进而在最大程度上应对数据泄露安全事件。

2.1 DeepSeek 的数据泄露安全风险

(1)基于外部的数据泄露安全风险

作为开源大模型,DeepSeek 对外公开的一系列代码极有可能为恶意行为者所利用,显著升高由外部攻击所导致的数据泄露安全风险。这种情况实际上已经在当下发生。奇安信 XLab 实验室曾对外发布报告称,在 2024 年 12 月 1 日至 2025 年 2 月 3 日期间共出现 2650 个仿冒 DeepSeek 的网站。这些仿冒域名的使用用途主要为钓鱼欺诈、域名抢注、流量引导,其中钓鱼欺诈主要通过窃取用户登录凭证、利用相似域名和界面误导用户、诱骗用户购买虚拟资产等手段实施诈骗。这些仿冒 DeepSeek 的域名中有 60% 解析 IP 位于美国,其余主要分布在新加坡、德国、立陶宛、俄罗斯和中国,仿冒域名呈现出明显的全球化特点。同时,除普通的平台用户,掌握专业技术的开发人员也难逃来自外部的恶意攻击和侵扰。多家威胁情报公司发现,恶意攻击者利用 DeepSeek 的热度,在 Python 软件包索引(PyPI)存储库上发布虚假软件包,旨在窃取开发人员系统中的敏感信息。这些软件包名为 deepseeek 和 deepseek-ai,伪装成 DeepSeek 的 Python API 客户端,在 2025 年 1 月 29 日被删除之前已被下载至少 222 次。可以预见的是,随着开源大模型以及人工智能辅助恶意软件开发的兴起,此类攻击手段因其实效性和实施难度不断降低而愈发猖獗,攻击者甚至能够精心策划个性化且复杂的攻击来突破传统安全措施,目前的数据安全合规制度体系由此岌岌可危。

(2)基于内部的数据泄露安全风险

DeepSeek 所采用的开源创新模式还意味着,即便无人从外部恶意攻击系统,公开的代码也可能会因善意开发人员的仔细审查而意外暴露弱点。再者,生成式人工智能训练的本质昭示着,即使代码是安全的,训练中使用的大量数据集也可能成为漏洞点。综合这两点可知,除源自外部的恶意侵袭,相当一部分数据泄露系由内部简单的配置错误或者人员疏忽造成,而人工智能系统的复杂性会加剧和放大这种意外泄露的风险。原因在于,以 DeepSeek 为代表的更先进的生成式人工智能可能会在交互过程中无意间存储或泄露更详细的个人数据[7]。同时,开源大模型开发的互联性及其对各种库和框架的依赖会产生多个潜在的危害点,而 DeepSeek 系列模型在全国乃至全球范围内的广泛本地化部署亦会促使这种意外泄露风险呈指数级增长态势。例如,据全球云安全巨头 Wiz Research 报告称,其在 2025 年 1 月就发现一个属于 DeepSeek 的可公开访问 ClickHouse 数据库,该数据库允许对数据库操作进行完全控制,包括能够访问内部数据。这不仅使攻击者能够轻而易举地获取敏感日志和明文聊天消息,还可能会让其利用 ClickHouse 配置,通过类似“SELECT * FROM file(filename)”的查询直接从服务器窃取明文密码和本地文件。尽管 DeepSeek 在收到 Wiz Research 的通知后迅速进行修复,对数据库予以安全加固,但长时间的暴露已为恶意行为者提供充足的机会来访问并窃取数据。类似地,当前在本地化部署过程中所经常使用的开源大模型服务工具“Ollama”,在默认情况下并未提供安全认证功能,这容易导致许多部署 DeepSeek 的用户忽略必要的安全限制,意外诱使数据泄露。据奇安信平台监测发现,有 8971 个运行了 Ollama 大模型框架的服务器中,有 6449 个活跃服务器,其中 88.9% 的服务器“裸奔”在互联网上,导致任何人不需要任何认证即可随意调用、在未经授权的情况下访问这些服务。

2.2 DeepSeek 的数据泄露安全风险预防机制

(1)技术层面的预防机制

DeepSeek 的数据泄露风险根源于所采取的技术措施不当,因而技术治理应处于主导地位。技术治理具有增强治理手段、扩大治理边界和优化治理结构的优势[8],“监管技术化的根本要求在于强调信息化、数字化技术在监管中的有效运用”[9]。面对 DeepSeek 的内外部数据泄露安全风险,依靠固定程式的管理治理已无法满足多样化、复杂化、系统化的数据泄露安全风险治理需求,“用魔法对抗魔法”的路径更值得期待。一方面,从数据本身的防泄漏属性上看,数据脱敏和隐私计算技术使得数据本身的敏感性在输入阶段即被剔除,并以可用而不可见的方式分析和处理数据,即便发生数据泄露事件也

不会威胁数据主体的人身和财产权益。在数据脱敏技术所导致的数据可用性丧失和数据价值损耗的矛盾问题上,动态脱敏技术、差分隐私技术和智能化脱敏技术等可提供解决方案。另一方面,从数据存储的防泄漏上看,针对 DeepSeek 开源大模型的特殊性,在分布式设备上通过联邦学习(Federated Learning)和分布式训练优化技术训练大模型并避免原始数据集中存储,可显著提升数据安全性,避免大范围数据泄露事件的发生。同时,DeepSeek 部署者可选择适用对称加密技术、非对称加密技术等进行数据加密储存,并辅之以逻辑隔离(访问控制列表)、虚拟化隔离(虚拟机隔离、容器隔离)等技术限定不同人员所能直接接触的数据范围,以防止内部不当泄露数据事件。这一模式与 DeepSeek 的混合专家模型所确立的多样数据的模型分配机制和不同专家模型的协作机制完全契合。

