TC260-PG-2026NA

网络安全标准实践指南

——人工智能训练数据清洗安全指南

（征求意见稿 v1.0-202601）

全国网络安全标准化技术委员会秘书处

2026 年 01 月

本文档可从以下网址获得：
www.tc260.org.cn/

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

前　言

《网络安全标准实践指南》（以下简称《实践指南》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织制定和发布的标准相关技术文件，旨在围绕网络安全法律法规政策、标准、网络安全热点和事件等主题，宣传网络安全相关标准及知识，提供标准化实践指引。

本文件起草单位：上海人工智能创新中心、中国电子技术标准化研究院、北京中关村实验室、国家计算机网络应急技术处理协调中心、杭州网易智企科技有限公司、上海市信息安全测评认证中心、北京火山引擎科技有限公司、上海稀宇科技有限公司、中国移动通信集团有限公司、北京百度网讯科技有限公司、中国联合网络通信集团有限公司、华为终端有限公司、中国网络空间研究院、阿里云计算有限公司、广西电网有限责任公司、北京小米移动软件有限公司、深圳市腾讯计算机系统有限公司、OPPO 广东移动通信有限公司、深圳昂楷科技有限公司等。

本文件起草人：王迎春、孟令宇、刘勇、贺敏、乔兴格、王广宇、喻佳、李薇、郑佳琪、费凡芮、张妍婷、王锟、苗晴晴、何极、王寒生、郭建领、沈俊成、徐阳、徐艺澈、马梦娜、刘栋、李慧芳、刘源、赵高华、徐浩、刘凯杰、方强、武杨、李根、涂利平等。

声　明

本《实践指南》版权属于网安标委秘书处，未经秘书处书面授权，不得以任何方式抄袭、翻译《实践指南》的任何部分。凡转载或引用本《实践指南》的观点、数据，请注明“来源：全国网络安全标准化技术委员会秘书处”。

摘　要

训练数据是人工智能发展的核心要素之一。训练数据清洗是保证模型训练质量的关键步骤。通过保障训练数据的清洗活动安全，确保所有直接用于模型训练的数据的质量，不包含违反社会主义核心价值观的、歧视性的内容，不存在商业违法违规、侵犯他人合法权益的现象，提升模型抵御对抗性风险、满足特定服务类型的安全需求的能力。本文件给出了训练数据清洗活动的安全原则、风险识别维度、清洗方法和实施流程，适用于各类需要对训练数据进行清洗活动的主体。

目　录

1 范围 1
2 术语定义 1
3 训练数据清洗安全原则 2
4 风险识别维度 3
5 清洗方法 7
6 实施流程 10
附录 A 过滤方法应用示例（资料性） 16
附录 B 数据质量指标（资料性） 18
附录 C 风险评估筛查示例（资料性） 19

1 范围

本文件给出了训练数据清洗活动的安全原则、风险识别维度、清洗方法和实施流程。

本文件适用于各类需要对训练数据进行清洗活动的主体，包括人工智能模型开发方、训练数据供应商等，也可为工程师、相关管理人员及主管部门提供参考。

2 术语定义

2.1 训练数据 training data

所有直接作为模型训练输入的数据。

注：包括预训练数据和优化训练数据。

[来源：GB/T 45654—2025,3.4]

2.2 训练数据清洗活动 training data cleansing activities

在数据输入模型训练前，对原始数据集进行错误检测、修正、转换和规范化的系统化过程。

2.3 训练数据清洗安全 security of training data cleansing process

通过技术和管理措施确保训练数据清洗活动的安全，防止模型因数据问题产生不安全的输出。

2.4 偏见 bias

对待特定对象、人员或群体时，相较于其他实体出现系统性差别的特性。

注：对待指任何一种行动，包括感知、观察、表征、预测或决定。

[来源：GB/T 41867—2022,3.4.10]

2.5 公平性 fairness

尊重既定事实、社会规范和信仰，且不受偏袒或不公正歧视影响的对待、行为或结果。

注1：对公平性的考虑是与环境高度相关的，并且因文化、代际、地理和政治观点而异。

注2：公平不等于没有偏见。偏见并不总是导致不公平，不公平可能是由偏见以外的因素引起的。

[来源：GB/T 41867—2022,3.4.1]

