AI大模型生成内容风险监测与可信治理白皮书

——大模型内容认知治理的总框架

发布单位

国家广告研究院

四川省人工智能研究院

弗若斯特沙利文

智慧星光

编撰团队

王 昕　徐振廷　饶庆昇　陈 妮

贺照峰　王 静　吴文博　王举伟　秦 予

2026年6月

目录

编制说明

生成式人工智能正在由单一技术工具演变为信息基础设施。大模型生成的答案、摘要、推荐和解释，正在进入公共服务、企业经营与个人决策链条。

这意味着，内容治理的对象正在发生结构性变化：过去主要治理“人发布的内容”，现在还必须治理“机器生成的答案”；过去主要关注网页、短视频、广告、评论和新闻，现在还必须关注 AI 搜索摘要、智能客服话术、AI 导购结论、自动报告、企业知识问答以及智能体的工具调用结果。

本白皮书围绕“AI 大模型生成内容风险监测与可信治理”展开，旨在提出一套能够被政策研究、产品治理、内容运营、合规审计和行业评估共同使用的方法框架。它既关注违法违规与有害内容，也关注事实真实性、来源权威性、信息时效性、解释充分性、证据可追溯性以及错误对公共认知和市场秩序的影响。

边界声明	本白皮书属于总纲型研究成果，主要依据公开政策法规、国际治理框架、行业实践和前期监测经验形成。由于尚未纳入统一口径下的大规模多模型实测数据，文中不对国内外模型风险比例作统计性结论，不构成法律意见、监管结论或对具体模型的性能背书。

本白皮书重点回答

- AI 大模型生成内容风险与传统互联网内容风险有何结构性差异；

- 如何从模型、应用、内容、传播与组织治理五个层面识别风险；
- 如何建立可复用的问题库、监测流程、评分规则与证据包；
- 如何形成“发现—核查—留痕—处置—复测—改进”的可信治理闭环；
- 政府、平台、企业与研究机构如何划分责任并推进落地。

适用对象与使用方式

适用对象	典型使用方式	主要产出
政府与公共机构	公共信息准确性监测、政策问答评估、舆情与城市认知治理	风险清单、核查报告、整改建议
大模型与平台企业	输出安全评估、来源质量治理、标识与投诉机制	模型评估、策略规则、证据日志
行业企业	客服、营销、知识库、导购、报告等 AI 应用治理	场景分级、审核流程、供应商要求
研究与第三方机构	跨模型测试、专项评估、标准研究与能力验证	测试集、指标体系、评估报告

执行摘要

核心判断	大模型时代，可信内容不再只取决于“谁发布了信息”，还取决于机器如何读取、理解、生成、引用和传播信息。治理重点需要从单条内容审核升级为对答案、信源、流程和责任链的系统治理。

六项关键结论

01. 治理对象已经改变。AI 回答具有即时生成、动态变化、多轮交互和个性化特征，固定页面抽检无法完整覆盖。

02. 风险不只来自模型。基础模型能力只是起点，提示词、RAG 知识库、工具调用、审核流程和发布机制都可能引入风险。

03. 事实与来源成为核心。在政务、医疗、金融、教育和商业推荐等高影响场景，来源是否可靠、是否真正支持结论，与答案措辞同等重要。

04. 风险需要分级治理。场景影响、用户脆弱性、传播范围、可逆性和证据充分度应共同决定治理强度，避免“一刀切”。

05. 证据链决定可治理性。没有提示词、模型版本、输出全文、来源快照、核查记录与复测结果，风险便难以复现、问责和持续改进。

06 .治理目标是可信基础设施。最终目标不是只删除错误内容，而是让权威信息可被识别、可被引用、持续更新，并让错误能够被纠正。

总体框架

本白皮书将治理体系概括为“一套分层风险结构、两类评价结果、七步运行闭环、四类组织主体”。一套分层结构用于定位风险来源；两类评价结果分别回答“内容质量如何”和“错误后果多大”；七步闭环把监测转化为治理行动；政府、平台、企业与第三方机构共同构成责任体系。

图 1　AI 大模型生成内容五层风险结构

框架的四个支点

- **分层风险：**定位问题发生在模型、系统、内容、传播还是组织环节；
- **双轨评价：**同时输出可信质量分和影响风险等级；
- **闭环运行：**把发现、核查、处置与复测连接为持续改进；
- **多方共治：**明确政府、平台、部署企业与第三方评估机构的责任边界。

第一章　时代背景：内容治理进入 AI 生成答案时代

本章要点：说明治理对象为何变化，以及“答案层”和“认知分发层”带来的新风险。

1.1 从用户发布内容到机器生成答案

互联网早期的内容治理主要面对用户上传、机构发布和平台分发的信息。网页、论坛、新闻、短视频、直播、电商评价和社交媒体内容具有相对固定的发布主体、页面地址与传播轨迹，治理通常围绕发布前审核、发布后巡查和违规处置展开。

生成式人工智能改变了这一结构。大模型根据训练数据、系统提示、用户上下文、检索增强结果和工具调用状态即时生成新的内容。它不仅检索信息，还会归纳、重组、比较、推断并给出建议，由此形成新的治理对象——AI 生成答案。

关键差异	传统治理多针对“可定位的内容对象”，大模型治理则必须同时处理“动态的生成过程”和“上下文相关的答案结果”。

1.2 从搜索结果到生成结论

搜索引擎通常呈现多个链接，用户需要自行比较来源与结论。大模型倾向于直接输出结构化答案，并常常包含排序、推荐、风险提示和行动建议。信息获取效率由此提高，但用户核查来源的动力和机会也可能降低。

当 AI 回答错误时，错误不再只是某个网页中的错误，而可能成为用户直接采信的判断依据。在政务服务中，它可能错误解释办理条件；在医疗健康中，可能遗漏就医边界；在金融消费中，可能淡化风险；在企业声誉场景中，可能错误描述资质、处罚或争议；在广告营销中，可能生成夸大或不当承诺。

1.3 AI 生成答案的六项结构性特征

特征	具体表现	治理含义
动态性	同一问题在不同时间、版本或温度参数下输出变化	必须留存时间、模型版本与完整输出
交互性	答案随多轮追问和上下文累积而改变	测试需覆盖对话链，而非单轮提示
归纳性	模型把多来源压缩为一个结论	需核验来源与关键断言的对应关系
个性化	地区、账号、语言、画像和权限影响输出	样本应控制变量并记录环境
不稳定性	同条件重复测试仍可能出现差异	需要复测与置信区间思维
可扩散性	生成内容可能被转载、收录并反向进入数据生态	治理需覆盖传播与信源生态