(2)管理层面的预防机制

“在新兴技术的发展过程中,促进技术向善最有效的手段始终是通过法律将标准和原则嵌入技术构架的底层行动逻辑,诱导出立法者期望的结果”[10]。技术治理的优越性需要组织性合规建设予以支持,DeepSeek 部署者的内部人员和资源配置机制是实现技术治理的基础条件。同时,基于数据全生命周期视角,在 DeepSeek 输入和输出全过程诸多场景中采用何种技术防范数据泄露安全风险,需以技术标准指南进行判断。管理层面的预防机制由事前设定的若干管理性事项并予以实时贯彻所构成,这些管理性事项围绕数据安全保护技术而展开,为其提供方向性指引,包括域名警示机制、数据分类分级保护机制、数据安全组织与评估机制、数据泄露事件应急处置机制等基本内容。

首先,域名警示机制要求 DeepSeek 系列模型在各端口采取显明的方式标注网站或软件的真实性和可信性。DeepSeek 部署者应与各大浏览器和应用商店等建立良好的沟通机制和附条件的授权机制,令 DeepSeek 网站入口或软件下载通道加注“官方”标志,且使用者在搜索 DeepSeek 时,官方网站或软件的排列应处于优先地位。在使用者点击 DeepSeek 网站或软件后,除系统验证使用者是否为真人外,还应以特征词等形式提醒其确认其所浏览的网站或软件是否为官方正版,这一点目前 DeepSeek 并未予以充分关注。同时,DeepSeek 部署者应定期查找并公布钓鱼网站、仿冒软件清单,并要求浏览器和应用商店等进行相应的处置措施。

其次,数据分类分级保护机制强调,针对不同类型和等级的数据应采取差异化的技术保护措施以防止泄露。理论上对于数据分类和数据分级的关系存在数据分类包含数据分级、数据分类决定数据分级、数据分类与数据分级并列存在等不同的观点[11]。但是,数据类型只是在初始意义上划分数据存在的场景,真正决定数据安全保护措施的仍然是数据本身的重要程度。从此意义上看,数据分类固然包含数据分级,但数据分级较之于数据分类具有更强的重要性。新近开放源代码促进会(Open Source Initiative, OSI)发布了开源 AI 定义的 1.0 正式版,认为只有源代码、数据集、训练过程等全部面向公众开放,允许自由使用、修改和分发的模型才是开源大模型。DeepSeek 并非此种意义上的开源大模型,其未完全公开训练数据,而仅通过技术报告详细披露了关键算法。因此,针对 DeepSeek 的数据分类分级保护机制应予以区分化设计,同时坚持数据全流程技术覆盖的基本原则。一方面,针对开源数据中可能涉及其他数据主体合法权益的部分采取最严格的数据安全保护技术;另一方面,针对未开源数据应依照一般数据、重要数据和核心数据的重要程度划分制定与之相匹配的数据安全保护技术。

再次,数据安全组织与评估机制指的是 DeepSeek 所属机构需建立常态化运作的数据安全负责人和管理机构,并进行安全评估与对抗攻击演练。我国《数据安全法》第 27 条第 2 款规定了重要数据处理者的数据安全负责人和管理机构设置责任。DeepSeek 所属机构自然属于重要数据处理者,其设置数据安全负责人和管理机构具有法律依据。《网络数据安全管理条例》第 30 条对数据安全负责人和管理机构的职能进行了具体化规定①,但具体到数据泄露安全风险上仍需予以一定程度的修正。数据安全负责人和管理机构职能设置的核心在于风险

① 《网络数据安全管理条例》第 30 条第 1 款规定:重要数据的处理者应当明确网络数据安全负责人和网络数据安全管理机构。网络数据安全管理机构应当履行下列网络数据安全保护责任:(一)制定实施网络数据安全管理制度、操作规程和网络数据安全事件应急预案;(二)定期组织开展网络数据安全风险监测、风险评估、应急演练、宣传教育培训等活动,及时处置网络数据安全风险和事件;(三)受理并处理网络数据安全投诉、举报。

导向,即根据数据泄露安全风险的特殊性设置具有针对性的制度并保持长久的制度活力。除内部工作人员泄漏事件可通过储存和操作权限等技术条件予以解决外,数据泄露安全风险主要是因外部攻击所致。对此,需制定完善的数据泄露安全风险评估机制并通过引入第三方专业评估机构的方式实现评估的科学化,包括模型本身、开源过程和本地或私有化部署的评估,在攻击对抗过程中实现 DeepSeek 防御能力的进化。为防止"纸面合规"现象的出现,确保数据安全负责人和管理机构的持续、有效运作,应设置轻微、一般和严重组织性合规违反行为的处罚,以反向激励形成梯度化的激励机制。同时,包括 Meta CEO 扎克伯格等在内的头部开源大模型厂商均提到,模型的开源会是保障大模型技术安全,解决安全漏洞的有效措施。因此,DeepSeek 所属机构不仅需受理并处理相关投诉和举报,还应合理吸收模型开源过程中相关技术人员等提出的安全改进意见。

