生成式引擎优化 · 方法论研究

生成式引擎优化（GEO）全链路方法论白皮书

Generative Engine Optimization
A Full-Chain Methodology Whitepaper

8 个中文生成式引擎 · 9 个行业场景 · 1,061 个监测任务
3,155 条完整引擎回答 · 56,469 条原始信源条目

信向XGEO
沈阳信向智能科技有限公司
GEO 方法论研究组

版本 2.1 · 2026 年 11 月
xgeo.vip · 地球 XGEOVIP

CONTENTS

目录

前言

本版更新（2.1） 4
数据规模（截至 2026 年 11 月 14 日） 5
摘要 6

正文

01 **引言** 8
1.1 背景与问题 8
1.2 研究目的 8
1.3 研究范围与边界 8
1.4 报告结构 9
1.5 术语约定：AI 原生、AI 数字化营销与 AI 获客 9

02 **研究方法** 10
2.1 研究设计 10
2.2 文献研究 10
2.3 实证数据采集 10
2.4 分析方法 11
2.5 研究局限 11

03 **理论基础与文献综述** 13
3.1 GEO 的概念界定 13
3.2 生成引擎的工作机制 13
3.3 现有研究的证据分布 14
3.4 本研究的理论定位 14
3.5 2026 年行业现状：从概念验证到合规化 14
3.6 2026 年 10-11 月新增文献与证据修正 15

04 **可见性的生成机制** 17
4.1 从排名可见性到引用可见性 17
4.2 四阶段模型 17
4.3 可见性的概率属性 18
4.4 从可见性到 AI 获客：作用边界 18

05 **中文生成式引擎的实证发现** 20
5.1 引用行为的整体特征 20
5.2 品牌提及与情感分布 20
5.3 品牌提及位置的两种模式 21
5.4 信源分布特征 22
5.5 引擎信源偏好差异 23

06 **内容要素与引用可见性的关系** 25
6.1 主题相关性 25
6.2 产品参数与品牌身份 25
6.3 来源权重 25
6.4 内容结构与数据要素 26
6.5 内容形态 26

07 引擎风格差异与内容适配 27
7.1 风格差异的量化依据 27
7.2 八个引擎的风格特征 27
7.3 适配原则与边界 28

08 可见性诊断框架 29
8.1 框架设计原则 29
8.2 六个维度 29
8.3 维度间关系 29
8.4 应用方法 30

09 测量方法论 31
9.1 测量原则 31
9.2 抽样与重复 31
9.3 时效管理 31
9.4 常见测量偏差 32
9.5 长尾提问的测量设计 32
9.6 数据边界与可复核性 32

10 风险、治理与合规 34
10.1 操纵性技术 34
10.2 监管框架 34
10.3 合规边界 34
10.4 服务模式与责任边界 35

11 方法论的工程化落地 36
11.1 为什么方法论必须工程化 36
11.2 从四阶段模型到三层诊断 36
11.3 从六维框架到可读的读数 36
11.4 从策略库到内容生成 37
11.5 从测量方法到批次监测 37
11.6 AI 原生：产品形态与数据边界 37
11.7 长尾提问的经济性 38
11.8 榜单类内容的写法 38
11.9 软件不替代什么 38

12 结论 40

REF 参考文献 41
新增文献（2026-10 下旬至 2026-11，作者信息待全文入库后补齐） 42

APX 附录 44
A 术语表 44
B 数据说明 44
C 文献索引 45
D 方法论到产品的落地对照 46
E 对外表述边界 46

报告信息

版本	2.1（2026-11-15 数据快照更新）
发布日期	2026 年 11 月
编制单位	沈阳信向智能科技有限公司 GEO 方法论研究组
文献基础	105 篇文献（90 篇已完成 PDF 入库，2023 年 11 月至 2026 年 10 月；15 篇为 2026 年 10 月下旬至 11 月新增高相关文献）
实证基础	8 个中文生成式引擎、9 个行业场景、1,061 个监测任务、3,155 条完整引擎回答、3,155 条结构化分析、56,469 条原始信源条目、13,725 条引用记录
行业观察	信向XGEO 行业观察账号 1,002 条回答，4,015 条竞品提及，831 个去重名称（2026-09-25 至 2026-11-14）
数据快照	2026-09-19 至 2026-11-14；2026-11-15 复核
替代版本	2.0（2026 年 10 月）；1.0（2026 年 6 月，基于 22 篇文献）
引用格式	本报告参考文献参考 APA 第 7 版格式；新增文献作者信息待补

本版更新（2.1）

2.1 版在 2.0 版的基础上做了十类更新。

1. 行业环境：补充 2026 年 3·15 曝光、四套标准体系的实施进展，以及 2026 年 11 月的市场观察。
2. 作用边界：新增「从可见性到 AI 获客」一节，区分目标与手段，明确 GEO 不能承诺什么。
3. 测量方法：新增长尾提问的测量设计与数据可复核性要求。
4. 工程落地：新增「方法论的工程化落地」一章，说明四阶段模型、六维诊断、策略库与测量方法如何变成可执行的软件流程。
5. 表述规范：新增对外表述边界附录，统一「AI 原生」「AI 数字化营销」「AI 获客」等词的使用方式。
6. 新文献补录：新增 3.6 节，纳入 2026 年 10 月下旬至 11 月的 15 篇高相关文献，并据此修正三条边界。
7. 数据规模面板：新增真实数据规模表，区分原始条目、去重记录与结构化样本。
8. 数据快照更新：将实证数据统一到 2026-11-14 生产库快照，更新回答级、信源级与引擎级统计。
9. 数据合理性复核：补充表内加总、分子分母与口径说明；公开版本以 9 个行业场景表述覆盖范围，不列品牌账号数。
10. 信向**XGEO** 判断：在关键章节穿插信向XGEO 对投入顺序、诊断顺序、引擎适配与数据边界的判断。

> 本版不改变 2.0 版的核心实证结论，只补充行业背景、边界条件与工程化路径。

数据规模（截至 2026 年 11 月 14 日）

本报告的实证数据来自 XGEO 商业监测环境的生产数据快照，查询复核日期为 2026 年 11 月 15 日。按统计口径分为：

口径	数量	说明
监测任务	1,061 个	按任务去重
完整引擎回答	3,155 条	回答级主样本
有效回答文本	3,145 条	回答文本非空；10 条为空或失败
结构化分析记录	3,155 条	品牌提及 512 条，含情感标注
自然提问	479 个	实际执行过的提问去重
题库提问	499 个	当前有效题库
中文生成式引擎	8 个	豆包、DeepSeek、通义千问、腾讯元宝、文心一言、智谱、Kimi、小米 MiMo
覆盖行业场景	9 个	医疗体检、中医健康、餐饮、教育、装修、美业、影视、婚庆摄影、AI/GEO 软件
行业观察信源	22,808 条	信向XGEO 账号，通用口径
行业观察引用	3,556 条	信向XGEO 账号，通用口径
原始信源条目	56,469 条	含跨任务重复，未去重
去重信源 URL	33,497 个	URL 去重
去重信源域名	6,956 个	域名归一化后去重
引用记录	13,725 条	按「回答-URL」去重
引用 URL	9,096 个	URL 去重
竞品提及	11,162 条	回答文本中的竞品/名称提及
去重竞品/名称	1,924 个	跨行业
信源详情记录	36,500 条	图文、视频与未分类
域名库	10,470 个	可检索域名
策略结果	6,332 条	策略匹配与生成结果
内容生成结果	365 条	内容生产记录

> 口径提醒：原始信源条目含跨任务重复，不能与去重信源 URL 相加；引用记录按「回答-URL」去重，与原始引用条目不同。对外引用时，以「3,155 条完整回答、56,469 条原始信源条目、13,725 条引用记录」为主口径；去重后为「33,497 个信源 URL、9,096 个引用 URL」。不同口径的数字不能互相替代。
> 计划扩样：完整回答拟扩至 10,400 条（500 个核心提问 × 8 个引擎 × 2 轮 + 100 个长尾提问 × 8 个引擎 × 3 轮），详见《数据扩样方案》。在完成采集与复核前，本报告不将计划值计入实证结果。

摘要

生成式引擎正在成为越来越多用户获取信息的重要入口，并在部分场景中替代传统搜索。这一转变使内容可见性的评价对象从「搜索结果排名」迁移到「生成答案中的引用」，并由此产生了生成式引擎优化（Generative Engine Optimization，GEO）这一新的研究与实践领域。

本报告在 105 篇文献基础（90 篇已完成 PDF 入库，其中 85 篇完成全文解析；15 篇为 2026 年 10 月下旬至 11 月新增摘要级文献）上，结合对 8 个中文生成式引擎、9 个行业场景、1,061 个监测任务的大规模实证采集，尝试回答三个问题：生成引擎的可见性由哪些阶段构成；哪些内容要素与引用可见性存在稳定关联；中文市场的引用行为呈现何种特征。本报告同时讨论这些发现如何工程化落地，以及「AI 原生」「AI 数字化营销」「AI 获客」三个商业词与方法论的关系。

主要发现如下。

第一，**GEO 是多阶段过程，各阶段证据强度差异显著。** 依据 Martinez（2026）对 45 项研究的系统综述，内容改写对「已被检索内容的引用与吸收」具有较强因果证据，对「自然可发现性」证据中等，对「流量与商业转化」证据薄弱。这一梯度意味着，把条件效应等同于端到端效果，是当前实践中最常见的推论错误。

第二，**可见性呈概率分布特征，单次观测不具代表性。** Schulte 等（2026）与 Żatuchin（2026）的实证均表明，在完全相同条件下重复运行，最优推荐结果存在显著波动。本报告的测量框架据此建立了重复测量与分布报告原则。

第三，**中文生成式引擎的信源分布高度分散，但存在明确的引擎级偏好。** 为避免单一行业样本占比过高影响通用结论，本报告在信向XGEO 行业观察账号的 22,808 条信源记录中统计：最高单一域名占比为 4.3%，分引擎观察时，豆包、文心一言、腾讯元宝等部分引擎呈现出与其所属生态高度相关的来源偏好：字节系引擎的信源中抖音内容占 24.2%，百度系引擎的百家号占 26.2%，腾讯系引擎的微信公众号占 12.1%。但这一规律并非普遍适用：通义千问的首位来源是 ithome.com（18.0%），属于科技媒体而非阿里自有生态；小米 MiMo、DeepSeek、Kimi、智谱的首位来源也以博客园、知乎专栏、行业媒体为主。生态归属能解释一部分偏好，不能解释全部。这一发现为内容分发策略提供了可操作依据。

