

ICS 35.240.01

CCS L76

团 体 标 准

T/CAACCHINA 004-2026

生成式引擎优化 (GEO) 第 4 部分：可信语料规范

Generative Engine Optimization (GEO) — Part 4: Trusted corpus specification

202X-XX-XX 发布

202X-XX-XX 实施

中国商务广告协会 发 布

目次

目次..... II

前言.....IV

引言..... V

1 范围 1

2 规范性引用文件 1

3 术语和定义 2

 3.1 语料 Corpus 2

 3.2 语料元数据 Corpus Metadata 2

 3.3 语料分级 Corpus Grading..... 错误! 未定义书签。

 3.4 可信语料 Trustworthy Corpus 2

 3.5 信源 Information Source 2

 3.6 语料生命周期 Corpus Lifecycle 错误! 未定义书签。

 3.7 语料投毒、语料污染 Corpus Poisoning & Corpus Contamination 3

4 可信语料与信源判断的基本原则 3

 4.1 真实性原则 3

 4.2 权威性原则 3

 4.3 专业性原则 3

 4.4 可追溯原则 3

 4.5 动态更新原则 3

 4.6 中立性原则 4

5 信源与语料标准 4

 5.1 信源准入要求 4

 5.2 语料分类与语料可信等级 4

 5.3 语料基础质量 5

 5.4 语料元数据..... 错误! 未定义书签。

6 信源与语料标注规范 6

6.1 标注流程6

6.2 标注维度与字段统一标准 7

6.3 综合可信等级标识 7

7 可信等级标识7

附 录 A.....9

 A.1 高风险行业定义9

 A.2 总体准入原则9

 A.3 高风险行业信源分级9

 A.3.1 最高可信信源9

 A.3.2 次级合规信源9

 A.3.3 绝对禁止信源10

附 录 B..... 11

 B.1 主要流程 11

 B.2 分级定义 11

 B.3 归档期限 11

 B.4 问题语料召回、下线、替换机制11

 B.5 销毁机制 12

附 录 C..... 13

 C.1 总体原则 13

 C.2 通用语料抽样量级 13

 C.3 高风险行业语料抽样量级 13

 C.4 分层抽样规则 14

 C.5 抽样结果判定阈值 14

参考文献.....15

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》给出的规则起草。

本文件是《生成式引擎优化（GEO）》系列标准的第 4 部分。该系列标准包括以下部分：

第 1 部分：通论及术语与定义；

第 2 部分：服务商评估规范；

第 3 部分：计价、评估与测量规范；

第 4 部分：可信语料规范；

第 5 部分：服务流程合规与安全要求。

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。

本文件由中国商务广告协会提出并归口。

本文件起草单位：北京知乎网技术有限公司、北京搜狐新媒体信息技术有限公司、上海市新榜信息技术股份有限公司、北京明略昭辉科技有限公司、上海智推时代科技有限公司、北京清蓝智汇科技有限公司、广州欢网科技有限责任公司、利欧集团数字科技有限公司、武汉当夏科技有限公司、钛镁时代(上海)人工智能科技有限公司、北京特比昂科技有限公司、北京优力互动数字技术有限公司、上海喜马拉雅科技有限公司、微梦创科网络科技(中国)有限公司、薯道电子科技(上海)有限公司、芜湖小象园教育科技有限公司、杭州盖立克思人工智能有限公司、海知见信息技术(北京)有限公司。

本文件主要起草人：张炎、吴哲、宋明玉、桑田、李玉博、王子冠、马记生、刘文明、王立新、毛桂荣、邓礼帆、周崧弢、张浩然、张豪、王健、侯艺、曾杰、吉晨亮、庞成博、张龙飞、冯羽健、展萌萌、傅海波、牛婷婷、侯燕、温雅、陈敏刚、周斌斌、周蕾、谭北平、李浩宇、张慧。

本文件主要审查人：丁俊杰、陈徐彬、吴明辉、吴默、陈忠伟、任鹏、全佳欣、李丹、严禧汶、唐兴通。

引 言

随着生成式人工智能技术的快速发展，GEO 服务中语料的质量和可信度成为影响 AI 生成内容质量的关键因素。可信语料是 GEO 服务的底层基础设施，直接关系到品牌在 AI 生成答案中的呈现效果和用户信任。

当前 GEO 行业在语料管理方面存在以下突出问题：语料质量参差不齐，缺乏统一的质量判定标准；信源准入缺乏规范，低质量、不可信信源影响 AI 生成内容的可信度；语料标注和管理流程不规范，导致语料质量难以保证；高风险行业语料缺乏特殊管理要求，存在合规风险。

