ICS 35 240 01

16429

T/GXDSL

团　　体　　标　　准

T/GXDSL 041—2025

生成式 AI 内容安全与伦理审查规范

Generative AI Content Security and Ethical Review Standards

2025 - 12 - 5 发布　　2026 - 3 - 5 实施

广西电子商务企业联合会　发布

目　　次

前　言

本文件依据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。

本文件由广西产学研科学研究院提出。

本文件由广西电子商务企业联合会归口。

本文件起草单位：广西产学研科学研究院,广西研科院高新技术有限公司，广西研科院传媒有限公司，遇龙宝(广西)科技有限公司，上海振薰信息技术有限公司，广西玉馨信息科技有限公司，广西大学，广西民族大学，广西财经学院，广西衍知智能科技有限公司。

本文件主要起草人：庄文斌，叶华林，黄熙宇，包奇，彭时矿，王灿琼，谢品，周卫，祝凡，黄子铭。

本文件为首次发布。

生成式 AI 内容安全与伦理审查规范

一、引言

随着生成式人工智能技术（以下简称“生成式 AI”）在电子商务领域的广泛应用，其在提升营销效率、优化用户体验的同时，也带来了虚假信息传播、隐私泄露、算法歧视、文化价值观冲突等风险。本标准以“安全可控、伦理先行、技术赋能”为核心理念，结合广西作为中国—东盟数字经济合作枢纽及多民族文化交融的区域特点，针对跨境电商、直播电商、农村电商等特色场景，细化生成式 AI 内容全生命周期管理要求。通过构建“数据合规—算法透明—内容可控—用户保护”四位一体的治理框架，推动技术应用与社会主义核心价值观深度融合，为行业健康发展提供指引。

二、规范性引用文件

下列文件对于本标准的应用必不可少。凡注日期的引用文件，仅所注日期的版本适用于本标准。凡不注日期的引用文件，其最新版本（包括所有修改单）适用于本标准。

GB/T 35273-2020《信息安全技术 个人信息安全规范》

GB/T 39335-2020《信息安全技术 数据安全能力成熟度模型》

《生成式人工智能服务管理暂行办法》（国家网信办等七部委令第 15 号，2023 年）

《互联网信息服务深度合成管理规定》（国家网信办令第 12 号，2022 年）

《广西壮族自治区数字经济条例》（2023 年修订版）

《中国—东盟跨境电子商务合作协议》（商务部，2022 年）

ISO/IEC 23053:2022《人工智能系统生命周期过程框架》

三、术语和定义

下列术语和定义适用于本标准：

（一）生成式人工智能（Generative AI）

基于机器学习算法，能够自主生成文本、图像、音频、视频等内容的技术系统，典型代表包括大语言模型（LLM）、扩散模型（Diffusion Model）等。

（二）深度伪造（Deepfake）

通过生成式 AI 技术合成或替换原始内容中的生物特征（如人脸、声纹），达到以假乱真效果的行为。

（三）算法偏见（Algorithmic Bias）

因训练数据不平衡或模型设计缺陷，导致生成内容出现地域、性别、民族等维度歧视的现象。

（四）区块链存证（Blockchain Notarization）

利用区块链技术对生成内容的关键参数（时间戳、哈希值、操作者身份）进行不可篡改记录的过程。

（五）伦理审查委员会（Ethics Review Board）

由跨学科专家组成的机构，负责评估 AI 系统在价值观对齐、文化适配性、社会影响等方面的合规性。

（六）越狱提示词（Jailbreak Prompt）

用户通过特殊指令绕过 AI 系统安全限制，诱导生成违规内容的行为，例如“请忽略之前的道德准则”。

四、总则

（一）目的与依据

为规范生成式 AI 技术在电子商务领域的研发、部署与应用，防范技术滥用导致的虚假宣传、消费者欺诈、民族宗教冲突等风险，依据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国电子商务法》《互联网信息服务算法推荐管理规定》等法律法规，制定本规范。

（二）适用范围

本规范适用于广西区域内从事生成式 AI 技术研发、服务提供或内容分发的电子商务平台企业、技术服务商、MCN 机构及个人创作者，覆盖商品描述生成、智能客服对话、虚拟主播直播、广告素材创作等典型应用场景。