第四，**被监测品牌的提及位置可区分为「首位集中」与「中段可见」两类模式。** 这是品牌级观测：查询特定品牌名后，看它在回答推荐名单中排第几，不是对回答体裁的分类；换一个品牌重查，位置分布可能不同。这里的「首位」指被监测品牌在名单中的排位，不是名单里只有一个品牌。首位集中型中，品牌一旦出现，几乎都排在第一位；中段可见型允许品牌在中段与末段被带入。两类模式对内容策略的要求不同。

第五，**内容要素中，主题相关性与来源权重的影响力高于品牌身份。** Chu 与 Hou（2026）的回归分析显示，产品参数可解释推荐排名方差的 82.4%，品牌身份仅解释 1.2%。Song 等（2026）进一步表明，来源偏好的影响可以超过内容本身的质量。

第六，**行业已从概念验证进入合规化阶段。** 2026 年 3·15 晚会曝光了 GEO 领域的虚假信息投放与数据投毒乱象；T/CAPT 026—2026 于 2026 年 8 月 12 日实施，T/CAAAD 046/047/048-2026 于 2026 年 10 月 23 日实施，T/CAACCHINA 001-005-2026 于 2026 年 9 月 14 日公布。合规不再是加分项，而是服务能否持续的前提。

补充：**2026** 年 **10** 月下旬至 **11** 月的新文献收紧了三条边界。 2023 年的 GEO 效果量不能直接用于今天；跨引擎一致率与顺序敏感性必须单独报告；引用不等于被说对，也不等于抓取成功。

本报告的结论是：GEO 能够改善「已被检索内容的利用效率」，但不足以稳定地改变「内容是否被检索」。清晰区分这两个层次，是制定合理投入策略的前提。2026 年的行业变化进一步说明，GEO 的竞争焦点正在从「能不能被引用」转向「能不能在合规边界内被持续验证」。

> 信向**XGEO** 判断：GEO 的投入顺序应当是「先保证被检索，再优化被引用，最后才谈转化」。这三步是排查顺序，不是「把同一篇文章改三遍」，也不是「只能改原文、不能发新文」：先确认页面是否真的进入检索池；已进入但未被引用，优先优化被实际检索或引用的那一页；已被引用但不转化，再谈承接页与转化路径。发新文章属于扩大覆盖，解决的是「可被检索的页面太少」，不能替代对已检索页面的引用优化。把顺序倒过来做，是 2026 年最常见的预算浪费。

关键词：生成式引擎优化；AI 搜索可见性；引用行为；信源分析；内容策略；AI 原生；AI 获客

01
引言

1.1 背景与问题

传统搜索引擎以链接列表的形式呈现结果，用户通过点击进入目标页面获取信息。在这一范式下，内容方的优化目标是提升页面在结果列表中的排序位置，搜索引擎优化（Search Engine Optimization，SEO）由此形成完整的产业与方法体系。

生成式引擎改变了这一流程。用户以自然语言提出完整问题，引擎自行完成检索、阅读、综合与生成，直接输出自然语言答案，并在答案中标注引用来源。用户不再需要点击链接，内容方也不再以获得点击为目标。

这一变化带来的核心问题是：当答案由引擎生成而非由用户选择时，「可见性」如何定义与测量？

1.2 研究目的

本报告旨在建立一套可用于实践的 GEO 分析框架，具体目标包括：

- 明确 GEO 的作用边界，区分有证据支持的环节与证据薄弱的环节
- 系统梳理内容要素与引用可见性之间的关联强度
- 提供中文市场的实证数据，补充以英文市场为主的现有研究
- 建立可复现的测量方法，避免单次观测导致的错误结论

1.3 研究范围与边界

研究对象：中文生成式引擎的答案生成与引用行为。

数据范围：文献覆盖 2023 年 11 月至 2026 年 11 月公开发表的 105 篇文献，其中 90 篇已完成 PDF 入库，15 篇为 2026 年 10 月下旬至 11 月新增高相关文献；实证数据采集自 8 个中文生成式引擎，采集截止时间为 2026 年 11 月 14 日。

不在本报告范围内的内容：

- 传统 SEO 的技术实现
- 生成引擎的模型训练与架构
- 具体的商业定价与服务方案
- 对特定引擎未来行为的预测

1.4 报告结构

第 2 章说明研究方法与数据来源；第 3 章梳理理论基础与既有研究；第 4 至 7 章呈现本报告的核心发现；第 8 至 9 章给出诊断框架与测量方法；第 10 章讨论风险与合规；第 11 章说明方法论的工程化落地；第 12 章为结论。

1.5 术语约定：AI 原生、AI 数字化营销与 AI 获客

本报告使用三个容易混淆的词，先做约定。

AI 原生：产品从设计第一天就围绕 AI 引擎的回答机制展开，而不是给传统 SEO 工具加一个 AI 模块。

AI 数字化营销：在本报告中，特指用户通过 AI 搜索了解、比较、选择产品或服务这一环，不包括投流、私域运营与广告投放。

AI 获客：上述过程中，品牌进入 AI 的回答与推荐，从而获得询盘与成交机会。它是目标，GEO 是手段。

三个词的关系是：**AI** 原生是产品设计方式，**AI** 数字化营销是场景，**AI** 获客是目标，**GEO** 是方法。四者不能互相替代，也不能用其中一个词掩盖另一个词的限制。

02
研究方法

2.1 研究设计

本报告采用文献研究与实证研究相结合的设计。

文献研究用于建立理论框架并界定证据边界。实证研究用于获取中文市场的第一手数据，补充现有研究以英文市场为主的不足。

两类研究在报告中承担不同功能：文献研究回答「什么已被证明」，实证研究回答「中文市场的实际情况如何」。 两者结论相互印证时，可信度较高；出现分歧时，本报告会明确标注。

2.2 文献研究

文献来源：arXiv 预印本平台、ACM 与 ACL 系列会议论文集。

时间范围：2023 年 11 月（GEO 概念提出）至 2026 年 11 月。

筛选标准：

一 主题与生成式引擎的可见性、引用、内容优化直接相关
二 具备可量化的实证结果或明确的方法论贡献
三 排除仅涉及传统 SEO、纯 RAG 检索技术、与内容策略无关的研究

文献规模：符合标准的文献共 105 篇。其中 90 篇已完成 PDF 入库，85 篇已完成中文解析；15 篇为 2026 年 10 月下旬至 11 月新增高相关文献，待完成 PDF 入库与全文解析。

综述方法：以 Martinez（2026）的系统综述作为证据分级的主要依据，该综述覆盖 45 项研究，提出了按「条件效应、总量效应、商业结果」三层划分的证据框架。本报告沿用这一框架。

2.3 实证数据采集

采集对象：8 个中文生成式引擎，分别为豆包、DeepSeek、通义千问、腾讯元宝、文心一言、智谱、Kimi、小米 MiMo。

采集方式：通过客户端在真实使用环境下发起自然语言提问，记录引擎返回的完整答案、引用来源、回答结构与生成时间。

样本规模：

数据类别	数量
监测任务	1,061 个
完整引擎回答	3,155 条
有效回答文本	3,145 条
结构化分析结果	3,155 条
原始信源条目	56,469 条
去重信源 URL	33,497 个
信源详情记录	36,500 条
引用记录	13,725 条
自然提问	479 个
当前有效题库	499 个
覆盖行业场景	9 个

样本性质说明：本报告实证数据来源于商业环境下的真实监测记录，覆盖医疗体检、中医健康、餐饮、教育、装修、美业、影视、婚庆摄影、AI/GEO 软件等 9 个行业场景。样本并非随机抽样，不具备推断全行业总体分布的能力。报告中所有基于该数据的结论，均应理解为「在采集样本中观察到的现象」，而非行业普遍规律。

2.4 分析方法

信源分析：对每条信源记录的 URL 进行域名归一化处理（去除 www 前缀、统一小写），按引擎分组统计域名分布与集中度。

引用分析：统计同一 URL 在样本中被引用的次数分布，区分「单次露出」与「重复引用」。

结构分析：对每条答案标注被监测品牌的提及位置（首位、中间、末尾、未提及）与描述倾向（情感得分，满分 100），据此识别引擎的回答结构特征。

内容特征分析：对答案文本计算参数化描述占比（参数比）与叙述性描述占比（叙事比），以及是否包含价格信息、统计数据等二元特征。

2.5 研究局限

本报告存在以下局限，阅读结论时应予考虑。

第一，样本非随机。 实证数据来自商业监测，品牌构成受采集范围影响，不能代表全行业。

第二，采集时点集中。 数据快照截至 2026 年 11 月 14 日，而生成引擎的模型版本与检索策略持续更新，结论具有时效性。

第三，无法做因果识别。 本报告为观察性研究，未设置随机对照组，因此只能报告相关性，不能确立因果关系。文献部分引用的因果结论，均来自原研究的受控实验。

第四，中文文献占比有限。 90 篇已入库文献中以英文发表为主，15 篇新增文献同样以英文为主；中文市场的专门研究仅有少数几篇，理论框架的适用性仍需更多本土研究验证。

第五，引擎行为不可完全观测。 生成引擎的检索与排序机制未公开，本报告只能从输出端反推，无法验证内部机制。

03 理论基础与文献综述

3.1 GEO 的概念界定

GEO 概念由 Aggarwal 等（2024）在 KDD 2024 发表的论文中系统提出，定义为「以内容创作者为中心、通过优化内容提升其在生成引擎回复中可见度的方法框架」。

该研究将生成引擎形式化为一个函数：接收用户查询与个性化信息，返回综合答案并附内联引用。在此框架下，内容方追求的目标从「排序位置」转变为「在生成答案中被引用的程度」。

与 SEO 的核心差异可归纳如下：

维度	SEO	GEO
优化目标	提升搜索排名	提升生成答案中的引用可见度
可见性度量	排名位置（线性列表）	引用词数、位置、影响力（多维）
优化对象	关键词、反向链接、元数据	内容结构、证据密度、语义对齐
效果稳定性	相对稳定	随生成轮次概率性波动
可观测性	有第一方工具	黑箱，需代理测量

来源：Aggarwal 等（2024）。

3.2 生成引擎的工作机制

Grossman 等（2026）基于 11,500 条查询的实证研究揭示，生成式搜索与传统搜索在信息来源上存在系统性差异：

> AI Overview 在 51.5% 的真实用户查询中被生成；传统搜索结果页、AI Overview 与 Gemini 三者检索来源的 Jaccard 相似度仅为 0.11 至 0.18；生成式搜索更偏好小众来源，传统搜索更依赖流行与权威网站。