本标准旨在建立 GEO 场景下可信语料的统一规范，明确语料可信判断的基本原则，建立信源准入要求和语料质量标准，规范语料标注流程和管理要求，为 GEO 服务提供高质量的语料基础保障。

本标准与 GEO 团体标准体系内其他标准的关系如下：与通论及术语与定义（第 1 部分）的术语与定义保持对齐；与服务商评估规范（第 2 部分）的服务商语评估要求对接；与计价、评估与测量规范（第 3 部分）的计价规范衔接；与服务流程合规与安全标准（第 5 部分）的数据安全与合规要求对齐。

生成式引擎优化（GEO）第 4 部分： 可信语料规范

1 范围

本文件规定了 GEO（生成式引擎优化）场景中可信语料的术语定义、分类分级、信源要求、内容质量、合规要求、元数据管理、全生命周期管理、评测方法与标识要求。

本文件适用于适用于以下主体：

- a) GEO 服务商
- b) 广告营销机构
- c) 内容/资讯平台
- d) 广告主/GEO 客户
- e) 第三方审核与评估机构

本文件不适用于：

- a) 科研实验、学术研究、内部模型训练测试，未对外开展 GEO 商业化服务、未面向公众输出生成内容的非经营性语料场景；
- b) 涉密信息、党政内部专属资料、未公开涉密语料及法律法规禁止公开使用的专属数据资源；
- c) 纯个人自用、无传播属性、无商业赋能属性的私人语料及碎片化个人数据内容；
- d) 硬件底层数据、纯机器日志数据、原始传感器数据等不参与大模型语义生成、检索推理与内容输出的非语义类原始数据；
- e) 专项行业监管体系下独立合规运营的专属语料场景，含金融、医疗、政务、司法等行业专属涉密及合规隔离数据。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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 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

GB/T 41867—2022 信息技术 人工智能 术语

GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求

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)

通过系统性的技术与内容策略，增强数字内容在生成式 AI 搜索引擎结果中的可见性、准确性和可信度的优化活动。

注 1：GEO 概念由 Aggarwal 等人于 2023 年在论文《GEO: Generative Engine Optimization》(arXiv:2311.09735) 中首次系统性提出。

注 2：GEO 的核心目标是让内容在生成式引擎响应中被发现、被引用、被推荐。

[来源：T/CAACCHINA 001-2026，通论及术语与定义 3.1.1]

3.2 语料 corpus

支撑生成式人工智能模型训练、调优、评测与推理运行的原始数据集合。

注 1：该类数据为生成式人工智能模型完成知识学习、语义理解、内容生成提供基础输入载体。

注 2：按用途可分为预训练语料、微调语料、评测语料和推理语料。

[来源：T/CAACCHINA 001-2026，通论及术语与定义 3.2.5]

3.3 语料元数据 corpus metadata

供 AI 引擎解析与溯源，用来描述语料内容、来源、审核状态、时效等属性的结构化数据。

注：语料元数据不包含语料的实际内容，而是描述语料的来源、质量、合规性等关键特征。

3.4 可信语料 trusted corpus

经权威审核、来源可追溯、内容真实准确，适用于生成式人工智能检索、理解、推理与生成引用的结构化信息素材。

注：可信语料的判定依据包括信源权威性、内容真实性、来源可追溯性等，

[来源：T/CAACCHINA 001-2026，通论及术语与定义 3.2.6]

3.5 信源 information source

产生或承载语料信息的原始载体。

注：信源包括但不限于品牌官方网站、权威媒体、行业数据库、官方认证文档、经审核的用户生成内容（UGC）等。

[来源：T/CAACCHINA 001-2026，通论及术语与定义 3.2.4]

3.6 语料投毒 corpus poisoning

通过批量注入虚假、误导性、低质量或恶意操纵性内容，干扰生成式 AI 模型认知与输出的违规行为。

3.7 语料污染 corpus contamination

通过批量注入虚假、误导性、低质量或恶意操纵性内容，对生成式 AI 模型认知产生的误导结果。

4 可信语料与信源判断的基本原则

4.1 真实性原则

语料内容基于可验证的事实，而非虚构、夸大、误导性表述或虚假数据。

4.2 权威性原则

优先采用高公信力信源，建立权威背书与分级准入机制。

4.3 专业性原则

由第三方机构或具有相应专业资质的专家进行评测、专业评估。

4.4 可追溯原则

语料全生命周期留痕，支持来源溯源、责任倒查与版本管理。

4.5 动态更新原则

建立语料时效性评估与定期更新机制，确保 AI 引用信息的及时性。如无定期更新机制，则分发平台/内容提供方应提供内容的发布日期，其他援引该内容的第三方可清晰知道该内容的发布时间。

