大模型训练数据安全研究报告

指导单位：

中国移动通信集团有限公司网络与信息安全管理部

编制单位：

中国移动通信有限公司研究院

中移湾区（广东）创新研究院有限公司

天翼数智科技（北京）有限公司

联通支付有限公司

中国财富研究院网络安全研究中心

专家名单(排名不分先后)：

何申、温暖、粟栗、李春梅、耿慧拯、余智、周莹、杨亭亭、郝留瑶、刘大洋、魏小珊、贺伟、杨雨菡、张萌、范世晔、刘向东、李曦明、刘颖卿、孙奥、马燕、李宽、马晶燕

前　言

数据是大模型训练的基础，是确保大模型可靠运行且释放最大价值的基础保障。随着大模型技术的快速演进，大模型训练数据安全的重要性不断提升。大模型训练数据面临投毒攻击、隐私泄露等多重挑战，对模型的攻击结果将造成行业应用方的持续影响。因此，训练数据的体系性安全研究与保障在各行业智能化转型与发展中更加重要。

本研究报告聚焦探讨大模型训练数据的特点、类型、风险、未来发展趋势等，提出了大模型训练数据全生命周期安全管理框架及技术防护对策、管理运营体系等，促进数据准备、模型构建、系统应用、数据退役等环节更加合规、透明、可控。报告号召产业链各主体共同关注大模型训练数据的安全，加强合作并实现资源共享、优势互补，共同推动大模型技术健康可持续发展。

目录

1 研究背景与目标

1.1 研究背景与意义

1.1.1 大模型在各领域的蓬勃发展态势

近年来，以深度学习为核心的大模型技术呈现爆发式增长，成为推动各行业数字化转型的关键力量。在自然语言处理领域，各类语言大模型能够完成文本生成、智能问答、语言翻译等复杂任务，广泛应用于智能客服、内容创作、智能写作等场景，显著提升了信息处理效率。在计算机视觉领域，大模型助力图像识别、目标检测、视频分析等技术不断突破，在安防监控、自动驾驶、医疗影像诊断等行业发挥重要作用。例如，在自动驾驶场景中，大模型通过对海量道路图像、传

感器数据的学习，实现精准的环境感知与决策控制；在医疗领域，基于大模型的影像分析系统能够辅助医生快速识别病变，提高诊断准确率。

此外，大模型在金融、教育、制造业等领域也展现出强大的应用潜力。在金融行业，大模型用于风险评估、信用评级、投资决策等环节，优化金融服务流程；在教育领域，个性化学习系统借助大模型分析学生学习数据，实现精准的学习推荐与辅导；在制造业，大模型支持智能生产调度、设备故障预测，推动智能制造升级。随着各行业对大模型需求的不断增长，其应用场景持续拓展，逐渐成为数字经济发展的重要引擎。

1.1.2 大模型训练数据安全的重要性

大模型训练数据安全的重要性体现在模型性能、法规合规和用户信任三个关键层面。大模型的核心能力构建在高质量且安全的数据基础之上，数据质量与安全性直接决定模型学习的准确性。在自然语言处理领域，未被污染的文本数据能帮助语言模型精准掌握语法规则、语义逻辑及语言习惯，生成符合人类表达习惯的内容，而掺杂错误拼写、语法混乱或偏见性内容的数据，会导致模型输出错误或价值观偏差的内容。从泛化能力看，安全的数据能让模型在面对未知数据时保持良好的适应性和预测能力，如图像识别模型若拥有多场景图像样本，就能准确识别各类目标，反之则易出现“过拟合”。数据安全更是模型稳定性的保障，医疗诊断模型数据被篡改可能危及患者生命，金融模型数据泄露会造成经济损失，维护数据安全是模型可靠运行的必要条件。

在法规合规方面，全球数据保护法规日益完善，企业开展大模型业务必须遵循相关要求。欧盟 GDPR 对数据主体权利、处理原则等严格规定，违规最高可处全球年营业额 4%或 2000 万欧元罚款；美国 CCPA 赋予消费者更多数据控制权；中国《数据安全法》《个人信息保护法》构建了全面治理框架。企业若不合规，不仅面临高额罚款，还会因声誉受损流失用户，影响市场竞争力，合规是规避法律风险、保障业务可持续发展的必然选择。

用户信任是大模型技术广泛应用的前提，数据安全则是赢得信任的基础。用户为获取个性化服务需提供个人信息等敏感内容，若数据安全无保障，会直接导致用户对模型及企业失去信任。从行业发展看，只有用户确信数据安全得到保护，才会提供更多数据促进模型优化迭代，反之，数据安全事件会削弱公众对人工智

能技术的信心，阻碍行业创新，因此保护数据安全关乎企业短期利益和行业生态健康。

1.2 研究目标与范围

1.2.1 研究范围

本研究聚焦于大模型训练数据安全领域，研究范围涵盖大模型训练全生命周期:

数据准备阶段：审查数据来源合法性与质量，清洗脱敏处理，检测跨模态语义关联，标注准确性校验，伦理审查防范偏见，开源数据协议与版权隐私双核查，阻断污染源头。

模型构建阶段：存储介质加密与访问控制，传输协议安全优化，防御 DDoS 与中间人攻击，联邦学习梯度加密与差分隐私保护，对抗样本检测与对抗训练增强鲁棒性。

系统应用阶段：提示词合规性检测与违规输出过滤，模型反演攻击防御，增量数据时效性与准确性多维度校验，实时监控模型性能波动，RLHF 引导合规输出，A/B 测试验证增量效果。

数据退役阶段：多重销毁技术实施与流程验证，存储介质残留数据清除，联邦学习节点数据分片物理销毁，退役数据与现役模型关联解耦，溯源信息脱敏与销毁效果审计。

1.2.2 研究目标

本研究报告拟通过系统性梳理大模型训练数据安全的全链条要素，实现以下四点研究目标：

(一)解析法规政策与行业标准：

系统梳理国内外大模型训练数据安全相关法规政策，分析其对数据主体权利、数据处理原则的规定，以及在引导行业规范发展、增强企业安全意识、促进数据合理流通与共享等方面的积极作用，同时探讨法规在实际执行中面临的挑战。

(二)解构数据安全风险体系：

从数据准备、模型构建、系统应用、数据退役全生命周期出发，分析各阶段

可能存在的数据质量、隐私泄露、数据污染等安全风险，揭示这些风险的表现形式、潜在影响及各阶段间的风险传导机制，为风险防控提供系统性认知。

(三)构建技术防护与管理框架：

提出覆盖大模型训练数据全流程的技术防护对策，结合数据加密、访问控制、隐私保护等技术手段，建立数据安全组织架构、风险评估、审计与合规管理等制度体系，实现技术防护与管理措施的协同联动，形成全方位安全保障框架。

(四)前瞻技术与产业发展趋势：

展望大模型训练数据安全领域隐私保护、数据溯源、对抗性攻击防御等技术的发展方向，以及跨行业协同、合规治理体系、专业化服务模式、数据权益市场等产业演进趋势，为行业未来发展提供前瞻性参考。

2 大模型训练数据类型与特点

2.1 大模型训练数据类型

(一)结构化数据

结构化数据是指具有固定格式和明确逻辑关系的数据，通常以表格形式存储在关系型数据库中，如 MySQL、Oracle。在大模型训练中，结构化数据常用于构建规则引擎、统计分析和预测模型。例如，银行信贷数据包含客户基本信息(年龄、收入、信用评分)、贷款记录(金额、期限、还款情况)等结构化字段，可用于训练风控模型，预测客户违约概率；电商平台的订单数据包含商品 ID、价格、购买数量、时间等信息，可用于训练销售预测模型，优化库存管理。结构化数据的优势在于便于查询、分析和处理，但在表达复杂语义和非结构化信息时存在局限性。

(二)半结构化数据

半结构化数据兼具结构化和非结构化数据的特点，通常以 XML、JSON、YAML 等格式存储，通过标签或键值对组织数据，虽无严格的表格结构，但具有一定的自描述性。在大模型训练中，半结构化数据常用于数据交换和整合场景。例如，网页中的 HTML 文档包含标题、段落、列表等标签，可通过解析提取结构化信息用于网页分类和内容推荐；API 接口返回的 JSON 数据包含多种类型字段，可直接用于模型输入。此外，半结构化数据在知识图谱构建中也发挥重要作用，如通过 JSON-LD 格式描述实体和关系，实现知识的语义互联。

(三)非结构化数据

非结构化数据是指没有固定格式、难以用传统数据库表结构存储的数据，包括文本文件、图像、音频、视频等。此类数据在大模型训练中占据重要地位，尤其在自然语言处理、计算机视觉和语音识别领域。文本数据是最常见的非结构化数据类型，用于训练语言模型进行文本生成、情感分析和机器翻译；图像数据通过卷积神经网络(CNN)处理，实现图像分类、目标检测和图像生成；音频数据借助循环神经网络(RNN)或 Transformer 架构，用于语音识别、语音合成和音频事件检测；视频数据则需结合图像处理和时序分析技术，实现视频内容理解和生成。非结构化数据的处理需要先进的机器学习算法和数据预处理技术，以提取有价值的特征用于模型训练。

