数据集安全：内涵、挑战与防护策略

彭文华
（国家工业信息安全发展研究中心，北京　100040）

摘　要：结合政府公共服务、医疗监控、军事等多领域实践案例，系统剖析数据集采集、预处理、标注、训练应用及模型输出各阶段面临的安全风险；通过对比数据安全与数据集安全的核心差异，提炼针对性防控逻辑及具体手段。研究发现，数据集安全是数据安全在AI场景下的深化延伸，当前面临数据偏见、投毒攻击、对抗干扰等特有风险，且不同领域风险呈差异化特征。构建“技术—管理—合规”数据集全链路风险闭环管控体系，可有效提升AI领域数据集安全防护能力，为AI应用安全发展提供理论与实践指导。

关键词：AI；AI安全；数据集安全；数据集风险
中图分类号：TP13；TP309　　**文献标志码：**A　　DOI：10.19358/j.issn.2097-1788.2026.06.008

中文引用格式：彭文华. 数据集安全：内涵、挑战与防护策略[J]. 网络安全与数据治理，2026，45(6)：57-65.
英文引用格式：Peng Wenhua. Dataset security：connotation，challenges，and protection strategies[J]. Cyber Security and Data Governance，2026，45(6)：57-65.

Dataset security：connotation，challenges，and protection strategies

Peng Wenhua
（China Industrial Control Systems Cyber Emergency Response Team，Beijing 100040，China）

Abstract：Drawing on multi-domain case studies from government public services，medical surveillance，and military applications，it systematically analyzes security risks encountered during dataset collection，preprocessing，labeling，training/application，and model-output stages. By contrasting datasets security with traditional data security，this paper distills targeted prevention logics and concrete countermeasures. The research reveals that dataset security is an AI-specific extension of data security，currently facing unique risks such as data bias，poisoning attacks，and adversarial interference，with risks profiles varying significantly across sectors. Establishing a full-chain "technology-management-compliance" closed-loop risk-control system for datasets can effectively enhance AI dataset security and provide both theoretical and practical guidance for the secure development of AI applications.

Key words：artificial intelligence；AI security；dataset security；dataset risk

0　引言

AI技术的飞速发展，使得数据集成为驱动其进步的核心要素。数据集的质量与安全直接决定了AI模型的性能与可靠性。然而，随着AI应用的日益广泛，数据集安全问题随之凸显。数据泄露会侵犯个人隐私与公共利益，数据投毒、对抗攻击等恶意行为则可能导致模型决策失效，甚至引发国家安全风险。数据集安全防护的紧迫性与重要性日益凸显。

当前，各领域已逐步意识到数据集安全的核心价值，相关研究与实践探索陆续展开，但仍存在诸多不足。王鹏等提出：AI+数据集的开发和人工智能发展都需要坚实的安全合规专业能力的支持。目前，许多企业在数据资源流通和AI应用开发方面存在较为显著的安全合规短板[1]。

政府开放数据作为政务领域AI数据集的重要来源之一，其安全问题备受关注。周静虹等从零信任视角出发，提出了政府开放数据的安全风险治理模式[2]。但如何适配AI模型训练的特殊需求，实现安全与利用的平衡，仍需进一步探索。

在健康医疗领域，熊劲光等指出，健康数据的分类分级是数据安全治理的基础[3]，但其成果尚未完全转化为适配AI数据集全流程的防护方案。

军事领域同样面临严峻的数据集安全挑战。陆正之等研究表明，军事智能模型依赖海量数据驱动，但

深度学习的不可解释性使得对抗攻击技术能够对军事数据安全构成严重威胁[4]。

现有研究多聚焦单一领域或单一类型风险，缺乏对 AI 领域数据集安全内涵的系统界定，以及跨领域风险特征的整合分析，难以形成普适性与针对性兼具的防护体系。基于此，本文立足 AI 领域数据集全生命周期，系统梳理数据集安全的内涵与风险，结合各领域实践经验，提出全链路防护策略，以期填补现有研究空白，为 AI 数据集安全保障提供理论支持与实践参考。

1 数据集安全的内涵

当前学界与产业界对数据安全的探讨已较为充分，但针对数据集安全的专项研究仍显薄弱，相关安全认知尚未形成广泛共识。本节将从数据集的定义与特性出发，系统辨析数据安全与数据集安全的核心差异，界定 AI 领域数据集安全的核心概念，进而明确其关键防护维度，为后续安全风险分析与策略构建奠定理论基础。

1.1 数据集的定义与特性

从定义边界来看，数据集本质是数据的子集，高质量数据集是“经过采集、加工等数据处理，可直接用于开发和训练人工智能模型，能有效提升模型性能的数据的集合”[5]。这些数据集具有明确的应用导向性，其质量对 AI 模型的训练效果起着关键作用。

1.2 数据安全与数据集安全

数据安全关注的重点是数据本身的安全，旨在“守住数据本身安全底线”。其核心目标是保护数据的保密性、完整性、可用性和合规性，防止数据被泄露、篡改或无法正常使用。数据集作为数据的子集，其安全管控内容应继承数据安全的维度，还需针对 AI 应用场景的特点进行细化，甚至涵盖 AI 模型基于数据集输出的结果安全可控的维度，兼顾数据所有人与 AI 应用使用者的双重权益。