4.6 中立性原则

本文件仅对语料质量作客观定义与规范，不做商业价值判断或排名推荐。

5 信源准入与语料质量标准

5.1 信源准入要求

5.1.1 一般行业信源准入要求

信源作为 GEO 生成语料的提供方，需要有完善的内容质量保证机制，包括但不限于：

- a) 有完善的内容审核机制；
- b) 有真实身份背书的创作者体系；
- c) 内容可追溯到具体责任人；
- d) 承载语料的网络内容平台对虚假信息有主动治理机制；
- e) 承载语料的网络内容平台对 AI 生成信息有标识；
- f) 必须具备合法来源与完整授权链；
- g) 依据《中华人民共和国著作权法》要求，禁止使用侵权、盗版、无授权内容。

5.1.2 高风险行业信源要求

医疗医美、食品健康、金融理财、法律、母婴等高风险行业的语料信源，除应遵循 5.1.1 相关要求外，须额外满足相应行业主管部门的资质认定要求，禁止内容清单参考附录 A 各行业细则。

5.2 语料分类与语料可信等级

语料按类型分为：

- a) 输入语料：品牌方提供的原始素材；
- b) 生产语料：服务商生产/优化后投放到公开渠道的内容；
- c) 检索语料：AI 平台从公开内容生态中检索到的内容；
- d) 知识库语料：AI 平台自建语料库；
- e) 提示语料：Prompt 语料。

语料可信等级依据权威性、专业性、真实性、时效性和内容质量五个维度进行综合评定，分为 1 星至 5 星。

表 1 GEO 语料可信等级维度及标准

等级	权威性	专业性	真实性	时效性	内容质量
1 星	—	—	—	—	—
2 星	—	—	高	—	—
3 星	—	—	高	高	—
4 星	高或中	高或中	高	高	—
5 星	高	高	高	高	高

5.3 语料基础质量

5.3.1 语料基础质量包含的维度

基础质量包括以下几个维度：权威性、专业性、真实性、时效性、整体内容质量。

5.3.2 禁止内容清单

禁止内容包括：

- a) 涉政敏感、色情暴力、违法违规内容；
- b) 未授权个人信息、隐私数据、内部涉密信息；
- c) 虚假信息、谣言、夸大效果、绝对化用语、违规承诺、伪造数据、伪造案例、不当比较、商业诋毁、误导消费者；
- d) 侵犯肖像权、名誉权、知识产权内容。

5.3.3 纯文字内容

专属质量考核维度包括：

- 内容可读性：语句通顺、标点规范、无错别字；
- 信息密度：有效事实、数据、观点占比达标；
- 结构化规范：统一版式层级、适配 AI 解析与知识抽取；
- 原创度：明确标注来源，洗稿、全文搬运内容按低质处理；
- 格式规范性。

5.3.4 图文内容

专属质量考核维度包括：

- 图文匹配度：图片内容与正文主题强关联；
- 图片本身质量：清晰度、完整性、视觉规范；
- 图文结构化与标注：标准图注/说明文字；
- 版式与阅读体验。

5.3.5 视频内容

专属质量考核维度包括：

- 画面质量：清晰度、完整性、视觉合规；
- 伴音质量和字幕质量：准确性、规范性、完整性；
- 剪辑与内容逻辑；
- 附加规范：片头片尾、元数据、版权。

5.3.6 音频内容

专属质量考核维度包括：

- 语音音质：信噪比、音量、采样标准；
- 语音语义清晰度；
- 内容完整性与时长规范；
- 转写文本质量；
- 人声与主体规范。

6 语料标注规范

6.1 标注流程

对 GEO 服务商生成的用于 GEO 优化的内容语料进行评估时，需要遵循标注规范。主要流程包括：采集、清洗、标注、质检、入库。

6.1.1 采集

按照既定信源白名单、行业范围、主题词包，定向抓取/收录原始文本、图文、问答、参数、案例等原始内容，形成原始语料池。采集来源包括内容及资讯平台、官网、电商平台、论坛、网站。

6.1.2 清洗

对原始语料做无关信息清洗、去重、脱敏，输出标准化待标注语料。完全重复内容直接删除；高度相似内容（文本重合度>80%）保留信息最全、来源最优、时效性最新的 1 条。

