ICS 35.240
CCS L77

T/CAPT

团　　体　　标　　准

T/CAPT 026—2026

生成式引擎优化（GEO）可信信息传播与信息生态治理规范

Standard for Trustworthy Information Dissemination and Information Ecosystem Governance in Generative Engine Optimization (GEO)

2026－08－11 发布　　2026－08－12 实施

中国新闻技术工作者联合会　　发 布

目　　次

前　　言 III
引　　言 V
1 范围 6
2 规范性引用文件 6
3 术语和定义 6
4 基本原则 7
4.1 合法合规 7
4.2 诚信真实 7
4.3 以人为本 7
4.4 公平有序 8
4.5 安全可追溯 8
4.6 主体协同 8
5 服务受理要求 8
5.1 基础资质核验 8
5.2 营销主张粗筛 8
5.3 不予受理情形 8
5.4 受理记录 8
5.5 第三方协作 8
6 内容审核要求 9
6.1 内容合规审核 9
6.2 核心事实主张证据核验 9
6.3 来源可信度分级 9
6.4 知识产权与 AI 标识 10
6.5 科技伦理审查衔接 11
7 GEO 优化实施要求 11
7.1 三区分治 11
7.2 语料接入与发布 12
7.3 防注入与元数据控制 12
7.4 多模态内容安全 12
8 监测、追溯与风险处置要求 12
8.1 全链路追溯 12
8.2 信息生态监测 13
8.3 效果评估 13
8.4 异常熔断与纠偏 13
8.5 应急演练与红队测试 14
9 高风险场景要求 14
9.1 高影响传播场景 14

9.2 高风险行业场景 14
9.3 特殊保护群体场景 14
9.4 跨境与境外信源场景 14
10 服务提供方组织能力要求 14
10.1 基础能力 14
10.2 能力分级 15
10.3 L1 基础能力 15
10.4 L2 专业能力 15
10.5 L3 综合治理能力 15
10.6 能力说明 15
11 附则 16
11.1 实施过渡 16
11.2 法规和平台规则衔接 16
11.3 投诉与行业自律 16
11.4 修订机制 16
11.5 资料性附录使用 16
附录 A（规范性） 全链路追溯数据要素 17
附录 B（资料性） 合规审核核验清单 18
附录 C（资料性） 高风险场景与禁止行为速查 19
附录 D（资料性） 服务能力核验基准 20
附录 E（资料性） 效果评估指标 21
参考文献 22

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。

本文件由中国新闻技术工作者联合会新闻信息标准化分会归口。由新华网融媒体未来研究院、媒体融合生产技术与系统国家重点实验室、新华社技术局、唯识（北京）网络科技有限公司（新经济学家智库）、复旦大学马克思主义研究院、北京邮电大学网络安全学院、北京清蓝智汇科技有限公司、杭州盖立克思人工智能有限公司、悠易互通（北京）广告有限公司、元聚变（上海）科技股份有限公司、上海树和智能科技有限公司、浙江省医学影像人工智能重点实验室等单位提出。

本文件起草单位：新华网股份有限公司、媒体融合生产技术与系统国家重点实验室、唯识（北京）网络科技有限公司（新经济学家智库）、新华社中国广告联合有限责任公司、新华社技术局、中国科技新闻学会融媒体科技传播专业委员会、中国科技信息杂志社、中国信息协会数据要素专委会、复旦大学马克思主义研究院、北京邮电大学网络安全学院、上海交通大学智能传播研究院、中国传媒大学数据智能与网络安全研究院、香港中文大学（深圳）智能决策实验室、河北传媒学院、宁波诺丁汉大学、中国大百科全书出版社有限公司、北京北大方正电子有限公司、上海人工智能研究院数字化治理中心、北京清蓝智汇科技有限公司、北京伯乐互联科技发展有限公司（点亮AI）、江苏天闳低空数字技术研究院、杭州盖立克思人工智能有限公司、悠易互通（北京）广告有限公司、上海树和智能科技有限公司、元聚变（上海）科技股份有限公司、广东南方智媒科技有限公司、广州日报报业集团（广州日报社）、江苏新华日报大数据有限公司、长江日报社（长报传媒集团）、四川日报报业集团、四川封面传媒科技有限责任公司、纵览（河北）传媒有限公司、苏州市广播电视总台、新疆博尔塔拉融媒体中心、无锡日报报业集团、广西日报社、黑龙江日报报业集团、大众报业集团（大众日报社）、陕西云造智联信息科技有限公司。

本文件主要起草人：李凌、王鹏铭、张芮宁、古斯敏、杨溟、路海燕、徐才、蔡津津、李朝卓、于富堂、孙晶、潘菲、黄佩、陈昌凤、李本乾、林常乐、栾轶玫、胡延平、李安、杨曦沦、王小毅、李伦、孙峻峰、张志刚、翟雅楠、鞠靖、杨昊、王真峥、曹素妨、丁晓斌、周奇、熊秀鑫、刘鹏飞、纪艳燕、位艳玲、许汀汀、单文盛、陈孺新、张毓隆、黄晓丽、纪艳燕、沈丽红、李彤欣、百华睿、彭嘉昊、鲁扬、程秋生、蔡芳、杨林、罗小龙、龙文飞、孙超、邓剑峰、孙唐虎、赵杰、任科、杨道泉、张小逸、罗琼、朱橙、汤代禄、杨华、赵乐。

引　言

生成式引擎优化（Generative Engine Optimization，GEO）服务正在影响AI问答、智能搜索、智能体知识库和检索增强生成场景中的信息呈现方式。为防范虚假信息、语料投毒、不正当竞争、AI生成合成内容标识不合规、数据安全和信息生态失衡等风险，本文件规定GEO服务在项目受理、内容审核、优化实施、监测追溯、风险处置、高风险场景和组织能力方面的基本要求，推动GEO服务真实、合规、透明、可追溯地开展。

