ICS 35.240.01
CCS L 70

T/AIIA

团　　体　　标　　准

T/AIIA XXXX—XXXX

基于大模型生成内容的安全管理规范

Safety management specifications for content generated based on large models

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

2025－XX－XX 发布　　2025－XX－XX 实施

深圳市人工智能产业协会　　发 布

目　　次

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。

本文件由深圳市迪博技术有限公司提出。

本文件由深圳市人工智能产业协会归口。

本文件起草单位：

本文件主要起草人：

引　言

随着人工智能技术的飞速发展，大模型在数字政府领域的应用日益广泛，其生成内容在政务服务等场景中发挥着重要作用。然而，大模型生成内容可能存在的安全风险，如违法违规、泄露隐私等，不仅会影响政务服务的质量和效率，还可能对国家安全、社会稳定及公众利益造成潜在威胁。

为规范数字政府领域基于大模型生成内容的安全管理，确保生成内容的安全、合规、可控，保障政务服务的安全性和可靠性，维护公众的合法权益，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》等相关法律法规，结合深圳市数字政府人工智能公共支撑平台中模型内容安全风控系统的应用实际，特制定本文件。

基于大模型生成内容的的安全管理规范

1 范围

本文件规定了在数字政府领域基于大模型生成内容的安全管理要求，包括内容生成的敏感词管理、生成策略与控制机制、审查机制、拒答机制与策略、组织保障与人员管理等。

本文件适用于政务服务机构、大模型技术提供方及相关第三方在数字政府领域中基于大模型进行内容生成的系统或产品的安全管理活动。

2 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/T45288.1：2025 人工智能 大模型 第1部分：通用要求

ISO/IEC 27001 信息安全管理体系

中华人民共和国网络安全法

3 术语和定义

下列术语和定义适用于本文件。

3.1

大模型 large-scale modle

基于大量数据训练得到,具有复杂计算架构,能处理复杂任务,且具备一定泛化性的深度学习模型。

[来源：GB/T45288.1-2025,3.1]

3.2

拒答机制 refusal mechanism

大模型在识别到输入请求超出其能力范围、涉及不适当或敏感主题或无法提供可靠答案时，主动选择不生成实质性回复，而是返回明确拒绝或引导性提示的机制。

3.3

生成内容 generated content

由大模型根据给定的提示词、上下文、数据等输入，通过其内部计算和处理过程所产生的文本、代码、图像（对于多模态模型）或其他形式的数据输出。

3.4

敏感词 sensitive word

在特定语境下可能引发国家安全、社会稳定、个人隐私或公共道德等方面风险的关键词、短语或表达。

4 内容生成敏感词管理

4.1 敏感词识别与分类

应建立覆盖以下维度的敏感词分类体系，结合多模态内容特征实现精准识别：

a）违法违规内容：指违反《中华人民共和国网络安全法》《中华人民共和国数据安全法》等法律法规的内容，包括但不限于：

- 政治风险：涉政敏感人物、敏感事件、违规地图、颠覆国家政权言论等；
- 暴恐风险：恐怖组织宣传、暴力血腥画面/描述、武器制造与传播等；
- 违法风险：违禁品交易、非法团体活动、侵犯公民个人信息等。

b）伦理与社会风险：

- 色情低俗：低俗性行为描述、低俗音频等内容；
- 歧视与仇恨：地域/性别/种族歧视言论、煽动群体对立内容；
- 道德争议：自杀诱导、校园霸凌引导、毒品滥用引导等；
- 文化禁忌：宗教亵渎内容、历史虚无主义言论、破坏文化传承的表述。

c）专业误导性内容：

- 医疗健康：未经科学验证的偏方、虚假医疗广告、无资质诊疗建议；
- 金融财经：违规荐股、虚假投资承诺、非法集资宣传；
- 事实错误：篡改历史事件、伪造新闻信息、错误学术观点传播。