最后,数据泄露事件应急处置机制,即在不可避免地发生数据泄露安全事件的情况下,DeepSeek 所属机构能在最大程度上降低对相关数据主体的权益损害。这一应急处置机制大致包括 4 个方面:明确应急处置机制的负责人和所属负责机构,确保其可全程参与应急处置机制的制定和实施,并具有协调多部门应急处置工作的权限和能力;进行场景化的应急处置机制设计,针对不同程度的数据泄露安全事件制定具有差异化的处置措施;规范应急处置机制的流程,按照"数据泄露事件发现并上报—负责人员联合多部门禁止处置—数据泄露原因查找—数据安全保护制度修复—事件记录与责任追究"等流程及时、有效地展开工作;避免应急处置机制的虚置化,即便在不发生数据泄露事件的状态下,亦应定期进行应急处置机制的演练,形成高机动性的反应能力[12]。

3 DeepSeek 数据偏离安全风险的介入机制

数据偏离安全风险系指在人工智能系统使用过程中出现的偏离人们合理预期的各类风险。这既可以表现为某种基于机器幻觉而引发的数据质量问题,如数据不真实可信、不完整、不相关等,也可以表现为数据处理行为的侵害性问题,如在输入端收集数据时超过必要限度而侵犯隐私,或者是在输出端生成侵犯他人人格权、财产权、商业秘密等合法权益的不当内容。DeepSeek 的数据偏离安全风险可通过数据输入和输出的两端微调,以及输出过程的算法解释权予以克服。

3.1 DeepSeek 的数据偏离安全风险

(1)机器幻觉安全风险

由于 DeepSeek 使用的纯强化学习技术不依赖任何监督微调,其虽可有效降低数据批注成本,但输入端的数据质量和中间段的数据处理过程难以得到技术人员的及时检验和矫正,输出端的虚假或错误信息随之泛滥成灾。也许会有人认为,以前的大模型也存在类似的问题,这未必可称为新型数据安全风险。此种论断恰是在忽略 DeepSeek 所展现出的强大的数据自动化处理和逻辑推演能力。在 MLA 架构和 MoE 结构的多重作用下,DeepSeek 不仅会如往常般生成作为输出结果的终端内容,更会将自身与之有关的所有推论过程预先对外展示以增强结论的说服力。当用户就结论中的某一点或某一部分予以进一步追问时,DeepSeek 甚至会引经据典,对先前的结论进行跨学科式的学术化与专业化包装,编造出更多的虚假或错误信息来构建逻辑闭环的伪因果分析链条,从而达到成功诱骗他人的目的。面对这样的情景,普通用户实在无法察觉到问题,极易因轻信人工智能的生成内容而陷入认知幻觉。毕竟,在一系列新技术的加持下,DeepSeek 的幻觉输出早已从本质上区别于以往生成式人工智能的常识性错误,我们不能再奢求一个普通人类仅凭自己极其有限的知识储备和逻辑思考能力来与之抗衡。从更宏观的视角来看,若长久放任这种现象,那么整个信息生态就有遭受污染和恶化之虞,居于输入端的训练数据集最终也难免会充斥海量的虚假或错误信息,认知幻觉难题因之变得越发难解,人们的思想逐渐有被人工智能而非人类自我控制的危险。

(2)数据侵害安全风险

监督微调的缺失意味着当人们利用 DeepSeek 来生成恶意内容时,人工智能模型自身并不能够迅速意识到这一点并采取相应的阻断措施。尽管研发人员预先在 DeepSeek 模型中设置有一定的安全限制以保证输出内容的合规性,但恶意行为人完全有可能开发出新的越狱技术来回避 DeepSeek 内置的自我审查机制,在人工反应前借 DeepSeek 顺利获取恶意内容。这种设想其实已经在近来的模型越狱事件中得到验证。如 Palo Alto Networks 的 Unit 42 研究人员对 DeepSeek 模型进行测试,发现了 3 种有效

的越狱技术,即 Bad Likert Judge、Crescendo 和 Deceptive Delight,这些技术能够绕过 DeepSeek 模型的安全限制,获取恶意输出。具言之,Bad Likert Judge 技术通过让模型使用 Likert 量表评估响应的危害性,然后促使模型生成与高评分相关的示例,从而获取恶意内容。测试中,研究人员使用该技术成功让 DeepSeek 模型生成了数据泄露工具、网络钓鱼邮件模板、社会工程优化建议等恶意输出。这显然会给人们的人格权和财产权等合法权益带来潜在的威胁。Crescendo 利用模型自身知识,通过逐步提供相关内容的提示,引导对话走向被禁止的话题,直至覆盖模型的安全机制。在对 DeepSeek 模型的测试中,研究人员从询问莫洛托夫鸡尾酒的历史开始,通过一系列相关提示,最终使模型生成了制作莫洛托夫鸡尾酒的详细指南。这显然已涉嫌侵犯他人商业秘密和不正当竞争。Deceptive Delight 则将不安全的话题嵌入良性话题的积极叙述中,先让模型创建连接这些话题的故事,再对每个话题进行详细阐述,以此绕过安全措施。测试中,研究人员使用该技术让 DeepSeek 模型生成了利用分布式组件对象模型(DCOM)在 Windows 机器上远程运行命令的脚本。这自然涉及到知识产权法领域的侵权问题。