生成式引擎优化（GEO）可信信息传播与信息生态治理规范

1 范围

本文件规定了GEO服务提供方在项目受理、内容审核、语料接入、优化实施、传播控制、监测追溯、风险处置、高风险场景和组织能力方面的要求。

本文件适用于GEO服务提供方、GEO服务使用方、第三方评估机构、行业组织、大模型平台方、AI搜索引擎、内容平台、独立第三方检索源和智能体知识库运营方，可参照本文件开展采购、评估、自律管理、风险协同或内部合规建设。

本文件不替代广告、人工智能生成合成内容标识、科技伦理审查、网络安全、数据安全、个人信息保护、数据出境以及行业监管规则项下的法定义务。

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 45438—2025 网络安全技术 人工智能生成合成内容标识方法

3 术语和定义

下列术语和定义适用于本文件。

3.1

生成式引擎优化 Generative Engine Optimization;GEO

通过内容结构、语义表达、可信度核验和上下文相关性优化，使目标信息更易被生成式AI理解、收录与合理引用的技术传播实践。

3.2

GEO服务提供方 GEO service provider

以商业或受托方式，为使用方提供GEO相关内容结构化、信源核验、语料接入、合规审核、监测评估、优化实施或风险治理服务的企业、机构或平台。

3.3

GEO服务使用方 GEO service user

通过GEO技术或服务开展品牌传播、产品营销、公共信息传播、知识服务、用户触达或声誉管理的组织或个人。

3.4

核心事实主张 Core factual claim

GEO服务中拟作为客观事实使用，并可能影响用户认知、交易决策、公共服务判断或模型引用结果的事实性表述。

3.5

可信语料 Trusted corpus

来源合法、内容真实、标注规范、可追溯且可核验的GEO服务相关语料。

3.6

来源可信度分级 Source credibility classification

依据信息来源的合法性、权威性、真实性、稳定性和可核验性，对来源进行分级并实施差异化使用的管理机制。

3.7

三区分治 Tri-zone governance

将品牌知识库内容按事实库、观点库、营销表达库分区存储、标识和使用的治理机制。

3.8

语料投毒 Corpus poisoning

蓄意向训练语料、检索数据源或可被模型解析的信息载体注入虚假、误导、冗余或偏见内容，干扰生成式 AI 正常输出的行为。

3.9

答案霸权 Answer dominance

特定主体在特定问题域中通过规模化内容干预，使其品牌或观点在生成式AI答案中持续占据异常高比例并降低信息来源多样性的现象。

3.10

伪共识 Manufactured consensus

通过多平台批量发布相同或高度相似的虚假信息，使大模型误将其视为客观事实或多数观点的现象。

3.11

提示词注入诱导攻击 Prompt injection inducement attack

通过网页、文档、元数据或其他信息载体嵌入隐蔽或误导性指令，诱使模型偏离正常生成逻辑的攻击行为。

3.12

异常熔断机制 Abnormal risk suspension mechanism

当发现高风险或持续性异常时，暂停相关接入、发布、优化或传播活动并启动核查、纠偏和复测的控制机制。

3.13

链路追踪标识 Trace ID

用于标识GEO服务项目、内容、语料、发布渠道、监测记录或纠偏记录之间关联关系的唯一追踪标识。

3.14

检索增强生成 Retrieval-augmented generation;RAG

在生成式AI回答过程中，通过检索外部知识库或信息源增强回答准确性的技术方法。

3.15

显式标识 Explicit label

在生成合成内容或交互场景中添加的、可被用户明显感知的标识。

［来源：GB 45438—2025］

3.16

隐式标识 Implicit label

采取技术措施在生成合成内容文件数据中添加的、不易被用户明显感知的标识。

［来源：GB 45438—2025］

4 基本原则

4.1 合法合规

服务提供方开展GEO服务应遵守法律法规、强制性标准、行业监管要求和平台规则，不应利用GEO技术规避法定审查义务。

4.2 诚信真实

服务提供方应诚实守信，如实说明服务能力、服务范围以及可能影响服务结果的风险和限制，以真实、准确、可核验的信息为基础开展GEO服务，不应伪造事实、资质、榜单、认证、评测、用户评价、专家意见或检测报告，不应夸大服务效果或作出虚假承诺。

4.3 以人为本

服务提供方应坚持以人为本，关注GEO服务对用户认知、市场秩序、公共利益和AI信息生态的影响，涉及特殊保护群体、敏感议题或重大公共利益的，应提高审核等级。

4.4 公平有序

服务提供方应维护公平有序的市场竞争和AI信息生态，不应通过语料投毒、伪共识制造、竞品抹黑、流量劫持、隐藏指令或其他组织化操纵方式干预模型输出。

4.5 安全可追溯

服务提供方应建立必要的审核、记录、追溯、复核和纠偏机制，使核心事实主张、语料来源、发布渠道和风险处置过程可核查。

4.6 主体协同

使用方提供素材、证据、授权文件或可信语料时，宜遵守诚信真实、来源可核验、最小必要和权益保护原则。大模型平台方、AI 搜索引擎、内容平台、独立第三方检索源和智能体知识库运营方，可参照本文件建立外部语料入库前安全扫描、污染反馈和溯源记录机制。