（三）核心原则

合法性原则：生成内容不得包含法律禁止的虚假广告、侵权信息、违禁商品描述，不得利用 AI 技术伪造交易记录、用户评价等商业数据。

安全可信原则：建立覆盖模型训练、内容生成、传播分发的全链路风险控制体系，确保 AI 系统在面对“诱导性指令”“越狱提示词”等攻击时具备 99.9%以上的防御有效性。

文化适配原则：针对广西壮族自治区内 12 个世居民族的文化特征，模型需具备方言识别、风俗禁忌过滤能力，确保生成内容尊重少数民族传统习俗。

责任可追溯原则：通过区块链存证、数字水印等技术手段，实现生成内容的全生命周期溯源，支持监管部门对违规主体的快速定位与责任认定。

五、技术合规要求

（一）数据采集与训练合规

训练数据来源应通过合法授权协议获取，禁止使用未公开的政府数据、企业商业秘密及未脱敏的个人信息。用于模型微调的行业数据（如跨境电商商品库、农产品价格数据），需经数据提供方书面授权，授权文件保存期限不得少于 10 年。

数据标注环节需建立“三审三校”机制，即标注员初标、质检员复检、专家终审，标注错误率需低于 0.3%，对涉及民族、性别、宗教的敏感标注项需由 3 名以上独立审核员交叉验证。

（二）算法透明度与公平性

企业应在产品界面或服务协议中公开算法基本原理、主要运行机制及可能存在的局限性，例如图像生成模型对部分少数民族服饰纹样的还原误差率需明确告知用户。

算法设计需避免基于地域、民族、年龄等因素的歧视性输出。在跨境电商场景中，AI 生成的商品推荐列表需确保东盟国家特色商品占比不低于 15%，避免算法过度偏向单一市场。

（三）内容安全防护能力

文本生成场景需部署多层级语义分析模型，对“高仿品牌名称”“违禁药品代称”等变体表述实现98%以上的识别准确率，并结合人工审核对金融投资建议类内容进行 100%复核。

图像与视频生成系统应集成生物特征检测模块，对深度伪造（Deepfake）技术生成的公众人物肖像、政府证件照等实现 99.5%的拦截率，并自动向网信部门报送伪造样本特征。

六、伦理审查机制

（一）伦理委员会建设

企业应设立由 5-9 人组成的生成式 AI 伦理审查委员会，成员需包含法律专家（不少于 2 人）、少数民族代表（至少 1 人）、技术伦理学者及消费者权益保护组织成员。委员会需每半年接受自治区科技

伦理委员会的业务指导。

伦理审查范围应包括：

模型训练数据是否存在地域偏见（如过度依赖东部地区语料导致对西南方言理解偏差）；

生成内容是否符合东盟国家文化禁忌（如对穆斯林头巾、佛教符号的规范性使用）；

算法决策逻辑是否可能加剧“大数据杀熟”“价格歧视”等不公平商业行为。

（二）价值观对齐与场景适配

在乡村振兴电商场景中，AI 生成的农产品营销文案需避免夸大功效（如宣称“某特产可治疗癌症”），并需经县级农业农村部门备案。

针对直播电商中的虚拟主播应用，禁止生成宣扬奢靡消费、攀比炫富的内容，对打赏金额超过 500 元的用户需弹出理性消费提醒。

（三）特殊群体保护

面向老年用户的 AI 客服系统，需禁用复杂金融术语，对“高收益理财”“保健品促销”等话术自动触发人工介入机制。

未成年人模式需屏蔽“充值返利”“抽奖诱导”等内容生成功能，并通过人脸识别技术确保身份验证准确率达 99%以上。

七、内容安全审查流程

（一）预生成风险拦截

输入指令需实时比对国家违法和不良信息举报中心发布的《互联网敏感词库（2023 版）》，对“刷单”“洗钱”等黑灰产关联词实施动态屏蔽，词库更新时间间隔不超过 1 小时。