1.4 从内容安全到可信认知

传统内容安全关注违法违规、有害信息、诈骗诱导、侵权和虚假广告等问题。AI 生成内容治理在此基础上增加了事实、来源、时效、完整、解释、标识、追溯和纠错等要求。内容即使没有触碰明确红线，也可能因为来源低质、条件遗漏或表达过度确定而造成实质性误导。

- 事实是否真实，关键断言能否被权威证据支持；
- 来源是否可靠、可访问且与结论匹配；
- 信息是否仍然有效，是否提示时间边界；
- 结论是否遗漏限制条件、适用范围和风险提示；
- 内容是否可识别、过程是否可追溯、错误是否可纠正。

1.5 治理原则

原则	内涵
以人为本	优先保护生命健康、财产安全、人格权益与弱势群体利益
风险相称	治理强度与场景影响、传播规模和错误可逆性相匹配
证据优先	以可复现的输出、权威来源和完整记录支撑判断
责任明确	模型提供方、应用部署方、内容发布方和使用方各负其责
持续改进	把投诉、处置和复测结果回流到问题库、规则库和知识库

第二章　全球治理趋势：风险分级、透明责任与组织治理

本章要点：梳理欧盟、美国、国际标准与中国大陆路径的共同方向和差异化重点。

2.1 共识正在形成，但治理路径并不相同

全球 AI 治理正在从原则倡议走向制度框架、组织管理和可审计实践。欧盟强调基于风险的法定义务与基本权利保护；美国 NIST 路径强调组织自愿风险管理与测量方法；国际标准强调管理体系和生命周期治理；中国大陆则突出内容安全、平台责任、算法治理、生成合成内容标识和分类分级监管。

这些路径虽不相同，却共同指向三项趋势：治理对象从模型能力延伸到模型输出，治理责任从技术团队扩展到组织管理，治理证据从原则声明转向可记录、可测试和可审计的过程。

2.2 欧盟路径：风险分级与透明义务

欧盟《人工智能法案》以风险为基础构建制度体系，并对特定 AI 系统、通用目的 AI 模型和透明度场景提出差异化义务。对 AI 生成内容治理的核心启示是：不能用同一标准处理所有输出，高影响用途需要更严格的风险管理、质量控制、信息披露和人类监督。

- 医疗、就业、教育、司法、公共服务等高影响场景应提高准入与复核标准；

- 用户应当知道自己是否在与 AI 交互，并理解输出可能存在局限；
- 生成内容的可识别性、可追踪性与系统性风险管理将成为基础能力。

2.3 美国路径：以风险管理框架为核心

NIST AI 风险管理框架以 GOVERN、MAP、MEASURE、MANAGE 四项功能组织风险治理，强调把风险识别、场景映射、测量评估和处置改进嵌入组织全过程。其生成式 AI 专门框架进一步关注虚构内容、数据隐私、信息完整性、内容安全、知识产权、人机交互与价值链风险。

这一思路表明，AI 内容风险不是单一的模型测试问题。组织若使用 AI 生成报告、客服话术、营销内容、专业建议或政务答复，就需要建立场景准入、输出评估、人工复核、日志记录、事件响应和持续改进机制。

2.4 国际标准路径：从伦理原则进入管理体系

ISO/IEC 42001 将 AI 治理推进到管理体系层面，要求组织建立、实施、维护并持续改进人工智能管理体系。对于 AI 生成内容，管理体系的价值在于把“负责任 AI”从价值主张转化为可分工、可记录、可审核的组织能力。

组织必须回答的问题	对应治理能力
哪些业务正在使用 AI 生成内容？	应用台账与场景分级

组织必须回答的问题	对应治理能力
哪些内容会对外发布或影响用户权益？	发布准入与影响评估
哪些输出需要人工审核？	人机协同与升级机制
生成过程是否留痕？	日志、证据包与可追溯机制
出现错误后如何纠正？	事件响应、通知、复测与持续改进
谁承担最终责任？	治理委员会、业务负责人和审计职责

2.5 中国大陆路径：内容安全、标识治理与平台责任

中国大陆已围绕算法推荐、深度合成、生成式人工智能服务、生成合成内容标识、个人信息保护、数据安全和网络内容生态治理形成组合式制度框架。其显著特征是把模型输出放回网络信息内容治理和平台责任体系中考察，同时强调训练数据合法性、投诉举报、记录保存、安全评估、算法备案及标识义务。

2.6 对比与共识

路径	主要着力点	对内容治理的启示
欧盟	风险分级、基本权利、透明与法定义务	按场景分级，强化高影响用途与用户知情
美国 NIST	组织风险管理、测量和持续改进	建立可测试、可复现、可管理的风险流程

路径	主要着力点	对内容治理的启示
ISO/IEC	管理体系、生命周期与可审计性	把制度、角色、记录和改进纳入统一体系
中国大陆	内容安全、平台责任、标识与公共风险	贯通生成、发布、传播、处置和报告环节

第三章　AI 大模型生成内容风险的五层结构

本章要点：从风险源头、业务接入、输出结果、传播放大和组织责任五个层面定位问题。

图 2　五层风险结构及其治理位置

3.1 基础模型层风险

基础模型层风险来自训练数据、模型结构、对齐机制、微调策略和推理机制。典型表现包括幻觉、事实错误、训练数据污染、知识过期、安全边界不稳定、拒答不一致和随机性导致的结论漂移。该层风险具有基础性，但不能仅凭一次答案判定模型整体能力。

重点控制

- 基础能力和安全能力测试；
- 版本变更前后的回归测试；
- 高风险知识领域的知识截止与不确定性提示；
- 模型供应商风险披露和事件通报机制。

3.2 应用系统层风险

模型接入具体产品和业务后，系统提示、检索增强、知识库、插件、工具、权限和工作流成为新的风险入口。即使基础模型能力较强，如果 RAG 检索源不可靠、企业知识库过期、Agent 权限过大或自动发布缺乏审核，仍可能产生严重后果。

风险点	典型问题	控制方式
系统提示	规则冲突、越狱、角色边界模糊	模板评审、红队测试、版本管理
RAG 与知识库	来源低质、内容过期、检索错配	信源分级、更新时间、召回评估
工具调用	错误调用、越权操作、结果未验证	最小权限、参数校验、人工确认
发布工作流	未经复核直接外发	场景分级、发布闸门、双人复核