5 服务受理要求

5.1 基础资质核验

服务提供方受理项目时，应核验使用方主体身份、业务资质、产品或服务合法性、内容发布授权，以及与项目相关的资质、备案、许可或其他合法经营依据。涉及医疗、金融、教育、食品、药品、化妆品、汽车、房地产、网络游戏、直播带货、跨境服务、特殊保护群体或重大公共利益的，应开展相应增强审查。使用方无法提供必要资质、授权或证明材料，且项目内容可能触及违法违规风险的，服务提供方不应受理。

5.2 营销主张粗筛

服务提供方应在受理阶段对营销主张、核心卖点、比较性表述、功效性表述、排名性表述、认证性表述和安全性表述进行基本证据审查。涉及绝对化用语、虚假或引人误解的商业宣传、无法证明的功效承诺、竞品贬损、虚构用户反馈、违法违规广告内容的，不应进入后续服务流程。5.2 侧重判断营销主张是否具备基本证据基础和项目可受理性；6.2 侧重发布、接入或入库前核验证据有效性、要素完整性、版本与时效性及可追溯性。

5.3 不予受理情形

不予受理情形包括：

a) 使用方主体身份、资质、授权链或产品服务合法性无法核验；
b) 项目内容涉及违法违规广告、虚假宣传、诈骗、传销、赌博、色情低俗、血腥暴力、侵犯隐私、公序良俗风险或其他明显违法违规事项；
c) 项目目的在于实施语料投毒、提示词注入、元数据操纵、伪共识制造、答案霸权或竞品抹黑；
d) 项目要求伪造政府、媒体、行业协会、学术机构、专家、用户或第三方评估机构背书；
e) 项目要求绕过平台规则、访问控制、技术措施或以不正当方式抓取受保护内容；
f) 项目涉及高风险行业或特殊场景但使用方拒绝提供必要资质、证据或审核配合；
g) 项目要求删除、篡改、伪造 AI 生成合成内容标识或隐瞒 AI 参与情况；
h) 项目存在其他足以影响公共利益、市场秩序、用户权益或信息生态安全的重大风险。

5.4 受理记录

服务提供方应留存受理审查记录。记录宜包括使用方主体信息、项目范围、资质材料、授权文件、营销主张、风险初判、不予受理理由、审核责任人和审查时间。

5.5 第三方协作

项目涉及内容生产、审核、数据处理、技术接入、监测评估等关键第三方协作主体的，服务提供方应识别其角色、能力和责任边界，并核验与项目风险等级相适配的资质、授权状态和重大违规记录。

6 内容审核要求

6.1 内容合规审核

服务提供方应在内容发布、语料接入或知识库入库前，对拟使用内容进行合法性、真实性、广告合规性、权益保护和平台规则适配审核。

a) 内容不得违反法律法规、强制性标准、行业监管规定或平台规则；
b) 内容涉及广告宣传的，应符合广告法、互联网广告及行业广告审查要求；
c) 内容涉及新闻信息、公共政策、专业知识或特殊保护群体的，应提高审核等级；
d) 内容涉及第三方作品、商标、字体、肖像、数据、案例或评价的，应核验授权或合法使用依据。

6.2 核心事实主张证据核验

内容发布、语料接入或入库前，应对核心事实主张完成证据要素绑定。每条核心事实主张至少应绑定：

a) 来源等级；
b) 来源名称；
c) 原始链接或原始文件；
d) 采集时间；
e) 最近核验时间；
f) 审核责任人；
g) 有效期；
h) 版本号；
i) 证据文件编号。

同一来源、同一文件、同一批次或同一参数表中的同类事实主张，可采用批量证据绑定方式，但应能够追溯至具体事实项、来源文件和版本。核心事实主张记录宜与证据编号、来源等级、内容版本、发布或接入渠道、AI标识状态和追溯标识建立对应关系。价格、监管状态、资质有效期、市场数据等时效性事实，应设置有效期、复核频率或触发式更新条件。已完成证据要素绑定和来源分级核验的，可作为服务提供方履行内容审核义务的证据材料；不当然免除使用方依法应承担的内容真实性责任。

6.3 来源可信度分级

服务提供方应根据信息来源的合法性、权威性、真实性、稳定性和可核验性，对来源进行A、B、C、D四级管理，见表1。

表1 来源可信度分级

等级	来源类型	使用规则
A	政府部门、司法机关、监管机构等具有法定职责的主体在职责范围内依法发布的信息（如气象、地震、公共卫生等专业领域信息），强制披露平台、官方统计数据，纳入国家网信部门公布的《互联网新闻信息稿源单位名单》的单位依法发布的原创新闻信息，符合本文件来源管理要求并纳入GEO可信媒体名单的媒体单位依法发布的原创新闻信息，依法公开的生效裁判文书和仲裁裁决文书，依法开放利用的档案材料、权威史志档案、国家权威数据库等	可作为核心事实性引用的优先来源，仍应完成单条事实核验

表 1（续） 来源可信度分级

等级	来源类型	使用规则
B	经同行评议的学术论文、独立科研报告、行业协会白皮书、主流媒体原创报道、公安机关发布的警情通报等动态性、过程性执法信息、权威科普平台、专业数据服务机构等	可作为重要证据来源，必要时结合 A 级来源或多源交叉核验
C	企业官网、产品手册、商业报告、用户反馈、普通媒体报道、公开网页等	可作为辅助来源，涉及核心事实主张时应提高核验强度
D	匿名来源、无法追溯来源、低质聚合站、历史违规来源、明显利益冲突且无法验证来源等	原则上不应用于核心事实主张；确需使用的，应说明理由并加强复核

注1：来源等级反映信息源整体可信程度，不当然等同于该来源发布的每一条具体信息均为事实。司法文书中的当事人陈述、媒体报道中的转述内容、专家观点、用户评价和预测性判断，应结合具体事实主张单独核验。