这一发现的方法论含义是：**GEO 不能直接沿用 SEO 的经验。** 两类系统在来源选择机制上存在实质差异。

3.3 现有研究的证据分布

Martinez（2026）对 2023 年 11 月至 2026 年 7 月期间的 45 项研究进行了系统综述，提出的核心判断是本报告证据分级的基础。

该综述将 GEO 形式化为一个随机、部分可观测的流水线，涵盖搜索激活、抓取与索引、检索、重排序与上下文分配、引用、显著性、事实吸收、忠实度与用户行为等环节。并指出，各环节的证据强度存在显著梯度：

以上下文为条件	证据强	内容改写可因果性改变引用与吸收
某些信息属性	证据中等	结构、数据、时效存在一定作用
传导至流量	证据薄弱	缺乏稳定的纵向、跨平台因果证据

该综述同时指出了基础论文结论被误用的具体方式：Aggarwal 等（2024）报告的「最高 40% 可见度提升」，其成立条件是在模拟器环境中、且五篇文档已被预先置于上下文，测量的是相对可见度增益，而非端到端效果。

3.4 本研究的理论定位

在上述框架下，本报告的定位如下：

接受分层证据观。 不将条件效应推广为端到端承诺，在各章结论中明确标注证据层级。

补充中文市场实证。 既有研究以英文市场为主，中文生成式引擎的引用行为研究仅有 Zhen 等（2026）一项大规模实证。本报告提供的商业环境数据可作为补充。

强调可复现性。 第 9 章的测量框架遵循 Martinez（2026）提出的协议建议，包括前置定义估计目标、多引擎多复述设计、重复测量与原始证据留存。

3.5 2026 年行业现状：从概念验证到合规化

截至 2026 年 11 月，GEO 行业呈现四个特征。

其一，需求已经出现，预算尚未大规模落地。 CTR 2026 年调研显示，70.4% 的广告主规划了 GEO 预算，14% 视为刚需，近四成因「效果评估难」观望。市场不是没有需求，而是缺少可复核的测量。

其二，传统搜索预算在迁移，但 **AI** 搜索的商业化尚未成熟。 百度在线营销收入连续九个季度同比负增长，2026Q2 同比 -19%；同期中国互联网广告投放仍保持增长（2026H1 3,853.4 亿元，同比 +7.1%）。QuestMobile 2026 年 5 月数据显示，AI 原生 App 月活 4.99 亿、同比增长 85.4%，月人均使用 92.7 次；传统搜索人均次数下降 19.1%、时长下降 13.5%。钱没有消失，是在改路；但流量迁移不等于商业闭环已经形成。 平台自身的 AI 营销收入增速也从高位回落，说明变现模式仍在探索。

其三，服务形态高度分散。 2026 年 9 月 19 日至 11 月 14 日，我们累计采集 3,155 条完整引擎回答，对应 56,469 条原始信源条目、13,725 条引用记录；其中信向XGEO 行业观察账号的 1,002 条回答对应 22,808 条信源记录、3,556 条引用记录，共出现 4,015 条竞品提及、831 个去重名称，覆盖全国全案服务商、工具与 SaaS 监测平台、区域本地服务商、海外监测工具、云厂商与平台型产品等形态。竞品提及只统计回答文本；信源级统计见第 5 章。市场认知高度分散，尚无统一的服务标准与效果口径。

其四，合规成为分水岭。 2026 年 3·15 晚会曝光了 GEO 领域的虚假信息投放与数据投毒乱象；中国信通院将滥用手段归纳为虚假内容制造、批量内容轰炸与隐性操纵优化三类。与此同时，T/CAPT 026—2026 于 2026-08-12 实施，T/CAAAD 046/047/048-2026 于 2026-10-23 实施，T/CAACCHINA 001-005-2026 于 2026-09-14 公布。合规不再是加分项，而是入场条件。

对从业者的含义： 把 GEO 当作「批量发稿」或「给 AI 洗脑」的做法，既面临监管风险，也面临引擎侧识别与降权风险；可持续的路径是把事实基础、内容质量与测量可复核性做好。

> 信向**XGEO** 判断：2026 年的分水岭不是「谁会做 GEO」，而是「谁能把数据来源、测量口径和责任边界写清楚」。没有这三样，规模越大，合规风险越高。

3.6 2026 年 10-11 月新增文献与证据修正

截至 2026 年 11 月，文献库完成 PDF 入库 90 篇（入库截止 2026-10-08）。2026 年 10 月下旬至 11 月，arXiv 上又出现 15 篇与本报告直接相关的高相关文献。两者合计 **105** 篇，构成 **2.1** 版的文献基础。 15 篇新增文献没有改变本报告的基本框架，但收紧了三条边界。

编号	文献	核心发现	对本报告的影响
2611.01126	LIVE-GEO	活网页 GEO 评测：CORE Top-1 66.3%–76.3%；三个引擎只有 48.8%–66.3% 选中同一款产品；ChatGPT 对候选顺序敏感（73.8% vs 93.8%）	跨引擎一致率与顺序敏感性必须单独报告
2611.00884	When 200 OK Is Not an Answer	10,307 个 URL 中 20.9% 的 200 响应没有正文；静默引用漂移率被低估约 3.8 倍	采集判活不能只看状态码
2611.02643	SearchInject	语言模型搜索中的间接提示注入与检索暴露控制	检索侧安全边界需纳入治理
2611.01582	Good Abandonment	53.2% 的「未点击」其实是成功；点击只有 6.5% 不成功	不能把点击率当唯一成功指标
2611.04225	IntentGEO	按意图分层的 GEO 推荐	长尾提问需按意图分层
2611.04005	Implicit Brand Injection	隐性品牌注入式对抗 GEO	品牌提及可能被操纵
2611.07614	The Popular Get More Popular	生成搜索中的品牌偏好审计	品牌身份仍有偏差效应

编号	文献	核心发现	对本报告的影响
2611.08804	Generative Search and Recommendation at Shopify	工业级生成搜索与推荐实践	行业落地已进入平台内部
2611.11955	CHO-Bench	上下文提示会改变 LLM 广告排名	商业上下文是独立变量
2611.10715	User Satisfaction with AI Overviews	从行为到满意度的判断	成功指标需多维
2611.11396	Search-by-Answer Bias	检索模型中的「按答案搜索」偏差	来源选择存在结构性偏差
2610.38090	Predicting What Answer Engines Cite	10 引擎同候选集：引擎间一致性 Spearman 0.704；最强打分器 0.575–0.578	引用份额可预测，但上限仍低
2610.39339	Does the Citation Matter?	3,611 条健康问题：AIO 出现 96.4%；传统前五仅 44.4% 被引用；56.7% 引用语句被支撑；20.8% 事实句无引用；准确性与支撑度不相关	「被引用」不等于「被说对」
2610.38361	Manifest-Assisted Retrieval	llms.txt：45 次工具试验零自动抓取；97% 域名一个月无请求；21/24 对抗响应被采纳	llms.txt 不是可见度杠杆，且构成注入面
2610.40164	Who Owns Digital News	77 国、9,761 家媒体、7 个 AI 系统；AI 引用集中度最高为流量 2.08 倍	引用分配具有市场结构效应

修正一：**2023** 年的 **GEO** 效果量不能直接用于 **2026** 年。 2609.07559（补录文献，不在上表 15 篇内）把 2023 年那批 GEO 改写效果量搬到十族现代引擎上重测，发现这些杠杆在任何一个引擎上都没有移动引用；无查询条件下的内容分与引用结果的 within-query Spearman 仅 0.11。这进一步支持第 4 章的边界：不要把旧论文的效果量当作现期承诺。

修正二：跨引擎一致率与顺序敏感性必须单独报告。 LIVE-GEO 的三个引擎只有 48.8%–66.3% 选中同一款产品；2610.38090 的引擎间一致性为 0.704，但最强打分器也只有 0.575–0.578。一个综合分不能代表多个引擎的分歧。 报告应至少给出分引擎读数、引擎间一致率、同一候选集在不同顺序下的波动。

修正三：引用不等于正确，也不等于抓取成功。 2610.39339 显示，AI 概览引用了来源，但只有 56.7% 的引用语句被所引页面支撑，20.8% 的事实性语句完全没有引用，准确性与支撑度不相关；2611.00884 显示，20.9% 的 200 响应没有正文。「引了没」「引对没」「抓到没」是三个不同的问题，必须分开测。

> 说明：本节 15 篇新增文献以 arXiv 摘要与每日情报记录为准，PDF 入库与全文精读待后续版本补齐；105 篇文献基础中，90 篇的入库统计口径不变。

04 可见性的生成机制

4.1 从排名可见性到引用可见性

在生成式引擎中，内容方追求的并非单一指标，而是多个可以独立变化的变量：

是否被检索到	内容是否进入引擎的候选上下文
是否被引用	引擎是否在答案中标注该来源
被引用多少	引用的强度与篇幅
引用位置	在答案结构中的显著程度
是否被吸收	内容是否实际影响了答案的表述

上述变量之间存在递进关系，但不存在必然的传递。一篇内容可以被检索到但未被引用，可以被引用但未被吸收，可以在答案中被广泛采纳却未标注来源。

对这一多维结构的认识，是本报告后续所有分析的前提。

4.2 四阶段模型

综合 Martinez（2026）的流水线模型与 Grossman 等（2026）的实证发现，本报告将 GEO 归纳为四个阶段。

第一阶段：可发现性。 引擎爬虫能否抓取页面、索引能否收录。这一阶段的机制与传统 SEO 高度重叠。

第二阶段：检索。 用户提问时，内容能否进入引擎的候选上下文。影响因素包括主题相关性、来源权重与时效性。

第三阶段：引用与吸收。 进入上下文后，引擎是否引用、是否采纳其表述、引用在答案中的位置。

第四阶段：行为转化。 用户接触答案后是否产生后续行为。

四阶段划分的意义在于证据强度的差异。 第二、三阶段存在较多受控实验支持，第四阶段缺乏可信的因果识别。将四个阶段混为一谈，是实践中产生过度承诺的主要来源。

排查顺序与干预手段。 四个阶段对应不同的瓶颈与不同的动作，不能混用。第一阶段与传统 SEO 高度重叠，不应把索引与抓取问题包装成 GEO 的成果。

观察到的情况	瓶颈判断	优先动作
目标页面未被索引或抓取	可发现性	修索引、抓取、站点结构与平台可检索性；此时改稿或加发同类内容通常无效
已进候选上下文但未被引用	检索后的选择	优化被实际检索到的那一页：补事实、补结构、补直接答案、统一口径、增加可核验来源
被引用但未被吸收或位置靠后	引用质量	调整表述方式、答案结构与证据密度；不是继续堆量
已被引用但不转化	承接与信任	修品牌表述、承接页、报价与行动路径；不是继续发文章

