|  |  |
| --- | --- |
| ICS | 35.240.30 |
| CCS | |  | | --- | | D:\000000部门项目\09标准化插件开发\程序源代码\StandardEditor_ShanDongKeXieYuan\团标首页面字母T.pngD:\000000部门项目\09标准化插件开发\程序源代码\StandardEditor_ShanDongKeXieYuan\团标首页面字母T后面的反斜杠.png XJBX |   L 67 |

西安市计量标准检测认证协会团体标准

T/XJBX 0065—2025

数字媒体生成式AI内容安全治理技术规范

Specification for security governance of generative AI content in digital media

2025 - XX - XX发布

2025 - XX - XX实施

西安市计量标准检测认证协会  发布

目次

[前言 III](#_Toc205745819)

[引言 V](#_Toc205745820)

[1 范围 1](#_Toc205745821)

[2 规范性引用文件 1](#_Toc205745822)

[3 术语和定义 1](#_Toc205745823)

[4 基本原则 2](#_Toc205745824)

[5 风险分类与分级管理 3](#_Toc205745825)

[6 技术控制要求 4](#_Toc205745826)

[7 运行监测与应急处置 5](#_Toc205745827)

[8 数据与模型管理 6](#_Toc205745828)

[9 质量评估 7](#_Toc205745829)

1. 前言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由西安市计量标准检测认证协会提出并归口。

本文件起草单位：合肥财经职业学院。

本文件主要起草人：陈琳。

1. 引言

近年来，生成式人工智能（Generative Artificial Intelligence, GAI）技术在数字媒体内容生产领域快速发展，广泛应用于文本、图像、音视频、三维建模等多种内容的自动生成与编辑。这类技术在提升生产效率、丰富创作形式的同时，也带来了虚假信息传播、版权侵权、隐私泄露、内容滥用等安全风险，甚至可能被用于生成有害、违规或不当信息，对社会稳定、网络生态与公共利益构成挑战。

目前，国内外在生成式AI内容安全治理方面已开展了一系列探索，涵盖算法可解释性、训练数据合规性、内容检测与溯源、风险分级管理、伦理审查及应急处置等方面。然而，行业内缺乏统一的技术规范，导致不同平台、机构在风险识别、处置机制与责任划分方面存在差异，难以形成协同治理合力。

本文件旨在为数字媒体生成式AI内容的生产、发布、传播和存储等全链条环节提供统一的安全治理技术要求，指导相关企业、机构和平台建立健全风险防控体系，提升内容安全审查的自动化、智能化与可追溯能力，推动生成式AI技术在合规、安全、可控的框架内健康发展。

本文件的制定坚持“安全优先、技术赋能、责任明确、协同治理”的原则，兼顾技术先进性与可操作性，结合国内法律法规和国际通行标准，适用于算法研发机构、数字媒体平台、内容生产企业以及其他涉及生成式AI内容的相关组织。

数字媒体生成式AI内容安全治理技术规范

* 1. 范围

本文件规定了数字媒体生成式人工智能内容安全治理的基本原则、风险分类与分级管理、技术控制要求、运行监测与应急处置、数据与模型管理以及质量评估等内容，适用于涉及生成式AI内容生产、编辑、发布、传播、存储和应用的各类组织与平台。

本文件适用于从事生成式AI算法研发、模型训练、内容生成及分发的企业、科研机构、平台运营方以及为其提供相关技术支持和服务的第三方机构。主要包括但不限于文本生成、图像生成、视频合成、语音生成、虚拟人生成、三维内容建模等数字媒体内容的安全治理活动。

本文件不适用于纯内部测试、且不对外公开发布的生成式AI内容生产活动，但其中涉及的数据采集、模型训练和安全管理可参照本规范执行。

* 1. 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求

GB/T 25069—2022 信息安全技术 术语

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 41867—2022 信息技术 人工智能 术语

* 1. 术语和定义

下列术语和定义适用于本文件。

生成式人工智能 generative artificial intelligence (GAI)