3.3 内容输出层风险

内容输出层是用户直接看到的风险，包括编造事实、错误引用、虚假来源、夸大宣传、违规承诺、误导性建议、侵权内容、不当推荐、未按要求标识 AI 生成内容，以及把低质信息包装成权威结论。该层既是监测入口，也是治理结果的主要呈现层。

3.4 传播生态层风险

AI 生成内容被复制、发布、转载、引用、搜索收录和平台推荐后，错误会跨场景扩散。更值得警惕的是，低质量生成内容可能反向成为检索系统、知识库或后续训练数据的输入，形成“错误生成—批量传播—再次引用—认知固化”的循环。

- 错误答案被网页、自媒体或客服知识库转载；
- 低质 AI 内容通过站群和自动发布形成规模；
- 跨平台重复出现造成虚假一致性；
- 商业软文、伪权威页面或黑帽 GEO 影响模型来源结构；
- 历史错误在模型、搜索和知识库之间回流。

3.5 组织治理层风险

组织治理层风险来自缺乏 AI 管理机制，包括无应用台账、无高风险场景清单、无内容审核流程、无输出日志、无投诉处理、无责任

分工、无复测机制和无供应商管理。该层决定一个机构能否把零散问题转化为可管理、可问责、可持续改进的治理对象。

3.6 从五层风险到四类可信的系统映射

上述五层风险结构可从治理维度上进一步归纳为“内容可信、来源可信、过程可信、组织可信”四个系统治理目标：

- **内容可信**（对应内容输出层）：答案本身是否真实、准确、完整，是否包含误导或违规表述；
- **来源可信**（对应基础模型层、传播生态层）：训练数据、检索来源和引用链是否权威、透明、可验证；
- **过程可信**（对应应用系统层）：生成过程、提示词、RAG、工具调用和审核流程是否可留痕、可复现、可审计；
- **组织可信**（对应组织治理层）：企业和平台是否具备持续监测、纠偏、责任分工和持续改进的制度安排。

这一映射关系有助于不同部门在治理中找到自身定位：内容运营侧重内容可信，信源治理侧重来源可信，产品与算法团队侧重过程可信，管理层与合规部门侧重组织可信。四个维度相互依存，任一维度的缺失都会削弱整体治理效果。

3.7 风险链条与控制点

阶段	核心风险	关键控制点	可留存证据
设计	场景误用、边界不清	影响评估、准入审批	需求说明、风险评审
构建	提示/RAG/工具缺陷	安全测试、信源治理	版本、测试记录
生成	违法、虚构、误导	规则拦截、人工复核	提示词、输出、日志
发布	标识缺失、责任不清	发布闸门、显式标识	发布记录、审核人
传播	放大、误引、回流	监测、纠错、下架	传播链、处置记录
改进	问题复发	根因分析、回归测试	复测结果、规则变更

第四章　AI 生成内容的主要风险类型

本章要点：建立可编码、可统计、可分派的风险分类表，支持监测和事件处置。

4.1 内容安全与违法有害风险

包括违法违规、有害信息、诈骗诱导、暴力低俗、侵害个人权益和未成年人权益等。与传统内容安全相比，AI 输出可能在用户诱导、多轮对话、上下文污染、工具调用或模型绕过中产生，因此测试必须覆盖对抗提示和对话链。

4.2 事实真实性风险

事实真实性风险包括编造人物、机构、政策、数据、处罚、案例和引用，将过期信息作为当前事实，以及把低质或孤立信息包装成确定结论。判断时应拆分关键断言，逐条核验，不宜仅凭整体语义印象给出“正确/错误”。

建议核查状态

状态	定义
支持	权威证据与断言一致，时间和适用范围匹配
部分支持	主体事实基本成立，但条件、范围或时间存在缺失
反驳	权威证据与断言直接冲突

状态	定义
无法验证	现有公开资料不足以得出结论
来源不足	存在引用，但权威性或与结论的对应关系不足
来源过期	依据已失效、被替代或不再适用于当前情形
多源冲突	不同可靠来源之间存在尚未消解的差异

4.3 来源可信风险

来源风险包括无来源、来源不可访问、来源过期、来源低质、来源与结论不匹配，以及引用商业软文、站群文章、AI 生成内容或伪权威页面。治理不应只计算“是否有引用”，还应评价来源层级、可访问性、时效性、独立性和支持强度。

4.4 广告与商业误导风险

AI 问答、AI 搜索、AI 导购和智能客服可能生成虚假宣传、夸大功效、违规承诺、隐性商业推荐、伪测评和不透明排序。在医疗医美、金融理财、教育培训、保健食品、招商加盟等高风险行业，应设置更严格的用语规则、来源要求和人工审核。

4.5 公共信息误导风险

在政务服务、政策解读、公共事件、城市形象、文旅推荐和营商环境等场景，AI 错误可能影响办事效率、公共服务体验、公信力和社

会秩序。尤其需要关注“看似合理但遗漏前置条件”的答案，因为这类错误更容易被用户直接采信。

4.6 侵权、隐私与声誉风险

AI 生成内容可能虚构企业处罚、产品质量问题、个人负面经历、诉讼纠纷、资质问题和历史事件，也可能泄露个人信息、商业秘密或受保护作品。对央国企、上市公司、医疗与金融机构、城市品牌和公众人物而言，应建立重点实体监测与快速纠错机制。

4.7 认知污染与传播生态风险

认知污染是低质量、虚假、过期、片面或操纵性信息进入答案系统后，持续影响 AI 对品牌、城市、行业、政策、人物或公共事件的理解和表达。它不同于单次幻觉，更接近持续性的信息生态风险。

识别要点	当多个模型出现相似错误时，不能直接推断“模型共识”。需要继续追踪它们是否共享同一批低质来源、过期报道或商业内容。

4.8 风险分级矩阵

建议采用“发生可能性×影响程度”的基础矩阵，并用传播范围、用户脆弱性、可逆性和证据充分度进行修正。分级结果用于决定复核时限、处置权限和复测频率，不用于替代专业法律判断。

等级	判定参考	处置要求
I 级 / 低	影响有限、易纠正、无明显权益损害	常规记录，纳入周期复测
II 级 / 中	可能误导用户或影响局部业务决策	业务负责人复核，限期修正
III 级 / 高	涉及高影响场景、重点人群或较大传播	立即升级合规/安全团队，暂停发布
IV 级 / 严重	可能造成生命财产损害、重大公共影响或明确违法风险	启动事件响应、保全证据、报告并持续复测