「发新文章」解决的是覆盖问题，不是引用问题。当页面数量不足、长尾问法没有对应内容时，扩量是对的；当已有页面被检索却不被引用时，继续发同类文章只是增加未被引用的库存。先看引用 URL，再决定改哪一页——引擎实际引用的页面，不一定是你最初发布的那一篇。

4.3 可见性的概率属性

生成引擎的输出具有内在随机性。Schulte 等（2026）明确指出，AI 搜索的可见性是概率分布而非单点值，单次观测不可靠。

Żatuchin（2026）提供了量化的证据。该研究在四个出口国家、六种查询语言条件下进行 234 次受控运行，每个条件重复六次：

> 6 个 prompt 中有 4 个在重复运行中首位推荐发生了变化。该变化率在浏览器界面与 API 之间、在开启与关闭联网搜索之间均保持一致，表明不稳定性源于系统本身，而非访问界面。

该结论对测量实践的直接含义是：任何基于单次查询的判断都不具备统计意义。第 9 章据此建立重复测量原则。

2026 年 11 月的 LIVE-GEO 把这一结论推进到跨引擎层面：三个生成引擎只有 48.8%–66.3% 选中同一款产品；ChatGPT 对候选顺序敏感，两种相反顺序下 Top-1 分别为 73.8% 与 93.8%。报告不能只给一个综合分，至少要给分引擎读数与引擎间一致率。

4.4 从可见性到 AI 获客：作用边界

在商业语境中，GEO 经常被表述为「AI 获客」。这一表述有价值，但必须区分目标与手段。

AI 获客，指用户通过 AI 搜索了解、比较、选择产品或服务时，品牌进入 AI 的回答与推荐，从而获得询盘与成交机会。它是目标。

GEO，是影响这一过程的方法与工具。它是手段。

从第 4.2 节的四阶段模型看，GEO 的作用主要集中在第二、三阶段：让内容进入候选上下文，并提高被引用与吸收的概率。第四阶段的行为转化，受产品、价格、渠道、销售与信任等多种因素影响，现有研究尚缺乏可信的因果识别。

因此，本报告对「AI 获客」的表述设定三条边界：

可以写	品牌在 AI 回答中的可见度、描述准确度与引用来源
可以写	从监测、诊断、策略、内容到回测的闭环
可以写	长尾提问覆盖带来的潜在机会
不要写	保证获客数量、询盘量、转化率或成交额
不要写	把「被引用」直接等同于「被选择」
不要写	用单次案例推断整体效果

一句话：**GEO 能提高被看见、被正确理解的概率，不能替客户完成成交。**

05
中文生成式引擎的实证发现

5.1 引用行为的整体特征

Zhen 等（2026）对中文生成式引擎的研究是目前该领域规模最大的实证。研究设计如下：

平台	四个中国主流大模型平台
界面	每平台 Web 与 App 两个界面，共 8 个界面
查询	614 个受控查询
重复	每个查询与界面组合固定重复 3 次
原始记录	214,119 条
主数据集	160,860 条引用级记录

该研究的主要发现包括五项。

其一，引用池到答案的选择率极低。引用池中的品牌被纳入答案的整体选择率为 8.3%；含联系方式的来源中，仅 12.4% 将联系方式贡献给答案。

其二，传统 **SEO** 指标不具备预测力。研究以 5118 与百度站长平台的站点评级构建综合质量分，结果显示该变量不是任何被考察结果的主要预测因子。

其三，引用存在半衰期，且与查询时效性相关。高时效性查询的被引页面拟合半衰期约为 39 天，低时效性查询约为 68 天。

其四，相当比例的曝光无法归因。约 13% 的品牌曝光无法匹配至同期引用池，约 71% 的联系方式曝光无法匹配至抓取正文。

其五，访问界面构成独立分层维度。同一平台的 App 与 Web 界面，来源集合存在系统性差异。

5.2 品牌提及与情感分布

本报告采集的 3,155 条结构化分析结果显示，各引擎在品牌描述倾向上存在显著差异。

引擎	样本量	平均情感得分	首位提及率
豆包	585	67.6	14.2%
通义千问	574	56.8	8.5%
腾讯元宝	569	64.0	13.9%
DeepSeek	500	55.8	9.8%
文心一言	269	71.7	14.9%
小米 MiMo	260	49.6	13.1%
智谱	256	47.8	13.3%
Kimi	142	56.3	13.4%

注：情感得分为 0 至 100 的标准化评分，反映答案对品牌的描述倾向；平均情感得分仅统计品牌被提及的样本（全库 512 条），未提及样本不参与情感均值；样本量为该引擎的回答总数；首位提及率指品牌出现时位于推荐序列首位的样本占该引擎样本量的比例。各引擎样本量不等（142–585 条），引擎间比较以各自样本为限，不做事后加权。

两点观察值得注意。

情感得分与首位提及率并不同步。 文心一言的首位提及率为 14.9%，情感得分 71.7，两项均居前列；智谱的首位提及率为 13.3%，处于中游，情感得分 47.8，位列末位；通义千问的首位提及率仅 8.5%，情感得分 56.8，处于中游。这表明「被推荐」与「被正面描述」是两个独立维度。

引擎间的描述倾向差异明显。 情感得分最高者（71.7）与最低者（47.8）相差约 24 分，最高者比最低者高约 50%，说明同一品牌在不同引擎中获得的评价存在实质性差异。

> 信向**XGEO** 判断：情感分和首位提及率必须分开看。一个引擎把你放在第一位但描述负面，比不提及更危险；反过来，中段提及但情感正面，仍有转化价值。

5.3 品牌提及位置的两种模式

这是对被监测品牌（信向XGEO）的品牌名查询结果，不是对回答体裁的分类；换一个品牌名重新查询，位置分布可能不同。下面的统计以一条回答为单位。每条回答通常会列出多家品牌；本表统计的不是名单里所有品牌的排位，而是被监测品牌在这份名单中的位置：「未提及」指名单中没有该品牌，「首位」「中间」「末尾」分别指该品牌排在名单第一、中部、最后。举例：通义千问 574 条回答中，522 条没有出现被监测品牌，52 条出现；出现的 52 条里，49 条排在第一位、2 条在中部、1 条在末位。名单本身仍有其他品牌，只是它们不计入本表。在有竞品名称抽取记录的 1,833 条回答中，单条回答列出的品牌数中位数为 5–7 个，均值为 5.1–7.3 个。这说明两类模式的差别不在「名单长短」，而在「被监测品牌能否在中段被带入」。

引擎	未提及本品牌	本品牌首位	本品牌中间	本品牌末尾
豆包	419	83	52	31
通义千问	522	49	2	1
腾讯元宝	471	79	17	2
DeepSeek	442	49	8	1
文心一言	222	40	6	1
小米 MiMo	225	34	0	1
智谱	220	34	0	2
Kimi	122	19	1	0

豆包的被监测品牌在非首位（中间加末尾）出现 83 次，其余七个引擎合计 42 次。单一引擎的非首位提及量约为其余七者之和的两倍，构成明确的结构性差异。

据此可将被监测品牌的提及位置区分为两类。

首位集中型：被监测品牌被提及时，高度集中于第一位，中间与末尾位置几乎不出现。通义千问、小米 MiMo、智谱、Kimi 的中间位置提及量均为 0 或接近 0；DeepSeek 与文心一言以首位集中为主，但分别有 8 次和 6 次中段提及，可视为过渡型。首位集中不等于名单里只有一家企业——回答通常列出 5–7 个品牌，只是被监测品牌一旦进入名单，通常就在第一位。

中段可见型：被监测品牌可以在回答的中段或末段被提到，不限于第一位。豆包是样本中唯一呈现明显中段可见特征的引擎（就本品牌而言）。

两类模式的实践含义不同。 首位集中型中，被监测品牌要么排在首位、要么根本不出现；排在中间或末尾的样本极少。因此策略上要先解决「进入推荐集合」，再解决「排在第一位」。中段可见型中，中段提及仍具价值。

> 信向**XGEO** 判断：首位集中型引擎先解决「进入推荐集合并排第一」，中段可见型引擎再解决「被带到」。同一套内容策略不可能同时打好这两类位置模式。

5.4 信源分布特征

为避免单一行业样本占比过高影响通用结论，本节使用信向XGEO 行业观察账号的 22,808 条信源记录（14,860 个去重 URL、3,727 个域名）进行统计。

通用行业观察样本中，占比最高的十个域名如下：

域名	记录数	占比
baijiahao.baidu.com	973	4.3%
cnblogs.com	895	3.9%
iesdouyin.com	850	3.7%
ithome.com	743	3.3%
sohu.com	516	2.3%
zhuanlan.zhihu.com	516	2.3%
mp.weixin.qq.com	490	2.1%
blog.csdn.net	422	1.8%
cet.com.cn	341	1.5%
toutiao.com	275	1.2%

注：占比 = 记录数 ÷ 22,808（通用行业观察样本信源总数）。

最高占比为 **4.3%**，表明中文引擎的信源结构高度分散，不存在单一主导来源。这一结论与 Zhen 等（2026）关于来源集合差异性的发现方向一致。

按信源类型划分，全库 36,500 条信源详情记录中，图文类占 77.9%（28,450 条），视频类占 3.9%（1,416 条），未分类占 18.2%（6,634 条）。图文仍是主要形态。

引用强度方面，通用行业观察样本的 3,556 条引用记录中，同一 URL 被引用一次的比例为 68.4%，两次为 15.4%，五次及以上仅 4.8%。多数信源为单次露出，未形成持续引用。

> 信向**XGEO** 判断：信源高度分散，意味着「押注一个大站」的策略在中文引擎里不成立。更有效的做法是按引擎生态配置来源组合，而不是追求单一高权重域名。

5.5 引擎信源偏好差异

在通用行业观察样本（1,002 条回答、22,808 条信源）中分引擎观察，信源分布呈现出结构性差异：部分引擎偏好自有生态内容，部分引擎偏好科技媒体与开发者社区，不能一概而论。