AI 应用依赖于数据集的全链路处理，从数据采集到模型输出，每个环节都存在独特的安全风险。在采集阶段，不仅要考虑数据的合法性与合规性，还需同步考量数据适配 AI 训练的质量合规性，必须符合相关隐私法规，对公共场合下采集到的车牌号、人脸信息等进行模糊化处理。在预处理、标注阶段，需防范数据污染、标注偏差等问题。数据污染可能导致模型学习到错误的模式，标注偏差则会使模型的输出偏离真实情况。在训练与应用阶段，需管控数据集携带的偏见风险，避免模型输出歧视性、有害性内容。在模型输出环节，需追溯结果与数据集的关联关系，确保负面结果可溯源、可追责。

这种细化本质上是围绕数据所有者→数据集→模型→AI 用户的特有链路进行延伸，与传统数据安全关注重点有所不同。

1.3 数据集安全的核心定义与关键维度

以金融行业为例，光大银行将数据安全与隐私保护融入数据集建设的全流程[6]。由此可见，AI 领域数据集安全是指在数据集的采集、预处理、标注、增强、训练和使用等全生命周期过程中，采取必要的技术与管理措施，确保数据集不被未授权访问、泄露、篡改、破坏，同时保障数据集的完整性、可用性和机密性，从而支撑 AI 模型的安全可靠运行。其关键维度及主要风险如图 1 所示。

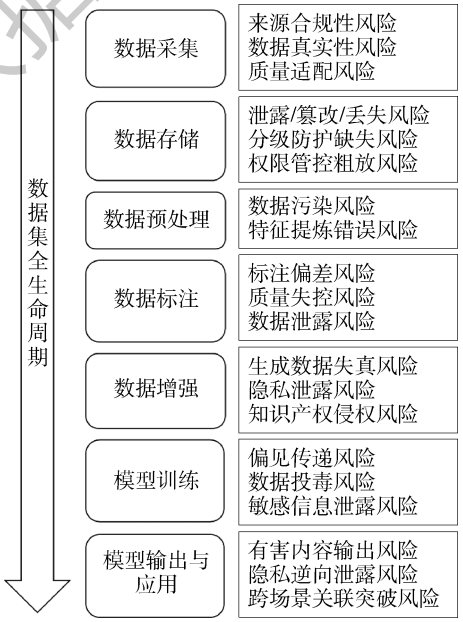

图 1　数据集全生命周期安全风险图谱

数据存储安全：数据存储是数据集安全保障的重要环节，此阶段面临着数据泄露、篡改、丢失等安全威胁。在存储过程中，需结合数据分类分级结果，对不同敏感程度的数据集采取差异化防护措施；建立数据版本控制机制，以追踪数据集的变更历史，为数据的使用提供可靠保障[7]；此外，需同步设置严格的访问权限管控机制，从权限管控层面切断恶意篡改、注入污染数据等数据投毒行为的实施路径。

数据预处理安全：数据预处理是对原始数据进行加工处理的步骤，其安全性直接影响数据后续使用价值与模型训练效果。在预处理过程中，需防范数据污染风险，避免因操作失误或外部干扰导致数据偏离真

实特征。在进行数据离散化、图片数据裁剪等操作时处理不当，会导致提炼错误的特征信息，干扰模型对数据规律的捕捉。

数据标注安全：数据标注是为数据赋予 AI 可识别标签的环节，标注的准确性与一致性对模型训练效果起着决定性作用，其安全重点在于防范标注偏差与质量失控风险。在标注过程中，需避免因人工操作失误导致标注结果失真。徐伟等通过扎根分析发现，数据标注阶段存在人工标注风险、运营模式风险和监管不足风险[8]。

数据增强安全：数据增强是通过生成、合成、模拟等方式扩充数据集规模的重要手段，其安全性需兼顾数据真实性、合规性与质量可控性，避免生成幻觉数据、含隐私数据。在增强过程中，需防范生成数据失真的风险，一旦处理不当，可能导致增强数据与真实数据特征偏差过大，引发模型过拟合或决策偏差。同时，需确保增强过程合规，若未进行充分脱敏处理，增强数据可能间接泄露个人隐私；合成数据时还需避免侵犯原有数据的知识产权，避免引发法律纠纷。

模型训练安全：模型训练是将数据集转化为 AI 模型决策能力的核心环节，需重点防范三类风险：一是偏见传递风险，若训练数据集中隐含性别、种族、地域等偏见，模型会学习并放大此类倾向，最终导致应用阶段出现歧视性决策；二是数据投毒风险，需警惕攻击者向训练集中注入恶意数据，避免模型决策边界偏移，丧失预测与判断的可靠性；三是数据泄露风险，需防范因模型漏洞等原因导致敏感信息通过模型参数、中间计算结果间接外泄。

模型输出与应用是数据集价值落地的最终环节，其衍生的数据集安全风险涉及有害内容输出风险、隐私逆向泄露风险等。攻击者可借助模型输出结果逆向推导原始敏感数据，甚至存在不法主体利用模型对数据集信息进行跨场景关联分析，突破数据使用边界。