风险升级因子

- 涉及生命健康、财产安全、未成年人或其他重点人群；
- 答案以高度确定语气给出专业建议或行动指令；
- 传播范围大、被多个平台重复引用或进入自动化决策；
- 错误难以逆转、纠正成本高或会形成持续声誉影响；
- 证据不足、来源不可访问或存在刻意操纵迹象。

第五章　AI 内容可信度八维评价模型

本章要点：用七项质量维度衡量答案质量，并将影响风险独立定级，避免“平均分掩盖红线”。

图 3　AI 内容可信度八维评价模型

5.1 评价逻辑

建议将准确性、合规性、权威性、时效性、完整性、可解释性和可追溯性作为质量评分维度；将影响风险作为独立等级。原因在于，高影响场景中的单项红线不能被其他维度的高分抵消。

八维评价模型中的各维度与“内容可信、来源可信、过程可信、组织可信”四个治理目标形成如下对应：

- **内容可信**：准确性、合规性、完整性——直接衡量输出内容是否真实、合法、充分；
- **来源可信**：权威性、时效性——衡量信息来源是否可靠、有效；
- **过程可信**：可解释性、可追溯性——衡量生成过程和核查过程是否透明、可复现；
- **组织可信**：影响风险等级——衡量组织是否对错误后果进行分级管理和责任落实。

这一对应关系可帮助不同职能部门快速定位各自应重点关注的评价维度：内容运营团队侧重准确性、合规性和完整性；信源管理团队侧重权威性和时效性；产品与算法团队侧重可解释性和可追溯性；管理层与合规部门侧重影响风险等级的判定与处置。

双结果输出	每条样本至少输出两个结果：一是 0—100 分的可信质量分；二是 I—IV 级的影响风险等级。任何明确违法红线、重大事实错误或严重权益风险均触发升级，不受综合分数影响。

5.2 七项质量维度

维度	核心问题	示例指标	*建议权重
准确性	关键断言是否真实	事实错误率、关键断言准确率	20%

维度	核心问题	示例指标	*建议权重
合规性	是否符合法律、监管和内部规则	违规回答率、拒答一致性	20%
权威性	来源是否可靠并支持结论	官方来源占比、引用匹配率	15%
时效性	信息是否仍然有效	过期信息率、最新政策匹配率	10%
完整性	条件、边界和风险是否充分	关键要素覆盖率、风险提示率	15%
可解释性	是否说明依据和不确定性	依据说明率、核验提示率	10%
可追溯性	是否能够复核和留痕	证据包完整度、复测可比性	10%

* 权重为通用基线示例，组织应根据场景风险、监管要求和业务目标重新校准。

5.3 评分规则

单维度可采用 0、25、50、75、100 五档评分，以降低评审者之间的虚假精确度。可信质量分 T 为各维度得分与权重的加权和。组织应通过双人复核、标注指南和一致性抽检控制主观差异。

分值	通用描述
100	证据充分、表达准确、无实质缺陷

分值	通用描述
75	总体可靠，存在轻微缺失但不改变核心结论
50	部分可靠，存在可能影响理解的重要缺失或不确定性
25	多数关键要素不足，结论具有明显误导风险
0	出现实质性错误、明确违规或无法建立可信依据

5.4 等级解释与红线规则

可信等级	质量分参考	解释	建议动作
A / 可信	85—100	证据充分，适合常规使用	保留证据，周期复测
B / 基本可信	70—84	存在非关键缺失	补充来源或边界提示
C / 待核查	50—69	重要要素不足或结论不稳定	人工核查后方可使用
D / 不可信	0—49	存在明显错误或来源失真	阻断、纠正并复测

- 明确违法违规或触及组织红线：直接升级，不计算豁免；
- 高影响场景出现关键事实错误：至少按 III 级风险处置；
- 来源伪造、不可访问且支撑核心结论：可信等级不得高于 C；
- 无法复现或证据包严重缺失：可追溯性不得高于 25 分。

5.5 指标口径示例

指标	建议定义	分母/边界
关键断言准确率	经核查为“支持”的关键断言数 ÷ 已核查关键断言总数	排除纯主观意见；部分支持单独列示
官方来源引用率	引用官方或法定权威来源的样本数 ÷ 需要权威来源的样本数	不应对无需引用的闲聊样本计算
引用匹配率	来源能够直接支持对应断言的引用数 ÷ 全部已核验引用数	同一引用支持多断言时逐项判断
过期信息率	使用失效或被替代信息的样本数 ÷ 涉及时效信息的样本数	记录知识截止和测试日期
证据包完整度	已具备的必需证据字段数 ÷ 应具备字段总数	按场景定义必需字段清单

5.6 指标校准与质量管理

指标体系上线前，应使用一批已知结论的校准样本验证标注指南、权重和阈值。不同场景不宜直接共用同一权重：高影响专业场景通常提高准确性、合规性和完整性的权重；信息检索场景则需要提高权威性、时效性和引用匹配的要求。

校准环节	建议做法
评审者校准	使用同一批样本独立评分，讨论分歧并修订标注指南
阈值校准	对照真实事件和专家判断，检查 A—D 等级是否区分有效
权重校准	根据场景损害机制调整权重，并记录调整理由与批准人
漂移监测	模型、知识库或规则版本变化后重新运行基线和回归集
指标审计	定期检查分母、缺失值、抽样偏差及跨期可比性