引擎	回答数	信源数	域名数	首位来源	首位来源条数	占本引擎	占全库
腾讯元宝	146	3,875	1,044	mp.weixin.qq.com	468	12.1%	2.1%
DeepSeek	147	3,717	796	zhuanlan.zhihu.com	263	7.1%	1.2%
豆包	147	3,519	813	iesdouyin.com	850	24.2%	3.7%
文心一言	112	3,453	565	baijiahao.baidu.com	906	26.2%	4.0%
Kimi	88	2,584	803	cnblogs.com	181	7.0%	0.8%
智谱	104	2,254	892	cet.com.cn	95	4.2%	0.4%
通义千问	146	1,776	226	ithome.com	319	18.0%	1.4%
小米 MiMo	112	1,630	276	cnblogs.com	300	18.4%	1.3%

注：本表回答数为通用行业观察样本（1,002 条），与 5.2 的全库样本量（3,155 条）不是同一口径，两表不能按引擎直接相除或相加。占本引擎 = 首位来源条数 ÷ 该引擎信源数，是各引擎内部的来源集中度；占全库 = 首位来源条数 ÷ 22,808，是单一域名在通用样本中的份额。两列都不是八个百分比相加为 100% 的口径，不能横向相加。

三项规律可归纳如下。

其一，生态归属影响信源偏好，但不是唯一解释。 字节系引擎的第一来源为抖音内容（24.2%），百度系引擎的第一来源为百家号（26.2%），腾讯系引擎的第一来源为微信公众号（12.1%）。这三个引擎的第一来源均为各自生态内的内容平台。但其余五个引擎不符合这一规律：通义千问的第一来源是 ithome.com（18.0%），小米 MiMo 是博客园（18.4%），DeepSeek 是知乎专栏（7.1%），Kimi 是博客园（7.0%），智谱是 cet.com.cn（4.2%）。它们更依赖科技媒体、开发者社区与行业站点，而非母公司自有内容平台。

其二，技术型引擎偏好技术内容。 Kimi 与小米的信源中，博客园、CSDN、知乎专栏等技术社区位居前列；Kimi 的前六来源中包含火山引擎开发者文档与 CSDN。

其三，信源集中度与引擎定位相关。 小米（276 个域名）与通义千问（226 个域名）的信源池最集中，首位来源占比分别为 18.4% 和 18.0%；腾讯元宝（1,044 个域名）与智谱（892 个域名）最为分散，智谱首位来源仅占 4.2%。

> 信向**XGEO** 判断：八个引擎不是八个投放渠道，而是八套引用机制。内容分发要按引擎分别配置，不能用一份通稿覆盖全部引擎。

06

内容要素与引用可见性的关系

6.1 主题相关性

Vishwakarma 等（2026）基于 25.2 万次对照实验，测定了五类内容要素对引用偏好的影响强度。

要素	回归系数
主题相关性	-0.58
价格信息	-0.27
时效性	-0.26
规格参数	-0.15
对比分析	-0.14

主题相关性的系数绝对值超过第二位要素的两倍，在五类要素中居首。

需说明的是，上述为回归系数，反映要素对引用结果的影响方向与相对强度，不代表「增加某要素可带来固定比例的提升」。

6.2 产品参数与品牌身份

Chu 与 Hou（2026）对 LLM 推荐系统中的品牌竞争机制进行了回归分析，结论具有较高的实践价值：

```
产品参数所解释的推荐排名方差      82.4%
品牌身份所解释的方差                1.2%
权威性语言带来的优势                相当于 0.17 个评级
营销性语言的作用                    接近于零
```

参数的解释力约为品牌身份的 **68** 倍。 该结果表明，在 LLM 推荐场景中，产品信息的完整度与准确度是主导因素，品牌知名度的影响远低于普遍预期。

同时，语言风格的作用存在明确分化：权威性表述有效，营销性表述无效。

6.3 来源权重

Song 等（2026）在 12 个 agent 模型、三个领域上的实验表明，来源偏好具有跨模型的稳定性：

> 每个模型都偏好某些来源、避开另一些，且各模型在哪些来源更优上基本一致。这种偏好可以压过需求满足度：少满足一项需求的项目，若来自被偏好的来源，被选中的概率约为三分之二；而需求满足更完整的项目，若来自不被偏好的来源，几乎从不被选中。

该结论的适用边界需明确说明： 研究场景为 agent 决策（商品选择、酒店预订、论文引用），并非直接针对 AI 搜索答案生成的 GEO 实验。将其应用于 GEO 实践时，应理解为来源权重的重要性提示，而非定量结论。

6.4 内容结构与数据要素

Liu 与 Xu（2026）通过特征级消融实验，测定了各内容特征对引用可见性的独立贡献：

```
统计数据          +0.0252
外部来源引用       +0.0127
```

二者为所测特征中贡献最大的两项。同前述，该数值为回归系数，应表述为「统计数据是影响引用可见性最大的特征之一（β = +0.0252）」，而非「可使引用率提升 2.52%」。

Yu 与 MuFeng（2026）研究了页面结构对引用行为的影响，结论是结构具有作用，但其影响力低于主题相关性与来源权重。

时效性方面，依据 Zhen 等（2026）的半衰期发现，内容更新频率应与查询的时效属性匹配，而非采用统一周期。

6.5 内容形态

Kumar（2026）基于 10 万次以上提示追踪的统计显示：

> 榜单类文章占全部 AI 引用来源的 21%，占内容类引用来源的 35.7%，为引用率最高的内容形态。

该数据为形态占比统计，反映的是「已被引用的来源中榜单所占比例」，不等同于「采用榜单形态即可获得引用」。

此外，Martinez（2026）指出，榜单类内容易被判定为推广性内容而遭过滤。因此榜单的撰写需遵循中立性原则：己方在文中的表述标准应与其他被提及方完全一致。

07
引擎风格差异与内容适配

7.1 风格差异的量化依据

本报告的实证数据显示，各引擎在答案的表述方式上存在可量化的差异。

引擎	参数比	叙事比	含价格比例	含统计比例
智谱	50.7	32.5	3.1%	16.2%
小米 MiMo	45.2	34.7	4.4%	10.9%
腾讯元宝	43.0	38.6	5.3%	10.7%
文心一言	41.9	39.0	4.9%	12.7%
Kimi	39.0	36.2	12.3%	13.7%
豆包	38.1	42.2	5.3%	5.8%
通义千问	37.5	39.0	1.9%	3.7%
DeepSeek	36.6	40.6	1.1%	5.8%

注：参数比与叙事比为答案文本中相应表述所占比例；含价格比例与含统计比例为答案中出现该类信息的样本占比。

三项差异值得关注。

参数化与叙述化的倾向差异。 智谱的参数比为 50.7、叙事比为 32.5，呈现明显的参数化倾向；豆包参数比 38.1、叙事比 42.2，叙述化倾向更为显著。

信息密度的差异。 Kimi 的含价格比例为 12.3%，DeepSeek 为 1.1%，相差逾十倍；智谱的含统计比例为 16.2%，通义千问为 3.7%。

信源类型与内容形态的关联。 Kimi 与通义千问的信源中技术文档占比较高，文心一言为百家号，豆包为抖音内容。内容形态与来源平台的适配，与引擎风格的适配具有同等重要性。

7.2 八个引擎的风格特征

基于持续监测数据，各引擎的答案风格可概括如下。

豆包：直接、简洁，偏好结论前置与要点列举，段落较短，语气亲和。适用于消费者决策类问题。

DeepSeek：技术深度较高，偏好长段分析与数据支撑，段落长度约 150 至 250 字，倾向给出明确建议。

Kimi：篇幅较长，结构层次清晰，注重来源可溯性，多引用可验证信源。

通义千问：结构清晰，偏好分点论述与对比呈现，语言中立，覆盖多个子角度。

文心一言：权威性取向明显，偏好中文权威信源（政府机构、百科、官方媒体、学术机构），措辞规范。

腾讯元宝：通俗易懂，多用类比与对比，语言平实，常见「误区纠正」式结构。

智谱：逻辑严密，偏好论据支撑与量化结论，避免夸张表述。

小米 **MiMo**：简洁实用，语言口语化，篇幅较短，侧重可操作步骤。

7.3 适配原则与边界

内容适配应遵循两项原则。

原则一，同一事实基础，差异化表达。 适配的实质是调整结构组织与表述方式，而非改变事实内容。

原则二，不得违反内容合规要求。 以「批量内容轰炸」方式实现的多引擎覆盖，已明确列入监管关注的滥用行为。中国信通院将 GEO 滥用的典型手段归纳为虚假内容制造、批量内容轰炸与隐性操纵优化三类。其中批量内容轰炸具体指：同一内容更换标题批量发布、针对单一提问生成大量相似稿件、承诺铺量霸屏。

此外，Chu 等（2026）的研究表明引擎侧已具备识别优化内容的能力，该类内容存在被降低权重的风险。

08
可见性诊断框架

8.1 框架设计原则

诊断框架的设计基于两项原则。

原则一，分维度而非单一评分。品牌在生成引擎中的表现不佳可能源于多种原因，单一综合评分无法指导改进行动。

原则二，区分前置条件与质量维度。部分维度是其他维度成立的前提，诊断时需先行判断。

8.2 六个维度

D1 可见度：品牌在引擎答案中被提及的比例。参照 Aggarwal 等（2024）报告的引用率上限，以 40% 作为满分基准。

D2 情感倾向：答案对品牌的描述倾向，以 0 至 100 标准化评分表示。本报告样本的观测区间为 47.8 至 71.7。

D3 提及深度：品牌出现时在回答推荐名单中的位置。判断标准需依据被监测品牌的提及位置模式（首位集中或中段可见）分别设定。

D4 跨引擎一致性：同一事实主张在不同引擎中的表述是否一致。不一致通常反映事实来源的表述不统一。

D5 品牌形象：答案所呈现的品牌定位与特征描述。本报告 3,155 条样本中，512 条（16.2%）包含品牌描述，398 条（12.6%）包含品牌优势描述。

D6 竞争位势：同一提问下，本品牌与竞品的相对提及情况。

8.3 维度间关系

六个维度并非平行关系。**D1** 构成前置条件：当品牌完全未被提及时，D2、D3、D5、D6 均失去分析对象。

D4 具有次级优先性。当各引擎对同一事实的表述存在冲突时，表明事实基础尚未统一，此时推进内容优化可能放大偏差。

D2、D3、D5、D6 为质量维度，在前置条件满足后逐项改善。

8.4 应用方法

诊断结果的解读应遵循以下顺序：

```
第一步　判断 D1 是否为零
第二步　D1 非零时，检查 D4 是否存在事实冲突
第三步　前述条件满足后，依据短板维度确定改进方向
```

改进方向应与维度对应：D1 不足需检视可发现性与检索环节；D2 偏低需检视信源中的负面内容；D3 不足需检视在首位集中型引擎中的排序；D4 不一致需统一事实表述；D5 错位需补充准确的事实描述；D6 落后需分析竞品的可见优势来源。