利用机器学习模型（特别是深度学习模型）生成与训练数据分布相似的新数据内容的人工智能技术，包括但不限于文本、图像、音视频、语音和三维模型生成。

数字媒体 digital media

通过数字化方式生成、存储、传播和展示的各类信息内容，包括文本、图片、音频、视频、动画和虚拟现实等形式。

内容安全 content security

在数字媒体生产与传播过程中，确保内容符合法律法规、伦理规范及平台政策的综合管理与技术保障措施，防止违法、违规、有害或不当信息的产生与传播。

风险分级管理 risk-based classification management

根据生成式AI内容的潜在安全风险程度、传播范围与影响力，将内容划分等级并实施差异化的管理与处置策略。

数据溯源 data provenance

记录、追踪和验证生成式AI模型训练数据及其来源、加工过程、版本变化的技术与方法，确保数据合规与可追溯。

模型可解释性 model interpretability

AI模型在决策或生成内容时，其内部运行机制和输出结果能够被人类理解和解释的特性。

深度合成 deep synthesis

利用深度学习等人工智能技术对原始数据进行生成、替换或合成，从而生成具有高度真实感的图像、音频或视频内容的过程。

有害内容 harmful content

可能对个人、群体或社会造成负面影响的内容，包括但不限于虚假信息、暴力、色情、歧视性言论、恐怖主义宣传等。

* 1. 基本原则
     1. 合法合规

生成式AI内容的生产、发布、传播应严格遵守国家法律法规、行业规范及平台管理政策，确保数据采集、模型训练、内容生成和分发的全过程合规。

* + 1. 安全可控

在技术架构和运行机制中，应纳入内容安全防护措施，确保生成内容的可控性和可预测性，防止出现不可预期的违规、有害或误导性信息。

* + 1. 责任可追溯

建立从训练数据、算法模型到最终生成内容的全链路记录与追踪机制，确保在出现违规或有害内容时能够明确责任主体，并实现快速溯源与处置。

* + 1. 技术与管理并重

在内容安全治理中，应将技术手段与管理制度相结合，通过算法过滤、人工审核、用户反馈等多重机制实现风险防控。

* + 1. 风险分级与动态管理

根据生成式AI内容的潜在风险等级，实施差异化管理措施，并根据风险变化动态调整安全策略与处置方案。

* + 1. 透明与可解释性

在保障商业秘密和用户隐私的前提下，应适度公开生成式AI的运行机制、数据使用情况及内容审核标准，提升用户与公众的信任度。

* + 1. 尊重伦理与人权

确保生成内容尊重社会公德、文化多样性与个人合法权益，不得生成歧视性、侮辱性、侵犯隐私或侵犯知识产权的内容。

* 1. 风险分类与分级管理
     1. 风险分类

生成式AI内容的风险可根据其性质和潜在危害分为以下主要类别：

1. 违法违规内容：违反法律法规、危害国家安全、扰乱社会秩序的内容，如反动言论、恐怖主义宣传等；
2. 有害不当内容：虽不触犯法律，但违反社会公德、损害青少年身心健康的内容，如色情、暴力、恶意诽谤等；
3. 虚假与误导性信息：包含事实错误或误导公众的信息，如假新闻、虚构事件、篡改数据等；
4. 侵权内容：侵犯他人知识产权、肖像权、名誉权等合法权益的内容；
5. 隐私泄露内容：未经许可披露他人个人信息或敏感数据的内容。
   * 1. 风险分级