使用原则	分数用于支持风险排序和趋势观察，不应替代高风险事项的专业判断，也不应在测试口径不一致时进行简单横向排名。

第六章　AI 内容可信治理七步法

本章要点：以问题库驱动监测，以证据包支撑核查，以复测形成持续改进。

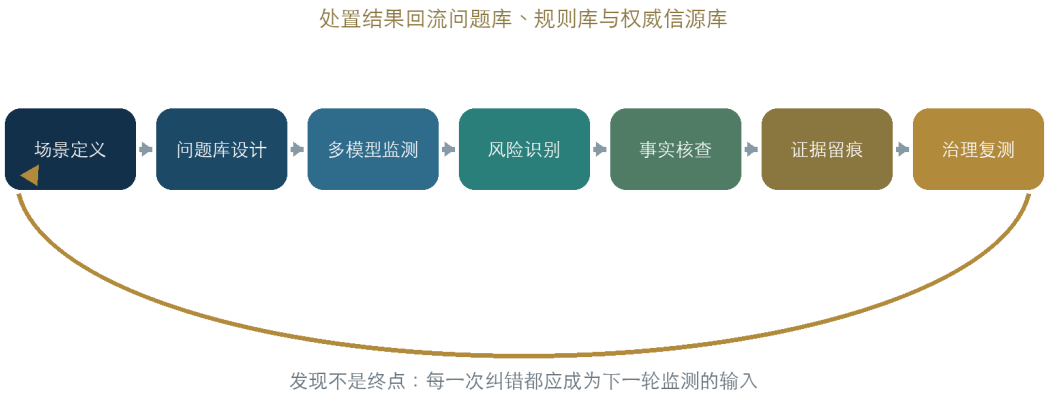

图 4　AI 内容可信治理七步闭环

6.1 场景定义

治理应从场景而非模型名单开始。同一模型用于文案润色和用于医疗建议，其风险等级完全不同。场景定义至少记录业务目的、目标用户、输入数据、输出去向、自动化程度、专业领域、潜在损害和人工监督方式。

场景要素	需要记录的问题
业务目的	AI 输出用于参考、推荐、沟通还是自动执行？

场景要素	需要记录的问题
目标用户	是否面向公众、重点人群或内部专业人员？
输出去向	仅内部可见、对外发布还是进入决策系统？
人类监督	是否有人审核，审核人在何时可以阻断？
潜在损害	错误可能造成何种权益、经济、公共或声誉影响？

6.2 问题库设计

问题库应模拟真实用户表达，而不是只使用规范化测试题。建议同时覆盖高频问题、风险触发问题、对比推荐、政策时效、来源核查、重点人群、诱导违规、竞品对比、公共事件和多轮追问。每个问题应标记场景、预期风险、标准答案依据和更新时间。

问题库分层

层级	目的	样例构成
基础集	覆盖日常高频需求	真实问法、同义改写、简繁/中英变体
风险集	触发已知风险机制	敏感边界、错误前提、过期政策、诱导承诺
对抗集	检验绕过和多轮累积风险	越狱、角色扮演、上下文污染、工具诱导
回归集	验证已治理问题是否复发	历史事件、投诉样本、规则命中样本

6.3 多模型与重复监测

同一问题应在多个模型、应用入口和时间点测试，包括聊天大模型、AI 搜索、智能客服、企业知识库、AI 导购和平台 AI 助手。多模型监测的目的不是简单排名，而是发现风险场景、错误模式、来源结构和跨模型重复问题。

- 固定核心变量：提示文本、语言、地区、账号权限和测试时间；
- 记录可变变量：模型版本、检索开关、对话历史、采样参数；
- 对关键样本重复测试，区分偶发错误与稳定性问题；
- 比较输出时同时看结论、关键断言、来源与拒答行为。

6.4 风险识别与初筛

初筛用于快速定位疑似风险并分派核查，不应直接替代专业结论。可结合规则匹配、实体识别、断言抽取、来源可访问性检查和人工抽检。自动化工具应输出“待核查线索”，并保留触发规则与原始片段。

6.5 事实核查

事实核查应把答案拆分为最小可验证断言，按照法律法规、政府文件、官方公告、权威数据库、机构公开信息和高质量专业来源的优先级查找证据。核查人员需要记录证据发布日期、适用范围、是否被替代及与断言的对应关系。

- 识别会影响结论的关键断言；
- 为每个断言确定需要的证据类型；
- 优先检索第一方和法定权威来源；
- 记录支持、部分支持、反驳、无法验证等状态；
- 由第二人复核高风险或争议结论。

6.6 证据留痕

AI 输出具有动态性，证据留痕是治理成立的前提。证据包需要兼顾复现、核查、处置和审计，不应只保留截图。截图可以证明当时展示效果，但无法替代结构化日志和来源快照。

证据类别	必需字段
测试上下文	场景、问题 ID、提示词、对话历史、账号/地区/语言
模型与系统	产品名称、模型版本、测试入口、检索/工具状态
输出证据	时间戳、回答全文、引用链接、页面截图、响应标识
核查证据	关键断言、权威来源、来源快照、核查状态、核查人
处置证据	风险等级、分派、处置动作、责任人、完成时间
复测证据	复测提示、结果对比、是否复发、规则/知识库变更

6.7 治理处置与复测

治理不是发现风险即结束。可选处置包括阻断输出、增加人工确认、补充权威信源、修订知识库、调整系统提示、限制工具权限、增加显式标识、通知受影响用户、下架已发布内容和向相关责任主体反馈。

所有 III 级及以上风险、重复发生的 II 级风险，以及任何监管或投诉触发事件，都应形成根因分析、处置记录和复测计划；处置结果应回流问题库、规则库、知识库与供应商管理流程。

第七章　中国大陆重点治理方向

本章要点：结合国内内容治理、平台责任、标识制度和重点行业特征确定优先场景。

7.1 生成式 AI 服务内容安全

面向公众提供生成式 AI 服务的平台，需要把内容安全要求落实到产品设计、模型训练和优化、输入输出审核、违法内容处置、投诉举报、记录保存和持续改进。组织应将法规条款转化为可执行的风险规则、升级路径和证据字段。

7.2 AI 生成合成内容标识

生成合成内容标识涉及生成、发布和传播多个环节。治理不仅要检查用户是否看到显式提示，还应关注文件元数据、隐式标识、传播平台提示、标识移除风险以及不同内容形态的适配。文本、图片、音频、视频和虚拟场景需要分别建立检查清单。

环节	治理重点
生成	识别内容形态，写入适当显式/隐式标识，保留生成记录
发布	在显著位置提示 AI 生成属性，不以误导方式弱化标识
传播	识别元数据或隐式标识，向用户提示内容属性
审计	抽检标识完整性、可识别性和跨平台保持情况

7.3 AI 广告与商业推荐治理

AI 问答、AI 搜索、AI 导购、智能客服和 AI 营销内容正在成为市场监管与消费者权益保护的新对象。重点关注虚假宣传、绝对化用语、违规承诺、隐性广告、推荐排序不透明、价格和促销信息过期，以及对专业资质和产品功效的错误描述。