2.6 重要数据 key data

特定领域、特定群体、特定区域或达到一定精度和规模的，一旦被泄露或篡改、损毁，可能直接危害国家安全、经济运行、社会稳定、公共健康和安全的数据。

注：仅影响组织自身或公民个体的数据一般不作为重要数据。

[来源：GB/T 43697—2024,3.2]

2.7 个人信息 personal information

以电子或其他方式记录的与已识别或可识别的自然人有关的各种信息。

[来源：GB/T 45574—2025,3.1]

2.8 敏感个人信息 sensitive personal information

一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息。

注：敏感个人信息包括生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等信息，以及不满十四周岁未成年人的个人信息。

[来源：GB/T 45574—2025,3.2]

3 训练数据清洗安全原则

训练数据清洗应遵循安全可控、分布多样、透明可溯、持续迭代

的原则开展，具体如下：

a）**安全可控原则**：训练数据清洗应在保障数据合法合规与内容安全的前提下进行。

b）**分布多样原则**：训练数据清洗需兼顾数据来源和场景的多样性，以支撑模型获得良好的泛化能力。

c）**透明可溯原则**：记录数据清洗过程中所采用的规则、标注标准与关键决策，确保清洗操作可审计、过程可复现、结果可追溯。

d）**持续迭代原则**：数据清洗策略应随业务场景、模型反馈与安全环境的变化而动态调整。

4 风险识别维度

4.1 数据质量风险

训练数据本身存在质量问题，影响模型的训练效果，并可能放大其他的安全风险，主要风险包括：

a）完整性不足：训练数据中存在大量缺失数据，导致信息不完整。

b）准确性不足：训练数据中包含内容错误、不精确或不真实的数据。

c）时效性不足：训练数据过于陈旧，且未标明时间局限性，已与当前现实情况不符。

d）可用性不足：训练数据存在格式错误、严重缺失、严重损坏等问题，导致其无法被正常读取、解析或有效利用。

e）数据重复性：训练数据中存在大量重复记录，可能导致模型过拟合或资源浪费。

4.2 违反社会主义核心价值观的内容风险

训练数据中包含违反社会主义核心价值观的内容，可能导致人工智能模型生成或传播对个人、群体或社会造成负面影响的内容。主要风险包括：

a）煽动颠覆国家政权、推翻社会主义制度；

b）危害国家安全和利益、损害国家形象；

c）煽动分裂国家、破坏国家统一和社会稳定；

d）宣扬恐怖主义、极端主义；

e）宣扬民族仇恨；

f）宣扬暴力、淫秽色情；

g）传播虚假有害信息；

h）其他法律、行政法规禁止的内容。

4.3 歧视性内容风险

训练数据中因样本代表性不足或历史性原因而存在歧视性内容，导致人工智能模型对特定群体或类别产生不公平、不准确或歧视性判断或行为的系统性倾向。主要风险包括：

a）民族歧视内容；

b）信仰歧视内容；

c）国别歧视内容；

d）地域歧视内容；

e）性别歧视内容；

f）年龄歧视内容；

g）职业歧视内容；

h）健康歧视内容；

i）其他方面歧视内容。

4.4 商业违法违规风险

训练数据中包含受《知识产权法》《反不正当竞争法》等法律保护的客体，而未进行合理的采集、处理、使用，导致模型在训练、使用、商业化过程中存在商业违法违规风险。主要风险包括：

a）侵犯他人知识产权；

b）违反商业道德；

c）泄露他人商业秘密；

d）利用算法、数据、平台等优势，实施垄断和不正当竞争行为；

e）其他商业违法违规风险。

4.5 侵犯他人合法权益风险

训练数据中包含可能侵犯他人合法权益的信息（特别是敏感个人信息），导致人工智能模型在训练或推理过程中侵犯个人信息主体或其他主体的合法权利。主要风险包括：

a）危害他人身心健康；

b）侵害他人肖像权；

c）侵害他人名誉权；

d）侵害他人荣誉权；

e）侵害他人隐私权；

f）侵害他人个人信息权益；

g）侵犯他人其他合法权益。

4.6 无法满足特定服务类型的安全需求风险

人工智能模型被应用于安全需求较高的特定服务类型，例如关键信息基础设施、自动控制、医疗信息服务、心理咨询、金融信息服务等领域，具有相较于一般模型更高的安全要求。主要风险包括：