> 信向**XGEO**判断：诊断顺序比诊断分数更重要。D1 为零时，先不要做内容和情感优化；D4 有事实冲突时，先统一事实，再谈表达。

09 测量方法论

9.1 测量原则

Martinez（2026）提出的可复现测量协议构成本章的基础。该协议要求研究在数据采集前明确估计目标，并区分四类不同的估计量：

一 文档被注入上下文后的条件效应
二 可复现流水线中的总效应
三 与商业界面的观察性关联
四 生产环境中的商业结果

混淆上述四类估计量，将导致无效结论。本报告的前述各章均标注了结论所属的估计量类型。

9.2 抽样与重复

多引擎覆盖：覆盖多个具名引擎，记录模型版本与采集日期。

多意图分离：不同意图与领域的样本应分开分析，不可合并统计。

多复述设计：每个信息需求应准备三至五个复述版本。

重复测量：Schulte 等（2026）建议以七至八次重复为起点，并在正式采集前通过试点研究估计方差。Martinez（2026）进一步建议按所需置信区间确定最终样本量。

引擎间一致率：同一批提问在不同引擎上的 Top-1 是否指向同一对象。LIVE-GEO 的 48.8%–66.3% 可作为基线参考。

顺序敏感性：同一候选集以不同顺序呈现时，结果是否变化。ChatGPT 的 73.8% vs 93.8% 说明顺序是独立变量。

9.3 时效管理

依据 Zhen 等（2026）的半衰期发现，内容更新周期应与查询的时效属性匹配：

高时效性查询（如「2026 年哪家好」「最新推荐」）	约 30 至 40 天
低时效性查询（如「是什么」「如何选择」「原理」）	约 60 至 70 天

采用统一更新周期的做法，在高时效场景下不足，在低时效场景下冗余。

9.4 常见测量偏差

偏差一，单次观测。 前文已述，单次观测不具备统计意义。

偏差二，缺乏对照。 Watanabe 与 Nakayashiki（2026）基于 ChatGPT 推荐流量的日志研究表明，未经校正的显著增长可能完全源于平台自身的增长。缺乏对照组的测量无法区分干预效果与平台波动。

偏差三，丢弃异常输出。 未触发搜索、未产生引用、返回错误的样本属于结果的一部分，不应排除。

偏差四，独立样本假设。 查询、来源、日期均为聚类单元，将每次生成视为独立样本将低估不确定性。

偏差五，把点击当成功。 2611.01582 显示，53.2% 的未点击其实是成功（good abandonment）；只统计点击会低估效果。

偏差六，把 **200** 当内容。 2611.00884 显示，20.9% 的 200 响应没有正文；采集判活不能只看状态码。

9.5 长尾提问的测量设计

长尾提问是 GEO 最容易被低估的部分。它们的单条搜索量不大，但数量多、意图具体、竞争度低；对中小企业而言，长尾往往比头部大词更接近真实成交场景。

长尾测量应遵循四项设计原则。

第一，先分层，再合并。 长尾提问至少可以分为词根层、场景层、地区层、比较层与价格层。不同层级的提及率不能直接合并成一个「整体可见度」，否则会掩盖结构差异。

第二，每条提问保留重复测量。 长尾提问数量多，不等于可以降低重复次数。建议先按层级抽样，估计方差，再确定每条提问的重复次数。

第三，零提及也是结果。 未提及、未引用、未触发搜索的样本，不能当作无效数据删除；它们恰恰说明覆盖缺口在哪里。

第四，按月看趋势，不按单次看结论。 长尾提问的可见度变化通常较小，单次波动不足以判断策略是否有效。

成本约束： 长尾测量的主要成本不是提问数量，而是重复次数与引擎覆盖数。工程化时应把「提问数量 × 引擎数量 × 重复次数」作为总查询量估算，并优先保证分层代表性，而不是盲目扩大提问总数。

9.6 数据边界与可复核性

测量结论能否被信任，取决于三件事：原始数据是否留存、统计口径是否说清、第三方能否复核。

原始数据留存。 每条结论应能回到具体引擎、具体提问、具体时间与完整回答。只保留汇总分数、不保留原始回答的报告，无法复核，也无法解释异常值。

口径说明。 样本量、去重规则、域名归一化方式、情感评分方法、首位提及定义，都应在报告中写明。不同口径的数据不能直接比较，更不能跨报告拼接。

数据边界。 如果监测使用客户的登录态与模型额度，应明确哪些数据只保存在客户本机、哪些数据上传云端、服务商能接触哪些数据。边界不清，客户无法判断数据是否被完整记录，也无法判断服务商是否会复用其数据。

第三方复核。 理想状态下，客户或第三方可以抽取若干条提问，在相同引擎、相同界面、相同时间窗口内重复测量，并对比结果。不能接受重复测量的报告，只能作为参考，不能作为效果依据。

引用忠实性。 2610.39339 显示，AI 概览引用了来源，但只有 56.7% 的引用语句被所引页面支撑，20.8% 的事实性语句完全无引用；准确性与支撑度不相关。报告应区分「引了没」「引对没」「抓到没」。

> 信向**XGEO** 判断：不可复核的数据，不能作为效果依据；宁可报告不好看，也不要把不可复核的数字写进报告。

一句话：可复核性不是报告的美观程度，而是每条结论能否被重新走一遍。

10
风险、治理与合规

10.1 操纵性技术

GEO 领域已出现多种具有操纵性质的技术路径，均有公开研究记录。

排名操控：Jin 等（2026）证明生成引擎的输出排名可被系统性操控。

提示注入：Yin 等（2026）研究了提示注入攻击在 RAG 环境中的存活率。

分布式投毒：Chen 等（2026）研究了分散式投毒对生成结果的影响。

引用洗白：Guo（2026）研究了多源 RAG 环境中的引用洗白现象及其防御。

检测与识别：Chu 等（2026）提出了检测与度量经 GEO 优化网页内容的方法。该研究表明引擎侧已具备识别能力。

检索暴露控制与隐性品牌注入：2611.02643 研究语言模型搜索中的间接提示注入；2611.04005 研究隐性品牌注入式对抗 GEO；2609.02964 的防御方案把攻击成功率从 50.32% 降到 6.20%。

10.2 监管框架

截至 2026 年 11 月，国内与 GEO 直接相关的标准与规范如下。

编号	发布机构	时间	内容
T/CAPT 026—2026	新华网牵头	2026-08 发布，2026-08-12 实施	生成式引擎优化（GEO）可信信息传播与信息生态治理规范
T/CAAAD 046/047/048	中国广告协会	2026-10-14 发布，2026-10-23 实施	服务要求、效果监测验证、内容源分级
T/CAACCHINA 001-005	中国商务广告协会	2026-09-14 公布	术语、服务商评估、计价测量、可信语料、合规
GB 45438-2025	国家标准	已实施	AI 生成内容标识方法

T/CAPT 026—2026 第 7.1 条规定，内容应按性质分为事实库、观点库与营销表达库三类管理，且营销表达库不得替代事实库作为事实依据。

10.3 合规边界

基于上述研究与实践，可界定以下合规边界。

不得采用操纵性技术。 Nimase 等（2026）的研究表明，白帽权威性内容重写的排名提升效果与最优对抗性攻击持平（NRG 0.83），且无关键词违规、文本自然度更高。该结论为白帽路线提供了直接依据，但需注意其表述为「持平」而非「优于」。

不得采用批量轰炸方式。 该方式已列入监管明确的滥用行为，且存在被引擎侧识别与降权的风险。

不得作出效果承诺。 依据第 4 章的证据分级，从被引用到商业转化的因果证据薄弱，任何关于流量与转化的确定性承诺均缺乏依据。

10.4 服务模式与责任边界

GEO 服务至少存在三种模式，责任边界不同，客户选择时应先分清。

模式	交付物	客户承担	服务商承担	主要风险
软件工具	账号、客户端、报告与内容初稿	配置、审核、发布、合规	工具可用性、数据记录、功能更新	买了不用，或把初稿直接发布
全案代运营	策略、内容、发布、报告	预算、品牌授权、最终审核	执行、投放、部分效果口径	数据黑箱、效果归因不清、内容合规
咨询 / 培训	方法、诊断、培训	执行、发布、合规	建议质量	建议无法落地，或落地后无人持续

三条责任边界应写进合同或服务说明：

一　谁发布，谁对发布内容负责
二　谁生成，谁对生成内容的事实基础负责
三　效果指标如何定义、如何采样、如何复核

不得把「被引用」写成「保证获客」。 第 4.4 节已说明，从引用到成交之间缺少可信的因果识别；任何确定性承诺都缺乏依据。

软件模式的特殊责任。 软件提供的是能力与记录，不替客户做发布决策。客户仍需审核内容、选择渠道、承担合规责任。把软件宣传成「全自动获客」既不符合事实，也会放大客户的误用风险。

> 信向**XGEO** 判断：客户真正需要的不是「排名承诺」，而是一套能自己复核的监测记录。软件模式的价值在于把判断留在客户内部，而不是把黑箱外包给服务商。

11
方法论的工程化落地

11.1 为什么方法论必须工程化

一套方法论如果只停留在报告里，客户仍然要请人解读、请人执行。工程化的目标，是把专业判断变成可重复的流程，让没有 GEO 专职人员的团队也能跑完监测、诊断、策略、内容与回测。

工程化不是把论文做成按钮，而是把论文结论转成产品里的约束条件：哪些环节可以自动化，哪些环节必须由人审核，哪些结论不能由系统替客户承诺。

> 信向 **XGEO** 判断：方法论只有变成流程，才会被客户真正使用；停留在报告里的方法论，最终会退化成一次性的咨询。

11.2 从四阶段模型到三层诊断

四阶段模型回答「可见性是怎么产生的」。产品落地时，把它压缩成客户能理解的三个问题。

论文层次	产品里的问法	客户看到的输出
可发现性	引擎能不能抓到、收录	站点与内容是否进入候选
检索	提问时能不能进入候选上下文	哪些提问下完全没有出现
引用与吸收	进入上下文后有没有被引用、被采纳	提及位置、描述倾向、引用来源
行为转化	用户是否产生后续行为	不由软件承诺，交由客户自己归因

关键设计：不把第四阶段做成产品承诺。 软件可以记录发布地址与后续监测结果，但不能把「被引用」自动推导成「成交」。

11.3 从六维框架到可读的读数

六维诊断的价值不在分数本身，而在诊断顺序。产品落地时遵循第 8.3 节的维度关系：

```
第一步　D1 是否为零
第二步　D1 非零时，D4 是否存在事实冲突
第三步　按短板维度匹配策略
```