1.4 数据安全与数据集安全防控技术对比

1.4.1 数据安全防控技术

数据安全防控措施聚焦加密、脱敏、权限管控等方面，遵循“最小可用”原则合理使用数据。加密通过对数据进行编码处理，使未经授权的人无法读取数据内容，保障数据的保密性。脱敏则是对敏感数据进行替换、掩码等变形处理，在保护数据隐私的同时保留数据的可用性。权限管控通过设置不同用户的访问权限，仅授予满足业务需求的最小数据访问权限，从源头规避数据泄露、篡改的核心风险，适配全行业通用的数据处理场景。

1.4.2 数据集安全防控技术

数据集安全是数据安全在 AI 领域的深化延伸，其防控措施在覆盖加密、脱敏、权限管控等数据安全技术的同时，进一步拓展纳入检测、修正、过滤、审查、溯源、核验等复合方法。

常规数据安全体系较少关注的，恰是数据集安全所特有的防控重点。“检测”可精准识别数据集中的异常数据、投毒污染与隐含偏见；“修正”针对检出问题开展定向校准与优化；“过滤”能够拦截有害信息流入模型训练环节；“审查”通过人工或自动化方式对数据集及模型输出进行合规性、伦理性校验；“溯源”实现模型输出结果与源数据集的全链路关联追踪；“核验”则用于验证数据本身的真实性、可靠性与一致性。

上述技术可有效识别数据集污染与偏见、拦截模型有害输出，辅助构建全链路风险闭环管控体系，既保障数据集本身的安全合规，又确保模型输出的安全可控。

2 数据集全链路安全风险与挑战

AI 领域数据集的全链路安全贯穿采集、预处理、标注、训练至模型应用全流程，面临数据投毒、偏见传递、隐私泄露等多重风险，其防控流程、核心重点以及防控技术与常规数据安全差异显著。吴世忠认为，数据供应链任一环节的安全缺陷或恶意操作均可能成为攻击入口，危及数据完整性、模型可靠性乃至整个系统安全[9]。王鹏等提出，如何在保障用户隐私的同时高效利用高质量数据集，已成为数据安全与隐私保护领域亟须解决的重要问题[10]。

2.1 技术层面的挑战

2.1.1 对抗攻击威胁

对抗攻击是 AI 领域数据集安全面临的主要技术挑战之一。陆正之等指出，对抗攻击技术可在模型训练和推理阶段对数据集安全产生威胁。在训练阶段，数据投毒攻击通过向训练集投放污染数据，导致模型决策边界偏斜，降低模型可用性；在推理阶段，逃避对抗攻击通过给输入数据施加微小扰动，使模型输出错误结果[4]。

2.1.2 数据安全技术局限性

传统数据安全防控技术虽然能用在数据集安全防控领域，但当前相关技术仍存在一定局限性。以数据脱敏为例，沈琳岚在金融数据脱敏研究中发现，传统

的静态脱敏方法容易暴露待脱敏数据；动态脱敏技术鲁棒性要求高，脱敏效果不及静态脱敏[11]。此外，在AI领域，大规模数据集的脱敏处理不仅需要保证数据隐私安全，还需维持数据的可用性，以支撑模型训练。

2.1.3 模型反向推理风险

张振超等提出，攻击者可通过“模型反向推理”或“成员推理攻击”等技术，从训练后的模型中推测出部分原始数据[12]。在AI领域，模型反向推理风险使得数据集的敏感信息面临泄露威胁。黄锫研究发现，目前已经有相关技术可以抽取出大模型中的数以GB为计算单位的庞大原始训练数据[13]。

2.2 管理层面的挑战

2.2.1 数据集分类分级不清晰

科学合理的数据分类分级是数据集安全管理的基础，但当前AI领域数据分类分级存在不规范问题。熊劲光等指出，健康医疗数据因缺乏统一的分类分级标准，导致数据整合难度大，影响分级准确性[3]。虽然邮储银行等部分机构已开始基于大模型开展数据安全分类分级实践[14]，但整体而言，AI领域统一的数据集分类分级标准仍有待建立。此外，数据集分类分级实操中存在适配矛盾：原始数据在预处理环节经匿名化处理、去除敏感内容后，其安全级别应相应下调，此场景与传统数据安全“从高原则”相悖，无法直接套用。

2.2.2 数据集安全责任划分不清晰

在AI数据集的全生命周期过程中，涉及数据提供者、采集者、处理者、模型使用者等多个主体，但当前数据安全责任划分不清晰，导致安全管理存在漏洞。孙清白指出，大模型产业链中，各主体的数据风险管控义务和法律责任不明确，一旦发生数据安全事件，容易出现责任推诿现象[15]。以医疗场景为例，发达地区数据集训练的模型用于欠发达地区，极易因样本与当地患者临床特征差异大，导致模型诊断准确度降低，引发医疗责任风险。此时又如何确定各方责任？为此，王锋等认为，提炼高质量人工智能数据工程体系与能力建设路径，有助于清晰划分各环节责任，保障数据集安全[16]。