注2：官方来源优先不等于排除消费者投诉、媒体调查、公共监督和第三方检测等非官方证据。服务提供方应结合证据内容、核验目的和利益关系综合判断。

注3：来源等级应结合发布主体、发布内容和可核验性综合判定。同一主体发布的不同信息应分别完成事实核验；来源类型交叉时，可按较高等级管理，但不当然免除单条事实核验。

注4：公安机关发布的警情通报等动态性、过程性执法信息，虽由具有法定职责的主体发布，但因事实可能仍处于发展或核查阶段，宜按B级使用。

注5：纳入《互联网新闻信息稿源单位名单》的单位，以及符合本文件来源管理要求并纳入GEO可信媒体名单的媒体单位依法发布的原创新闻信息，可按A级使用，仍应完成单条事实核验。新闻转载、转述、评论、预测性内容或非新闻信息，不当然按A级使用。GEO可信媒体名单未发布前，相关媒体来源仍按照表1其他规则判定。

服务提供方应建立来源等级动态复核机制。来源发生撤销、处罚、重大更正、权威性变化或被发现存在虚假信息风险的，应及时降级、限制使用或重新核验。

为实施本文件关于来源分级、语料接入和追溯管理的要求，相关机构可建立并动态调整GEO可信媒体名单。该名单可将纳入国家网信部门公布的《互联网新闻信息稿源单位名单》的媒体单位作为基础范围，并结合本文件关于事实核验、内容授权、纠错更新和追溯记录的要求纳入其他媒体单位。该名单用于GEO服务中的来源管理、语料接入和项目适配参考，不当然免除单条事实核验，不构成行政许可、官方认证、平台背书或排他性资格。

6.4 知识产权与 AI 标识

6.4.1 知识产权与商业标识核验

服务提供方应核验拟发布、接入或入库内容的知识产权、商业标识、授权状态和竞品权益风险。

a) 不应未经授权使用他人作品、数据、图片、音视频、案例、报告或数据库内容；
b) 不应在隐藏标签、元数据、隐藏文本或提示词指令中嵌入竞争对手的商标、字号、字体及其他商业标识；
c) 不应违反 robots 协议、突破技术措施、绕过访问控制，或以影响网站正常运行、侵犯商业秘密或他人合法权益的方式抓取内容；
d) 内容存在版权争议、授权范围不明或权利人异议的，应暂停相关发布、接入或入库并启动核查；
e) 同一服务提供方承接同行业竞品项目的，不应将一方项目中的商业秘密、未公开产品信息、内部检测报告或受托资料共享给另一方。

6.4.2 AI 生成合成内容标识

服务提供方应依法履行自身标识义务；涉及使用方发布或使用AI生成合成内容的，应提示、协助并核验使用方按照适用规则履行标识义务。

a) 记录内容是否由 AI 生成、AI 辅助生成或人机协同编辑；
b) 记录显式标识、隐式标识、平台声明、元数据或水印状态；

c） 第三方生成合成服务调用方受技术能力限制无法自行添加或核验隐式标识的，应进行合理检查，不规避或删除工具自带标识，并配合按适用规则声明内容属性；
d） 平台未提供声明或标识功能的，服务提供方宜在内容中以适当方式提示 AI 生成或辅助生成情况。

人工主导、AI仅作摘要、排版、翻译、检索辅助的，不当然改变原信源等级；AI主导生成、人工仅作形式审核的，应按AI生成合成内容记录标识和使用边界。人机协同深度编辑的语料属性，宜结合最终审核发布责任主体、关键事实主张是否经人工逐条复核等因素判定。

6.4.3 标识禁止行为

标识禁止行为包括：
a） 不应删除、篡改、伪造或规避依法应有的 AI 生成合成内容标识；
b） 不应将 AI 生成合成内容伪装为权威机构、专家、用户或媒体的真实意见；
c） 不应利用未标识 AI 内容制造虚假用户评价、客户案例、使用效果对比图或第三方测评；
d） 不应通过技术处理使显式标识难以识别，或通过格式转换、截图、裁剪等方式规避标识要求；
e） 不应以内部追溯记录替代依法应向用户呈现的标识或声明。

本条不新增法定标识义务。与 GB 45438—2025 及相关上位规则不一致的，从其规定。

6.5 科技伦理审查衔接

6.5.1 适用性自评

服务提供方应依据人工智能科技伦理审查相关规定，对所承接项目是否属于需开展科技伦理审查的人工智能科技活动进行适用性自评。
a） 涉及生命健康、身心安全、特殊保护群体、重大公共利益或社会认知风险；
b） 可能影响就业、教育、金融、医疗、公共服务、舆论环境或公共政策理解；
c） 使用具有明显诱导性、沉浸式或难以撤回的信息传播方式；
d） 项目范围、行业属性、传播对象、数据类型或技术路径发生重大变化并可能触发新的科技伦理风险。

6.5.2 审查衔接与实施边界

项目可能触发科技伦理审查要求的，服务提供方应提示使用方或相关责任主体履行相应程序，配合准备必要材料并留存提示记录。委托方或相关责任主体拒绝履行相关程序，且经评估认为继续实施将导致相关科技伦理风险无法有效控制的，服务提供方不应继续实施相关GEO服务活动。

服务提供方应完成内容合规与广告审核、核心事实主张证据核验、来源可信度分级与证据链管理、知识产权与AI生成合成内容标识核验，以及科技伦理审查适用性自评与衔接。

6.1至6.4为内容级审核动作；6.5为项目级科技伦理审查适用性自评与衔接动作。二者并立，不可互替。前四项任一项未通过的，不应发布、接入或入库；经自评可能触发科技伦理审查要求的，应完成相应提示、配合和凭证留存后，方可进入实施环节。