同时,侧重将大模型知识迁移至小模型的数据蒸馏以及从其他模型输出中提取信息的采样学习技术,虽能显著增强不同用途的模型间的数据交互性,高效实现人工智能模型各方面能力的跃升,但也容易引起以下两类尚缺乏有效监管的数据偏离安全风险。第一,因隐蔽式获取未经他人同意的各种信息而涉嫌侵犯隐私的风险。假设 A 模型通过医疗数据训练得到诊断能力,B 模型通过采样学习 A 的输出来完成能力迁移,那么 B 模型在掌握相同能力的同时,极有可能间接获取患者的隐私特征,这显然违背个人信息授权使用中的“知情同意”原则。第二,因数据来源不明而易出现的归责困境。当多个模型借知识蒸馏共同参与数据流转时,数据来源的透明度实际上已不如意,监管者难以追溯到初始的数据泄露源头。例如,DeepSeek 若基于数据蒸馏和采样学习技术从第三方模型中获取语料,其生成内容由此可能包含未被授权的企业数据和知识产权数据等,但责任划分将受数据流转因果关系链条不清晰的影响而陷入僵局。

3.2　DeepSeek 的数据偏离安全风险介入机制

(1)数据输入和输出的两端微调

《数据安全技术数据安全风险评估方法》针对数据质量控制情况,强调采用人工检查、自动检查或其他技术手段对数据的真实性、准确性、完整性校验情况。其中提及的检查手段是对 DeepSeek 数据输入和输出两端进行微调的体现,但这并非意味着令其重归监督微调式的大模型发展道路,也不是利用一个高质量的数据集通过标准语言建模目标对模型进行微调,而只不过是在维持 DeepSeek 现有发展模式的基础上以模型外的特殊机制构造实现调整效果。

针对机器幻觉安全风险,其根源于 DeepSeek 纯强化学习过程中所获得的杂乱数据,因而需在数据输入端设置特殊机制构造,减少不良数据对 DeepSeek 的影响。一方面,DeepSeek 应避免使用其他大模型生成的数据或内容,否则可能出现机器幻觉叠加涌现的困境。较之于 DeepSeek 而言,相当一部分大模型的性能较弱,但其所生成的数据或内容经常出现在各大网站、软件中,包括文本、图像和视频,容易进入 DeepSeek 的训练数据和实时检索数据库中。对此,需引入合成内容检测工具用于识别和分类大模型生成或修改的数据或内容,如 GPTZero 即是用于检测大模型生成文本的工具。同时,美国国家标准技术研究院(NIST)发布的“NIST Trustworthy and Responsible AI: Reducing Risks Posed by Synthetic Content”提出将对抗性攻击和防御作为评估合成内容检测器质量的常用框架,这些攻击通常涉及向训练集或测试集添加精心设计的数据输入,以使系统在测试集中错误分类样本。在攻击和对抗之间,DeepSeek 对于大模型生成数据或内容的鉴别能力将显著提升。另一方面,DeepSeek 可使用诸如合成数据等高质量数据。合成数据可能完全来源于收集数据或不(直接)使用收集数据或间接使用收集数据,其可以帮助机器学习算法训练并提高它们的准确性[13]。这一过程可在数据预处理阶段通过嵌入相应算法予以实现。在高质量数据(集)的基础上,可通过引入检索增强生成(retrieval - augmented generation, RAG)技术和思维链,提高 DeepSeek 在推理能力养成过程中的准确性。

针对数据侵害安全风险,重要的不是人工监督微调,而是如何将不良内容截停于输出之前,或如何将机器价值对齐人类价值,后者在解决问题上更具有彻底性。一方面,DeepSeek 应内嵌内容过滤器于数据输出端口之上,直接阻断仇恨言论、色情暴力等

有害指令的输出。同时,DeepSeek 系列模型可建立黑名单制度,对已知和潜在的恶意内容进行记录和管理,从而在数据输出端起到屏蔽效果。在制度落地层面上,可参考电气和电子工程师协会关于预防算法偏差的标准 IEEE P7003 的具体框架。该标准旨在帮助开发者定义、测量和传达算法中的偏差,并提供在设计、开发、测试和评估过程中的最佳实践,以避免不合理的差异影响的产生,其中自然包括恶意内容生成的治理框架体系。另一方面,DeepSeek 应内嵌价值对齐机制于数据输出端口之上。价值对齐机制强调 DeepSeek 的发展需符合人类价值观而非纯粹的智能体,其被认为是解决机器幻觉、越狱问题、不当引导难题等的有效路径。价值对齐机制并非仅可由监督微调所实现,其至少还包括人类反馈强化学习、合作逆强化学习、宪法人工智能、红队测试等技术路径[14]。例如,ShieldLM 可通过收集大量中英文标注的"查询—响应"来增强与人类安全标准的一致性,支持在训练时加入自定义规则,并自动创建与人类标注一致的自然语言分析,帮助解释响应为何被认为是安全、不安全或有争议[15]。价值对齐机制可以"可信根"的形式嵌入 DeepSeek 系列模型中,之后"每次数据的运行都要回溯到'可信根'去检验是否符合运行原则,对可疑伪造的、虚假篡改的不达标准数据即时熔断运行并作'无害化'处理"[16],以此杜绝恶意内容生成等的模型推理因素。