2.2 大模型训练数据特点

(一)数据规模海量

大模型的卓越性能依赖于庞大的数据基础，其训练数据规模已达到 PB 级甚至更高量级。以 OpenAI 训练 GPT-3 为例，其训练数据包含超过 45TB 的文本，涉及书籍、维基百科、新闻文章等广泛来源，以覆盖人类知识的多元领域。海量数据能够支撑大模型学习复杂的语言模式、语义关系和逻辑推理规则，从而实现高质量的文本生成、问答交互等任务。在计算机视觉领域，用于训练图像识别模型的数据集，如 ImageNet 包含 1400 多万张图像，涵盖 2 万多个类别，确保模型能够学习到各类物体的特征，实现精准的图像分类和目标检测。若数据规模不足，模型可能出现“欠拟合”现象，无法充分学习数据中的潜在规律，导致性能显著下降。

(二)数据多样性高

大模型训练数据呈现出高度的多样性，涵盖文本、图像、音频、视频等多种格式，以及结构化、半结构化和非结构化数据形态。文本数据包括新闻资讯、社交媒体帖子、学术论文等，用于自然语言处理模型理解语义、语法和语境；图像数据如卫星遥感图像、医学影像、商品图片等，支持计算机视觉模型进行图像识别、分割和生成；音频数据包括语音指令、音乐、环境声音等，助力语音识别和音频生成模型的训练；视频数据则结合了图像与音频信息，可用于视频理解和生成任务。这种多样性不仅要求数据采集与存储技术具备兼容性，也对模型的多模态学习能力提出挑战，需通过跨模态融合技术实现不同类型数据的协同处理。

(三)数据时效性强

在快速变化的数字时代，大模型训练数据需具备强时效性，以捕捉最新知识和趋势。例如，金融领域的大模型需实时更新股票市场数据、经济政策动态，以准确预测市场走势；新闻推荐模型需及时获取最新新闻资讯，为用户提供实时信息。若使用过时数据训练模型，可能导致模型输出与现实脱节，如推荐陈旧新闻、做出错误的市场预测。此外，随着新技术和新应用场景的涌现，数据的时效性要求进一步提升。例如，生成式 AI 模型需不断学习最新的语言表达方式和用户需求变化，以生成符合当下语境的内容。

(四)数据关联性复杂

大模型训练数据中的关联性极为复杂，不同数据之间存在潜在联系，这些联系对模型的理解和生成能力至关重要。在知识图谱构建中，实体与关系数据相互关联，形成庞大的语义网络，帮助模型理解概念间的逻辑关系；在推荐系统中，用户行为数据(如点击、购买记录)与商品属性数据关联，使模型能够挖掘用户偏好，实现精准推荐。此外，跨模态数据之间也存在复杂关联，如视频中的画面、声音与字幕信息需协同分析，才能准确理解视频内容。模型需通过深度学习算法自动挖掘和学习这些复杂关联，以提升对数据的综合理解和应用能力，而数据关联性的复杂性也增加了数据管理和模型训练的难度。

3 大模型训练数据安全的法规政策

3.1.国外法规政策

(一)欧盟《人工智能法案》

欧盟《人工智能法案》于 2024 年 5 月正式获批，作为全球首部综合性人工智能立法，堪称人工智能监管领域的重要里程碑。该法案以风险分级为基础，将人工智能系统划分为“不可接受风险”“高风险”“通用”等类别，并分别制定了严格的监管规则。

在“不可接受风险”类别中，明确禁止利用 AI 技术对公民进行社会评分、基于生物特征识别的大规模监控，以及在执法中使用预测性警务算法等应用，从源头上杜绝严重侵犯人权和社会公平的人工智能应用。

对于“高风险”类应用，如自动驾驶汽车、医疗诊断 AI 系统、教育领域的智能评分系统等，法案提出了全面且细致的合规要求。在数据治理方面，要求数

据来源可靠、可追溯，数据收集遵循最小必要原则，并确保数据主体的知情权与控制权；在算法层面，需具备可解释性，开发者必须能够向监管机构和用户说明算法的决策逻辑，避免因算法黑箱导致的不公平决策和安全隐患。

在大模型训练场景下，若模型被认定为“高风险”或可能产生广泛社会影响，开发企业需履行严格的透明度义务。例如，公开模型训练数据的来源、规模、类别等关键信息，接受第三方审计，以验证模型在数据使用、算法设计等方面符合法规要求。

同时，法案对违规行为制定了严厉的处罚措施，罚款额度最高可达企业全球年营业额的 7%，以此形成强大的法律威慑，确保人工智能技术在欧盟市场的安全、可靠应用。

(二)美国相关法规进展

美国在人工智能监管方面采取联邦与州层面协同推进的模式。在联邦层面，2023 年 10 月拜登签署的《关于安全、可靠、值得信赖地开发和使用人工智能的行政命令》，明确了人工智能治理的国家战略方向。该行政命令聚焦于保护隐私、防范偏见、保障国家安全等核心议题，要求联邦机构在使用人工智能技术时，严格评估数据安全风险，采用隐私增强技术保护个人信息，避免算法偏见导致的不公平决策。

在州立法层面，加利福尼亚州走在前列。加州通过的《人工智能责任法案》，要求企业在使用人工智能进行决策时，需对可能产生的歧视性影响进行评估和披露。若企业的人工智能系统在招聘、贷款审批等关键领域造成不公平结果，将面临法律诉讼和高额赔偿。纽约州则推出《人工智能安全与责任法案》，对高风险人工智能系统的开发、部署和使用进行全生命周期监管，规定开发者需进行安全测试、风险评估，并向公众公开系统的关键信息，如算法原理、数据来源等。

此外，美国国家标准与技术研究院(NIST)发布了《人工智能风险管理框架》，为企业和机构提供了一套可操作的指南，帮助其识别、评估和减轻人工智能应用中的风险，涵盖数据质量、算法稳健性、隐私保护等多个维度，促进人工智能技术在安全可控的轨道上发展。

(三)其他国家和地区法规

英国：英国政府发布《促进创新的人工智能监管方法》，采取“基于原则”的治理思路，在保障数据安全、隐私保护和公平性的基础上，鼓励人工智能创新发展。同时，通过《数据保护和数字信息法案》，进一步规范数据使用规则，为人工智能发展营造良好的数据生态，促进数据在合法合规前提下的自由流动与共享，增强人工智能模型训练的数据支撑。

加拿大：《人工智能和数据法案》设立专门的咨询委员会，为人工智能监管政策制定提供多元化、独立的意见输入。该法案强调敏捷治理，要求受监管者对高影响人工智能系统开展事前风险评估，识别、评估并减轻潜在的伤害或偏见输出风险，确保技术应用符合公共利益和伦理准则，在创新与安全之间寻求平衡。

日本：2025 年 5 月通过的《人工智能相关技术研究开发及应用推进法》，旨在综合且有计划地推进人工智能研发与应用，提升国民生活水平和经济竞争力。该法注重从国家战略层面引导资源投入，促进产学研协同创新，同时强调在技术发展过程中保障数据安全、个人隐私，推动人工智能技术与社会伦理规范相融合。

3.2 国内法规政策

(一)综合性立法推进

我国在人工智能安全立法领域持续发力，已初步构建起一套多层次、多维度的法律制度框架。《中华人民共和国网络安全法》作为网络空间安全的基础性法律，为人工智能系统的网络安全保障提供了坚实基石。其明确要求关键信息基础设施运营者对重要数据进行本地存储，若因业务需求确需跨境传输，则必须通过严格的安全评估流程，以此防止数据泄露及被恶意利用，保障人工智能训练数据在网络传输环节的安全性。

《中华人民共和国数据安全法》进一步强化了数据安全管理的主体责任，规定企业等各类数据处理者需建立健全数据分类分级保护制度。在人工智能大模型训练场景下，企业需依据数据的敏感程度，如将数据划分为核心数据、重要数据和一般数据，实施差异化的安全防护策略。对核心数据采用高强度加密算法，确保数据存储与传输的保密性；针对重要数据，设置严格的访问控制权限，限制特定人员在特定场景下的访问，降低数据被不当获取或篡改的风险。同时，企业还需定期对数据处理活动开展全面的风险评估与监测，及时发现并处置潜在的数据安全隐患。

《中华人民共和国个人信息保护法》着重聚焦个人信息保护，在人工智能数据安全治理中发挥着关键作用。该法明确禁止过度收集个人信息行为，严格规范个人信息处理规则。对于敏感个人信息，如生物识别、医疗健康数据等，企业在处理前必须取得用户的 “单独同意”，且在大模型训练过程中，要求企业对涉及的个人信息进行匿名化处理，切断信息与特定个人的直接关联，最大程度保护用户隐私。此外，法律强制规定数据处理者需设立个人信息保护负责人制度，定期开展合规审计，确保个人信息处理活动全程符合法律规范。