d）其他风险：包括但不限于垃圾广告、恶意谩骂、敏感人物关联内容等。

4.2 敏感词等级定义

内容安全敏感词根据危害程度分为以下三级：

a）高危害敏感词：指可能危害国家安全、破坏社会稳定、煽动民族仇恨、宣扬恐怖主义、传播色情暴力、泄露国家秘密的内容；可能引发重大舆情或法律责任的生成内容，包括涉政敏感言论、暴恐方法、侵害未成年人言论、违禁品制作指南等。

b）中危害敏感词：指存在偏见歧视、事实性错误、专业误导风险的内容，包括如未经验证的科研结论、金融投资建议、敏感文化议题、涉及伦理争议的生成内容等。

c）低危害敏感词：指通用生活信息或娱乐类内容，但需基础事实性校验，包括如日常咨询、科普知识、非专业领域建议等。

4.3 敏感词评估方法

结合多维度数据与动态分析技术，实现全流程敏感词评估，具体包括：

a）场景分层评估法：根据社交聊天、图片人脸等关联场景的敏感度分级，动态调整审核阈值。例如，针对未成年人场景的侵害言论阈值应高于普通场景；

b）名单库匹配机制：利用预设名单和自定义名单，实现关键词、人脸、图片的快速匹配与风险标记。

5 内容生成策略与控制机制

5.1 提示词（Prompt）设计与安全策略控制

通过系统化的Prompt设计规范、模板管理机制及输入预处理流程，结合平台权限控制与内容安全组件，从源头阻断敏感内容生成路径，确保用户输入与模型交互全程处于安全约束框架内。具体措施包括：

a）安全Prompt设计原则

- 角色限定：在系统层应植入强制身份声明指令，明确AI服务边界。
- 模板分级管控：基于Prompt工程模块的模板分类建立分级管理机制：预置模板由平台管理员统一维护，并嵌入不可修改的安全约束条款；自定义模板需经过内容安全校验，关联敏感词校验功能。
- 动态适配与优化：通过平台内置的Prompt优化工具，对用户输入的自定义Prompt进行结构校验，自动补充边界约束。

b）输入预处理

- 敏感词过滤：支持模糊匹配，包括拆分字、形似字、音似字、简繁体、大小写、大写数字等形式的相似词。
- 预处理日志与审计：所有输入内容的过滤记录应同步至用户行为日志模块，包括：被拦截的敏感词/短语及关联用户信息、模板匹配与权限校验结果等内容。

5.2 敏感内容生成的前置规避机制

5.2.1 组织应建立和维护动态更新的内容安全策略与规则库，用于指导输入预处理、生成过程控制与输出审查。

5.2.2 规则库应至少涵盖 4.2 中定义的各类风险内容，并明确其对应的等级（4.3）。

5.2.3 应指定专门团队负责规则库的维护，该团队应定期或不定期对规则库进行评审和更新。
5.2.4 规则库的更新应有明确的流程和版本控制。

5.3 输出控制策略

在政务服务场景中，通过token级屏蔽、输出内容重写、实时内容打分与拦截等技术，确保大模型生成的政务内容符合法律法规、政策要求及政务服务规范，避免敏感信息、错误表述或违规内容输出：

a）token级实时控制
- 黑名单强制屏蔽：基于关键词名单与图库名单，在解码阶段对涉密政务术语、诋毁政务机关表述等敏感 token 进行替换。
- 白名单精准引导：针对特定场景，限制输出词汇至政务白名单，避免生僻术语或歧义表述。

b）生成后处理技术
- 无害化改写：对高风险句子调用政务专属改写模型，转化为合规表述，确保保留核心信息且符合政务规范。
- 可追溯标记：在生成内容中嵌入隐式标识，关联识别统计中的记录，实现“模型版本－审核人员－生成时间”全链路追溯。

5.4 多模态生成内容的安全校验要求

5.4.1 图像内容安全校验

a）目标与场景识别：利用目标检测技术识别武器、毒品等违禁品，通过场景识别定位色情、涉政等敏感场景；
b）文本与匹配校验：采用OCR技术提取图片中的文本，与关键词名单比对；通过图文匹配技术检测“图片涉政+文本合规”等不一致风险；
c）图库匹配：与政治人物、暴恐图片等预设图库和自定义图库进行相似度匹配，命中则标记风险。