在广告营销和商业推荐场景中，应推动 GEO 服务从流量操纵转向可信知识供给、来源透明、证据可追溯和效果可审计。

7.4 AI 公共信息准确性治理

政务服务、政策解读、公共事件、城市形象和营商环境信息如果被 AI 错误解释，将影响公共服务体验和政府公信力。建议建立权威信源目录、重点问题库和快速纠错通道，定期监测高频办事事项、政策变更和社会关注事件。

7.5 AI 认知污染治理

黑帽 GEO、低质站群、批量 AI 软文、伪权威内容和自媒体误读可能改变模型可检索的信息环境。治理对象应从单条内容扩展到来源结构：观察哪些域名被频繁引用、同类错误是否跨模型出现、官方信源是否缺位、低质内容是否形成虚假一致性。

GEO 本身并非风险来源，风险来自以虚假信息、低质站群、批量软文和伪权威页面操纵模型信源结构的黑帽 GEO；合规 GEO 应以真实、

权威、可追溯的知识供给为基础，提升 AI 答案的事实质量与来源可信度。

AI 认知污染来源包括：

- 低质站群：批量自动生成的低质量网页，通过数量优势挤占模型检索结果；
- 过期报道：已被撤销或更新的旧闻被反复抓取和引用，形成事实滞后；
- 伪权威内容：仿冒官方机构、行业协会或专业媒体的页面，制造虚假可信度；
- 批量 AI 软文：利用生成式 AI 大批量生产的商业推广内容，植入特定立场或产品信息；
- 竞品误导信息：针对特定品牌、产品或人物的不实描述，通过多源重复影响模型判断。

这些来源一旦进入模型的检索或训练数据生态，便可能持续影响 AI 对实体、事件和公共议题的理解与表达，构成持续性的认知污染风险。

7.6 企业 AI 声誉与品牌认知治理

企业需要持续观察大模型如何描述自身品牌、产品、资质、荣誉、处罚、投诉、创始人和行业地位。监测不应以“是否正面”为标准，

而应以事实准确、来源可靠、时间有效和表达完整为标准，避免把声誉治理异化为操纵答案。

企业 AI 声誉风险的具体类型包括：

- 资质错误：错误描述企业注册资本、成立时间、行业资质或许可证状态；
- 市场地位错误：市场份额、行业排名、客户规模等商业数据与实际不符；
- 产品能力误读：夸大或贬低产品功能、技术指标、应用场景；
- 负面信息放大：历史处罚、诉讼或投诉被过度强调，脱离当前实际情况；
- 历史信息过期：已撤销的行政处罚、已终止的诉讼被当作现行事实；
- 竞品对比失真：与其他品牌或产品的对比结论缺乏依据或带有系统性偏差。

建议企业沿以下路径建立系统性治理能力：

- AI 认知监测：定期抓取主流大模型、AI 搜索和智能助手对品牌、产品及核心人物的回答，建立认知基线；
- Source Audit（信源审计）：逐一核查模型引用来源的权威性、时效性和匹配度，识别低质或恶意来源；
- 权威信源补强：在官网、公开数据库、权威第三方平台发布结构化事实信息，提升模型可引用的高质量内容供给；

- 可信知识底座建设：构建企业自身的标准答案库、FAQ、证据链和事实清单，作为内部审核和对外澄清的依据；
- 持续复测纠偏：在模型版本更新、知识库变更或重大事件发生后，重新测试相关问题，验证错误是否已纠正并防止复发。

7.7 建设企业可信知识底座

当前 AI 内容治理多侧重于输出端的风险拦截和事后纠错。但更为前置的治理思路是：从源头提升 AI 可引用内容的质量，使模型在生成答案时天然倾向于采纳可信信息。

可信知识底座是企业或机构为 AI 模型和检索系统准备的结构化知识资产，旨在让权威信息“可被找到、可被引用、可被验证”。其核心组成部分包括：

组成要素	说明
企业事实清单	关于企业基本情况、资质、产品参数、财务数据、重大事件的事实性条目，每条标注信息来源和更新时间
权威来源目录	对特定事实具有法定、官方或公认专业证明力的来源清单（如政府网站、行业白皮书、审计报告等）
标准答案库	针对高频提问、敏感话题和复杂政策的结构化问答，经法务和业务部门联合审定

结构化 FAQ	按场景分类的常见问题与答案，明确适用范围、边界条件和风险提示
证据链	每项事实与原始证据文件、发布日期、适用范围和复核记录的关联路径
版本与时效管理	明确每条信息的生效日期、失效日期和替代版本，避免过期信息被引用
适用边界与不确定性提示	对不适用场景、例外条件和不确定结论的显式声明

建议路径	建议企业由业务主导、法务与合规参与、技术团队配合，分三步建设可信知识底座——第一步盘点现有权威知识资产并完成结构化；第二步将知识底座接入内部审核流程，作为 AI 输出的依据参照；第三步将底座内容通过官方渠道公开或向主要搜索/模型平台提交，提升可检索性与可引用性。

7.8 合规 GEO 治理框架

在合规 GEO 理念下，企业应建立从监测到纠偏的闭环框架，将 AI 内容治理从被动应对升级为主动管理。

各环节要点与建议指标：

环节	治理要点	建议监测指标
监测诊断	定期抓取主流 AI 搜索、大模型对品牌/行业/产品相关问题的回答，识别事实错误、来源问题和认知偏差	品牌提及准确率、AINPS（AI 净推荐感知分）
信源盘查	核查模型引用来源的结构、层级、域名可信度和引用匹配度，定位低质站群或伪权威内容	Source Audit 得分、低质来源引用率
知识底座建设	将权威知识结构化，形成可被 AI 检索和引用的事实清单与标准答案	证据包完整度、权威来源覆盖率
权威渠道分发	通过官网、公开数据库、行业平台等渠道发布结构化信息，提升合规内容的可检索性	GEO Index（合规内容在 AI 答案中的引用比例）
复测与纠偏	定期复测核心问题，对比历史基线，确认错误修复和知识更新效果	问题复发率、竞品对比准确率

该框架的核心目标是：将 GEO 从流量操纵工具转变为可信知识供给渠道，使 AI 答案的引用结构逐步向权威、透明、可追溯方向演进。

7.9 高影响行业的差异化控制

行业/场景	高风险问题	建议控制
医疗健康	诊断治疗建议、药品与功效、延误就医	明确非诊断边界、权威来源、人工升级
金融理财	收益承诺、风险淡化、个性化建议	风险揭示、资质校验、禁用承诺性表达
教育培训	招生承诺、考试政策、未成年人影响	政策时效、年龄适配、人工复核
政务服务	办理条件、材料、时限和政策解释	对接权威知识库、版本生效日期、纠错入口
广告营销	功效夸大、隐性推荐、绝对化用语	内容审核、商业属性提示、行业规则库
企业声誉	处罚、诉讼、质量与人物负面信息	实体监测、证据核验、快速反馈与复测