a）内容不准确，严重不符合科学常识或主流认知；

b）内容不可靠，虽然不包含严重错误的内容，但无法对使用者形成帮助。

4.7 对抗性风险

训练数据中存在的恶意样本或攻击行为，可能干扰或破坏模型的训练过程，降低模型的性能和可靠性。主要风险包括：

a）对抗样本攻击：训练数据中存在对抗性样本，影响人工智能模型的训练过程，使其在部署后产生预期外的错误行为；

b）后门攻击：训练数据存在被植入的隐藏漏洞，影响人工智能模型的行为；

c）标签翻转攻击：训练数据中某类样本的标签被篡改，影响人工智能模型的准确率。

5 清洗方法

5.1 数据质量处理

数据质量处理方法包括但不限于:

a）完整性处理：识别并根据策略（删除、填充、插值）处理训练数据中不完整或缺失的数据。

b）准确性处理：识别并根据策略（删除、截断、转换、替换）处理训练数据中显著偏离正常模式的异常数据。

c）时效性处理：移除或修正过时或不再相关的数据，定期更新数据集，确保数据反映最新的信息和趋势。

d）可用性处理：识别并删除训练数据中不可用的数据。

e）重复性处理：识别并删除或合并训练数据中完全相同或近似重复的数据。

5.2 数据来源控制

数据来源控制方法包括但不限于:

a）数据来源筛选：严格筛选数据来源，优先选择权威、经过验证的数据源，如专业数据库、机构发布的数据等，避免使用未经验证的低质量数据源。

b）数据来源审查：检查数据来源的交易合同、开源许可协议、相关授权文件等。特别对其中所涉及的主要知识产权侵权风险进行识别，尤其是对于包含文学、艺术、科学作品的数据，应重点识别其中的著作权侵权问题。

5.3 内容安全审查

内容安全审查方法包括但不限于:

a）规则体系过滤：结合关键词、短语、上下文和语义模式，构建规则体系进行识别和过滤。

b）内容审核模型：利用成熟的内容审核模型，对训练数据的内容进行自动识别和审核。

c）人工审核：对于复杂、模糊或高风险的训练数据内容，引入专业人工团队进行最终判断、删除或修正。

5.4 个人信息保护方法

个人信息保护方法包括但不限于:

a）匿名化：对训练数据中的个人信息进行处理，使得个人信息主体无法被识别或者关联，且处理后的信息不能被复原。匿名化方法可参考TC260-PG-AAAABB《网络安全标准实践指南——个人信息保护个人信息匿名化指南》。

b）去标识化：通过去标识化技术使得数据无法对应到特定个人，但保留通过额外信息重新识别的可能性。个人信息去标识化方法可参考GB/T 37964-2019《信息安全技术 个人信息去标识化指南》附录A、附录B以及TC260-PG-AAAABB《网络安全标准实践指南——个人信息保护 个人信息去标识化指南》。

5.5 偏见缓解方法

偏见缓解方法包括但不限于:

a）重采样：通过增加代表性不足群体样本或减少过度代表群体样本，平衡训练数据分布以缓解偏见。

b）重加权：为训练数据中不同群体或样本分配不同权重，以平衡其在模型训练中的影响力。

c）数据增强：为少数群体生成新的、多样化的合成数据，以增加数据多样性并减少刻板印象。

5.6 安全验证方法

安全验证方法包括但不限于：

a）数据验证：依据预先设定的、明确的规则（如格式、类型、值域），对输入数据进行校验，隔离不符合规则的异常数据。

b）异常检测：运用统计或模型，识别并隔离那些表面合规但在特征分布上表现出显著异常的恶意样本，及时隔离异常数据。

表 1 安全风险与清洗方法映射表

安全风险	缓解方法
数据质量风险	5.1 数据质量处理 5.2 数据来源控制 5.6 安全验证方法
违反社会主义核心价值观的内容风险	5.2 数据来源控制 5.3 内容安全审查
歧视性内容风险	5.3 内容安全审查 5.5 偏见缓解方法
商业违法违规风险	5.2 数据来源控制 5.3 内容安全审查
侵犯他人合法权益风险	5.2 数据来源控制 5.3 内容安全审查 5.4 个人信息保护方法