(二)专项政策与法规出台

战略规划引领：《新一代人工智能发展规划》作为我国人工智能领域的纲领性文件，为产业发展制定了清晰的战略蓝图。规划中明确将人工智能安全发展纳入重点任务，强调在推动技术创新与产业应用的同时，同步加强安全风险防范与治理，为后续一系列政策法规的制定提供了宏观指导方向。

算法与应用规范：《关于加强互联网信息服务算法综合治理的指导意见》以及《互联网信息服务算法推荐管理规定》，针对人工智能算法的设计、开发、应用等环节，提出了全面的治理要求。算法开发者需确保算法的公平性、透明性和可解释性，避免算法歧视与偏见，防止算法被滥用导致安全风险。而《互联网信息服务深度合成管理规定》和全球首部生成式人工智能专门立法—《生成式人工智能服务管理暂行办法》，则紧密围绕人工智能合成技术与生成式应用，从技术发展、服务规范、监督检查等维度进行详细规制。规定服务提供者需对训练数据来源的合法性、真实性负责，保障数据质量；明确生成内容的标识义务，防止虚假信息误导公众；建立安全评估与投诉举报机制，及时处理各类安全问题，全方位促进人工智能应用的健康、规范发展。

内容标识新规：将于 2025 年 9 月 1 日起施行的《人工智能生成合成内容标识办法》，进一步完善了人工智能内容管理规范。该办法明确界定了人工智能生成合成内容的范围，要求服务提供者针对文本、图片、音频、视频、虚拟场景等生成合成内容，添加清晰可辨的显式标识，或在文件元数据中嵌入隐式标识，帮助用户准确识别信息来源与性质。同时，对内容传播者也提出核验与二次标识要求，有效遏制虚假信息传播扩散，维护网络信息传播秩序。

产业支持与规范：《“十四五” 大数据产业发展规划》《关于促进企业数据资源开发利用的意见》《关于促进数据标注产业高质量发展的实施意见》以及《关于促进数据产业高质量发展的指导意见》等政策文件，从大数据产业整体布局出发，深入到数据资源开发、数据标注等细分领域，为人工智能发展所需的数据资源提供了充足的政策支持与规范引导。通过鼓励数据要素流通共享、提升数据质量、加强数据标注产业标准化建设等举措，既保障数据资源的丰富供给，又确保数据在收集、处理、使用过程中的安全合规，夯实人工智能发展的数据基础。

安全治理与伦理审查：全国网络安全标准化技术委员会发布的《人工智能安全治理框架》1.0 版，针对模型算法安全、数据安全和系统安全等内生安全风险，以及网络域、现实域、认知域、伦理域等应用安全风险，提出了包容审慎、风险导向、技管结合、开放合作的治理原则，并给出具体的技术应对与综合防治措施，为人工智能安全开发应用提供了重要的技术指引。此外，正在推进的《科技伦理审查办法(试行)》，致力于从伦理审查角度，规范人工智能技术研发与应用活动，确保技术发展符合社会伦理道德准则，防范因技术滥用引发的伦理风险，推动人工智能在安全、伦理的双重约束下稳健前行 。

4 大模型训练数据安全风险分析

4.1 安全风险体系

大模型训练数据安全风险以数据生命周期为脉络，形成环环相扣的系统性风险网络。

数据准备阶段，训练数据偏见风险源于数据集中歧视性内容、偏差标注或群体表征失衡，使模型学习到错误价值观；跨模态数据关联风险因语义映射篡改或噪声干扰，导致模型习得错误关联逻辑；开源数据合规风险则来自协议条款复杂、数据来源不可控，易引发版权纠纷与法律追责。

进入模型构建阶段，训练过程数据泄露风险因训练日志和中间参数保护不当，导致数据特征与算法逻辑暴露；联邦学习隐私风险源于梯度更新信息可被逆向分析，还原原始数据敏感特征；对抗样本污染风险使模型学习错误决策边界，对正常数据判断产生系统性偏差。

系统应用阶段，提示词注入数据污染风险通过恶意提示诱导模型生成违规内容，污染训练数据；模型反演数据泄露风险利用模型输出逆向推导训练数据敏感

信息；增量训练数据失控风险因未验证的新增数据携带过时或错误信息，干扰模型知识体系，降低模型性能。

数据退役阶段，训练数据溯源残留风险因退役数据含数据处理全流程细节，泄露后导致技术优势丧失;联邦学习数据残留风险由未彻底清除的分片数据引发，可拼凑还原原始数据集；模型迭代数据关联风险源于退役数据与现役模型的隐性联系，即使脱敏也可能成为信息泄露突破口。

这些风险相互交织、层层传导，任一环节的风险失控都可能引发跨阶段的几何级放大效应，形成覆盖大模型训练数据全生命周期的安全威胁。

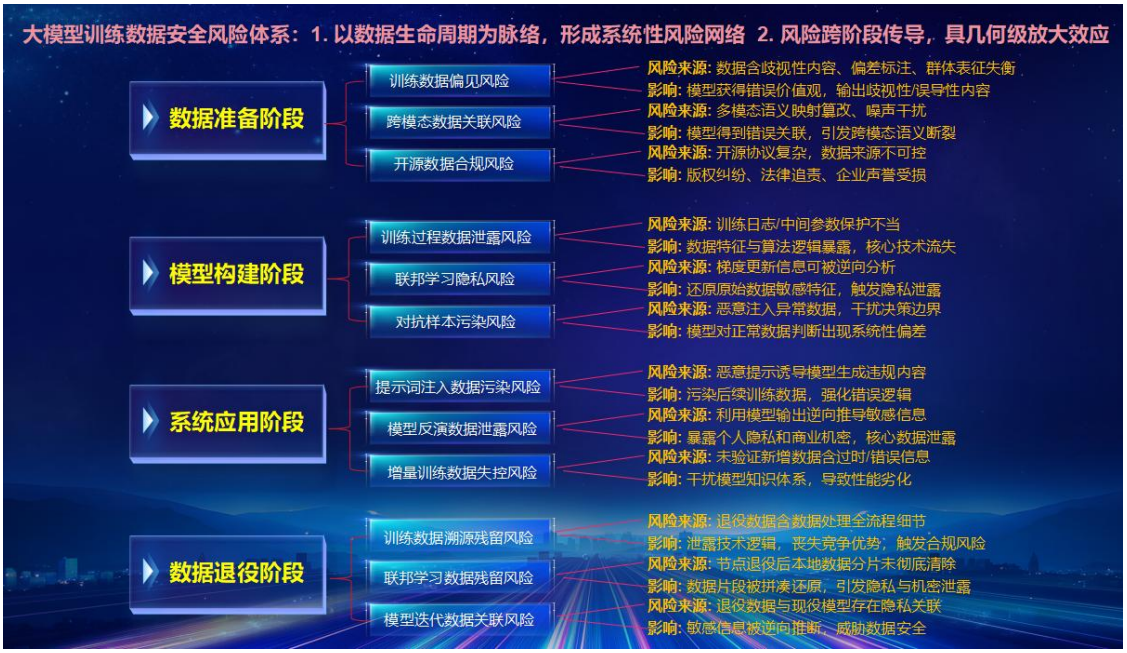

图 1 大模型训练数据安全风险体系

4.2 数据准备阶段安全风险

4.2.1 训练数据偏见风险

在大模型训练的数据准备阶段，训练数据偏见风险是不容忽视的关键隐患。数据集中若包含偏见性内容，如歧视性文本、偏差性标注或失衡的群体表征，会如同隐藏的“病毒”般渗透进模型的学习过程。例如，文本数据中对特定职业、性别、种族的刻板描述，图像数据中对不同文化符号的片面呈现，或是标注体系中隐含的主观价值判断，都可能导致模型在学习语言规律和语义关联时，无意识地吸收并强化这些偏见。当模型输出内容时，这些潜在的偏见会以价值观偏移的形式显现，可能生成带有歧视性、误导性或违背公序良俗的回答，影响公众认知

甚至引发社会争议。这种风险不仅源于数据采集时的样本选择偏差，也可能来自数据标注流程中的人工主观干预，且一旦进入训练环节，偏见的修正需付出极高成本。因此，在数据准备阶段建立严格的伦理审查机制、采用自动化工具检测偏见倾向，并结合人工校准剔除违规内容，是阻断训练数据偏见风险传导至模型输出的关键手段，关乎大模型应用的社会价值导向与伦理底线。

4.2.2 跨模态数据关联风险

在大模型数据准备阶段，跨模态数据关联风险源于多模态数据间语义映射被篡改或扭曲，导致模型学习到错误关联逻辑。文本、图像、音频、视频等模态需通过语义关联构建统一表征，但若数据处理环节中模态对应关系被人为操纵或引入噪声，模型会误将异常关联视为有效规则。例如，文本主题与图像内容背离、音频情感与视频场景冲突等隐性矛盾，可能被模型捕捉为“合法”模式，致使生成任务中出现跨模态语义断裂，如严肃文本匹配荒诞图像、欢快音频对应悲伤场景等。