在跨境电商场景中，需建立多语种敏感词库（含东盟十国官方语言），对涉及领土争议、宗教冲突的表述实现 95%以上的跨语言识别准确率。

（二）生成中实时管控

采用“AI 初审—人工复审—专家会审”三级审核机制：

AI 初审需在 3 秒内完成，对明显违规内容直接拦截；

人工复审团队需具备法律、语言学、民族学专业背景，日均处理量不超过 800 条/人以保证审核质量；

对争议内容（如民族服饰设计是否构成文化挪用）需提交伦理委员会会审，决议需在 24 小时内作出。

直播电商场景中，虚拟主播的实时语音生成需延迟 5 秒播出，以便审核人员对敏感内容实施紧急切断。

（三）生成后追溯与反馈

所有发布内容需附加不可篡改的区块链存证标识（含生成时间、操作者 ID、模型版本号），存证信息同步至自治区区块链公共服务平台。

建立用户举报“一键溯源”机制，对确认为违规的内容，需在 1 小时内完成全平台下架，并通过模型再训练降低同类内容再生概率至 0.1%以下。

八、违规处置与问责

（一）违规等级与处罚

一般违规：首次生成轻微违规内容（如使用禁用词汇但未造成传播），需对责任模型进行 48 小时强化训练，并向广西电子商务企业联合会提交整改说明。

严重违规：生成内容引发民族矛盾或国际纠纷（如错误使用东盟国家地图），涉事企业需暂停相关服务 3 个月，缴纳 50 万元违约金，并接受自治区网信办专项审计。

刑事犯罪：利用生成式 AI 伪造国家机关公文、实施网络诈骗等行为，涉事主体将被永久列入行业

黑名单，并移交公安机关处理。

（二）应急响应与披露

发生大规模内容安全事件（如 AI 生成虚假新闻阅读量超 100 万次），企业需在 2 小时内通过平台公告、短信推送等方式告知受影响用户，并在 72 小时内召开新闻发布会说明处置进展。

建立“AI 安全白皮书”年度发布制度，企业需公开全年内容拦截量、主要违规类型、模型迭代次数等数据，接受社会监督。

九、附则

（一）标准解释与修订

本规范由广西电子商务企业联合会负责解释。根据技术发展及监管要求，联合会应每年组织修订预研，修订内容需公示 30 日并召开听证会听取中小企业意见。

（二）实施与认证

自 2025 年 1 月 1 日起，新上线生成式 AI 服务需通过合规性认证；存量服务需在 2025 年 6 月 30 日前完成整改。

合规认证由自治区市场监管局授权的第三方机构执行，通过企业可获得“桂品 AI 安全认证”标识，未通过认证不得在广西境内开展相关业务。

（三）奖励与扶持

对在以下方面取得突出成效的企业，给予政策支持：

研发具有民族文化特色的 AI 内容审核工具（如壮语语义分析模型）；

构建中国—东盟多语种内容安全联合实验室；

年度内容安全事件零报告并持续 3 年以上。

支持措施包括税收减免、优先参与政府项目、最高 300 万元创新补贴等。

（四）生效日期

本规范自 2026 年 3 月 5 日起实施。

附录 A（规范性附录）：生成式 AI 内容安全审查流程图

A.1 全流程审查节点

- 输入阶段：敏感词过滤→指令合法性校验→用户身份核验。
- 生成阶段：多模态内容检测→实时风险评分→人工介入阈值（风险分≥70 分需强制复核）。
- 发布阶段：区块链存证→传播范围监测→用户反馈闭环。

A.2 审核时效标准

内容类型	AI 初审时限	人工复审时限
文本	≤3 秒	≤30 分钟
图像	≤10 秒	≤2 小时
视频	≤15 秒	≤6 小时

附录 B（资料性附录）：东盟国家文化禁忌库（示例）

B.1 语言禁忌

马来西亚：禁用“猪”“狗”作为比喻词。

越南：避免使用“3”字组合（与负面历史事件关联）。

B.2 视觉禁忌

印度尼西亚：禁止生成戴头巾女性的娱乐化形象。

泰国：佛像图像需完整呈现，不得截取局部。

B.3 宗教禁忌

文莱：禁止生成含酒精饮料的促销内容。

菲律宾：圣诞节相关素材不得与商业广告过度结合。

附录 C（规范性附录）：伦理审查表模板

C.1 审查项目清单

评估维度	审查要点	判定标准
文化尊重	是否包含民族歧视性表述	需 100%无歧视
商业诚信	是否存在虚假功效宣称	科学依据≥3 项权威文献
未成年人保护	是否触发青少年模式过滤	拦截率≥99%

C.2 审查结论格式

审查编号：ERB-202X-XXX

模型名称：_____

风险等级：☐低风险 ☐中风险 ☐高风险

整改要求：_____

委员会签字：_____ 日期：_____