2.2.3 数据集安全监管滞后

AI技术的快速发展使得数据集安全面临的威胁不断变化，但数据集安全监管却存在滞后性。徐伟等研究发现，生成式AI数据安全风险具有动态性、瞬时性特点，现有监管手段难以及时发现和应对新型数据安全风险[7]。针对生成式AI模型的对抗样本攻击技术不断更新，而监管部门尚未建立完善的检测与防御标准，导致监管措施无法有效应对新型攻击手段，难以保障数据集安全。许身健指出，在提高数据要素市场化配置能力与效率的征途上，共享有活力、交易有秩序、监管有保障的数据治理机制建设任重道远[17]。

2.3 法律与伦理层面的挑战

2.3.1 法律法规不完善

当前与AI数据集安全相关的法律法规仍存在不完善之处，难以全面覆盖数据集安全的各个环节。周静虹等指出，政府开放数据领域存在立法层级低、规制不足等问题，这一问题在AI领域同样突出[2]。对于AI模型训练数据的合法性审查、数据泄露后的赔偿标准等问题，现有法律法规尚未作出明确规定，导致企业在数据集安全管理中缺乏明确的法律依据，难以有效防范法律风险。

2.3.2 数据隐私保护与数据利用平衡难题

在AI领域，数据集的广泛利用与数据隐私保护之间存在矛盾。刘嘉琪指出，政府数据开放共享需在安全与利用之间寻求平衡，这一平衡难题在AI数据集应用中更为突出[18]。一方面，为提升模型性能，需要大量高质量的数据集进行训练，要求数据尽可能开放共享；另一方面，数据开放共享可能导致用户隐私泄露，侵犯用户的合法权益。在征信领域，合规与安全是数据集建设的生命线，需平衡好隐私保护与行业发展的关系[19]。

2.3.3 算法偏见导致的公平性问题

数据集的质量直接影响AI模型的公平性，若数据集中存在偏见，将导致模型输出不公平结果。孙清白指出，训练数据代表性不足会导致模型存在偏见[15]。这种算法偏见不仅会影响模型的可靠性，还可能引发歧视性问题，侵犯特定群体的合法权益，违背数据伦理原则，对数据集安全的伦理层面构成挑战。

3 构建数据集安全管控模型

本节立足前文对AI领域高质量数据集面临的全链路安全风险与核心挑战的分析，借鉴姜春宇等提出的“盘、建、研、管、用”五步走高质量数据集建设路径[20]，梳理出以流程、防控、能力为核心的三维立体数据集安全管控模型，为AI领域数据集安全管理提供可落地的框架支撑。

3.1 三维立体管控体系的整体架构

数据集安全管控的三维立体管控模型，以过程维为基础脉络，以防控维为核心手段，以能力维为支撑保障，三者相互渗透、相互作用，共同构成闭环式安全防护网络，如图2所示。

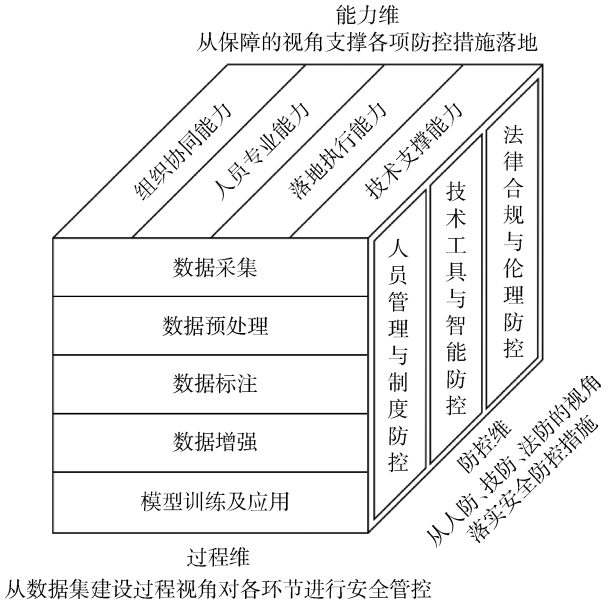

图 2　数据集安全防控模型

3.2　过程维：覆盖数据集全生命周期关键环节

过程维聚焦数据集从源头到应用的全生命周期，划分为采集、预处理、标注、增强、应用五大关键环节，每个环节均对应差异化的安全管控重点，形成环节、风险、措施的精准匹配。

采集环节：重点管控数据来源合规性与质量风险，防范非法数据引入、数据污染等风险，从源头保障数据集安全。

预处理环节：围绕数据格式适配与隐私保护，平衡数据可用性与敏感信息防护。

标注环节：防控标注错误与恶意投毒风险，避免错误标注或恶意标注影响数据集质量。

增强环节：针对数据扩充过程中的安全隐患，确保增强数据与原始数据的特征一致性，防范因算法偏差导致的数据集特征失真。

训练及应用环节：重点防范模型输出风险与数据滥用，确保数据集在模型训练与推理中的安全应用。

3.3　防控维：多元手段协同化解安全风险

防控维作为安全防护的核心手段，按人防、技防、法防的逻辑构建防护体系，针对过程维各环节的风险特征，提供针对性解决方案，形成技术防风险、管理压责任、法律划边界的协同格局。