此类风险隐蔽性强，传统质量检测难以及时识别深层语义异常。模型一旦习得错误关联，可能在推理中引发连锁反应，导致多模态内容语义一致性崩塌，甚至在关键场景引发误导。防范需构建跨模态语义校验机制：利用联合嵌入模型量化评估模态对语义相似度，设置阈值过滤异常关联；建立时序逻辑校验规则，确保模态间时空对应与物理规律一致。通过全流程语义一致性管控，阻断错误关联注入路径，保障多模态数据映射真实可靠，避免模型因数据污染产生决策偏差。

4.2.3 开源数据合规风险

在大模型数据准备阶段，开源数据合规风险如同隐藏的“法律地雷”，随时可能因引用不当引发严重后果。开源数据集虽为训练提供便利，但协议条款的复杂性与数据来源的不可控性，使其成为合规隐患的高发区。许多开源协议对数据使用、修改、再分发有严格限定，若训练方未完整遵循协议要求，如超出授权范围使用数据、未按规定保留版权声明、随意修改数据后对外发布，或未履行数据共享、溯源义务，都可能构成违约，面临法律追责。

更棘手的是，部分开源数据集在采集、整合过程中，可能混入未获授权的受版权保护数据，或包含侵犯个人隐私、违反伦理道德的内容。若训练方未对开源

数据进行严格筛查，将违规数据纳入训练，不仅会导致模型存在法律瑕疵，还可能引发舆论危机，损害企业声誉。因此，建立完善的开源数据合规审查机制至关重要，需逐一对数据集协议条款进行解析，通过技术手段与人工审核结合，筛查数据来源合法性，确保所有引用行为符合法律规范与伦理要求，从源头规避潜在风险。

4.3 模型构建阶段安全风险

4.3.1 训练过程数据泄露风险

在大模型模型构建阶段，训练过程数据泄露风险是威胁数据安全的重要隐患。模型训练日志、中间参数等信息若未被妥善保护，可能因系统漏洞、人为操作失误或权限管理失控等原因意外泄露，导致数据特征与算法逻辑暴露。训练日志详细记录了数据预处理流程、特征工程细节及训练迭代过程，中间参数则包含模型在学习过程中捕获的关键数据模式与权重分布，二者均蕴含数据内在特征与算法实现逻辑。

此类泄露可能使竞争对手或恶意主体通过分析日志内容，逆向推断训练数据的敏感属性(如用户隐私字段、行业敏感指标)，或通过解析中间参数，还原模型的核心算法逻辑与优化策略，导致企业核心技术资产流失。更严重的是，若训练数据涉及国家安全、商业机密或个人信息，泄露事件可能触发合规风险，面临监管处罚与法律追责。因此，需构建全流程数据防护体系：对训练日志与中间参数实施加密存储，限制访问权限至最小必要范围；采用联邦学习、差分隐私等技术手段，在保障模型性能的同时阻断数据特征与算法逻辑的泄露路径；建立实时监控与应急响应机制，对异常访问行为与数据流动实施动态预警，确保训练过程数据安全可控，避免因信息泄露引发技术优势丧失与合规危机。

4.3.2 联邦学习隐私风险

在大模型构建的联邦学习场景中，隐私风险如同潜伏的“数据猎手”，暗藏于梯度更新的传输与交互过程。联邦学习虽以“数据不动模型动”为核心，避免原始数据跨域传输，但训练中节点上传的梯度更新信息，仍可能成为隐私泄露的突破口。由于梯度包含模型参数的变化趋势与数据特征的统计信息，恶意参与者

可通过精心设计的逆向分析攻击，如梯度反演攻击、成员推理攻击等，从梯度更新数据中还原出原始数据的敏感特征。

这种风险在多节点协作训练时尤为突出，单一节点的梯度更新看似无关紧要，但随着训练迭代，多个节点的梯度信息相互关联，攻击者便有机会拼凑出完整的原始数据特征。若原始数据涉及个人身份信息、医疗记录、商业机密等敏感内容，隐私泄露不仅会损害数据所有者权益，还可能触发法律合规风险，引发监管处罚。因此，需强化联邦学习的隐私保护机制，通过同态加密、安全多方计算、差分隐私等技术，对梯度更新数据进行脱敏与混淆处理，阻断攻击者从梯度反推原始数据的路径，确保联邦学习过程中数据隐私安全。

4.3.3 对抗样本污染风险

在大模型构建阶段，对抗样本污染风险源于攻击者向训练数据注入精心设计的异常数据，致使模型学习到错误的决策边界。此类样本通过细微调整数据特征分布或添加针对性噪声，干扰模型正常学习过程，迫使模型将错误分类、异常输出纳入决策逻辑。

训练过程中，模型基于包含对抗样本的数据集进行参数优化，会逐渐将错误模式识别为有效特征。随着迭代次数增加，被污染的模型会固化错误决策逻辑，导致对正常数据的判断出现系统性偏差。例如在图像识别任务中，对抗样本可能使模型混淆不同类别图像；在文本处理中，少量特征修改即可误导模型误判语义倾向。若应用于金融风控、医疗诊断等关键领域，错误决策将造成严重后果。

防范对抗样本污染需构建全流程防御机制。在数据预处理阶段，采用统计分析与异常检测算法识别潜在恶意样本；训练过程中，通过对抗训练、正则化等技术增强模型鲁棒性；同时建立动态监控体系，实时检测模型性能波动，及时发现并阻断污染风险，确保模型决策边界符合真实数据分布与业务需求。

4.4 系统应用阶段安全风险

4.4.1 提示词注入数据污染风险

在大模型系统应用阶段，提示词注入数据污染风险主要体现在用户通过构造恶意提示词，诱导模型生成违规内容，并污染后续训练数据。攻击者利用模型对

输入提示词的响应机制，精心设计包含错误信息、敏感内容或恶意指令的提示，迫使模型输出违背伦理规范、法律法规或业务规则的内容。

当模型的输出数据被纳入后续训练集时，这些受污染的内容会被模型重复学习，导致错误模式、偏见认知或有害逻辑在模型中不断强化。随着迭代训练的持续进行，模型将逐渐偏离预定的安全输出范围，产生更多错误或有害的响应。例如在文本生成任务中，恶意提示词可能诱导模型输出虚假信息、煽动性言论或侵权内容；在知识问答场景下，使模型给出错误知识或误导性回答。

为防范此类风险，需建立严格的提示词与输出内容审核机制。通过自然语言处理技术对输入提示词进行合规性检测，识别潜在恶意指令；同时对模型输出内容实施实时监控与过滤，阻断违规内容进入训练数据。此外，可采用强化学习从人类反馈(RLHF)等技术，引导模型生成符合规范的内容，降低被恶意提示词污染的可能性。

4.4.2 模型反演数据泄露风险

在大模型系统应用阶段，模型反演数据泄露风险源于攻击者利用模型输出结果逆向推导训练数据中的敏感信息。由于模型参数与训练数据存在潜在关联，攻击者可通过构造特定输入、分析输出响应，逐步挖掘出原始训练数据中的隐私内容。攻击者通常采用成员推理攻击或属性推理攻击等手段。成员推理攻击通过多次调用模型，分析输出差异判断特定数据是否存在于训练集中；属性推理攻击则基于模型输出结果，推断训练数据中特定个体的敏感属性。在人脸识别、医疗诊断等涉及个人隐私或机密数据的应用场景中，攻击者可通过反复测试模型，获取训练数据中包含的身份信息、健康记录、商业机密等敏感内容。

随着模型复杂度增加，攻击者还可利用生成对抗网络(GAN)等技术，通过模型输出重建原始数据特征。此类攻击不仅威胁用户隐私安全，还可能导致企业核心数据泄露。防范模型反演风险需从技术和管理层面共同发力，采用差分隐私、同态加密等技术对模型参数和输出结果进行脱敏处理，同时限制模型调用权限，建立访问审计机制，降低数据泄露风险。

4.4.3 增量训练数据失控风险

在大模型系统应用阶段，增量训练数据失控风险主要源于未经严格安全验证的新增数据引入模型训练流程。随着应用场景的拓展与数据持续积累，大量实时生成的增量数据若未经过滤、清洗和有效性验证，易携带过时知识、错误信息或低质量内容，导致模型性能出现显著波动。

由于增量数据的时效性、真实性难以把控，陈旧或错误的信息可能与原有模型知识体系产生冲突，干扰模型的正确学习。当模型在训练过程中盲目吸收这些低质量数据时，不仅无法提升性能，反而会破坏已有的稳定知识结构，造成模型泛化能力下降、准确率降低等问题。在金融预测、医疗诊断等对数据时效性和准确性要求极高的领域，增量数据中过时的市场趋势、错误的诊断标准等信息，可能导致模型输出严重偏离实际的错误结果。