本条对可能引发较大社会影响的GEO服务项目规定科技伦理审查适用性自评与衔接要求；该自评不替代人工智能科技伦理审查相关规定项下的法定伦理审查。本章不替代广告审查、行业专项审查、科技伦理审查、平台规则审核及其他依法或依合同应履行的义务。

内容发布与语料接入核验清单参见附录B。

7 GEO 优化实施要求

7.1 三区分治

服务提供方应对品牌知识库或项目知识资产实施事实库、观点库、营销表达库分区管理，并分别设置入库门槛、内容属性标识和使用规则。
a） 事实库应存放已完成证据绑定、来源分级和审核的事实性内容，包括经核验的主流媒体原创新闻报道中的事实性内容；

b） 观点库可存放评论、社论、专家观点、媒体评论、行业分析、场景解释等主观性内容，并标明观点来源和时间；
c） 营销表达库可存放口号、创意文案、卖点包装和场景化表达，不得替代事实库作为事实依据；
d） 营销表达中的事实主张应以事实库版本为准。

7.2 语料接入与发布

服务提供方应优先采用官方接口、授权渠道、公开合法内容源或经审核的知识库接入语料，不应通过异常频率、批量低质发布、站群污染或伪装用户行为影响模型判断。

官方接口不可用、限速或封闭时，服务提供方宜留存平台官方说明、不可用证据和替代路径选择依据。媒体机构依规开放的API、稿源系统或授权内容接口，可作为官方或授权接口使用。

服务提供方宜结合来源可信度分级、历史违规记录和项目风险等级，建立可接入来源的白名单、灰名单和黑名单管理机制，并定期复核。

语义重复度检测主要面向同一品牌、同一主题或卖点、同一评估时间窗口内的内容。服务提供方应结合项目风险等级设置重复度、发布频率、渠道分布和版本更新控制要求。

7.3 防注入与元数据控制

服务提供方应排查网页、文档、图片、PDF、元数据、隐藏文本、OCR识别文本、智能体指令和其他可被模型解析的信息载体中的隐藏指令、误导性指令、异常元数据和诱导性内容。

a） 不应通过隐藏标签、CSS 隐藏文本、文档元数据、图片 OCR 识别文本、PDF 文字层或文档隐藏文本嵌入误导模型的指令；
b） 不应伪造发布时间、作者、来源、审核状态、引用关系、权威机构标识或 AI 标识；
c） 不应诱导智能体或其他系统替用户执行违背其真实意愿的检索、比较、推荐、购买、授权或数据提交行为；
d） 应对疑似语料投毒、伪共识、竞品抹黑、商标劫持或流量劫持和异常传播行为进行记录、核查和必要处置。

7.4 多模态内容安全

涉及图片、音频、视频、直播切片、PDF、PPT、长图、交互页面或其他多模态内容的，服务提供方应核验可见内容、隐藏文字层、字幕、封面、OCR识别结果、元数据、水印和AI生成合成内容标识状态。无法核验的，应降低信任等级、补充人工复核或暂停接入发布。

8 监测、追溯与风险处置要求

8.1 全链路追溯

8.1.1 追溯节点和要素

服务提供方应建立覆盖项目受理、客户准入、材料审核、内容生成或优化、内容审核、语料接入、发布传播、监测分析、风险处置和纠偏复测的追溯机制。最小追溯要素宜包括：

a） 谁委托；
b） 谁提供原始资料；
c） 谁生成、优化、审核、确认、接入、发布；
d） 何时生成、审核、接入、发布、替换或下线，以及下线或替换原因；
e） 发布或接入渠道；
f） 内容版本、证据编号、来源等级和 Trace ID；
g） AI 标识状态、授权状态和纠偏状态。

全链路追溯数据要素的具体规定见附录 A。

8.1.2 平台型与非平台型控制

平台型服务提供方宜通过系统日志、流程节点、权限控制、审核记录、发布记录和导出报告形成自动化或半自动化追溯。非平台型服务提供方可采用项目台账、审核表、证据清单、发布记录和人工复核

记录等方式形成等效控制。等效控制措施宜具备可被独立第三方核查验证的条件，并留存可供核查的记录。

8.1.3 数据安全与防篡改

追溯数据在采集、传输、存储、导出和归档过程中，应采取与风险等级相适应的访问控制、权限管理、哈希摘要、数字签名、时间戳、日志留痕或其他防篡改措施。对具有长期复用价值、公共引用价值或后续知识资产化管理价值的可信事实材料，服务提供方宜加强证据材料完整性、形成时间、版本变更和授权状态记录，可根据项目需要采用哈希摘要、可信时间戳、电子签名、分布式存证或第三方存证等增强存证方式。相关存证不当然证明事实内容真实。

涉及个人信息处理的，应遵守个人信息保护相关法律法规；涉及依法适用网络安全等级保护制度或密码管理要求的，从其规定。

8.1.4 脱敏、隔离与保存

追溯记录涉及个人信息、商业秘密、未公开产品信息、内部检测报告或第三方受托资料的，应按最小必要原则，结合数据敏感程度采取访问控制、去标识化或脱敏处理、权限隔离等措施。承接同行业竞品项目的，应建立项目墙、人员权限隔离、数据访问控制、日志留痕和利益冲突披露机制。一般项目核心追溯数据保存期限不宜少于3年。高风险行业或特殊场景项目全部追溯数据保存期限不宜少于3年；法律法规、行业监管、合同约定或争议处理另有更长要求的，从其规定。