人防：以制度流程为抓手，通过规范管理约束人员行为，压实责任主体，确保操作不违规、责任可界定。

技防：依托各类技术手段，对数据集全生命周期开展精准防控，为数据集安全做好技术保障。

法防：以法律法规与伦理准则为刚性边界，守住法律红线，规避合规风险。

3.4　能力维：夯实安全管控的基础支撑

能力维作为体系落地的保障，从组织协同、人员专业、落地执行、技术支撑四个维度构建支撑体系，确保有足够的能力支撑过程维与防控维的各项措施有效落地。

组织协同能力：建立垂直管理组织，提高跨部门协同管理能力，避免零散管理导致的管控盲区。

人员专业能力：建设专业化团队，强化数据集安全相关人员的专业能力，提升人员风险识别与应对能力。

落地执行能力：建立健全管理制度体系，足额配发资源，做好监督管理工作，确保各项管理要求与防控措施落地见效。

技术支撑能力：加强相关技术工具研发，为过程维与防控维的措施做好技术保障，提升安全管控的效率与精准度。

3.5　三维维度的协同联动机制

过程维、防控维、能力维三者并非独立运作，而是通过多维度协同联动形成闭环管控，如图 3 所示。

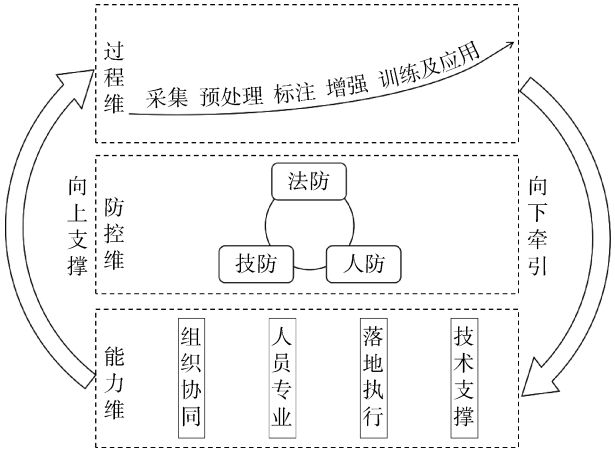

图 3　数据集安全三维模型协同联动机制

过程维中各阶段的具体工作，需防控维中的各项手段予以落地，各种防控手段又依赖能力维提供全方位支撑保障，三者相辅相成，相互依赖，缺一不可。

在实施过程中，宜通过建立定期协同会议、风险联动处置、数据共享互通等机制，实现各项工作的动态适配，确保数据集安全管控模型持续有效运行。

程烁等认为，高价值公共数据集安全治理需要从数据要素制度体系和技术支撑两个维度同步展开[21]，这一思路与数据集安全三维模型不谋而合，各维度协

同发力，提升数据集安全管控效果。

4 数据集全链路安全防控

基于数据集安全管控模型，本节立足数据集建设实践视角，提炼各环节核心安全风险点，制定针对性防控操作策略，以实现风险的精准识别、及时处置与持续迭代优化。数据集全链路安全风险防控重点如图 4 所示。

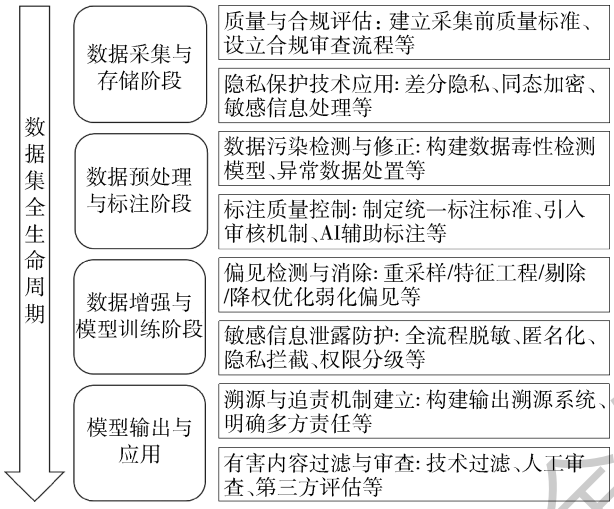

图 4　数据集全链路安全风险防控重点

4.1 采集阶段的风险防控

4.1.1 质量与合规评估

建立严格的数据采集质量评估机制，在采集前明确数据质量标准和规范，并对采集设备、采集方法进行验证。同时，设立合规审查流程，在采集前对采集活动进行合法性和合规性审查，确保采集行为符合法律法规和相关政策要求。如有必要，可引入第三方合规审计机构，对采集过程进行监督和评估。丁浩等提出的“输入质量—操作质量—训练质量”三维评测框架，可作为数据集质量评估的参考[22]。

4.1.2 隐私保护技术应用

采用差分隐私、同态加密等先进隐私保护技术，在数据采集环节对个人敏感信息实施防护。差分隐私通过向数据中添加噪声来模糊个体信息，同时保留数据集的统计特征，使攻击者无法从数据中精准推断出特定的个人信息。同态加密支持在密文状态下直接进行计算，无需对数据进行解密，可在全流程数据处理中持续保障数据隐私性。