5.4.2 视频内容安全校验

a）帧级图像校验：提取关键帧生成图像序列，复用图像校验技术（目标识别、OCR等）；
b）行为与面部分析：通过行为识别检测暴力、性暗示等动作；
c）跨模态一致性校验：对比音频文本与画面内容，通过音画同步技术检测语音与嘴型、场景与音频的匹配度。

6 生成内容的审查机制

6.1 审查日志与记录留存机制

6.1.1 审核行为记录

每次审核行为均应记录，包括以下要素：
a）基础信息：审核事件唯一标识、审核时间、生成内容类型（文本/图片/音频/视频）、关联的安全策略版本；
b）主体信息：审核人员账号（主账号/子账号）、用户标识（脱敏处理）、关联应用及场景；
c）内容信息：原始生成内容、敏感信息标记、审核结论（通过/违规/疑似）及理由；
d）流程信息：审核环节、各环节耗时、流转路径。

6.1.2 技术要求

审查日志的管理应满足以下技术要求：
a）应提供日志可通过明细查询功能查看，支持按时间、账号、风险类型筛选；
b）日志访问权限受权限管理控制，仅主账号及指定管理员可查询，操作记录同步留存；
c）应支持支持日志数据导出，格式适配识别统计分析需求。

6.2 生成内容合规性要求

生成内容必须严格遵守国家现行法律法规，对涉政、暴恐、色情、违法等违法违规内容实施零容忍策略。系统需基于预设名单与自定义名单动态更新关键词库和审核规则，通过策略配置确保内容符合安全规范。

7 拒答机制与策略

7.1 高风险问题拒答规则

基于关键词名单、预设图库等设置敏感问题黑名单，实行强制拒答策略。黑名单涵盖法律禁止的涉政、暴恐、色情、违法等内容，具体要求如下：

a）针对涉及违反法律法规及危害国家安全的提问，系统检测到此类问题后，立即终止响应并记录提问内容、时间及用户标识至审查日志，直接拒答。

b）黑名单采用动态更新机制，依据预设名单与自定义名单的更新规则定期调整；通过系统层熔断机制确保高风险问题被稳定拦截，保障拒答执行的刚性。

7.2 拒答策略稳定性与一致性要求

通过支持拆分字、形似字、音似字等形式的模糊匹配和上下文关联检测保障一致性，确保相同意图问题触发相同拒答，具体实现方式如下：

a）模糊匹配：对输入问题进行变体检测，识别通过字符替换、谐音等方式规避检测的提问，确保相似表述的问题触发相同拒答。

b）上下文关联检测：实现会话级上下文追踪，当用户通过多轮对话分步提问，系统识别关联性后触发全局拒答。

c）所有拒答决策均记录至审查日志，供审计核查。

8 组织保障与人员管理

8.1 内容安全责任团队与岗位职责

关键岗位权限通过权限管理严格控制，确保职责分离，基于权限管理建立角色分工体系：

a）超级管理员（主账号）：负责整体安全策略制定、账号权限分配；

b）审核负责人：处理高级审核环节的争议内容，监督审核流程；

c）审核员（子账号）：执行初级审核，标记高风险内容并流转。

8.2 审查人员权限管理与安全操作流程

实行“最小权限”原则：子账号仅能查看所属场景的审核内容，主账号可查看全量数据。所有审核操作记录明细查询，包含操作人、时间、结果，支持异常行为追溯。

8.3 审查人员培训与能力建设

8.3.1 组织应对所有涉及内容安全的管理、技术和审核人员进行岗前培训和定期继续教育。

8.3.2 培训内容应至少包括：相关法律法规、本规范的要求、内容安全风险识别能力、审核标准与流程、应急处置流程等。

8.3.3 应对关键岗位人员的胜任能力进行评估和考核。

参 考 文 献

［1］ 全国人民代表大会.中华人民共和国网络安全法.2016 年

［2］ 全国人民代表大会.中华人民共和国数据安全法.2021 年

［3］ 国家互联网信息办公室、工业和信息化部、公安部、国家市场监督管理总局.互联网信息服务算法推荐管理规定.2021 年