(2)数据输出结果的算法解释权

《生成式人工智能服务管理暂行办法》第 19 条第 1 款规定了大模型部署者的算法解释义务,这是大模型数据安全风险应对的巨大法治进步②。只不过该文件将算法解释权的请求主体限定在有关主管部门,导致权利主体范围过于限缩,可能使得相关数据主体无法主张权利。针对 DeepSeek 系列模型,算法解释权的请求主体应扩充至使用者,当 DeepSeek 使用的数据对使用者存在法律上或经济上的显著影响时,使用者有权向 DeepSeek 部署者提出异议,要求其予以解释并进行数据修改或更正。算法解释权尤其适合 DeepSeek 数据输出过程中所涉及的信息未经授权、数据来源不确定、数据处理不合理等情形。例如,针对已公开个人信息,我国《个人信息保护法》第 13 条第 1 款第 6 项和第 27 条确立了合理处理规则,只要 DeepSeek 在合理范围内处理相关个人信息便已合规。但事实上,判断数据获取行为的合理性依据是,是否遵循已公开个人信息的公开目的和用途[17],这种目的和用途在数据的场景混杂之间很容易被逾越。如果在目的和用途被逾越时点即对 DeepSeek 部署者施以处罚,有过于预防化保护的法治风险。设定使用者的算法解释权,以 DeepSeek 算法解释义务的履行程度判断其施以处罚的必要性,是平衡人权保障和法益保护的中间合理路径。

在算法解释权的行使条件上,首先,应进一步明晰适用场景。依据《个人信息保护法》第 24 条的规定,DeepSeek 输出何种信息内容才构成对使用者个人权益的重大影响尚有疑问。对此,可在数据分类分级的基础上,根据数据自身的重要程度以及后续可能的危害后果来认定大模型输出数据的过程是否涉嫌对个人权益的重大影响,以达到准确廓清算法解释权的行使空间以平衡技术进步与权利保障的根本目的。其次,界定解释内容的应然范围。从制度设计初衷来看,赋予使用者算法解释权显然不是让其成为技术专家,知晓人工智能运行的底层原理,而是要将处理数据的因果过程展示给用户,以使其明了个人信息的由来与去向究竟是怎样的,为个人信息权益的自我保护奠定必要的初始基础。遵此逻辑,生成式人工智能的研发者或部署者应集中说明模型训练过程中的数据来源、数据处理方式以及数据类型等关键应用信息,数据输出对个人和社会等各方面可能产生的影响,以及模型整体的安全性能[18]。最后,明确解释内容的呈现形式。出于方便使用者理解的考虑,应要求企业通过可视化的方式展现以上各种解释内容。

4 DeepSeek 数据删除安全风险的核验机制

大模型的数据删除安全风险是被既往文献普遍遗忘的数据安全风险,这一风险在 DeepSeek 系列模型中体现得尤为明显。针对 DeepSeek 所面临的客观删除不能或删除损害等困境,实现数据删除的去绝对化,并以动态核验机制抗衡逻辑复现现象是较为科学的举措。

4.1 DeepSeek 的数据删除安全风险

② 《生成式人工智能服务管理暂行办法》第 19 条第 1 款规定:有关主管部门依据职责对生成式人工智能服务开展监督检查,提供者应当依法予以配合,按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明,并提供必要的技术、数据等支持和协助。

我国《个人信息保护法》第 47 条明确个人信息的主动删除和被动删除义务,即便技术上难以删除相关个人信息的,也应进行最小化储存且禁止用于其他数据处理活动。《数据安全技术数据安全风险评估方法》将数据删除安全风险进行独立规定,内容包括数据删除管理和存储介质销毁,核心在于数据删除的及时性、有效性、彻底性验证,并同步删除多副本、缓存数据、到期备份数据、第三方委托数据,销毁存储介质。在数据删除义务普遍被认可的情况下,DeepSeek 存在若干数据删除困境。

采用 MoE 结构设计的 DeepSeek 系列模型在参数量方面已经达到惊人的 6710 亿个,这一超大参数规模会骤然升高模型本身的复杂程度,急剧削弱数据的透明性。相应地,模型内部的生成过程和外在的输出结果往往难以得到周密的解释和验证。这就会限制人们对模型工作方式的理解和审查,即便是模型原初的开发者,也无从确定目标删除数据身在何处,以及修改参数行为的系统性影响究竟是怎样的。退一步说,即使明确知道目标数据点所在地,在很多情况下,我们也不能彻底删除相关数据。一方面,DeepSeek 所具备的强大的推理能力使其完全能够对用户已手动删除的数据予以逻辑复现,让数据“起死回生”。通过模型解释性工具也可发现,微调过程中残留的隐私数据总会形成一定的幽灵特征,即便删除原始数据,这些特征仍会影响后续输出。另一方面,考虑到 DeepSeek 系列模型在自主学习和预训练上的纯粹性,主张删除特定数据恐会不当影响整个模型的运行性能,甚或诱使全系统的严重瘫痪以及与之存在紧密联系的可信网络空间的崩坏,继而造成不成比例的巨大经济损失。须知,在纯强化学习模式下,即使是研发人员自己,也难以彻底知悉各数据间的逻辑关联。即便删除目标数据未影响 DeepSeek 的性能,重新收集数据训练模型或者利用剩余数据重新训练模型的成本也是不可估量的,明显不符合经济原则[19]。