根据内容的危害程度、传播范围和处置难度，将风险等级划分为高风险、中风险和低风险（见表1）。

1. 风险等级划分

| 风险等级 | 特征描述 | 处置要求 |
| --- | --- | --- |
| 高风险 | 涉及违法违规、严重有害或重大社会影响的内容；一旦传播可能造成不可逆损害 | 必须立即拦截或下架，启动应急处置机制，并上报相关主管部门 |
| 中风险 | 内容存在不当、侵权或误导性，但影响范围有限且可修正 | 暂停发布或传播，经审核整改后可恢复 |
| 低风险 | 内容轻微偏离规范，存在潜在争议，但危害可控 | 可通过标注、提示等方式引导用户辨别，无需立即下架 |

* + 1. 风险管理措施

高风险内容应建立实时识别与拦截机制，并由专人负责处置与报告。

中风险内容应进行人工复核，并在整改后重新评估发布。

低风险内容应定期监测，并通过用户教育、提示标注等方式降低风险影响。

* + 1. 动态调整

风险等级应根据事件发展、舆情变化和监管要求动态调整，确保管理措施与实际风险水平相匹配。

* 1. 技术控制要求
     1. 模型训练数据合规性

模型训练所用数据应来源合法、授权充分，不得包含违法违规或未经许可的受保护信息。

对来源不明或存在争议的数据，应在使用前进行合法性与伦理风险评估，并保留评估记录。

对包含个人信息的数据，应进行匿名化或去标识化处理，并符合个人信息保护规范。

* + 1. 内容生成前置过滤

在内容生成阶段引入风险词库、违规样本库等过滤机制，对输入提示词（prompt）及模型生成候选内容进行实时比对。

对触发高风险关键词的生成请求，应直接拒绝或进入人工审核环节。

过滤规则应根据最新法律法规、行业标准及舆情变化定期更新。

* + 1. 内容生成后检测

在生成内容输出前，使用自动化检测工具对其进行语义分析、图像识别、音视频分析等综合安全审查。

对检测结果为高风险的内容，应自动拦截并启动人工复核。

检测模型应具备持续学习与更新能力，以应对新型违规内容的出现。

* + 1. 溯源与水印技术

对生成内容应嵌入不可见的数字水印或元数据标签，以标识其生成来源、时间和模型版本。

水印应具备防篡改和可验证性，确保在内容传播过程中仍可追溯。

平台应提供内容溯源验证接口，供监管机构和授权用户使用。

* + 1. 标注与提示

对AI生成内容，应在用户可见位置添加明显标识或提示信息，说明内容由人工智能生成。

对可能引起误解的合成内容，应附加风险提示或来源说明。

标注格式和位置应保持统一，并符合行业通用可读性标准。

* + 1. 安全策略与更新

应建立技术策略动态更新机制，确保过滤规则、检测算法、风险词库等与法律法规、舆情环境保持同步。

技术更新过程应进行版本管理和回溯记录，确保可追踪与可验证。

* 1. 运行监测与应急处置
     1. 实时监测

平台应部署实时监测系统，对生成式AI内容生产、发布和传播过程进行全链路监控。

监测范围应涵盖生成请求、生成结果、用户反馈、传播路径及外部引用情况。

监测数据应定期存储和备份，保存期限不得少于法律法规规定的时限。

* + 1. 异常告警

系统应根据风险分级设定多级告警阈值，并支持自动化告警推送至安全管理人员。

对触发高风险告警的内容，应自动锁定并限制进一步传播。

告警信息应包含事件时间、内容摘要、风险类别、影响范围等关键要素。

* + 1. 事件响应

高风险事件应在发现后第一时间启动应急响应机制，包括内容下架、传播封锁、账户限制等措施。

中风险事件应在限定时间内完成复核，并采取整改或限制措施。

对重大事件应同步上报至相关主管部门。

* + 1. 处置流程

处置流程如下：

1. 识别：通过监测系统、人工审核或用户举报发现异常内容；
2. 评估：安全团队对风险等级和影响范围进行快速评估；
3. 处置：根据风险等级实施相应的管控措施；
4. 复盘：事件结束后进行原因分析和流程改进；
5. 记录：完整记录事件信息、处置过程及结果，存档备查。
   * 1. 用户举报与反馈机制