4.2 预处理与标注阶段的风险防控

4.2.1 数据污染检测与修正

构建数据毒性检测模型，利用机器学习算法和数据挖掘技术，对预处理过程中的数据进行实时监测，识别并处理可能存在的有毒、污染数据。同时，建立数据集安全风险指标体系，对数据集安全风险进行量化评估，及时掌握数据集安全情况[23]。

4.2.2 标注质量控制

制定统一、明确的标注标准和指南，并对标注人员进行培训，确保标注标准的一致性。引入标注质量审核机制，对标注结果进行随机抽查和审核，及时发现并纠正标注偏差。可以采用多人标注、AI 标注、交叉验证等方式提高标注的准确性。

4.3 模型训练阶段的风险防控

4.3.1 偏见检测与消除

开发偏见检测工具，利用自然语言处理、数据分析等技术，对数据集潜藏的偏见开展系统性评估与量化分析。一旦识别出偏见问题，应采用数据重采样、特征工程优化等方法对数据集进行针对性调整，以消除或弱化偏见因素对模型训练的负面影响。同时，在模型训练流程中嵌入公平性约束条件，确保模型在不同群体上的预测表现具备一致性与公平性。

4.3.2 敏感信息泄露防护

除对数据本身进行脱敏之外，还应对训练日志、中间计算结果中的敏感字段进行加密或匿名化处理，避免原始信息直接或间接暴露。优化模型架构设计，采用数据匿名化相关技术，同时设置隐私问题拦截机制，降低因模型参数逆向推导导致的信息泄露风险。建立训练环境安全管控机制，对模型训练节点进行权限分级管理，全程记录参数调用、数据传输等操作日志，确保泄露风险可追溯、可定位。

4.4 模型输出与应用阶段的风险防控

4.4.1 溯源与追责机制建立

构建模型输出溯源系统，通过区块链等技术记录模型训练过程中使用的数据处理步骤、模型参数等信息，以便准确追溯到数据集具体内容。同时，明确数据所有者、数据处理者、模型开发者等各方在数据安全和模型输出结果方面的责任，建立健全追责机制。当出现问题时，能够依据相关法律法规和合同约定，对责任方进行追究。

4.4.2 有害内容过滤与审查

建立有害内容过滤审查机制，依托文本分类、图像识别等技术手段，对模型输入数据及输出结果实施监测，精准拦截有害内容。同时，设立人工审查专项环节，聚焦模型输出的关键结果开展合规校验，确保其符合伦理准则与安全规范要求。

5 数据集安全的防护策略

基于数据集安全管控模型，本节立足全流程风险防控视角，构建技术、管理、法律与伦理协同发力的防护体系，明确各维度核心防护策略与实施路径，以筑牢数据集全生命周期安全屏障。数据集安全系统防护策略体系框架图如图 5 所示。

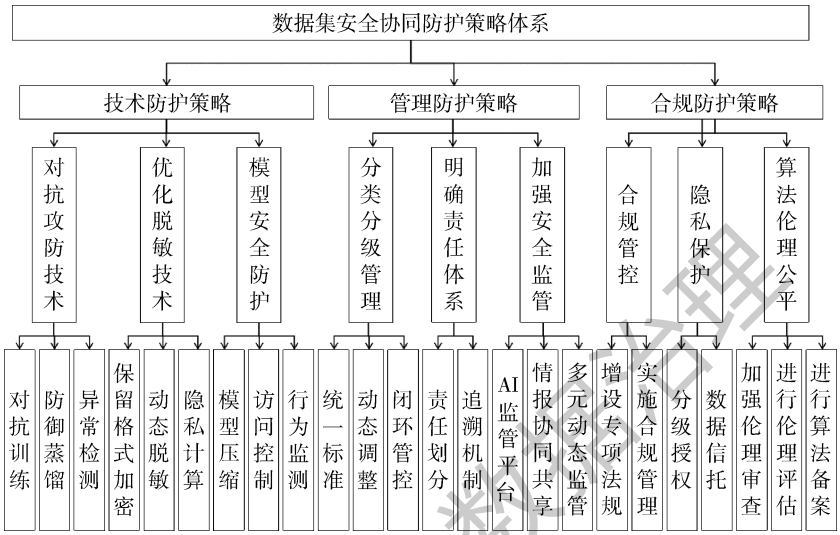

图 5　数据集安全协同防护策略体系框架图

5.1 技术防护策略

5.1.1 对抗攻击防御技术

针对对抗攻击威胁，可采用对抗训练、防御蒸馏等技术提升模型对对抗样本的抵抗能力。对抗训练的核心原理是通过“攻防对抗”重构训练范式，在正常训练样本中融入对抗样本，迫使模型同时学习正常数据特征与对抗扰动特征，避免决策边界因扰动偏移，从而提升抗攻击能力。陆正之等提出，通过在训练过程中引入对抗样本，使模型学习到对抗扰动的特征，从而提高模型的鲁棒性[4]。其关键应用流程需兼顾领域适配性，先结合场景筛选攻击算法，生成符合要求的对抗样本；再按比例（如 1∶4）混合正常样本与对抗样本构建训练集，优化损失函数引入对抗损失项；最终通过攻击测试集验证鲁棒性，确保防御效果。