防范此类风险需构建完善的增量数据安全验证机制。通过设置严格的数据筛选规则，对新增数据的时效性、准确性、合规性进行多维度校验；利用数据质量评估算法识别低质量数据，并建立动态反馈机制，实时调整数据筛选策略。同时，对增量训练过程进行严格监控，确保新增数据能够安全、有效地融入模型知识体系，避免因数据失控引发模型性能劣化。

4.5 数据退役阶段安全风险

4.5.1 训练数据溯源残留风险

在大模型数据退役阶段，训练数据溯源残留风险对数据安全构成直接威胁。退役数据中的溯源信息涵盖数据采集路径、标注规范、特征工程算法、清洗规则等全流程技术细节，完整保留了数据处理的核心逻辑。

当这些包含溯源信息的退役数据脱离安全管控，无论是存储介质不当流转，还是数据迁移操作疏漏，都可能导致敏感信息泄露。恶意主体获取数据标注规则后，可逆向解析数据分类标准；掌握特征工程参数，便能复刻数据预处理算法；而清洗流程记录一旦泄露，数据筛选阈值、异常值处理策略等关键技术细节将暴露无遗。

数据处理逻辑的泄露，不仅会使企业丧失技术竞争优势，还可能导致竞争对手直接复制数据处理流程，削弱模型差异化竞争力。若溯源信息涉及敏感数据处理方式，更可能违反数据安全法规，引发监管处罚。同时，数据处理逻辑作为模

型构建的核心资产，其泄露将动摇企业数据安全根基，影响业务连续性，甚至可能引发法律纠纷与声誉危机。

4.5.2 联邦学习数据残留风险

在大模型数据退役阶段，联邦学习数据残留风险源于联邦学习节点退役后，本地存储的数据分片未得到彻底清除。联邦学习通过多个节点协同训练，各节点仅保留本地数据分片，避免原始数据集中传输，以保障数据隐私。然而，当节点完成任务进入退役流程时，若未对本地存储的数据分片进行完全擦除或物理销毁，残留的数据片段便成为潜在的安全隐患。

这些残留的数据分片可能包含原始训练数据的关键特征、敏感属性或业务核心信息。一旦存储介质被非法获取，或是因管理疏漏导致数据泄露，恶意主体可通过收集多个节点残留的数据片段，拼凑还原出完整或接近完整的原始数据集。在涉及医疗健康、金融交易等敏感领域的联邦学习场景中，残留数据若包含患者病历、交易记录等信息，将直接导致用户隐私泄露，严重侵犯个人权益。

此外，数据分片残留还可能引发知识产权纠纷与商业机密泄露风险。企业或机构在联邦学习项目中积累的数据资产与训练成果，若因节点退役处理不当而外泄，不仅会造成数据价值流失，还可能削弱参与方在市场竞争中的技术优势，甚至引发法律诉讼，对企业声誉和运营造成重大打击。

4.5.3 模型迭代数据关联风险

在大模型数据退役阶段，模型迭代数据关联风险源于退役数据与现役模型间存在的隐性联系，这种联系可能导致敏感信息被逆向推断。大模型在迭代过程中，新模型的训练通常基于历史数据和先前模型的优化结果，使得退役数据与现役模型在数据特征、参数分布、训练逻辑等方面存在内在关联。

当退役数据脱离受控环境后，恶意主体可通过分析现役模型的输出特征、参数权重等信息，结合公开的模型架构和训练方法，逆向推断退役数据的关键特征和敏感内容。即使退役数据经过脱敏处理，模型迭代过程中形成的关联关系仍可能成为信息泄露的突破口。例如，通过观察模型对特定输入的响应模式，攻击者能够推测出退役数据集中的敏感属性分布；利用模型参数的变化趋势，可反推训练数据中的关键特征选择逻辑。

这种风险在涉及个人隐私、商业机密或国家安全等敏感领域尤为严重。一旦敏感信息被逆向推断，不仅会导致用户隐私泄露、商业机密曝光，还可能使企业核心技术资产遭到窃取，严重削弱市场竞争力。此外，模型迭代数据关联风险还可能引发法律合规问题，违反数据安全相关法规，给企业带来监管处罚和声誉损失。

5 大模型训练数据安全防护对策

5.1 安全防护对策体系

大模型训练数据安全防护的四个阶段对策并非孤立存在，而是通过大模型训练数据生命周期形成紧密的逻辑链条与协同关系，共同构筑起系统性防护体系：

层层递进，环环相扣：数据准备阶段作为起点，通过防范数据偏见、校验跨模态语义、核查开源数据合规，从源头上确保数据质量与合法性，为后续阶段提供安全可靠的数据基础。模型构建阶段基于此，通过守护训练数据隐私、加固联邦学习安全、阻断样本污染，保障数据在训练过程中的安全性和模型的可靠性。系统应用阶段则聚焦数据在实际使用中的安全，拦截提示词污染、增强反演防御、优化增量训练，防止数据在交互过程中被污染或泄露。数据退役阶段通过保障溯源信息安全、阻断联邦数据泄露、解耦数据关联风险，彻底消除退役数据带来的潜在隐患，四个阶段依次推进，形成完整的数据安全防护链条。

前序为基，后序反馈：前一阶段的防护对策是后一阶段的基础，若数据准备阶段未能有效消除偏见或确保数据合规，后续模型构建和应用阶段将面临更大风险。同时，后一阶段会向前反馈优化需求，如系统应用阶段发现的提示词污染问题，可促使数据准备阶段加强对敏感内容的检测；增量训练中暴露出的数据质量问题，能够推动数据筛选和质量评估规则的完善。这种反馈机制使防护体系不断优化升级。

技术交叉，协同增效：各阶段防护对策在技术手段上相互交叉融合。例如，差分隐私技术在模型构建阶段用于保护联邦学习隐私，在系统应用阶段也可增强模型反演防御能力；加密技术在数据准备阶段用于保护开源数据中的敏感信息，在模型构建阶段用于训练数据加密，在数据退役阶段用于处理溯源信息等。多种技术的协同使用，实现了对数据安全风险的多维度、全方位防控。

目标统一，动态防护：四个阶段的防护对策都围绕大模型训练数据安全这一核心目标，共同致力于保障数据的完整性、保密性和可用性。随着数据在不同阶段的流动和变化，各阶段防护对策动态调整、相互配合，共同应对数据在全生命周期中面临的各类安全风险，形成一个有机的整体防护体系。

图 2 大模型训练数据安全防护对策体系

5.2 数据准备阶段安全防护对策

5.2.1 全流程防范训练数据偏见

针对训练数据偏见风险，可从数据采集、标注、审查等环节构建全流程解决方案。在数据采集阶段，需制定严格的数据采样标准，确保样本覆盖不同群体、地域和文化背景，避免因采样范围狭窄导致的数据失衡；同时建立数据多样性评估指标，量化衡量数据集中各特征分布的均衡程度。

在数据标注环节，通过标准化标注流程和明确的标注指南，减少标注人员的主观偏差。采用多人交叉标注、标注结果一致性校验等方式，降低个体认知差异带来的影响。引入主动学习技术，优先标注易产生分歧或存在模糊性的数据，提升标注质量。

在数据审查环节，利用自动化检测工具对数据集进行偏见识别，通过自然语言处理技术和图像分析算法，识别文本中的歧视性表述和图像中的刻板呈现。同时，组建多学科背景的伦理审查团队，从社会、法律、伦理等多角度对数据进行人工复核，对存在偏见的数据进行修正或剔除。通过技术手段与人工干预相结合，最大限度降低训练数据中的偏见风险，保障大模型输出内容的客观性与公正性。

5.2.2 联合校验跨模态语义关联

针对跨模态数据关联风险，可从数据清洗、关联校验、动态监测三方面构建解决方案。在数据清洗环节，采用多模态预训练模型对原始数据进行初步筛查，利用其在大规模数据上学习到的通用语义表征能力，识别各模态数据间明显的语义冲突与异常关联。同时，通过人工标注与自动化工具相结合的方式，对潜在风险数据进行深度核查，剔除语义不匹配的跨模态数据对。

在关联校验阶段，构建跨模态联合表征学习框架，将不同模态数据映射到统一语义空间，通过余弦相似度、欧氏距离等度量方法量化模态间语义一致性，对低于设定阈值的数据关联进行标记与处理。引入逻辑规则引擎，基于常识知识与领域规则，验证跨模态数据在时空逻辑、物理规律等方面的合理性，如验证视频场景与音频内容的情感匹配度。

在动态监测阶段，建立实时监控系统，在模型训练与应用过程中持续监测跨模态数据的关联情况，一旦发现异常关联趋势及时预警。通过不断优化语义校验模型与规则体系，提升对复杂语义异常的识别能力，确保跨模态数据在整个生命周期内的语义一致性与可靠性。