8.2 信息生态监测

服务提供方应根据项目风险等级监测信息准确性、来源多样性、负面偏差、伪共识线索、答案霸权趋势、语料投毒风险和纠偏效果。风险分级宜结合发生概率、影响范围、受众敏感度、纠正难度、传播可逆性、合规后果和社会影响，将风险划分为高、中、低三级。

8.3 效果评估

服务提供方开展效果评估时，应区分以下五类指标：内部风控指标、过程履约指标、效果观察指标、AI引用类指标和生态观察指标。具体指标参见附录E；其中，生态观察指标用于观察GEO服务活动对信息生态的影响。

服务提供方不应承诺或暗示“保证第一”“稳定前三”“永久收录”“必然被大模型引用”“官方推荐合作”“排他性平台合作”等绝对化或误导性效果。

服务提供方可约定过程性目标，如事实库完善率、证据链完整率、标识合规率、监测频次、纠偏时效和交付节点。过程性指标不应被表述为对模型输出结果的绝对承诺。

效果观察结果不得抵消合规风险。出现下列情形之一的，不应将项目评估为合格交付：

a) 核心事实主张未经证据核验或证据链严重缺失；
b) 存在虚假宣传、伪造背书、竞品抹黑、语料投毒、伪共识制造或答案霸权等行为；
c) AI 生成合成内容标识、知识产权、个人信息或数据安全要求存在重大违规；
d) 高风险行业或特殊场景项目未完成必要增强审核；
e) 发现重大风险后未按约定时限采取暂停、核查、纠偏或复测措施。

8.4 异常熔断与纠偏

服务提供方发现下列情形时，应暂停相关接入、发布或优化活动，并启动核查、纠偏或升级审核：

a) 疑似虚假信息、事实造假、重大过期信息或竞品抹黑；
b) 疑似语料投毒、提示词注入、伪共识趋势、答案霸权或元数据异常；
c) 命中禁止行为、黑名单来源、重大投诉或监管风险；
d) 客户提供材料与已发布内容出现实质性不一致；
e) 涉及高风险行业、特殊保护群体或重大公共利益且风险无法排除。

高风险事件宜自发现或者收到明确风险反馈起1小时内启动初步响应，并自该起算点起4小时内完成初步止损；中风险事件宜在24小时内完成初步核查；一般纠偏宜在7个工作日内完成并形成记录。属于高风险情形的，按高风险事件时限优先处理。服务提供方的可控纠偏范围限于自有或受托内容源、受托

发布渠道、项目知识库、提交材料和相关记录；对第三方独立内容源或模型参数化记忆不可直接控制的，应留存通知、反馈、复测和风险提示记录。

8.5 应急演练与红队测试

L2及以上服务提供方宜结合项目风险建立应急演练机制，对语料投毒、提示词注入、伪共识、答案霸权、重大事实错误、AI标识违规和平台反馈异常等情形进行演练并留存记录。L3服务提供方宜定期开展红队测试或对抗性验证，检查隐藏指令、异常元数据、多模态注入、智能体诱导、竞品抹黑、商标劫持或流量劫持和追溯链断裂等风险，并将测试发现纳入整改和复测。

9 高风险场景要求

9.1 高影响传播场景

涉及重大时政、党和国家领导人重要表述、意识形态安全、资本市场、公共服务、突发事件、重大公共政策或大范围社会传播的项目，应实施增强审核、权威信源优先、敏感期监测和人工升级审批。

a) 受理阶段应核验项目目的、传播对象、传播范围、敏感时间窗口和授权依据；
b) 内容审核阶段应核验事实来源、发布口径、表述边界和可能引发的公共影响；
c) 发布后应提高监测频次，记录模型引用偏差、事实错误、负面扩散和纠偏结果。

涉及重大时政、意识形态安全等内容的，宜以权威发布口径、正式文件、主管部门公开信息或具有相应资质的新闻机构审定内容为主要核验依据，并提高人工复核等级，不应仅依赖自动化流程。

9.2 高风险行业场景

涉及医疗健康、药品、医疗器械、金融投资、保险、教育培训、食品保健、儿童产品、汽车安全、房地产、网络游戏、直播带货等行业的，应核验行业资质、广告审查、产品许可、风险提示和禁止性宣传要求。

a) 应核验经营主体、产品服务、许可备案、检测报告、广告审查或监管状态；
b) 应识别保本收益、治愈率、绝对安全、升学就业保证、价格库存失实等高风险表述；
c) 涉及时效性事实的，应设置有效期、复核频率或触发式更新条件。

9.3 特殊保护群体场景

面向未成年人、老年人、残疾人、患者、求职者、学生、消费者等特殊保护群体的项目，应加强诱导性、欺骗性、情绪化、对立化和沉浸式内容风险审查。

a) 不应利用年龄、疾病、认知能力、教育焦虑、就业压力或信息不对称实施诱导；
b) 应提高人工审核等级，核验内容是否存在歧视、污名化、恐吓、过度承诺或不当比较；
c) 应留存特殊保护群体风险识别、审核结论和纠偏处置记录。

9.4 跨境与境外信源场景

涉及跨境传播、境外信源、境外平台、境外模型或数据出境的项目，应核验数据安全、个人信息保护、版权授权、境外平台规则和国家安全相关要求。

a) 应识别数据来源、处理目的、传输路径、境外接收方和授权状态；
b) 涉及个人信息、重要数据或受委托处理数据的，应按适用法律法规履行评估、合同、认证或其他程序；
c) 境外信源用于核心事实主张的，应核验来源权威性、版本、发布时间和本土适用边界。

对公开网页可能被境外搜索引擎或AI系统依法依规收录的情形，服务提供方宜结合是否主动向境外传输数据、是否包含个人信息或重要数据、是否受委托处理等因素，判断是否构成本文件项下的数据出境活动，并留存判断依据。本说明不替代个人信息保护、数据安全和数据出境相关法律法规的适用。