4.2　DeepSeek 的数据删除安全风险核验机制

(1)数据删除的去绝对化

对于数据删除义务的行使限度,理论上存在不同的理解,包括绝对删除和相对删除两种形式[20]。相对删除强调的是特定数据在人工智能系统空间内并非一定要达到完全不可再用的状态,而只需实现“不再可用”即可。换言之,在相对删除语境下,特定数据在人工智能系统空间内仍可继续存在,只是无法被检索和使用。绝对删除数据的观点是前数据时代的不当遗留,彼时一般数据尚未呈现出其聚合之后的高价值,因而数据的绝对化保护占据主导地位。然而,DeepSeek 内数据的绝对删除不仅在客观上无法实现,亦压缩了数据创新的空间。在数据的绝对化保护逻辑下,DeepSeek 的数据分析和使用过程需遵循原初数据主体设定的目的和用途,只要这一目的和用途发生改变,便应重新取得授权同意。毫无疑问,数据创新来源于众多数据隐藏价值的分析和挖掘,DeepSeek 的数据处理过程随时随地均可能发生无预期的使用场景变动之情形[21],对其进行强限制与创新相违。同时,数据的绝对化保护天然地偏向个体权益的保护,但若干数据不仅蕴含个体权益,尚且包括社会权益;不仅涉及人格权益,亦指向财产权益,且后种权益的价值或许更高。由是观之,数据删除与否应是多项权益相互衡量的结果。面对着个体数据权利与 DeepSeek 的数据能力之间的不对等地位,付诸于自力救济的数据绝对化保护模式反而落入下乘,需由公权机关自上而下地介入系统性、动态性、复杂性和不确定性的数据分析和利用过程中,并据此实现对相关数据主体的权益保护。此时,只要 DeepSeek 将其数据处理过程的风险控制在可被容许的范畴之内,便认为其取得了数据处理的合法性。在风险控制路径的指导下,只要 DeepSeek 的数据删除行为不至于发生不被容许的风险即可,这是数据删除去绝对化的理论依据。

(2)数据相对删除的可履行性

数据删除去绝对化后,指向特定数据的相对删除义务具有事实上的可履行性。一方面,在数据溯源层面上,大模型以外的其他领域利用时间戳、审计表、快照、事件日志、版本控制和区块链等技术能够记录数据的创建、修改和删除过程,帮助定位和删除特定数据。这些技术并非不可被用于 DeepSeek 之上。例如,谷歌公司给 AI 助手 Gemini 推出了“全局记忆”功能,订阅用户可授权 AI 永久存储并调用所有历史对话记录。同时,针对大模型的数据删除难题,机器遗忘(Machine Unlearning)方法旨消除特定训练数据对已完成预训练模型的影响,同时保持该模型的实用性。新近有研究者基于权重显著性提出一种可满足高效、精准和稳定三维特征的机器遗忘框架——SalUn(Saliency Unlearn),证明了其在机器遗忘领域的创新性和重要性[22]。另有研究指出,机器遗忘既可通过事后擦除的方式实现,亦可在源头

上防止隐私泄露。前者主要包括精确遗忘(分解模型并局部遗忘,如 SISA 训练方法)和近似遗忘(完整模型的影响消除)两种方法,后者则运用差分隐私(Differential Privacy)并在模型训练时引入随机噪声的方式解决数据的敏感性[23]。这些方法均可有效解决 DeepSeek 数据删除时训练过程的随机性、训练过程的增量性、灾难性遗忘等难题。另一方面,在逻辑复现层面上,由于数据删除的相对化不会损坏 DeepSeek 整个模型的运行性能,只要其合理运用过滤机制即可避免应被删除的数据重新出现。过滤机制通常被适用于大模型生成不良或有害内容的场景中,并依据相应的技术措施实现阻止或消除效果。其原理在于,通过运用数据输入过滤分类器,当特定数据包含在输入提示中时,基于与阻止列表的匹配性,其会阻止图像的生成。同时,数据跟踪技术在防范逻辑复现问题上亦有其作用。数字水印工具不仅可通过在内容中嵌入信息来帮助验证内容的真实性和来源,其还可与已相对删除的数据相匹配,从而达到禁止特定数据再次使用的效果。

(3)数据相对删除的动态核验

DeepSeek 数据删除义务的履行需经历根据删除数据的请求确认数据范围并在综合评估的基础上予以删除的流程。在数据删除后,如果该请求系由数据主体主动发起,则需通知其删除结果;如果该请求系由 DeepSeek 主动发起,则无需履行通知义务。DeepSeek 数据删除义务是否履行彻底是评价的核心环节,目的是防止出现非法留存、违规恢复或伪删除相关数据的行为。例如,在数据主体主动发起数据删除请求的场合,DeepSeek 可能经确认和评估后进行了表面的数据删除行为,数据主体经通知后发现确实无法再次检索到特定数据,但其事后可以“解锁”等形式轻松恢复特定数据。因此,数据删除的动态核验机制建设尤为关键。在这一机制中,核验主体主要不是请求删除数据的主体而是 DeepSeek。增设数据删除的动态核验机制必须考虑到特定主体的履行义务可能性以及成本与收益的比例性。就前者而言,动态核验机制的履行难度显著低于前一阶段的数据删除行为,因为寻找特定数据这一最艰难的环节已被克服,所谓核验只不过是按图索骥再次验证数据是否可用,因而 DeepSeek 不存在特殊履行障碍;就后者而言,DeepSeek 为此所付出的成本尚且在可接受的范围之内,其动态核验机制的展开是采取了类似使用者的视角对已相对删除的特定数据进行搜索并查看搜索结果,且这种核验行为仅需在固定的时间跨度上定期进行即可。与之相对的是,这项付出轻微成本的行为却可有效防止特定数据二次复现,进而拥有显著的合规收益。当然,这一机制的运作并非绝对的排斥提出删除请求的数据主体的参与。相关数据主体若再次搜索到已被删除的特定数据,仍可要求 DeepSeek 予以删除或上报给相关公权机关。公权机关接到举报后,既应督促 DeepSeek 部署者履行数据删除义务,亦可施以相关行政处罚。