6.1.3 标注

基于 GEO 语料质量规范，对清洗后的合格语料进行打标、补充元数据、分类归档，完成属性定义与结构标准化。采取盲审制度，标注员不应准确知晓语料内容的来源、投放账号等信息，保证标注的中立、客观。

6.1.4 质检

对标注完成的语料做全维度合规及质量抽检/全检：

——常规批次（≤500 条）：全量检查；

——大批量批次（>500 条）：按 10%比例抽检，问题率>3%则扩大至 50%，>5%执行全检；

——高风险领域：100%全检。

6.1.5 入库

将质检合格的标注结果留存入库，支持后续查询、索引，并对本次标注进行结果输出。全流程的操作人员、记录、时间等留痕。

6.2 标注维度与字段统一标准

需覆盖语料来源、监管合规、质量判断、风险评估、营销合规等方面的标注维度与字段。

6.3 综合可信等级标识

基于 5.2 标准进行标注和综合评分。基于本次标注样本语料的可信等级占比情况，判断标注结果，给出综合可信等级标识。

7 可信等级标识

基于标注样本语料可信等级占比情况，给出综合可信等级标识。综合可信等级分为 1 级至 5 级，对应不同的样本量要求和星级样本占比要求。

表 2 不同综合可信等级对应的样本量及星级样本占比

综合可信等级	标注样本量	星级样本占比	标注机构
1 级	≥ 400	$< 50\%$	服务商自测
2 级	≥ 400	$\geq 50\%$	服务商自测
3 级	≥ 600	$\geq 65\%$	第三方检测机构、服务商自测
4 级	≥ 600	$\geq 75\%$	第三方检测机构
5 级	≥ 1000	$\geq 85\%$	第三方检测机构

附 录 A

(资料性)

高风险行业信源要求

A.1 高风险行业定义

包含但不限于：时政舆情、法律法规、金融理财、医疗健康、教育升学、食品安全、安全生产、公共应急、房地产、医美、婚恋心理咨询、特种设备、军工、政务公示等涉及公共安全、人身财产、生命健康、社会稳定、公众重大利益的垂直领域。

A.2 总体准入原则

高风险行业语料不允许低等级、UGC、匿名信源入库训练，实行信源白名单准入、最小范围采集、100%人工复核、全链路可追溯机制。高风险领域语料禁止自媒体、个人账号、匿名社群、小道消息、无资质自媒体内容作为训练信源。

A.3 高风险行业信源分级

A.3.1 最高可信信源

- a) 国家部委、省级主管部门、地方政务公开平台官方发布内容（政策、法条、标准、公告、批复、公示）；
- b) 国家级权威媒体、中央重点新闻网站一类新闻资质平台刊发内容；
- c) 行业主管机构、监管总局、标准化委员会、官方智库发布的行业标准、规范、指南、白皮书、诊疗方案、合规条文；
- d) 具备官方备案、资质牌照、行政许可的行业权威机构公开数据。

A.3.2 次级合规信源

- a) 省级权威媒体、行业协会、学会、官方研究机构公开内容；
- b) 头部持牌合规机构（持金融/医疗/教育资质企业）官方公示内容；
- c) 已被国家级权威信源正式引用、背书、佐证的行业公开资料。

A.3.3 绝对禁止信源

- a) 所有自媒体、UGC 平台、个人博主、匿名账号、社群言论；
- b) 未经核实的行业传闻、解读、二次转述、拼凑整合内容；
- c) 无资质机构发布的科普、建议、攻略、测评、分析结论；
- d) 过期政策、失效标准、作废条文、旧版行业规范；
- e) 带有营销导向、夸大宣传、商业引流、诱导消费的行业内容；
- f) 无法溯源、无发布主体、无时间戳、无资质背书的模糊内容。

附 录 B

(资料性)