高风险场景核验重点和核心禁止行为参见附录C。

10 服务提供方组织能力要求

10.1 基础能力

服务提供方应根据自身业务范围、项目风险等级和服务复杂度，至少具备本章规定的相应等级组织能力。能力等级与业务承接范围的对应关系如下：

a) 建立项目受理、内容审核、证据核验、来源分级、发布接入、监测处置和纠偏复盘流程；
b) 配置具备相应业务知识的审核人员或法务、合规角色；
c) 留存项目方案、审核记录、证据材料、发布台账、监测报告和纠偏记录；
d) 对第三方协作主体进行识别、授权核验和必要监督。

10.2 能力分级

服务能力按照适用主体和核心要求，分为L1、L2、L3三级。服务能力分级要求见表2，服务能力核验基准参见附录D。

表 2 服务能力分级要求

等级	适用主体	核心要求
L1	小型、非平台型或低风险项目服务提供方	可采用文档化台账、人工审核和基础记录方式，满足受理、事实核验、来源分级、AI 标识核验、发布台账和异常记录要求。
L2	具备结构化流程和稳定服务能力的服务提供方	应具备结构化追溯、双人审核、到期提醒、专项复核、监测报告、权限管理和可导出记录能力。参与高风险或特殊场景项目时，宜接受 L3 能力主体支持、复核或监督。
L3	平台型、规模化或承接高风险项目的服务提供方	应具备自动化监测、异常熔断、红队测试、跨项目隔离、第三方评估、应急演练和综合治理能力。

10.3 L1 基础能力

L1基础能力包括：

a) 至少配备能够完成基础内容审核、事实核验和记录管理的人员；
b) 可采用人工台账、表格、文档和截图方式留存项目受理、证据核验、来源分级、AI 标识核验和发布记录；
c) 不应由内容生成或优化人员单独完成其自身内容的最终审核；确需兼任的，应记录利益冲突说明和复核安排；
d) 应能够提供基础交付物清单，包括受理记录、核心事实清单、证据材料、发布台账和异常纠偏记录。

10.4 L2 专业能力

L2专业能力包括：

a) 应建立结构化项目管理、双人审核、证据链台账、到期提醒和专项复核机制；
b) 应支持按项目、时间、责任人、内容版本或 Trace ID 导出不可编辑的原始记录；
c) 参与高风险或特殊场景项目的，宜在 L3 能力主体支持、复核或监督下进行；
d) 应开展定期自查、人员培训和异常处置复盘。

10.5 L3 综合治理能力

L3综合治理能力包括：

a) 应具备自动化或半自动化监测、异常熔断、风险预警、纠偏复测和报告能力；
b) 应建立语料污染反馈、红队测试、应急演练、跨项目隔离和第三方评估机制；
c) 应能够承接第 9 章所列高风险行业、特殊场景或规模化项目，并形成专项审核、会签和留痕材料；
d) 宜具备可信事实资产管理、增强存证、接口治理和多平台监测能力。

10.6 能力说明

服务提供方可基于本章进行能力自我说明或第三方评估。能力说明不构成行政许可、市场准入、强制认证或排他性经营条件，不应被用于暗示官方认证、平台背书或行业排他资格。

11 附则

11.1 实施过渡

采用本文件时，新增项目宜按照本文件要求执行；存量项目可在6个月基础过渡期内补充关键记录，在12个月内逐步完善追溯、监测和能力建设要求。过渡期安排为采用本文件时的实施建议，不构成行政强制义务或市场准入条件。

11.2 法规和平台规则衔接

本文件与法律法规、强制性标准、行政规章或平台规则不一致的，按法律法规、强制性标准、行政规章和依法有效的平台规则执行。

11.3 投诉与行业自律

行业组织可结合本文件建立投诉、提示、风险通报和自律管理机制。投诉处理过程中，宜保障被投诉方陈述、提交补充材料和申请复核的机会，并记录处理依据、处理结论和纠正措施。

11.4 修订机制

本文件可根据法律法规、监管政策、平台规则、技术发展和行业实践变化进行修订。实施中形成的典型案例、工具模板、符合性评价方法和配套指南，可另行发布。

11.5 资料性附录使用

本文件资料性附录用于提供实施参考。相关主体采用资料性附录时，宜结合项目实际技术条件、分工安排、预算水平、风险等级和合同约定进行调整，不宜僵硬套用为统一违约责任、强制罚则或市场准入条件。

附 录 A
（规范性）
全链路追溯数据要素

表A.1规定GEO服务全链路追溯的最小数据要素。服务提供方可根据项目规模和系统能力扩展字段。

表 A 追溯数据要素

字段	约束	说明
project_id	必填	项目唯一标识
client_id	必填	使用方主体标识
provider_id	必填	服务提供方主体标识
content_id	必填	内容或语料标识
trace_id	必填	链路追踪标识
core_claim_id	条件必填	核心事实主张标识
source_level	条件必填	来源可信度等级
evidence_id	条件必填	证据文件编号
version_no	必填	内容或知识库版本号
ai_label_status	条件必填	AI 标识状态
authorization_status	条件必填	授权状态
reviewer	必填	审核责任人
published_channel	条件必填	发布或接入渠道
risk_status	推荐	风险状态
correction_status	推荐	纠偏流程状态
correction_status_extended	推荐	纠偏后观察状态、复测结果或残余风险
deposit_method	推荐	哈希摘要、可信时间戳、电子签名、分布式存证、第三方存证等
deposit_certificate_id	推荐	存证凭证编号
trusted_fact_asset_flag	推荐	是否纳入可复用可信事实资产管理范围
material_provider	条件必填	资料提供主体，涉及使用方或第三方提供原始材料时填写
optimizer_editor	条件必填	优化或编辑责任人，发生内容优化、改写、编辑等动作时填写