5.2.3 开源数据版权隐私双核查

针对开源数据合规风险，可构建覆盖全生命周期的合规管理体系。在数据引入阶段，建立开源数据集协议解析机制，由法律与技术团队协同审查 BSD、MIT、GPL 等不同协议的授权范围，明确数据使用、修改、分发的具体限制，对超出业务需求授权的数据集拒绝引入。同时，通过开源协议数据库自动匹配条款风险点，标记需特殊处理的权利保留事项。

在数据筛查阶段，采用知识产权检测工具扫描数据内容，比对版权登记库与公开资源，识别潜在侵权素材；运用隐私计算技术对个人信息字段进行脱敏处理，

确保数据符合 GDPR、《个人信息保护法》等法规要求。针对医疗、教育等敏感领域数据，额外开展伦理审查，杜绝包含歧视性、攻击性内容的数据进入训练流程。

在数据使用阶段，建立合规性日志记录系统，实时追踪数据调用、修改、输出的全流程，确保可溯源；对衍生模型或二次开发成果进行协议兼容性评估，避免因分发模式与开源协议冲突引发版权纠纷。通过定期开展合规培训、更新开源协议知识库、引入第三方合规审计等措施，形成动态化的风险防控机制，保障开源数据使用全程合法合规。

5.3 模型构建阶段安全防护对策

5.3.1 最小权限守护训练数据隐私

针对训练过程数据泄露风险，可从数据加密、权限管控、技术防护、监控响应四个维度构建解决方案。在数据加密层面，对训练日志、中间参数等敏感数据实施全生命周期加密，采用国密算法对存储介质加密，传输过程中启用 TLS 协议确保数据传输安全，防止因存储介质丢失或传输截获导致信息泄露。

在权限管理层面，建立基于角色的最小权限访问控制体系，严格划分训练、运维、审计等不同角色的权限边界，通过多因素认证强化身份核验，禁止跨角色越权访问。对关键数据操作记录完整日志，确保操作行为可追溯，杜绝内部人员违规访问导致的数据泄露。

在技术防护层面，引入联邦学习框架实现数据不动模型动，避免原始数据集中暴露；应用差分隐私技术对训练数据添加噪声，在不影响模型性能的前提下保护数据特征隐私。同时，采用安全计算环境隔离训练过程，通过容器化技术构建封闭训练空间，阻断数据泄露的技术路径。

在监控响应层面，部署实时数据安全监测系统，对训练日志与中间参数的访问、调用、导出等行为进行异常检测，设置流量阈值与行为基线，发现可疑操作立即触发预警并自动阻断。定期开展训练环境安全评估与渗透测试，持续优化防护策略，确保训练过程数据安全可控。

5.3.2 差分隐私加固联邦学习安全

针对联邦学习隐私风险，可构建多层级技术防护体系强化隐私保护。在梯度处理环节，采用同态加密技术对上传的梯度更新数据进行加密处理，使服务器在

密文状态下聚合梯度，避免明文梯度直接暴露；结合安全多方计算协议，将梯度分解为多个子秘密分发给不同节点，确保单一节点无法获取完整梯度信息，阻断梯度反演攻击路径。

在隐私增强层面，引入差分隐私机制向梯度更新中添加合适噪声，通过扰动数据特征的统计信息，降低攻击者从梯度中推断原始数据的可能性；同时优化噪声添加策略，在保证模型收敛精度的前提下，最大化隐私保护强度。针对联邦学习多轮迭代中的隐私累积风险，设置全局隐私预算管理机制，动态监控各节点隐私消耗情况，防止过度隐私泄露。

在系统架构层面，构建可信执行环境(TEE)作为联邦学习的安全计算基座，利用硬件隔离技术确保梯度聚合过程在加密容器内完成，外部无法窥探计算过程与中间结果。此外，建立联邦学习节点身份认证与行为监控机制，对异常梯度上传行为进行实时检测与阻断，通过技术防护与管理措施结合，形成全流程隐私保护闭环，保障联邦学习场景下的数据隐私安全。

5.3.3 实时监控阻断样本污染链条

针对对抗样本污染风险，可构建覆盖数据采集、训练、部署全流程的防御体系。在数据采集阶段，建立多维度异常检测机制，通过对比样本与历史数据分布的统计特征，识别数据特征突变、分布异常的潜在对抗样本。运用聚类分析算法对输入数据进行无监督学习，将偏离正常簇群的样本标记为可疑数据。

在模型训练阶段，引入对抗训练机制，通过生成对抗网络(GAN)生成模拟攻击样本，与原始训练数据混合训练，迫使模型学习鲁棒特征，提高对对抗扰动的免疫力。优化损失函数设计，添加正则化项约束模型对细微扰动的敏感性，避免模型过度拟合对抗样本引入的噪声。

在模型部署阶段，建立实时监控与响应系统，持续监测模型输出结果的稳定性与一致性，设置异常检测阈值，当模型对同类输入的输出出现大幅波动时触发预警。采用模型集成策略，通过多个独立训练的模型对同一输入进行交叉验证，当不同模型输出结果存在显著差异时，拒绝该输入并进行人工审核。通过全流程防御措施，确保模型在面对对抗样本攻击时仍能保持决策的准确性与可靠性。

5.4 系统应用阶段安全防护对策

5.4.1 双校验拦截提示词数据污染

针对提示词注入数据污染风险，可构建覆盖输入检测、输出过滤、反馈优化的全链条防护机制。在提示词检测环节，运用自然语言处理技术构建多维度合规性分析模型，通过关键词匹配、语义情感分析及违规模式识别，实时检测输入提示词中的恶意指令、敏感内容或错误引导倾向，对高风险提示词实施拦截或预警。

在输出内容管控层面，建立动态过滤系统，结合规则引擎与机器学习模型，对模型生成内容进行伦理合规性、法律风险性及业务规则一致性校验，自动识别并阻断虚假信息、煽动性言论、侵权内容等违规输出，防止污染数据进入后续训练流程。同时，设置人工复核环节，对系统标记的可疑内容进行深度审核，提升过滤精度。

在模型优化环节，强化基于人类反馈的强化学习(RLHF)机制，通过专业标注团队对模型输出进行质量评分与合规性评估，将反馈信号转化为训练信号，引导模型优先学习符合规范的响应模式。定期开展模型安全性评估，通过模拟恶意提示词攻击测试模型鲁棒性，持续优化检测过滤策略，形成“检测-过滤-优化”的闭环防护体系，阻断提示词注入导致的数据污染路径。

5.4.2 架构优化增强反演防御能力

针对模型反演数据泄露风险，需构建技术防护与管理控制相结合的立体防御体系。在技术防护层面，采用差分隐私技术对模型参数和输出结果进行处理，通过向输出添加精心设计的噪声，在保证模型可用性的同时，降低攻击者从输出结果推断训练数据的可能性。引入同态加密技术，允许在加密状态下对模型进行计算，使得攻击者即使获取输出结果，也无法解密其中的敏感信息。

优化模型架构设计，在满足业务需求的前提下，减少模型对特定训练数据特征的依赖，降低模型与训练数据之间的关联性。采用模型集成策略，通过多个独立训练的模型对同一输入进行处理，然后综合多个模型的输出结果，增加攻击者逆向推导原始数据的难度。

在管理控制层面，建立严格的模型访问权限管理机制，对模型调用进行细粒度的权限控制，仅授权可信用户或系统访问模型。同时，部署全面的访问审计系统，记录所有模型调用行为，包括调用时间、输入参数、输出结果等信息，以便

在发生数据泄露事件时能够快速追踪和定位问题。定期进行安全评估和渗透测试，模拟攻击者的行为对模型进行攻击测试，及时发现并修复潜在的安全漏洞，确保模型在面对各种攻击手段时都能保持数据安全。

5.4.3 闭环管理优化增量数据训练

针对增量训练数据失控风险，可构建覆盖数据筛选、质量评估、训练监控的全流程管理体系。在数据筛选环节，制定多维度筛选规则，通过时间戳比对验证数据时效性，结合业务规则库校验数据合规性，利用知识图谱交叉验证信息准确性，自动拦截陈旧、错误或违规数据。

在质量评估层面，建立动态数据质量评分模型，综合运用统计分析、异常检测算法评估数据完整性、一致性与可信度，对评分低于阈值的数据进行标记或剔除。同时，引入主动学习机制，优先选取高价值数据参与训练，并通过人工抽检与自动化验证相结合的方式，持续优化质量评估策略。

在训练监控阶段，部署实时性能监测系统，对比增量训练前后模型的准确率、召回率等核心指标变化，当性能波动超过预设阈值时触发预警。采用影子模式并行评估增量数据对模型的影响，通过 A/B 测试验证新增数据的有效性，确保训练过程可控。建立数据反馈闭环，根据模型性能变化反向优化数据筛选与质量评估规则，形成 “筛选-评估-监控-优化”的持续改进机制，保障增量训练安全、高效进行。