无法满足特定服务类型的安全需求风险	5.1 数据质量处理 5.2 数据来源控制 5.3 内容安全审查 5.4 个人信息保护方法 5.5 偏见缓解方法 5.6 安全验证方法
对抗性风险	5.1 数据质量处理 5.6 安全验证方法

6 实施流程

6.1 训练数据收集与来源审查

6.1.1 明确收集的范围与目的

根据业务需求说明书、模型设计文档等，明确训练数据的范围、用途、预期规模、更新频率、保存期限，避免过度收集。

6.1.2 数据来源审查

数据收集时，应记录并审查数据的来源，根据数据来源的可靠性和合法性进行初步筛选，排除明显不符合要求的数据源。

数据来源安全评价参考GB/T 45652-2025《网络安全技术 生成式人工智能预训练和优化训练数据安全规范》7.2.1、7.3.1。

6.2 数据质量初步处理与数据标注

6.2.1 数据质量初步处理

根据 5.1 对数据质量进行初步处理:

a）检查数据的完整性，识别并处理训练数据中不完整或缺失的数据。

b）检查数据的准确性，识别并处理训练数据中显著偏离正常模

式的异常数据。

c）检查数据的时效性，移除或修正过时或不再相关的数据。

d）检查数据的可用性，识别并删除不可用的数据，去除数据中的噪声和无关信息。

e）检查数据的重复性，识别并删除或合并完全相同或近似重复的数据。

6.2.2 数据标注

数据标注是数据准备的重要环节，应制定规范的标注流程、标注规范以确保数据标注的质量和效率。明确标注的目标和标准，以确保所有的标注工作都符合任务的需求。进行标注人员的培训，包括标注工具的使用和标注规则的理解，以确保标注的准确性。

在标注过程中，实施标注质量控制，包括定期的标注抽查和错误反馈，以及对标注结果的统一校验。建议采用内部搭建的数据标注系统进行标注，以保障数据的安全和提高标注的效率，并对标注数据进行备份和存储，以防止数据丢失。

数据标注安全评价参考GB/T 45674-2025《网络安全技术 生成式人工智能数据标注安全规范》9 数据标注安全评价方法。

6.3 风险识别与清洗实施

6.3.1 风险识别评估与清洗策略制定

依据第 4.2 节至第 4.7 节，可结合统计分析、域名黑名单过滤，基于关键词匹配的内容筛选方法、基于模型的筛选方法等技术手段进

行训练数据风险识别，并进行综合风险评估。风险评估维度可参考附录C 风险评估筛查示例。根据风险识别和评估的结果，制定清洗策略:

a）确定数据清洗的目标和优先级。

b）确认每类风险的容忍阈值。

c）制定具体的数据清洗规则，包括：技术路线、工具版本、参数配置、责任人、结果验证时间节点等。

6.3.2 数据清洗实施

依据第 5.2 节至第 5.6 节，进行数据清洗活动。

6.4 清洗数据风险二次评估

6.4.1 违反社会主义核心价值观的内容风险二次评估

评估方法如下:

a）采用人工抽检从每种训练数据中分别随机抽取不少于 4000 条数据，人工抽检合格率不低于 96%。

b）采用技术抽检从每种训练数据中分别随机抽取不少于总量 10%的数据，使用关键词、分类模型等技术方式检查有害内容的比例，合格率不低于 98%。

6.4.2 歧视性内容风险二次评估

评估方法如下:

a）对清洗后的数据进行分布分析和偏见指标计算，检查不同群体的代表性是否平衡。

b）使用清洗后的数据训练模型，测试模型对不同群体的判断是否公平。使用偏见检测工具对模型进行测试，检查模型是否存在系统性偏见。

6.4.3 商业违法违规风险二次评估

评估方法如下：

a）对清洗后的数据来源进行再次检查，检查数据来源的交易合同、开源许可协议、相关授权文件等，确保所有数据来源都符合法律法规和授权要求。

b）对数据中的文学、艺术、科学作品等进行详细审查，确保没有侵犯知识产权。随机抽样不少于1000条数据，样本不存在涉及主要知识产权侵权问题，或存在有主要知识产权侵权风险的，已进行合法、有效的处置。