附 录 B
（资料性）
合规审核核验清单

表 B 内容发布与语料接入核验清单

核验项	核验要点
主体和授权	核验使用方主体、资质、业务许可、发布授权和第三方协作关系
广告和内容合规	核验广告法、行业广告、互联网内容管理和平台规则要求
核心事实主张	核验证据绑定、来源等级、版本、有效期、审核责任人和证据编号
来源分级	核验 A/B/C/D 级来源判定、动态复核和降级恢复记录
商标、字号、字体及其他商业标识保护	核验商标、字号、字体、作品、肖像、数据和商业秘密授权状态
AI 标识	核验显式标识、隐式标识、平台声明、人机协作记录和标识禁止行为
科技伦理	核验适用性自评、提示记录、申报配合、重新自评和拒绝履行处置
防注入	排查隐藏标签、元数据、OCR 文本、PDF 文字层、隐藏文本和智能体诱导指令
监测纠偏	核验监测样本、风险分级、熔断记录、纠偏工单和复测结果

附 录 C
（资料性）
高风险场景与禁止行为速查

表 C.1 高风险场景核验重点

场景	核心核验事项
医疗健康、药品、医疗器械	主体资质、产品许可、广告审查、功效边界、禁用表述和风险提示
金融、投资、保险	金融资质、收益承诺、风险揭示、投资适当性和监管状态
教育培训、就业服务	办学资质、就业承诺、升学承诺、师资证明和特殊人群保护
食品、保健食品、化妆品、儿童产品	产品许可、功能宣称、检测报告、适用人群和安全风险
汽车、房地产、网络游戏、直播带货	比较性宣传、价格库存、测评依据、消费者权益和平台规则
涉政、涉军、涉史、意识形态和重大公共政策	权威发布口径、正式文件、主管部门公开信息和人工升级审批
跨境与境外信源	数据安全、个人信息、版权授权、境外平台规则和国家安全要求

表 C.2 核心禁止行为

类别	禁止行为
虚假信息	伪造事实、资质、认证、榜单、检测报告、用户评价、专家意见或媒体报道
不正当竞争	竞品抹黑、商标劫持、流量劫持、虚假比较、商业秘密侵犯
语料投毒	站群污染、批量低质发布、隐藏指令、恶意元数据和伪装来源
伪共识	跨平台批量发布虚假或高度相似内容，制造虚假多数意见
答案霸权	通过规模化操纵使单一主体在特定问题域中异常占据答案空间
标识违规	删除、篡改、伪造或规避 AI 生成合成内容标识
特殊保护群体权益损害	面向特殊保护群体实施诱导性、欺骗性、情绪化或对立化内容传播

附 录 D
（资料性）
服务能力核验基准

表 D 能力核验基准

能力项	L1	L2	L3
项目受理	人工记录	结构化受理表	系统化准入与风险初判
内容审核	人工审核	双人审核和专项复核	多角色会签和自动化辅助审核
证据追溯	文档台账	结构化证据链	Trace ID 全链路追溯
监测处置	人工巡检	定期监测和报告	自动化监测、熔断和复测
竞品隔离	基础保密	权限和项目隔离	人员、数据、日志、利益冲突综合隔离
应急能力	异常记录	处置流程和复盘	演练、红队测试和第三方评估

附 录 E
（资料性）
效果评估指标

效果评估指标用于观察项目履约和信息生态影响，不构成对模型输出结果的绝对承诺。

表 E 效果评估指标

类别	指标示例	说明
内部风控指标	证据链完整率、来源分级覆盖率、AI 标识合规率、授权核验完成率	用于评价服务过程合规性
过程履约指标	内容资产建设进度、监测频次、纠偏时效、报告交付节点	用于评价项目交付过程
效果观察指标	AI 答案提及率、有效提及率、核心卖点召回率、情感正向率、信息准确率	用于观察服务后的外部表现
AI 引用类指标	信源引用总量、信源提示词覆盖率、被引用页面数、页面级引用活跃度	用于观察目标信源在 AI 回答中的引用情况
生态观察指标	未证实主张比例、误导性表述比例、信源多样性变化、疑似伪共识线索数、品牌语境关联变化	用于观察 GEO 活动对信息生态的影响

高风险行业、高影响传播场景或特殊保护群体相关项目，宜适当提高独立采样轮次和复核强度。具备条件且项目确有需要的，可开展多平台、多时间点、多轮独立会话以及多账号、多地域或多终端采样。

参 考 文 献

[1] 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
[2] GB 45438—2025　网络安全技术　人工智能生成合成内容标识方法
[3] 中华人民共和国广告法
[4] 中华人民共和国个人信息保护法
[5] 中华人民共和国数据安全法
[6] 中华人民共和国网络安全法
[7] 互联网广告管理办法
[8] 人工智能生成合成内容标识办法
[9] 人工智能科技伦理审查与服务办法（试行）
[10] 生成式人工智能服务管理暂行办法
[11] Google Search Central. Spam policies for Google web search.
[12] OpenAI. Usage policies and model behavior documentation.
[13] Microsoft. Responsible AI Standard.
[14] Aggarwal P. et al. GEO: Generative Engine Optimization. arXiv:2311.09735.
[15] Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
[16] ISO 31000:2018 Risk management — Guidelines.
[17] ISO/IEC 42001:2023 Artificial intelligence management system.
[18] NIST AI Risk Management Framework.
[19] OWASP Top 10 for Large Language Model Applications.