语料全生命周期管理

B.1 主要流程

语料全生命周期管理的主要流程为：采集、清洗、标注、入库、使用、归档、销毁。

B.2 分级定义

核心语料：含国家秘密、关键基础设施、生物特征、精准位置、未成年人敏感信息、可直接识别个人的高敏感信息；模型核心训练集、不可再生授权语料。

重要语料：个人一般信息、商业机密、合同数据、标注数据集、付费授权语料、企业内部业务数据。

一般语料：公开网页、无敏感内容的新闻/百科、去标识化且不可复原的语料、公共版权到期内容。

B.3 归档期限

核心语料：归档保存 10 年；涉及诉讼/监管调查的，延长至案件终结后 1 年；法律另有更长要求的从其规定。

重要语料：归档保存 5 年；合同/授权到期后未续约的，最长不超过 3 年。

一般语料：归档保存 2 年；批量公开采集、无业务追溯需求的，可缩短至 1 年。

法律/审计/合规记录（元数据、审批单、销毁日志、授权文件）：永久归档（电子+离线备份）。

B.4 问题语料召回、下线、替换机制

GEO 服务商应建立：

- a) 错误更正机制：一旦发现已投放的语料存在事实性错误或被 AI 错误引用，在 72 小时内启动修正程序，同步更新官网、知识库及外部发布渠道，并记录变更日志；
- b) 应急响应机制：如发生因语料问题导致的重大舆情，如被大模型输出错误价格导致股价波动或消费者投诉，暂停相关 GEO 活动。

B.5 销毁机制

历史数据销毁前应保留一定时间，便于监管要求追责，原则上至少保留 6 个月以上。当满足以下任意条件时，可触发销毁：达到归档期限（到期前 30 天自动预警）；提前归档后无留存必要；内容违法、侵权、隐私泄露风险不可消除；授权/同意永久撤回；监管/法院指令销毁。

附录 C

(资料性)

语料抽样评估量级

C.1 总体原则

本文件语料抽样样本量设计，基于统计学样本量计算公式、SAC/TC260《生成式人工智能服务安全基本要求》、语言服务语料库抽样检验规范制定，兼顾统计置信度、评估精度、行业合规要求与工程落地性。

核心统计依据采用通用可信样本量公式。行业通用基准参数：置信水平 95%、允许误差 5%，该参数为国内外大模型语料质检、文本数据评估的通用基线标准。针对高风险行业语料，提升评估标准至置信水平 99%、允许误差 3%。

C.2 通用语料抽样量级

在 95%置信水平、5%允许误差、最大方差预估 ($p=0.5$) 条件下：

- 语料总量 $N \leq 1000$ 条：全量核查（100%抽样）；
- $1000 < N \leq 10000$ 条：抽样量 ≥ 200 条；
- $10000 < N \leq 100000$ 条：抽样量 ≥ 300 条；
- $N > 100000$ 条：固定抽样 385 条（95%置信、5%误差基准值）。

C.3 高风险行业语料抽样量级

在 99%置信水平、3%允许误差条件下：

- $N \leq 1000$ 条：全量 100%人工核验；
- $1000 < N \leq 10000$ 条：抽样量 ≥ 400 条；
- $10000 < N \leq 100000$ 条：抽样量 ≥ 600 条；
- $N > 100000$ 条：固定抽样 850 条（99%置信、3%误差科学定值）。

C.4 分层抽样规则

采用分层随机抽样方式，禁止纯随机无序抽样。按照语料来源、行业领域、发布时间、敏感等级、内容类型分层，各层级按占比比例抽样；

针对新增迭代语料：月度/季度更新语料<1 万条，固定抽样 200 条；更新语料 ≥ 1 万条，执行通用抽样标准；

问题复盘抽样：历史检出不合格、高风险疑点语料，实行 100%复检抽样。

C.5 抽样结果判定阈值

- a) 普通语料：抽样不合格率 $\leq 3\%$ ，判定为质量合格；不合格率 $> 3\%$ ，启动全量复检与整改；
- b) 高风险语料：抽样不合格率 $\leq 1\%$ ，判定为质量合格；检出任意严重违规内容，直接判定批次不合格，全量清查。

参考文献

- [1] GB/T 1.1—2020 标准化工作导则[S]. 北京: 中国标准出版社, 2020.
- [2] GB/T 41867—2022 信息技术 人工智能 术语[S]. 北京: 中国标准出版社, 2022.
- [3] GB/T 45654—2025 网络安全技术 生成式人工智能服务安全基本要求[S]. 北京: 中国标准出版社, 2025.
- [4] 《中华人民共和国广告法》[Z]. 全国人民代表大会常务委员会.
- [5] 《中华人民共和国个人信息保护法》[Z]. 全国人民代表大会常务委员会.
- [6] 《中华人民共和国数据安全法》[Z]. 全国人民代表大会常务委员会.
- [7] 《生成式人工智能服务管理暂行办法》[Z]. 国家互联网信息办公室.
- [8] 《标准化法》[Z]. 全国人民代表大会常务委员会.
- [9] 《团体标准管理规定》[Z]. 国家标准化管理委员会.
- [10] SAC/TC260 生成式人工智能服务安全基本要求[S].