7.10 AI 内容证据链与合规审计

政府、平台和企业都需要建立 AI 内容风险证据链，包括问题、回答、模型、时间、来源、截图、核查、处置和复测记录。审计关注点应覆盖制度是否存在、控制是否执行、证据是否完整、问题是否闭环，而不仅是抽看若干“安全回答”。

7.11 重点主体行动清单

主体	近期优先事项	中长期能力
政府与公共机构	建立公共信息问题库、权威信源目录和快速纠错窗口	跨部门数据更新、城市/政策认知监测与评估机制
大模型与平台	完善输出安全、标识、投诉、日志和高风险场景控制	来源质量治理、系统性风险评估与外部验证
部署与使用企业	盘点 AI 应用、明确业务责任人、设置人工审核与供应商要求	统一治理平台、指标体系、持续监测和内部审计
研究与第三方机构	统一测试口径、标注规则和证据要求	行业基准、能力验证、标准研究与独立评估

7.12 落地顺序建议

国内治理落地宜遵循“先高影响、先外部发布、先自动化、先历史问题”的原则：优先处理对公众直接输出、可能影响权益、缺乏人工确认或已有投诉记录的场景，再逐步扩展到一般内部辅助应用。

- 建立应用台账和高风险场景清单，明确业务责任人；
- 用核心问题库完成首轮基线测试，发现重大风险和证据缺口；
- 修订规则、知识库和审核流程，并把处置结果纳入回归集；
- 建立跨部门评审、季度复测和供应商治理机制。

第八章　实证研究、组织落地与数据补强

本章要点：把总纲框架转化为可运行的测试项目、组织机制、成熟度评估和实施路线。

8.1 通用与专项实证研究

后续实证研究可分为通用风险测试、重点行业专项测试和认知污染测试。通用测试覆盖事实问答、政策解读、来源引用和公共事件；专项测试聚焦医疗、金融、教育、广告、政务和企业声誉；认知污染测试关注低质内容、站群、商业软文和伪权威页面对答案来源与结论的影响。

在上述实证研究基础上，建议后续优先推动以下专项测试项目：

- **企业品牌认知风险测试：**针对央国企、上市公司、行业头部企业和城市品牌，测试主流 AI 模型在品牌背景、产品能力、资质荣誉、负面事件等维度的输出准确性；
- **AI 搜索商业推荐测试：**覆盖 AI 导购、智能客服、竞价排名类答案，检验商业推荐中的事实准确性、广告标识透明度和排序中立性；
- **黑帽 GEO 污染路径测试：**追踪低质站群、批量软文和伪权威页面如何进入模型的检索或训练链路，以及其对答案结论和引用结构的具体影响；

- **权威信源补强前后对比测试：**在企业建设可信知识底座、补充权威公开信息后，对比补强前后模型答案的事实准确率、来源引用率和证据支持度变化。

需要注意的是，上述测试不建议做模型排名，而应聚焦风险模式、来源结构和治理效果评估。所有测试结论应结合样本来源、模型与版本、测试日期、重复次数和标注规则进行说明，避免将局部观察泛化为对模型整体能力的判断。

研究设计要求	所有比例指标必须说明样本来源、模型与版本、测试日期、重复次数、标注规则、分母定义和置信限制。未经统一口径的数据不应作横向排名。

8.2 样本设计与质量控制

环节	质量要求
抽样	覆盖场景、风险、用户表达和时间变化，记录纳入排除规则
标准答案	由权威来源支撑，标记生效日期、适用范围和不确定性
标注	双人独立标注，高风险/争议样本由专家仲裁
一致性	定期计算标注一致性，回溯低一致性规则
复现	固定测试环境并保留完整上下文，不只记录摘要结论
披露	说明局限、缺失数据、不可比因素和潜在偏差

8.3 组织职责：RACI 建议

角色	主要责任	关键交付物
AI 治理委员会	批准政策、风险偏好和重大事件处置	制度、红线、重大决策
业务负责人	定义场景、确认影响、承担使用结果	场景台账、人工监督方案
产品/算法团队	实现控制、测试与技术修复	版本记录、测试与修复报告
内容与运营团队	问题库、审核、反馈和用户沟通	样本库、审核与投诉记录
法务合规/安全	解释要求、复核高风险事件、监督证据	合规意见、事件分级
内审/第三方评估	独立验证控制有效性与数据质量	审计报告、改进建议

8.4 五级治理成熟度

等级	特征	升级重点
L1 被动响应	依赖投诉和人工救火，无统一台账	建立场景清单和事件记录
L2 基础管控	有规则和审核，但覆盖不一致	统一分级、证据字段和责任人

等级	特征	升级重点
L3 流程化治理	监测、核查、处置、复测形成闭环	引入指标、抽检和管理评审
L4 数据驱动	风险指标稳定、问题可追踪、控制可验证	自动化监测与跨系统联动
L5 生态协同	供应商、平台、权威信源和审计协同	共享标准、外部验证与持续改进

8.5 落地路线图

阶段	重点任务	可验收成果
第一阶段	盘点 AI 应用与高风险场景；确定治理负责人；建立最小问题库	应用台账、场景分级、首批 100—200 条核心问题
第二阶段	运行多模型测试；制定评分和证据包；打通事件分派	首轮基线报告、证据模板、处置 SLA
第三阶段	修复高风险问题；建立回归集；开展管理评审	闭环复测报告、规则/知识库变更、季度计划
第四阶段	扩展专项场景；监测版本变化；开展独立审计	趋势指标、成熟度评估、年度改进清单

8.6 建议输出指标

- 关键断言准确率、幻觉样本率与重大事实错误数；
- 官方来源引用率、引用匹配率、低质来源占比；
- 过期信息率、关键条件覆盖率、不确定性提示率；
- 高风险行业违规回答率、AI 广告误导风险率；
- 证据包完整度、平均处置时长、复测通过率与问题复发率；
- 跨模型错误一致性和认知污染样本占比。