平台应提供便捷的用户举报入口，支持多种类型的违规内容分类选择。

对有效举报应在规定时限内反馈处置结果。

用户反馈信息应纳入风险分析数据库，用于优化监测与防控策略。

* + 1. 持续改进

应定期对监测与应急处置体系进行评估，结合事件复盘结果和外部监管要求优化流程。

结合技术发展趋势，提升自动化处置比例和智能化水平。

* 1. 数据与模型管理
     1. 数据合规管理

数据采集应遵循合法、正当、必要原则，不得超范围收集与使用。

对涉及个人信息、敏感信息的数据，应进行匿名化或去标识化处理，并符合个人信息保护相关标准。

数据来源、采集方式、使用范围及授权文件应留存完整记录，确保可追溯。

* + 1. 数据质量控制

训练数据应经过清洗、标注、审核，剔除违法违规、有害不当或低质量内容。

应建立数据质量评估机制，定期对数据集进行抽检和更新。

对引入外部数据集的，应进行合法性和安全性评估。

* + 1. 模型版本管理

模型应实行版本化管理，记录每次训练的时间、数据来源、参数配置及优化目标。

每个版本的模型应进行安全性和内容生成效果的评估，并保留评估报告。

对发现存在安全风险的模型版本，应立即停用并进行整改。

* + 1. 模型访问控制

对模型的访问应实行分级授权管理，确保不同角色的人员仅能访问其工作所需的模型和数据。

重要模型应采取多因素认证、密钥管理等安全措施，防止未经授权的调用与下载。

应记录所有访问操作，并定期审计访问日志。

* + 1. 模型安全评估

模型上线前应进行算法安全评估，覆盖对抗样本鲁棒性、有害内容生成倾向、可解释性等方面。

应定期对已部署模型进行再评估，发现安全隐患及时修正。

对高风险应用场景，应引入第三方评估机构进行独立审查。

* + 1. 备份与恢复

应建立数据与模型的备份机制，确保在发生系统故障或安全事件后可快速恢复。

备份数据应加密存储，并与生产系统隔离。

定期测试恢复流程，确保可用性。

* 1. 质量评估
     1. 评估目标

质量评估应以验证内容安全治理体系的有效性、覆盖性和响应性为目标，确保生成式AI内容在全生命周期内符合安全、合规、可控的要求。

* + 1. 评估指标

可根据平台或机构的实际情况，建立包括但不限于以下方面的指标体系：

1. 内容安全合规率：在抽检样本中，符合法律法规及平台规范的内容比例；
2. 违规内容处置及时率：从发现到处置完成的平均时间；
3. 风险识别准确率：自动化检测系统识别的准确性与漏报率；
4. 模型安全性指标：包括有害内容生成倾向评估、鲁棒性测试结果；
5. 用户举报反馈率：用户举报案件的有效回复比例与时效；
6. 事件复发率：同类安全事件在整改后再次发生的比例。
   * 1. 评估周期

日常监测应持续进行，并形成月度或季度报告。

综合评估应至少每年开展一次，涵盖技术、管理、培训等全方面内容。

在法律法规或技术环境重大变化时，应开展专项评估。

* + 1. 评估方法

结合自动化检测结果、人工审核抽检、第三方评估等多种方式开展评估。

采用量化指标与定性分析相结合的方法，确保评估结果客观全面。

评估过程应有完整记录，评估结论应经管理层审核。

* + 1. 持续改进

应根据评估结果制定改进计划，明确改进目标、措施、责任人和完成时限。

对发现的重大问题应立即整改，并验证整改效果。

应建立经验反馈机制，将事件处置经验和技术优化成果纳入培训与系统优化中。

* + 1. 外部监督与合作

鼓励与行业组织、科研机构、监管部门合作，开展联合评估与技术攻关。

应接受行业和社会的合理监督，定期公开治理成效报告，提升透明度与公信力。