5.5 数据退役阶段安全防护对策

5.5.1 介质销毁保障溯源信息安全

针对训练数据溯源残留风险，可构建覆盖数据退役全流程的安全处置体系。在数据识别阶段，建立溯源信息分级分类标准，对采集路径、标注规则等敏感溯源数据进行标记，明确退役数据中需重点保护的技术细节与处理逻辑。

在安全处置层面，采用数据脱敏与逻辑擦除技术，对退役数据中的溯源信息进行深度处理。通过模糊化标注规则、混淆特征工程参数、加密清洗流程记录等方式，阻断从退役数据逆向解析处理逻辑的路径。对存储介质实施物理销毁或安全格式化，确保溯源信息无法通过残留数据恢复。

在管理控制环节，制定退役数据处置审批流程，由技术、合规、安全团队联合审核退役方案，确保溯源信息处置符合法规要求与企业安全标准。建立退役数据台账，详细记录数据来源、处理逻辑、处置方式及责任人，实现全流程可追溯。定期开展退役数据安全评估，通过模拟攻击测试溯源信息残留风险，持续优化处置策略，防止因溯源信息泄露导致技术资产流失与合规风险。

5.5.2 数据清除阻断联邦残留风险

针对联邦学习数据残留风险，可构建覆盖节点退役全周期的安全处置机制。在退役评估阶段，制定节点数据残留风险分级标准，依据数据敏感程度、分片完整性等维度评估处置优先级，对医疗、金融等敏感领域的节点数据实施重点监控。

在数据清除层面，采用多重擦除技术对节点存储介质进行处理，通过多次覆写、加密粉碎等方式确保本地数据分片无法恢复；对关键业务节点的存储介质实施物理销毁，从硬件层面阻断数据残留风险。建立数据分片清除验证流程，通过哈希校验、磁盘深度扫描等技术手段，验证数据擦除的彻底性，确保无残留数据片段。

在管理控制环节，建立联邦学习节点退役审批制度，要求参与方提交数据清除报告与介质处置证明，由安全合规团队审核确认。构建节点退役台账，记录数据分片类型、存储位置、处置方式及责任人，实现全流程可追溯。定期开展节点退役安全审计，通过模拟数据恢复攻击测试残留风险，持续优化处置策略，防止因数据残留导致隐私泄露与商业机密流失。

5.5.3 深度解耦销毁数据关联风险

针对模型迭代数据关联风险，需构建覆盖数据生命周期的解耦防护体系。在数据预处理阶段，采用特征混淆技术对训练数据进行扰动处理，通过添加随机噪声、模糊化特征边界等方式，削弱数据特征与模型参数之间的直接关联。同时，建立特征重要性动态评估机制，定期识别并替换高风险关联特征，降低退役数据被逆向推断的可能性。

在模型架构设计层面，引入解耦训练机制，将模型迭代过程拆分为独立的知识提取与参数优化阶段。通过知识蒸馏技术，将历史模型中的有效信息提炼为通用特征表示，切断现役模型与具体退役数据的直接联系。在模型更新过程中，采

用增量学习与参数隔离策略，确保新模型的训练仅依赖必要的历史知识，而非具体的历史数据实例。

在数据退役环节，实施深度解耦销毁策略。对退役数据进行多轮次特征重构，通过生成对抗网络等技术生成替代数据，保留原有数据的统计特性但不包含真实敏感信息。同时，建立模型迭代审计机制，记录每代模型的关键参数变化和数据依赖关系，在数据退役后定期评估现役模型与已销毁数据的关联风险，及时发现并阻断潜在的逆向推断路径。通过全流程的解耦防护措施，确保模型迭代过程中数据关联风险可控，防止敏感信息泄露。

6 大模型训练数据安全的管理与运营

6.1 数据安全组织与人员管理

6.1.1 数据安全组织架构设计

大模型训练数据安全需构建权责清晰的三级组织架构，强化全流程协同治理。顶层设数据安全管理委员会，由高管及业务、法务负责人组成，负责制定安全战略，审批数据跨境、高风险业务等重大决策，协调跨部门资源，平衡安全与业务发展。中层设数据安全管理部门，统筹安全工作：依据《数据安全法》制定制度与流程，开展风险评估与合规审计，建立应急机制，协同研发将安全要求嵌入大模型训练架构，实现源头防控。基层组建执行团队，由安全工程师、运维人员等构成，负责部署加密、访问控制等系统，监控数据流动，阻断异常访问，更新安全设备；遇事件快速响应预案，并与管理部门联动落实整改。企业通过定期安全例会、跨部门小组及内部系统共享信息，消除协作壁垒，提升大模型训练数据安全的整体防控效能。

6.1.2 数据安全人员能力要求与培训

大模型训练数据安全要求构建差异化人员能力体系与分层培训机制。数据安全管理人员需具备战略协同能力，精通《数据安全法》等法规，将合规要求转化为企业制度，针对大模型训练场景制定数据跨境传输、敏感数据防护策略，平衡安全与业务。

技术人员需聚焦专业技能，掌握同态加密、差分隐私等算法，熟悉大模型分布式存储计算特性，基于云计算架构设计安全方案，运用联邦学习实现数据“可

用不可见”。操作人员需严格执行规范，数据管理员掌握分级分类规则确保存储访问合规，运维人员识别系统异常(如未授权操作)防范泄露。企业培训体系需分层实施：法规培训解读最新政策，技术培训按岗位定制(如加密技术、设备实训)，安全意识培训覆盖全员。引入 CISSP 等认证并关联考核，通过笔试、实操评估效果，确保人员能力匹配数据安全需求。

6.2 数据安全风险评估与管理

6.2.1 风险评估方法与流程

大模型训练数据安全风险评估需结合多元方法与全流程管控，实现风险的精准识别与分级处置。

评估方法可采用定性、定量与半定量结合模式：定性评估依赖专家经验，对风险等级进行主观分级(如高、中、低)，适用于快速识别新兴风险；定量评估通过数据建模量化风险，结合历史泄露频率与经济损失计算风险值，用于关键系统深度分析；半定量评估通过评分矩阵(如 1-5 分制)转化主观判断，生成风险优先级图谱，平衡效率与准确性。

完整评估流程涵盖三大核心环节：风险识别需梳理训练全流程风险点，包括数据采集的非法爬取、标注的隐私泄露、存储的未授权访问、传输的中间人攻击、训练中的算法后门与数据投毒等；风险分析结合渗透测试、业务影响分析(BIA)等手段，量化风险发生概率与对企业声誉、合规性的潜在冲击；风险评价依据企业风险偏好与法规要求，对风险排序，区分高风险事项(如核心数据漏洞)与低风险问题(如非敏感数据异常访问)，为处置策略提供依据。

6.2.2 风险应对策略与措施

基于大模型训练数据安全的风险评估结果，企业需采用差异化策略构建动态防护体系。风险规避通过终止高风险活动实现，如停用不合规数据源、拒绝在安全薄弱的第三方平台训练模型，从源头消除风险。风险降低依赖技术与管理手段，数据存储采用 AES-256 加密防止泄露,访问控制实施多因素认证与最小权限原则，定期更新系统补丁修复漏洞，并建立数据水印追踪流向以降低风险发生概率与影响。风险转移借助外部资源，如购买数据安全保险转嫁法律赔偿与恢复成本，将

非核心安全运维外包弥补技术短板。风险接受针对低影响风险，通过日志审计监控非敏感数据误操作并定期复盘，避免过度投入。

针对典型风险的具体措施包括：数据全生命周期加密结合零信任架构动态验证访问，防范泄露；训练数据去重、脱敏处理，利用模型逆向工程检测与发布前安全测试保障模型安全；建立数据合规审查机制，在采集、使用、共享环节验证合法性，并强化《数据安全法》等法规培训确保全流程合规。通过系统化策略与精准措施，提升大模型训练数据安全防护的有效性与适应性。

6.3 数据安全审计与合规管理

6.3.1 数据安全审计机制建设

数据安全审计是保障大模型训练数据安全的关键环节，需构建制度、技术、流程协同的审计体系。企业应建立覆盖数据全生命周期的审计制度，明确内部审计部门或第三方机构为审计主体，涵盖数据访问、处理、存储、传输等环节的安全性，实行定期审计与专项审计相结合的审计周期。在技术手段上，综合运用日志分析，采集并解析系统日志，结合机器学习算法识别异常访问行为并预警；漏洞扫描利用专业工具定期检测系统、网络及应用中的安全漏洞并生成修复建议；合规性检查依据法规与企业标准，核查数据处理流程与安全技术配置，如验证加密算法合规性、数据跨境传输安全性。审计过程需留存完整记录，形成报告提交管理层，为持续优化数据安全管理措施与技术防护手段提供依据，确保其有效性。