6.4.4 侵犯他人合法权益风险二次评估

评估方法如下：

a）对清洗后的数据进行匿名化检查，确保个人信息无法被识别或关联。检查数据的去标识化处理是否符合标准，确保数据在保留重新识别可能性的同时，无法直接对应到特定个人。

b）从全部训练数据中随机抽取不少于4000条数据，检查所抽取数据中是否包含个人信息或敏感个人信息；存在个人信息或敏感个人信息的，均已取得对应个人的授权同意以及符合法律、行政法规规定的其他情形。

c）从包含个人信息或敏感个人信息的训练数据中随机抽取不少于1000条或10%的数据，均已取得对应个人的授权同意以及符合法律、行政法规规定的其他情形。

6.4.5 无法满足特定服务类型的安全需求风险二次评估

评估方法如下:

a）对数据内容进行科学常识和主流认知的校验，确保数据内容准确无误。

b）对模型的输出进行人工测试，检查模型在特定服务类型中的表现是否符合专业要求。

6.4.6 对抗性风险二次评估

评估方法如下:

a）对清洗后的数据进行格式、类型和值域检查，数据应符合预定义的规范和标准。

b）使用统计或机器学习算法对数据进行异常检测，没有表面合规但特征分布异常的恶意样本。

c）用清洗后数据重新训练模型，使用对抗样本、后门攻击检测工具等对模型进行鲁棒性测试，测试宜在独立GPU集群完成。

6.4.7 版本控制

对清洗数据进行版本控制，每次清洗可在数据目录中标注“清洗批次号”。对于关键操作，宜写入日志管理工具，并规定保留的具体期限。

6.5 持续监控与迭代优化

6.5.1 建立数据清洗管道与自动化

设计标准化的数据清洗管道，集成适合的自动化工具与技术以提升效率与一致性。例如，可采用工作流调度平台构建管道、使用数据质量框架进行规则验证，并利用相关组件的监控告警功能，实现流水线异常的自动发现与通知。

6.5.2 持续监控与反馈机制

设定数据清洗的监控指标，企业内部可定期发布数据清洗误杀和漏杀案例等供员工学习。

6.5.3 定期审查与更新

审查与更新方法如下：

a）定期对现行清洗策略进行全面复盘。复盘中需对关键词库、审核模型、风险评估阈值等核心要素进行重新核定，核定的依据包括但不限于：关键词和规则的有效性、审核模型的性能、阈值设定的合理性以及业务需求与安全环境的最新发展。

b）若遇重大需求变更，则宜启动专项审查，依据专项审查的结论、变更需求的具体内容以及对新风险场景的评估结果调整清洗策略，确保清洗策略与变化后的目标保持一致。

附　录　A

（资料性）

过滤方法应用示例

A.1 关键词表示例

表 A.1 列举了包含违反社会主义核心价值观的内容、包含歧视性内容、侵犯他人合法权益 3 个维度的关键词表示例。

表 A.1 关键词表示例

维度	关键词示例
包含违反社会主义核心价值观的内容	违法物品制作：毒品制作、炸药制作、枪械改装等
	违法行为教程：自杀教程、暴力教程、黑客教程等
	违法内容传播：色情内容、暴力内容、恐怖主义内容等
	违法服务推广：代开发票、代孕服务、非法金融活动等
	极端言论：分裂国家、种族仇恨、宗教极端言论等
	色情内容：色情视频、色情图片、色情文学等
包含歧视性内容	地区+负面词汇
	性别+刻板印象
	年龄+偏见
	职业+偏见
	民族+歧视词汇
	宗教+偏见词汇
	身体特征+负面词汇
	性取向+负面词汇
侵犯他人合法权益	可参考 GB/T 42460-2023《个人信息去标识化效果评估指南》附录 A、附录 B、附录 C 设计标识符