5.1.2 优化数据脱敏技术

结合沈琳岚提出的动态脱敏技术与保留格式加密算法，优化 AI 领域数据集脱敏方案[11]。对于高敏感数据集，采用差分隐私脱敏算法，通过向数据中添加微小噪声，在保留数据统计特征的同时保护个体隐私，实现数据脱敏存储与传输。此外，引入隐私计算技术，利用数据分片计算原理，将原始数据集拆分为多个互不重叠的子数据集分片，各分片仅保留局部数据信息，各参与机构仅对本地数据分片进行计算并上传加密后的中间结果，由可信第三方或通过分布式节点聚合中间结果得到最终模型训练参数，即可在不泄露原始数据的前提下，实现跨机构数据联合训练，平衡数据隐私保护与数据利用。

5.1.3 模型安全防护技术

为防范模型反向推理风险，可采用模型压缩、知识蒸馏等技术，减少模型参数中的敏感信息。同时，建立模型访问控制机制，对模型的使用进行授权管理，防止未授权用户获取模型并实施反向推理攻击。张振超等提出可通过构建模型行为监测系统，实时跟踪模型的输入输出情况，识别异常的模型访问与推理行为，及时发现并阻断模型反向推理攻击[12]。

5.2 管理防护策略

5.2.1 规范数据分类分级管理

借鉴王蒙等提出的基于数据集与场景化的金融数据分类分级方法，建立 AI 领域统一的数据分类分级标准[14]。从数据特征、业务需求、应用场景等维度出发，制定数据集分类分级目录，明确不同级别数据的安全要求与防护措施。建立数据分类分级动态调整机制，及时更新数据分类分级结果。构建“治理—管理—技术—运营”四层协同管理与评价模型，有助于规范数据分类分级管理，实现闭环管控[16]。

5.2.2 明确数据安全责任体系

现阶段我国对于数据的产权归属、流通交易及收益分配等问题仍处于探索阶段[21]。应构建覆盖数据集

全生命周期的责任体系，明确数据提供者、采集者、处理者、使用者等各主体的安全责任。孙清白指出，需结合大模型产业链，明确上下游各环节相关主体的数据风险管控义务和法律责任[15]。同时，建立数据安全责任追溯机制，利用区块链等技术记录数据流转过程，实现数据安全事件的溯源与追责。

5.2.3 加强数据安全监管能力建设

建立适应 AI 领域数据集安全特点的监管体系，提升监管的及时性与有效性。徐伟等提出，引入敏捷治理理念，构建多元、包容、敏捷互动的监管体系[8]。一方面，加强监管技术研发，利用 AI 技术构建数据安全监管平台，实现对数据集全生命周期的实时监测与风险预警；另一方面，建立监管部门、企业、科研机构等多方协同监管机制，加强信息共享与合作，及时应对新型数据安全风险。

5.3 合规防护策略

5.3.1 完善法律法规体系，加强合规管控

政府层面可加快制定与 AI 数据集安全相关的法律法规，明确数据集安全的法律责任、监管标准、赔偿机制等。周静虹等提出，需制定专门性法律，统筹和细化数据治理规则[2]。同时，加强法律法规与行业标准的衔接，推动形成统一的数据集安全规范体系。企业层面应以《中华人民共和国个人信息保护法》等相关法律法规为红线，制定合规映射条款，在数据集建设全流程逐一贯彻落实。涉及数据跨境场景时，还应增设欧盟《GDPR》《AI 法案》等相关法律法规的合规内容，必要时可引入第三方合规机构加强合规管理。

5.3.2 平衡数据隐私保护与数据利用

各企业应建立数据分级授权使用机制，根据数据敏感程度与使用场景，实现数据的差异化开放共享。刘嘉琪等指出，政府数据开放共享需建立有条件开放和依申请公开两种模式，这一思路可应用于 AI 领域数据集共享[18]。企业可引入数据信托机制，由第三方机构负责数据集的管理与运营，在保障数据隐私的前提下，实现数据集的合理利用。

5.3.3 规范算法伦理与公平性

监管部门可加强对 AI 数据集的伦理审查，确保数据集中不存在偏见，保障模型的公平性。孙清白提出，需确保训练数据的代表性，消除数据偏见[15]。企业应建立算法伦理审查机制，对 AI 模型的训练数据集、算法原理、输出结果等进行伦理评估，履行算法备案、变更、注销备案手续，确保算法模型符合公平、公正、透明的伦理原则。伦理审查清单如表 1 所示。

表 1　AI 应用伦理审查清单（示例）

审查维度	审查内容
有毒有害	是否包含暴力、极端主义等有害内容
	是否存在片面描述、幻觉输出
法律法规	是否包含政治敏感内容
	是否违反法律条款
	是否存在隐私信息泄露风险
	是否侵犯专利、著作权等他人知识产权
	是否存在商业秘密泄露风险
偏差偏见	数据分布是否存在系统性偏差
	是否残留历史性偏见
	是否嵌入主观性偏见
伦理道德	是否违背公序良俗与道德伦理
	是否隐含性别、种族、地域等歧视性内容