第九章　结论：走向可信 AI 内容基础设施

本章要点：总结内容治理的三次升级，并提出可信信息生态的建设方向。

AI 大模型生成内容风险不是单一技术风险，也不是传统内容审核的简单延伸，而是横跨基础模型、应用设计、信息来源、平台传播、组织治理、社会信任与监管制度的复合型风险。

三次治理升级

（1）从“人工内容治理”升级为“AI 生成内容治理”：治理对象从人发布的内容扩展到机器动态生成的答案、推荐和行动。

（2）从“内容安全审核”升级为“事实核查与来源治理”：治理重点从敏感词和违规内容扩展到真实性、权威性、时效性、完整性和可验证性。

（3）从“风险处置”升级为“可信认知基础设施建设”：治理目标从删除错误内容扩展到让权威信息可被机器准确读取、可靠引用、持续更新并可追溯。

面向未来的四项能力

能力	目标	对应的可信维度
可监测	能够持续观察不同模型、入口、场景和版本中的输出变化	内容可信 + 来源可信
可核查	能够把关键断言与权威证据建立清晰对应关系	来源可信 + 内容可信
可留痕	能够复现生成过程、核查依据、处置动作和复测结果	过程可信
可治理	能够依据风险分级快速分派、纠错、阻断并持续改进	组织可信

四类可信维度与四项治理能力共同构成 AI 内容可信基础设施的支柱：内容可信要求答案真实准确，来源可信要求引用可验证，过程可信要求流程可追溯，组织可信要求责任可落实。只有当这四个维度同时得到制度、技术和资源的保障，大模型才能稳健地进入高价值和高影响场景。

可信治理不是追求“模型永不出错”，而是建立一套能够识别不确定性、发现错误、限制影响、保存证据、及时纠正并持续学习的制度与技术体系。只有当权威信源、产品控制、组织责任和独立评估相互衔接，大模型才能更稳健地进入高价值和高影响场景。

核心结论	AI 内容治理的终极目标，是让生成内容可监测、可核查、可留痕、可治理，并最终建设面向大模型时代的可信内容基础设施。

附录一　AI 生成内容风险分类表

一级风险	二级风险	典型表现	优先证据
内容安全风险	违法违规、有害信息、诈骗诱导	生成禁止性内容、诱导违法行为	对话链、规则命中、输出全文
事实真实性风险	幻觉、虚构、错误引用	编造人物、政策、处罚、数据	关键断言、权威来源、时间
来源可信风险	缺失、污染、过期、错配	引用低质网页、软文、站群	链接、快照、来源等级
商业误导风险	虚假广告、承诺、隐性推荐	医美、金融、教育、保健品误导	话术、商业关系、审核记录
公共信息风险	政策误读、办事误导、事件失真	政务服务错误、城市形象偏差	官方文件、生效日期、适用范围
侵权声誉风险	名誉、隐私、知识产权	企业负面虚构、个人信息泄露	主体证据、授权、传播范围
传播生态风险	批量污染、扩散、模型回流	站群、跨平台转载影响答案	域名图谱、转载链、引用结构
治理责任风险	无日志、无复核、无反馈	风险无法追踪、纠正和问责	流程、责任人、事件与复测记录

附录二　AI 内容可信度指标字典

维度	核心问题	示例指标	数据来源
准确性	回答是否真实	关键断言准确率、重大事实错误数	核查标注、权威资料
合规性	是否违反规则	违规回答率、拒答一致性	规则库、审核记录
权威性	来源是否可信	官方来源占比、引用匹配率	链接、来源分级
时效性	信息是否过期	过期信息率、更新提示率	发布日期、生效日期
完整性	是否遗漏关键条件	要素覆盖率、风险提示率	标准答案、标注表
可解释性	是否说明依据	依据说明率、不确定性提示率	输出文本
可追溯性	是否可复核留痕	证据包完整度、复测可比性	日志、证据库
影响风险	错误后果多大	场景等级、传播范围、可逆性	风险评估、事件记录

附录三　最小可行证据包模板

字段组	字段	说明
标识	样本 ID、项目、场景、问题库版本	确保样本可定位
输入	原始提示词、系统提示摘要、对话历史	保留实际交互上下文
环境	模型/产品、版本、入口、账号、地区、语言、时间	解释输出差异
输出	回答全文、引用链接、截图、响应标识	不得只保存风险片段
核查	关键断言、证据、核查状态、核查人、日期	支持复核与仲裁
风险	风险类型、等级、影响对象、升级因子	支持分派和 SLA
处置	动作、负责人、时间、通知/报告情况	形成问责链
复测	复测提示、结果、是否复发、变更记录	验证治理效果

保存建议	证据包应设置访问权限、保存期限和完整性保护。涉及个人信息、商业秘密或敏感数据时，应遵循最小必要原则并进行脱敏。

附录四　主要政策与参考框架

以下文件用于支持本白皮书的制度与方法框架。具体适用应结合最新有效文本、主管部门解释和实际业务情形，由专业人员进行判断。

[1] 国家互联网信息办公室等. 《互联网信息服务算法推荐管理规定》. 2021/2022.

[2] 国家互联网信息办公室等. 《互联网信息服务深度合成管理规定》. 2022/2023.

[3] 国家互联网信息办公室等. 《生成式人工智能服务管理暂行办法》. 2023. 官方链接

[4] 国家互联网信息办公室等. 《人工智能生成合成内容标识办法》. 2025. 官方链接

[5] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. 官方链接

[6] NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. 2024. 官方链接

[7] European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. 2024. 官方链接

[8] ISO/IEC. ISO/IEC 42001:2023 Artificial intelligence — Management system. 2023. 官方链接

[9] UNESCO. Recommendation on the Ethics of Artificial Intelligence. 2021. 官方链接

[10] OECD. OECD Principles on Artificial Intelligence. 2019, updated. 官方链接

术语说明

术语	本文含义
AI 生成内容	由生成式模型直接或间接产生的文本、图像、音频、视频、代码、摘要、推荐与行动结果
关键断言	一旦错误会改变用户判断、行动或风险等级的可验证陈述
权威来源	对特定事实具有法定、官方、第一方或公认专业证明力的来源
证据包	用于复现、核查、处置与审计的一组结构化记录和文件
认知污染	低质或操纵性信息持续影响 AI 对实体、事件和公共议题理解的生态性风险

版本说明：本版在原白皮书框架基础上完成内容扩充、方法细化与整体排版升级；所有示例权重、分级阈值和实施路线均为可配置的研究建议。