A.2 规则体系示例

表 A.2 列举了包含违反社会主义核心价值观的内容、包含歧视性内容、侵犯他人合法权益 3 个维度的规则体系示例。

表 A.2 规则体系示例

维度	规则类型	示例
包含违反社会主义核心价值观的内容	多模态规则	图像+文本联合风险，例如图片中出现武器且文本中提到“出售”或“价格”
	上下文规则	识别教唆行为，例如“如何制作炸药”或“如何走私毒品”
	关键词组合	检测暴力内容，例如对暴力行为、色情内容的传播描述
	溯源规则	阻断违法交易，例如“➕微信号：1234567890”或“QQ 群：123456789”
包含歧视性内容	正则规则	检测显性歧视组合，例如“某地区的人都是骗子”或“女性不适合做 xxx 职业”
	语义规则	识别职业性别偏见，例如“男性适合做 xxx 职业，女性适合做 xxx 职业”
	对抗规则	防御变体规避，例如将“女司机”替换为“女司机”或“女性的司机”以避免性别偏见
侵犯他人合法权益	结构化规则	身份信息组合，例如姓名+身份证号+手机号出现在同一段落
	声纹规则	例如音频中检测到特定声纹且文本中提到″我是 XXX″
	图像规则	例如人脸照片和身份证摆拍图同时出现
	行为规则	例如自然人的位置信息与活动记录同时出现

附 录 B

（资料性）

数据质量指标

表 B.1 列举了常见的数据质量指标。

表 B.1 数据质量指标

检查维度	具体检查项	检测方法	量化指标
完整性	关键字段缺失率	统计每列空值占比	缺失率 = 空值数 / 总样本数 × 100%
	文件完整性	验证文件能否正常打开/读取（图像、音频、视频）	损坏文件占比 = 损坏文件数 / 总文件数 × 100%
准确性	异常值检测	数值型字段：Z-score/IQR 离群值分析	异常值占比 = 异常样本数 / 总样本数 × 100%
		文本型字段：长度/字符异常检测	
	值域合理性	检查字段值是否在合理范围（如 年龄∈[0,120]）	超限值占比
	枚举值合理	检查分类字段值是否在预定义列表（如 性别∈['男','女','其他']）	非法枚举值占比
重复性	精确重复检测	识别所有字段完全相同的样本	重复率 = 重复样本数 / 总样本数 × 100%
	近似重复检测	文本：MinHash/Jaccard 相似度检测	近似重复样本占比
		图像：pHash 汉明距离	
基础统计	分布偏移检测	对比训练集与业务场景的分布差异（如地域分布）	KL 散度/JSD 距离
	描述性统计	输出数值字段的 min/max/mean/std	标准差 > 均值 或 min<0（非负字段）

附　录　C

（资料性）

风险评估筛查示例

表 C.1 列举了偏见风险、有害内容风险、隐私泄露风险、对抗性风险的评估筛查示例。

表 C.1 风险评估筛查示例

风险类型	筛查目标	量化指标	处置动作
数据质量风险	完整性	见附录 B B.1	删除、填充、插值
	准确性		删除、截断、转换、替换
	重复性		删除、合并
包含违反社会主义核心价值观的内容	危害国家安全	命中率 置信度 相似度	删除样本 + 记录源地址
	恐怖主义、极端主义		
	民族仇恨		
	暴力、淫秽色情		
	虚假有害信息		
	其他有害内容（比如未授权版权内容）		
包含歧视性内容	显性歧视内容	命中率	删除样本 + 标注污染源
	隐性分布偏见	群体分布统计	数据重采样 + 注入平衡样本
		均衡差异分数	
		差异影响（DI）	
		机会均等差（EOD）	
		均等错误率（EER）	
		均等机会（EO）	
		混淆矩阵差异	
侵犯他人合法权益	个人信息	出现密度 识别率	确保授权 + 匿名化、假名化处理
	敏感个人信息		
对抗性风险	恶意代码/链接	检出率	隔离文件
	数据投毒样本	离群度	样本删除 + 记录攻击路径 + 模型鲁棒性测试

参考文献

[1] 中华人民共和国个人信息保护法

[2] GB/T 35273-2020 信息安全技术 个人信息安全规范

[3] GB/T 37964-2019 信息安全技术 个人信息去标识化指南

[4] GB/T 41867-2022 信息技术 人工智能 术语

[5] GB/T 42460-2023 信息安全技术 个人信息去标识化效果评估指南

[6] GB/T 45652-2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

[7] GB/T 45654-2025 网络安全技术 生成式人工智能服务安全基本要求

[8] GB/T 45674-2025 网络安全技术 生成式人工智能数据标注安全规范

[9] TC260-PG-20244A 网络安全标准实践指南——敏感个人信息识别指南

[10] TC260-PG-AAAABB 网络安全标准实践指南——个人信息保护 个人信息匿名化指南

[11] TC260-PG-AAAABB 网络安全标准实践指南——个人信息保护 个人信息去标识化指南