6 结论

本文聚焦 AI 领域数据集安全，明确其以数据安全为基础，是 AI 场景下的细化延伸，需覆盖采集、预处理、标注等全生命周期，管控链路连锁风险，兼顾数据所有人与 AI 使用者权益。

当前面临多层面挑战：技术上有对抗攻击、数据脱敏局限及模型反向推理风险；管理上存在数据分类分级不规范、责任划分不清、监管滞后问题；法律与伦理层面则面临法规不完善、隐私保护与数据利用难平衡及算法公平性问题。防控上，数据集安全在数据安全加密、脱敏等基础手段外，还需检测、溯源等复合手段，构建全链路风险闭环管控体系。同时，本文从技术、管理、法律与伦理层面提出防护策略。

未来需应对合成数据、幻觉输出、多模态数据等新问题、新挑战，加强技术研发、法规完善、跨领域合作与人才培养，探索新兴技术应用，构建更安全的数据集生态。在技术研发方面，可进一步强化对大模型输出进行过滤、采用敏感数据遗忘技术等措施，加强数据集技术防护[24]；在跨领域合作方面，可借鉴公共数据、金融等领域的数据集建设经验[6,21]，推动 AI 领域数据集安全发展；在法规完善方面，结合高质量数据集建设需求，制定更具针对性的法律法规[13]，为数据集安全生态构建提供保障。

参考文献

[1] 王鹏,丁一桐. 高质量数据集赋能人工智能产业发展研究[J]. 特区实践与理论,2025(2):36-44.

[2] 周静虹,翁武良,毕崇武. 零信任视角下政府开放数据的安全

风险治理模式研究[J]. 图书情报工作,2025,69(15):3-13.
[3] 熊劲光,田平,黄春柳,等. 基于数据集与场景化的健康医疗数据分类分级创新研究与实践探索[J]. 中国数字医学,2025,20(6):37-42,54.
[4] 陆正之,黄希宸,彭勃. 军事智能数据安全问题:对抗攻击威胁[J]. 网络安全与数据治理,2024,43(11):23-28.
[5] 全国数据标准化技术委员会. 高质量数据集 建设指南[R]. TC609-5-2025-01,2025.
[6] 董波,潘学芳,李晓歌,等. 金融高质量数据集建设的研究和探索[J]. 中国金融电脑,2025(9):65-67.
[7] 林成创,张昱晟,关龙辉,等. 交通行业 AI 数据集建设[J]. 中国交通信息化,2025(6):95-98.
[8] 徐伟,何野. 生成式人工智能数据安全风险的治理体系及优化路径——基于 38 份政策文本的扎根分析[J]. 电子政务,2024(10):42-58.
[9] 吴世忠. 大模型时代高质量数据集建设的进展、挑战与治理路径[J]. 保密工作,2025(10):4-8.
[10] 王鹏,程思儒. 人工智能高质量数据集的发展趋势及热点——基于 CiteSpace 的知识图谱分析[J]. 技术经济与管理研究,2025(4):43-48.
[11] 沈琳岚. 面向金融模型的数据脱敏研究与实现[D]. 西安:西安电子科技大学,2024.
[12] 张振超,吴昊,谢颐. 生成式人工智能的数据安全风险及其应对策略分析[J]. 数字技术与应用,2025,43(8):69-71.
[13] 黄锫. 人工智能大模型训练数据的风险类型与法律规制[J]. 政法论丛,2025(1):23-37.
[14] 王蒙,韩冰洁,丰瑾. 邮储银行基于大模型的数据安全分类分级实践[J]. 邮政研究,2025,41(4):11-15.
[15] 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.
[16] 王锋,谭蕊,沙若男,等. 高质量数据集安全合规管理机制探讨[J]. 质量与认证,2025(12):44-46.
[17] 许身健. 人工智能监管中的数据治理层次论[J]. 探索与争鸣,2024(10):26-30,177.
[18] 刘嘉琪,任妮,刘桂锋. 政府数据开放共享政策的内容体系构建研究[J]. 数字图书馆论坛,2022(8):19-26.
[19] 陈国锋. 征信高质量数据集建设探析[J]. 金融纵横,2025(8):58-64.
[20] 姜春宇,白玉真,刘渊,等. 构建企业级人工智能高质量数据集:方法与路径[J]. 大数据,2025,11(6):47-56.
[21] 程铄,夏义堃. 高价值公共数据集价值释放的机理、实践与路径[J]. 图书情报工作,2025,69(18):30-40.
[22] 丁浩,张畅,顾乐,等. 面向 AI 全生命周期的高质量数据集评测体系研究[J]. 数字化转型,2025,2(8):87-97.
[23] 刘金龙. 生成式人工智能数据安全风险与应对策略研究[J]. 长江信息通信,2025,38(9):182-184.
[24] 徐明. 生成式人工智能大模型的安全挑战与治理路径研究[J]. 信息通信技术与政策,2025,51(1):10-19.

(收稿日期:2025-12-24)

作者简介:

彭文华(1984-),男,硕士,高级工程师,主要研究方向:数字化转型、AI、高质量数据集、大数据等。

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