6.3.2 合规管理体系构建

在全球数据安全法规趋严的背景下，企业需构建适配大模型训练的合规管理体系，防范法律风险与数据安全隐患。

企业需首先梳理国内外法规框架，包括中国《数据安全法》《个人信息保护法》、欧盟 GDPR、美国 CCPA 等，明确数据采集、存储、使用、共享全流程的合规要求与责任边界，并据此制定覆盖数据分类分级、跨境传输规范、用户权利响应等内容的管理制度，将法规要求转化为内部执行标准。

在大模型训练全流程嵌入合规审查节点，引入第三方数据前由法务与安全部门联合审查来源合法性；训练阶段验证数据处理逻辑合规性，确保算法设计不侵犯隐私；数据跨境传输前，评估方案并申请安全认证。定期通过内部自查与外部

审计开展合规性评估，排查训练流程及存储、共享环节的漏洞，针对问题制定整改计划并跟踪闭环，如优化跨境传输方案或升级加密技术。

建立分层合规培训机制，针对管理层、技术团队、操作岗位定制培训内容，强化《数据安全法》等法规理解；设立合规举报渠道，鼓励员工反馈风险。通过系统化合规管理，企业可有效满足监管要求，为大模型训练业务筑牢合法合规的安全底座。

7 发展趋势与对策建议

7.1 发展趋势

7.1.1 技术发展趋势

大模型训练数据安全的技术发展趋势呈现多维度演进特征。在隐私保护技术领域，联邦学习正从简单的横向联邦向更复杂的纵向联邦与联邦迁移学习拓展，实现跨机构、跨领域的数据协同训练而不共享原始数据，例如医疗领域中不同医院基于纵向联邦学习联合训练疾病预测模型,既保护患者隐私又提升模型准确性。同态加密技术则在计算效率上取得突破，全同态加密从理论走向实际应用，可支持大模型在密文状态下完成复杂神经网络计算，解决数据“可用不可见”的核心难题，未来有望在金融风控等对数据隐私要求极高的场景广泛应用。

数据溯源与水印技术成为保障数据主权的关键方向。区块链技术与数据指纹结合，可构建不可篡改的训练数据溯源链，精确记录数据采集、加工、使用的全流程操作，实现数据来源可验证、流转可追踪。例如，新闻媒体机构将采集的公开数据上链存储，当发现训练数据被非法使用时，可通过区块链记录快速定位侵权源头。数据水印技术则向语义级水印升级，不仅能在图像、文本中嵌入不可感知的标识，还能将版权信息编码到数据特征空间，即使数据经过脱敏、变换等处理，仍可通过提取水印确认归属，有效防范数据滥用与盗版问题。

对抗性攻击防御技术从被动响应转向主动免疫。对抗训练机制持续优化，通过生成多样化的对抗样本并注入训练过程，迫使模型学习鲁棒特征，提升对数据投毒、提示词注入等攻击的抵抗力。例如，自动驾驶模型在训练中加入精心设计的对抗性图像样本，可增强对恶劣天气、传感器噪声等干扰的适应性。同时，动态安全边界技术被引入，模型在推理阶段实时监测输入数据的分布异常，通过自

适应调整决策阈值或触发二次验证,阻断潜在攻击请求,形成“检测-防御-反馈”的闭环防护体系。

边缘计算与去中心化存储重塑数据处理架构。边缘节点的数据本地化处理能力显著增强，在工业物联网场景中，传感器采集的数据可在边缘端完成预处理与加密，仅向中心模型传输聚合后的特征数据，减少敏感数据的暴露风险。去中心化存储网络(如 IPFS)通过分片存储与共识机制，将训练数据分散存储于多个节点，避免集中式存储的单点故障与攻击面过大问题，同时利用智能合约自动执行数据访问权限管理，提升存储环节的安全性与可靠性。

跨模态数据安全融合技术成为研究热点。针对文本、图像、音频等多模态数据的联合训练需求，安全多方计算(MPC)被扩展至跨模态场景，实现不同模态数据在加密状态下的特征融合与模型训练。例如，在智能安防领域，结合视频图像与语音数据进行跨模态异常检测时，MPC 技术可确保两类数据在不泄露原始信息的前提下协同计算，提升模型对复杂场景的识别能力。此外，基于注意力机制的隐私保护技术，可动态识别多模态数据中的敏感区域并实施差异化加密，在保障安全的同时减少计算开销。

7.1.2 行业发展趋势

大模型训练数据安全的行业发展正朝着协同化、规范化、服务化与权益化方向迈进，重塑产业生态格局。在多方协同层面，跨行业数据安全联盟将成为新趋势。医疗、金融、科技企业会基于隐私计算与联邦学习框架组建联盟，建立统一的数据交互安全标准，打破数据孤岛的同时确保敏感数据可控共享。例如，医疗行业可联合保险企业，在保障患者隐私前提下，利用脱敏数据共同训练健康风险评估模型，实现数据价值与安全的双赢。区块链技术应用也将从单一企业扩展至全产业链，上下游企业通过智能合约构建数据存证联盟链，自动验证数据使用权限，保障训练数据从采集到应用全流程可追溯，提升数据供应链整体安全性。

合规治理领域正加速形成“法规约束+行业自律”的双重规范体系。跨国企业为应对全球法规差异，会设立区域性数据安全合规中心，根据不同地区法规特点，定制数据加密、权限管理和审计方案，实现动态合规。行业自律组织将发挥更大作用，头部企业联合制定生成式 AI 训练数据行业白皮书，明确版权归属、数据标

注伦理等细则，推动形成统一的行业实践指南，引导企业在数据安全合规层面进行良性竞争。

市场服务模式向专业化、定制化方向发展。数据安全服务市场将细分出数据合规咨询、安全架构设计、攻击场景模拟等垂直领域服务商，针对大模型训练全流程提供定制化解决方案。云服务巨头会持续完善“安全即服务”生态，推出模块化数据安全套件，企业可按需选择数据脱敏、访问控制等功能组件，快速搭建适配自身需求的安全防护体系，降低中小企业在数据安全领域的准入门槛。

数据权益市场逐步成型，数据主权概念从理论走向实践。随着公众数据安全意识觉醒，用户对数据的掌控需求日益强烈，企业与用户间的数据权益博弈将催生新型数据交易模式。用户通过去中心化身份技术，对自身数据的使用进行分级授权，并参与数据价值分配。例如，在个性化推荐场景中，用户可选择向企业有偿开放部分行为数据用于模型训练，企业则需建立透明的数据价值评估与回馈机制，这种模式将重塑数据生产关系，推动行业向更公平、可持续方向发展。

7.2 对策建议

7.2.1 构建全生命周期技术防护体系，强化数据安全风险防控

针对大模型训练数据准备、构建、应用及退役阶段的核心风险，整合加密算法、隐私计算与动态监控技术。数据准备阶段，通过多模态语义联合校验、开源数据版权隐私双核查机制，从源头阻断偏见数据与违规数据流入。模型构建阶段，运用联邦学习差分隐私技术保护梯度安全，结合对抗训练增强模型鲁棒性，防止训练过程数据泄露与样本污染。系统应用阶段，部署提示词合规检测与反演攻击防御模块，对增量数据实施多维度校验，避免实时交互中的数据污染。数据退役阶段，采用介质物理销毁与数据关联深度解耦技术，消除溯源信息残留与联邦学习节点数据泄露风险。通过技术手段全流程嵌入，形成环环相扣的风险防控链条。

7.2.2 完善数据安全管理运营机制，落实组织合规协同治理

建立三级数据安全组织架构，顶层管理委员会统筹安全战略与重大决策，中层管理部门制定制度流程并开展合规审计，基层执行团队落实技术防护与应急响应。同步构建人员能力体系，针对管理人员强化法规转化与战略协同能力，技术人员聚焦同态加密、联邦学习等专业技能，操作人员严格执行数据分级分类与访

问控制规范。风险评估需结合定性、定量方法，覆盖数据采集至退役全流程风险点，通过渗透测试与业务影响分析量化风险等级。合规管理对标国内外法规要求，在数据跨境传输、敏感数据处理等环节嵌入审查节点，定期开展内部审计与外部合规评估，确保训练流程符合《数据安全法》《个人信息保护法》等法规要求。

7.2.3 前瞻布局新兴技术与产业生态，推动安全能力迭代升级

聚焦联邦学习、同态加密等隐私保护技术的演进，推动纵向联邦与全同态加密在医疗、金融等敏感领域的应用，实现数据“可用不可见”。发展区块链与语义级数据水印技术，构建不可篡改的溯源链条，保障数据主权与版权归属。布局对抗性攻击防御技术，通过动态安全边界与自适应对抗训练，提升模型对投毒、反演等攻击的主动免疫能力。推动跨行业数据安全联盟建设，基于隐私计算框架建立统一交互标准，联合制定生成式 AI 训练数据行业规范。培育专业化数据安全服务市场，鼓励云服务商提供模块化安全套件，降低中小企业防护门槛。探索数据权益分配机制，通过去中心化身份技术实现用户数据分级授权，推动数据价值公平分配，构建可持续发展的产业生态。