每个读数尽量能回到具体引擎、具体提问、具体回答。不能回溯到原始回答的分数，不进入报告。 这是数据可复核性的最低要求。

11.4 从策略库到内容生成

策略库解决「该补什么」，内容生成解决「怎么写」。两者之间需要三道约束：

> 一　每篇内容必须对应一个具体提问
> 二　每篇内容必须对应一个具体短板维度
> 三　生成的是初稿，必须由客户审核后发布

这三道约束同时是合规约束。缺少提问来源与诊断依据的批量生成，容易滑向「批量内容轰炸」，触碰第 10 章的红线。

11.5 从测量方法到批次监测

第 9 章要求重复测量、多引擎覆盖、原始证据留存。产品落地时，这意味着：

- 监测按批次组织，而不是单次查询；
- 同一提问在多个引擎上分别记录，不合并统计；
- 原始回答、信源、时间戳留存，报告可回溯；
- 发布后回填地址，下一轮监测验证是否被引用。

单次查询只能作为线索，不能作为结论。 报告至少包含三项：分引擎读数、引擎间一致率、顺序敏感性。只给一个综合分，会把引擎分歧与呈现顺序的影响抹掉。

11.6 AI 原生：产品形态与数据边界

「AI 原生」不是给传统 SEO 工具加一个 AI 模块，而是从产品设计的第一天就围绕 AI 引擎的回答机制展开：用 AI 模拟真实用户提问，以引擎真实回答为测量对象，用论文研究支撑诊断与策略，用 AI 生成内容，再由客户发布并回测。

在工程形态上，本报告建议三条设计原则。

其一，监测尽量接近真实用户界面。 第 5.1 节引用的 Zhen 等（2026）发现，同一平台的 App 与 Web 界面来源集合存在系统性差异；Żatuchin（2026）则表明不稳定性源于系统本身，而非访问界面。两者并不矛盾：界面差异会影响「看到什么」，系统随机性会影响「每次看到的是否相同」。因此，监测应记录界面与时间，并在同一界面条件下重复测量。

其二，模型侧采用 **BYOK**，避免绑定单一模型厂商。 客户用自己的模型额度完成生成，工具不赚模型差价，也不把模型能力包装成自有能力。这样做的另一个好处是，模型价格与版本变化不会锁死客户。

其三，数据边界要可解释。 以信向XGEO 的桌面客户端形态为例，登录态与模型 Key 保存在客户本机；云端只处理引擎公开回答、报告链接等必要数据。哪些数据在本地、哪些上云、哪些服务商接触不到，应能用一句话说清。数据边界不清的「代运营黑箱」，是客户不信任效果数据的主要原因之一。

11.7 长尾提问的经济性

GEO 的机会大量存在于长尾提问：具体场景、具体地区、具体预算、具体人群的问法，单条搜索量不大，但数量多、意图明确、竞争度低。

长尾监测的难点不在技术，而在成本。传统人工监测无法覆盖数百至数千条提问；按 API 调用计费的工具，也会在规模化时迅速推高成本。工程化的方向是：

一 用词根与语义展开生成提问，而不是人工逐条编写
二 按批次监测，而不是单次查询
三 把内容生成成本控制在可覆盖长尾的水平
四 用回测数据淘汰无效提问，把预算集中到有效长尾

在 BYOK 模式下，单篇初稿的模型调用成本可以降到几分钱量级，具体取决于模型价格、文章长度与重试次数。成本下降的意义不是「可以无限发稿」，而是「可以用可接受的成本覆盖足够多的真实提问，再从中筛选出有效机会」。

边界： 长尾覆盖不等于长尾成交。提问覆盖只是把「被看见」的概率铺开，是否形成询盘，仍取决于产品、价格、渠道与销售。

11.8 榜单类内容的写法

第 6.5 节引用的 Kumar（2026）数据显示，榜单类文章占全部 AI 引用来源的 21%，占内容类引用来源的 35.7%。这是内容策略的重要依据，但也是最容易踩线的内容形态。

榜单写作应遵循四条规则：

一 入选标准先写清楚，再列名单
二 己方与其他被提及方使用同一套描述标准
三 不写「第一」「唯一」「最强」等绝对化表述
四 区分「AI 回答中的提及热度」与「真实能力评价」

榜单的目标是帮助用户做选择，不是替用户做结论。 己方在榜单中的篇幅、语气与信息密度，不应明显优于其他被提及方；否则容易被判定为推广性内容，反而降低被引用概率。

11.9 软件不替代什么

工程化能替代的是重复劳动与专业判断的封装，不能替代三件事：

一　不能替代客户做发布决策
二　不能替代客户承担内容合规责任
三　不能替代客户完成成交

软件交付的是能力，不是结果。 这句话既是产品边界，也是本报告对「AI 获客」的最终限定。

12
结论

本报告的主要结论如下。

第一，**GEO** 的作用范围存在明确边界。 内容优化对「已被检索内容的引用与吸收」具有较强证据支持，对「内容能否被检索」证据中等，对「商业转化」证据薄弱。将三者混同，是当前实践中过度承诺的主要来源。

第二，可见性必须作为分布而非单点来测量。 生成引擎输出的随机性使得单次观测不具备代表性。重复测量、多引擎覆盖与原始证据留存，是可复现测量的基本要求。

第三，中文生成式引擎的信源结构高度分散，但存在引擎级偏好。 通用行业观察样本中最高单一域名占比为 4.3%。分引擎看，豆包、文心一言、腾讯元宝的第一来源与其所属生态高度相关；通义千问、小米 MiMo、DeepSeek、Kimi、智谱的第一来源则是科技媒体、开发者社区与行业站点。这一分布差异说明：内容分发不能按「一个大站」或「一个生态」统一配置，而要按引擎逐一匹配来源类型。

第四，被监测品牌的提及位置可区分为首位集中与中段可见两类。 这是品牌级观测，不是回答体裁分类；换品牌重查，位置分布可能不同。两类模式对内容策略的要求不同，需分别处理。

第五，内容要素中，主题相关性与来源权重的影响力显著高于品牌身份。 产品参数可解释推荐排名方差的 82.4%，品牌身份仅 1.2%。这一结论对内容生产具有直接的指导意义：把产品讲清楚，比强调品牌知名度更有效。

第六，行业已进入合规化阶段。 2026 年 3·15 曝光与四套标准的实施，使 GEO 从「概念验证」进入「合规交付」。服务商能否说清数据来源、测量口径与责任边界，正在成为比「能不能被引用」更前置的门槛。

第七，**2026** 年 **10-11** 月的新文献收紧了三条边界。 2023 年的效果量不能直接用于今天；跨引擎一致率与顺序敏感性必须单独报告；引用不等于被说对，也不等于抓取成功。

> 信向**XGEO** 的最终判断：GEO 不是流量捷径，而是一项需要长期复测的工程。能把「监测—诊断—策略—内容—回测」做成闭环的团队，才会在 2026 年之后的合规化竞争中留下来。

本报告的实践建议是： 将 GEO 视为一项可测量的持续工程，而非一次性投放。以分维度诊断确定改进方向，以重复测量验证效果，以内容合规为不可逾越的边界，以可复核的数据边界建立信任。AI 获客是目标，GEO 是手段；两者不能混为一谈。

后续研究方向包括： 中文市场的大样本纵向研究；从引用到商业结果的因果识别；多引擎环境中竞争均衡的动态分析；以及引擎侧检测机制与内容优化策略之间的演化关系。下一步计划将完整回答样本扩至 10,400 条（500 个核心提问 × 8 个引擎 × 2 轮 + 100 个长尾提问 × 8 个引擎 × 3 轮），在完成采集与复核后更新本报告。

REFERENCES

参考文献

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. arXiv:2311.09735.

Chen, T., Wei, Y., Jin, W., & Jiang, Z. (2026). Divide and Doubt: Diverse Distributed Poisoning for Retrieval-Augmented Generation. arXiv:2609.27090.

Chu, J., Leng, Y., Li, M., Shen, Y., Shen, X., & Zhang, Y. (2026). GEO-Flag: Detecting and Measuring GEO-Optimized Web Content. arXiv:2608.16824.

Chu, X., & Hou, Y. (2026). Incumbent Advantage: Brand Bias and Cognitive Manipulation Dynamics in LLM Recommendation Systems. arXiv:2606.17443.

Grossman, R., Liu, S., & Chen, M. K. (2026). How Generative AI Disrupts Search: An Empirical Study of Google Search, Gemini, and AI Overviews. *ACM SIGIR Conference*. arXiv:2604.27790.

Guo, F. (2026). CiteShade: Citation Laundering in Multi-Source Retrieval-Augmented Generation and Its Counterfactual Defense. arXiv:2609.15660.

Ibrahim, H., & Zaki, Y. (2026). Understanding AI Provider Recommendations in Local Service Markets. arXiv:2609.18341.

Jin, H., Chen, R., Zhang, P., Luo, Y., Zeng, H., Luo, M., & Wang, H. (2026). Controlling Output Rankings in Generative Engines for LLM-based Search. arXiv:2602.03608.

Kumar, P. (2026). Generative Engine Optimization at Scale: Measuring Brand Visibility Across AI Search Engines. Ranqo. arXiv:2606.20065.

Liu, Z., & Xu, P. (2026). Think Before Writing: Feature-Level Multi-Objective Optimization for Generative Citation Visibility. arXiv:2604.19113.

Malthouse, E. (2026). Evaluating Brand Retrieval and Ranking in Large Language Model Recommendations. Northwestern University. arXiv:2609.16304.

Martinez, O. (2026). A Critical Survey of Generative Engine Optimization. arXiv:2607.14035.

Nimase, O., Chen, Z., Qi, G., Zhao, Y., & Hu, X. (2026). GEO-BENCH: Benchmarking Ranking Manipulation in Generative Engine Optimization. arXiv:2605.29107.

Schulte, J., Bleeker, M., & Kaufmann, P. (2026). Don't Measure Once: Measuring Visibility in AI Search (GEO). University of St. Gallen. arXiv:2604.07585.

Song, J., Park, H., Shim, J., Song, W., & Jo, Y. (2026). Source Preference in the Wild: How LLM Agents Favor Items by Source, and How to Reduce It. Seoul National University. arXiv:2610.03195.

Sun, X., Kong, L., Zhang, A., Zeng, L., & Wang, T. (2026). How LLMs Are Persuaded: A Few Attention Heads, Rerouted. arXiv:2605.09314.