5 结语

DeepSeek 类生成式人工智能的新技术特征衍生新型数据安全风险,包括数据泄露、数据偏离和数据删除安全风险。“技术向善”要求实现 DeepSeek 类生成式人工智能数据安全风险的科学治理[24],对此必须针对不同类型的数据安全风险推导相应的精准化治理机制,摆脱现有规定因类型化数据安全风险治理机制缺失而导致的治理失灵陷阱。针对数据泄露安全风险的预防机制、数据偏离安全风险的介入机制、数据删除安全风险的核验机制相结合,将数据保护和流通利用相协调的理念贯彻其中,并辅之以具体可行的技术措施或管理方法,最终有效避免 DeepSeek 类生成式人工智能数据安全风险损害的发生。

参考文献:

[1] 程乐. 生成式人工智能的法律规制——以 ChatGPT 为视角[J]. 政法论丛,2023(4):69 - 80. Cheng L. Legal regulation of generative Artificial Intelligence: A ChatGPT perspective[J]. Journal of Political Science and Law,2023(4):69 - 80.

[2] 郭小东. 生成式人工智能的风险及其包容性法律治理[J]. 北京理工大学学报(社会科学版),2023,25(6):93 - 105. Guo X D. Risks of generative Artificial Intelligence and its inclusive legal governance[J]. Journal of Beijing Institute of Technology(Social Sciences Edition),2023,25(6):93 - 105.

[3] 商建刚. 生成式人工智能风险治理元规则研究[J]. 东方法学,2023(3):4 - 17. Shang J G. Research on Meta - Rules for risk governance of generative AI[J]. Oriental Law,2023(3):4 - 17.

[4] 孔祥承. 国家安全视阈下生成式人工智能的法治应

对——以 ChatGPT 为视角[J]. 法治研究,2023(5):61 – 70. Kong X C. Legal governance response to generative AI from the perspective of national security—Taking ChatGPT as an example[J]. Research on Rule of Law,2023(5):61 – 70.

[5] Kaplan J, McCandlish S, Henighan Tom, et al. Scaling laws for neural language models[J]. arXiv preprint arXiv:2001.08361 (2020).

[6] Guo D, Yang D, Zhang H, et al. Deepseek – r1: Incentivizing reasoning capability in llms via reinforcement learning[J]. arXiv preprint arXiv: 2501. 12948 (2025).

[7] Bai Y, Pei G, Gu J, et al. Special characters attack: Toward scalable training data extraction from large language models[J]. arXiv preprint arXiv: 2405. 05990 (2024).

[8] 王帅. 生命治理维度下马尔库塞技术极权批判的价值与反思[J]. 科学技术哲学研究,2024,41(4):115 – 120. Wang S. The value and examination of marcuse's critique of technological totalitarianism in the dimension of life governance[J]. Studies in Philosophy of Science and Technology,2024,41(4):115 – 120.

[9] 关佳. 金融精准扶贫的数字普惠面向:核心动力与实现路径[J]. 现代经济探讨,2020(10):56 – 64. Guan J. Digital inclusive finance orientation in targeted financial poverty alleviation: Core driving forces and implementation paths[J]. Modern Economic Research,2020(10):56 – 64.

[10] 唐林垚. "元宇宙"的规制理论构建及中国方案[J]. 上海大学学报(社会科学版),2022,39(5):88 – 102. Tang L Y. Theoretical construction of Metaverse regulation and Chinese solution[J]. Journal of Shanghai University (Social Sciences Edition), 2022, 39 (5): 88 – 102.

[11] 熊波. 数据分类分级的刑法保护[J]. 政法论坛,2023,41(3):155 – 167. Xiong B. Criminal protection of data classification and grading[J]. Tribune of Political Science and Law,2023,41(3):155 – 167.

[12] 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. Tou X D. On the data security risks of generative AI and responsive governance[J]. Oriental Law,2023(5):106 – 116.

[13] Gal M S, Lynskey O. Synthetic data: Legal implications of the data – generation revolution[J]. IoWa L. rev. 109 (2023): 1087.

[14] 李思雯. 人工智能价值对齐的路径探析[J]. 伦理学研究,2024(5):99 – 108. Li S W. Exploration of paths for value alignment of Artificial Intelligence[J]. Studies in Ethics,2024(5):99 – 108.

[15] Zhang Z X, Lu Y, Ma J, et al. Shieldlm: Empowering LLMS as Aligned, Customizable and Explainable Safety Detectors[C]. In Finding of the Association for Computation Linguistics: EMNLP 2024,2024:10420 – 10438.

[16] 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. Li S. Governance paths for data security of generative AI from the perspective of risk prevention—Taking GPT – like models as an example[J]. Journal of Xizang Minzu University (Philosophy and Social Sciences Edition), 2023,44(6):139 – 145.

[17] 程啸. 论公开的个人信息处理的法律规制[J]. 中国法学,2022(3):82 – 101. Cheng X. On the legal regulation of public processing of personal information[J]. China Legal Science,2022(3):82 – 101.

[18] 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. Zhang X. Data risks and governance paths of generative AI[J]. Science of Law(Journal of Northwest University of Political Science and Law), 2023, 41 (5): 42 – 54.

[19] 林北征. 没有删除,只能遗忘:AI 大模型个人信息删除义务的解释与重构[J]. 西安交通大学学报(社会科学版),2024,44(6):152 – 165. Lin B Z. Not deletion, only forgetfulness: A purposeful interpretation to the deletion of personal in AI large models[J]. Journal of Xi'an Jiaotong University(Social Sciences),2024,44(6):152 – 165.

[20] 陈爱飞. 解释论视域下的区块链个人信息删除权[J]. 南京社会科学,2022(6):110 – 120. Chen A F. The right to erasure of personal information in blockchain from the perspective of interpretation[J]. Nanjing Journal of Social Sciences,2022(6):110 – 120.

[21] Hoofnagle C J, Van Der Sloot B, Borgesius F Z. The European Union general data protection regulation: What it is and what it means[J]. Information & Communications Technology Law, 2019, 28(1): 65 – 98.

[22] Fan C Y, Liu J, Zhang Y, et al. Salun: Empowering machine unlearning via gradient – based weight saliency in both image classification and generation[J]. arXiv preprint arXiv:2310.12508 (2023).

[23] Wang W Q, Tian Z, Zhang C, et al. Machine unlearning: A comprehensive survey[J]. arXiv preprint arXiv:2405.07406 (2024).

[24] 刘艳红. 生成式人工智能的三大安全风险及法律规制——以 ChatGPT 为例[J]. 东方法学,2023(4):29 - 43. Liu Y H. Three major security risks of generative AI and their legal regulation—Taking ChatGPT as an example[J]. Oriental Law,2023(4):29 - 43.

Governance of new type of data security risks in DeepSeek – like Generative Artificial Intelligence

ZONG Shao – hao[1], LUO Shi – long[2]

(1. The Institute for Data Law, China University of Political Science and Law, Beijing 100088, China;
2. Wuhan University School of Law, Wuhan 430072, China)

Abstract: The governance of data security risks in generative artificial intelligence is a critical issue currently facing us. The existing data security governance system for generative artificial intelligence remains in its embryonic stage, failing to establish effective typed governance mechanisms for different types of data security risks, and is out of sync with the practical development of DeepSeek – like generative artificial intelligence that has iterated to an advanced stage. The new technical features of the DeepSeek series of models—including multi – head latent attention mechanisms, mixture of experts models, and pure reinforcement learning—have led to quantitative changes or alienation of original data security risks, giving rise to data leakage security risks, data deviation security risks, and data deletion security risks with distinct operational logics. These risks require the exploration of matching typed governance mechanisms.

First, data leakage security risks stem from external attacks, simple internal configuration errors, or human negligence. To address these risks, preventive mechanisms at both the technical and management levels should be established respectively. Technical governance should take a leading role here, requiring the systematic application of technologies such as data masking and privacy – preserving computing, federated learning and distributed training optimization, as well as data encryption storage and isolation. Preventive mechanisms at the management level consist of pre – defined management matters that are implemented in real time, including basic elements such as domain warning mechanisms, data classification and grading protection mechanisms, data security organization and evaluation mechanisms, and data leakage incident emergency response mechanisms.

Second, data deviation security risks refer to various risks that deviate from reasonable expectations during the use of artificial intelligence systems, encompassing both data quality issues caused by machine hallucinations and invasive issues in data processing behaviors. In response, adjustments at both the input and output ends of data should be emphasized—distinct from the manual fine – tuning model—and the right to algorithmic explanation of data output results should be established. Here, the two – end adjustment mechanism does not mean reverting to the large – scale model development path of supervised fine – tuning; rather, it achieves adjustment effects through special mechanisms constructed outside the model while maintaining DeepSeek's current development model. For example, to address machine hallucination security risks, synthetic content detection tools can be introduced to identify and classify data or content generated or modified by large models, or high – quality data such as synthetic data can be used to reduce the impact of poor – quality data on DeepSeek. To address data infringement security risks, content filters can be embedded at data output ports to directly block the output of harmful instructions such as hate speech, pornography, and violence, or value alignment mechanisms can be embedded at data output ports. Meanwhile, the right to algorithmic explanation of data output results serves as a reasonable intermediate path to balance human rights protection and legal interest protection, particularly suitable for scenarios involving unauthorized information, uncertain data sources, and unreasonable data processing during DeepSeek's data output process.

Finally, regarding the data deletion security risks faced by DeepSeek—including objective impossibility of deletion or deletion – induced harm—de – absoluteizing data deletion and using dynamic verification mechanisms to counter logical reproduction phenomena are relatively scientific measures. After de – absoluteizing data deletion, the relative deletion obligation targeting specific data becomes factually fulfillable. On this basis, fulfilling DeepSeek's data deletion obligations involves a process of confirming the scope of data to be deleted based on deletion requests and carrying out deletion after comprehensive evaluation.

Key words: DeepSeek; ChatGPT; Generative Artificial Intelligence; data security risks; data governance