Vishwakarma, R., Kumar, S., & Jamidar, R. (2026). What Gets Cited: Competitive GEO in AI Answer Engines. Sprinklr. arXiv:2605.25517.

Watanabe, K., & Nakayashiki, K. (2026). Disentangling Answer Engine Optimization from Platform Growth: A Log-Based Natural Experiment on ChatGPT Referral Traffic. Glasp Inc. arXiv:2606.04362.

Yin, Y., Wang, S., Koopman, B., & Zuccon, G. (2026). Can It Reach the Generator? Investigating the Survival of Prompt-Injection Attacks in Realistic RAG Settings. arXiv:2605.28017.

Yu, J., & MuFeng, Y. (2026). Structural Feature Engineering for Generative Engine Optimization: How Content Structure Shapes Citation Behavior. arXiv:2603.29979.

Zhen, T., Liu, Y., Zhang, G., & Niu, Y. (2026). What Do Chinese-Language Generative Search Engines Cite and Surface? A Large-Scale Empirical Study. Aidso Wendao Research Institute. arXiv:2607.15771.

Żatuchin, D. (2026). The Language of the Question Selects the Market: Query Language and Exit IP as Separable Factors in Commercial Recommendations from a Generative Search Interface. arXiv:2608.30052.

新增文献（2026-10 下旬至 2026-11，作者信息待全文入库后补齐）

- Do GEO Gains Survive the Live Web? Evaluating In-the-Wild Visibility Gains of Generative Engine Optimization via LIVE-GEO. arXiv:2611.01126.
- When 200 OK Is Not an Answer: Silent Fetch Failures in Agentic Web Content Acquisition. arXiv:2611.00884.
- SearchInject: Retrieval Exposure Control for Indirect Prompt Injection in Language-Model Search. arXiv:2611.02643.
- Measuring User Success in Randomized Search Experiments: Interpreting Good Abandonment in Generative AI Search. arXiv:2611.01582.
- IntentGEO: Intent-aware Generative Engine Optimization for Recommendation. arXiv:2611.04225.
- Characterizing Adversarial GEO via Implicit Brand Injection. arXiv:2611.04005.
- The Popular Get More Popular: Auditing Brand Bias in Generative Search. arXiv:2611.07614.
- Generative Search and Recommendation at Shopify. arXiv:2611.08804.
- CHO-Bench: How Do Context Hints Affect Ad Rankings in LLM Advertising? arXiv:2611.11955.
- From User Behaviors to Satisfaction Judgments: Reasoning about User Satisfaction with AI Overviews. arXiv:2611.10715.

- Search-by-Answer Bias in Retrieval Models. arXiv:2611.11396.
- Predicting What Answer Engines Cite: Open Decision Models, Rerankers and a Distilled 4B Student on Ten Engines under Controlled Retrieval. arXiv:2610.38090.
- Does the Citation Matter? Quality, Grounding, and Physician-Rubric Accuracy of Google's AI Overviews for Medical Information. arXiv:2610.39339.
- Manifest-Assisted Retrieval in Artificial Intelligence Search: Evidence Gaps and a Security Evaluation. arXiv:2610.38361.
- Who Owns Digital News, and Who Gets Cited: Measuring Media Concentration Across 77 Countries and Seven AI Systems. arXiv:2610.40164.

APPENDIX

附录

附录 A 术语表

术语	英文	定义
生成式引擎优化	Generative Engine Optimization (GEO)	通过优化内容提升其在生成引擎答案中可见度的方法体系
答案引擎优化	Answer Engine Optimization (AEO)	与 GEO 高度重叠的术语，通常视为 GEO 的组成部分
可见性	Visibility	内容或品牌在生成答案中被呈现的程度，具有分布属性
引用池	Citation Pool	引擎检索到并进入候选上下文的文档集合
吸收	Absorption	被引内容实际影响答案表述的程度，区别于仅被列出
半衰期	Half-life	被引内容的效力衰减至初始值一半所需时间
首位集中型	First-position-concentrated	被监测品牌被提及时高度集中于第一位的模式；是品牌级观测，不代表回答里只有一家品牌
中段可见型	Mid-answer visible	被监测品牌可在回答中段或末段被提及的模式
参数比	Parametric Ratio	答案文本中参数化描述所占比例
叙事比	Narrative Ratio	答案文本中叙述性描述所占比例
可见性稳定性指数	Visibility Stability Index (VSI)	多次测量结果的离散程度指标
AI 原生	AI-native	产品从设计第一天就围绕 AI 引擎的回答机制展开，而非在传统 SEO 工具上叠加 AI 模块
AI 数字化营销	AI-powered digital marketing	本报告中特指用户通过 AI 搜索了解、比较、选择产品或服务这一环
AI 获客	AI-driven customer acquisition	品牌进入 AI 的回答与推荐，从而获得询盘与成交机会；是目标，不是效果承诺
长尾提问	Long-tail Query	单条搜索量不大、但数量多、意图具体的提问集合
可复核性	Reproducibility	结论能回到原始回答、信源、时间与统计口径，并可被第三方重复验证

附录 B 数据说明

本报告实证数据的采集与统计口径如下。

采集方式：通过客户端在真实使用环境下发起自然语言提问，记录引擎返回的完整答案、引用来源、回答结构与生成时间。

数据结构：

信源记录	每条记录对应一次「引擎在某个答案中引用了某个 URL」
引用记录	每条记录对应一条回答中的一个被引 URL 的引用强度与排序
结构化分析	每条记录对应一个引擎回答的品牌提及、情感、位置与内容特征标注

统计口径：

域名归一化	去除 www 前缀、统一小写后统计
占比计算	某域名记录数 ÷ 该引擎（或全库）信源记录总数
情感得分	0 至 100 标准化评分，由答案文本的语义倾向计算
首位提及	被监测品牌出现时位于推荐序列第一位的样本占比

样本局限：样本来源于商业监测环境，品牌构成受采集范围影响，不具备全行业推断能力。

客户数量口径：公开版本不列示品牌账号数，以避免把样本规模误读为客户数量；行业覆盖以 9 个行业场景表述。品牌账号数仅保留在内部审计与扩样方案中。

数据来源：XGEO 商业监测系统生产数据快照；查询复核日期 2026-11-15；原始记录留存于公司内部，可按需复核。本报告不列示内部表名、字段名、服务名与连接信息。

统计截止日期：2026 年 11 月 14 日。

口径说明：本报告不同表格因字段完整性不同，样本量存在差异。第 5.2、5.3 节使用 3,155 条结构化分析结果；第 2.3 节列出的完整引擎回答为 3,155 条；第 5.4、5.5 节使用信向XGEO 行业观察账号的 22,808 条信源记录（通用口径），避免单一行业样本占比过高影响通用结论；数据面板同时保留全库 56,469 条原始信源条目与 13,725 条引用记录。各表在自身口径内统计，不跨表相加，也不将不同口径的比率直接比较。

回答级与信源级口径：第 3.5 节的竞品提及观察使用信向XGEO 行业观察账号的 1,002 条回答（274 个任务、91 个提问），对应 22,808 条信源记录、3,556 条引用记录，其中出现 4,015 条竞品提及、831 个去重名称；全库 3,155 条回答共出现 11,162 条竞品提及、1,924 个去重名称。第 5.4、5.5 节使用行业观察样本的 22,808 条信源记录与 3,556 条引用记录；数据面板的全库信源/引用数字为总量口径。回答级与信源级不可相加。

附录 C 文献索引

本报告文献基础为 105 篇。下表为 90 篇已完成 PDF 入库文献的主题分布（2023 年 11 月至 2026 年 10 月）；15 篇 2026 年 10 月下旬至 11 月新增文献见第 3.6 节。

主题	篇数	代表文献
概念奠基与理论框架	3	Aggarwal 等（2024）；Martinez（2026）
测量与评估	12	Schulte 等（2026）；Zhen 等（2026）；Kumar（2026）
内容优化方法	10	Liu 与 Xu（2026）；Yu 与 MuFeng（2026）；Vishwakarma 等（2026）
技术机制	9	Sun 等（2026）；Chu 与 Hou（2026）
攻击与防御	11	Jin 等（2026）；Nimase 等（2026）；Chu 等（2026）
跨领域与垂直场景	14	Ibrahim 与 Zaki（2026）；Malthouse（2026）
治理与风险	4	Martinez（2026）
企业知识库与检索增强	15	—
其他	12	—

注：上表为 90 篇已完成 PDF 入库文献的主题分布；加上 2026 年 10 月下旬至 11 月的 15 篇新增高相关文献，本报告文献基础合计 105 篇。15 篇新增文献见第 3.6 节，未计入本表的主题分布。

附录 D 方法论到产品的落地对照

方法论	产品中的对应	客户能看到什么	不能替代什么
四阶段模型	三层诊断：可发现性 / 检索 / 引用与吸收	卡在哪一层	不能替代客户做转化归因
六维诊断	六维读数与诊断顺序	哪一维低、先修哪个	不能替代客户判断业务优先级
策略库	按短板匹配策略	这一轮该补什么	不能替代客户审核内容
内容生成	按提问、引擎、维度组装提示词	文章初稿	不能替代客户发布与合规审核
测量方法论	批次监测、原始回答留存、回测	按月趋势与可回溯证据	不能承诺排名或成交
数据边界	本地登录态、BYOK、云端最小化	哪些数据在本地、哪些上云	不能替客户承担数据合规责任

附录 E 对外表述边界

本报告及基于本报告的内容，应遵守以下表述边界。

可以写：

```
AI 原生：从第一天围绕 AI 引擎回答机制设计产品
AI 数字化营销：聚焦 AI 搜索这一环，不投流、不做私域工具
AI 获客：让品牌在 AI 回答中被看见、被理解、被推荐
GEO 软件：把监测、诊断、策略、内容生成与效果验证做成闭环
核心链路自研；模型侧 BYOK，不绑定模型厂商
```

不要写：

> 自有大模型；全栈自研
> 行业第一、领先、唯一、最强
> 保证排名、保证被引用、保证获客数量或转化率
> 通过信通院 GEO 可信测评（本报告编制单位未参加该测评）
> 服务过多少家客户、多少项专利（目前没有可公开数字）

一条总原则： 能回到原始回答与信源的结论，可以写；不能回到原始数据的形容词，不写。

利益冲突声明

本报告由沈阳信向智能科技有限公司编制。该公司提供 GEO 软件产品，因此第 11 章在讨论工程化落地时以自身产品形态为例。报告中的实证数据来自公司商业监测环境，样本非随机；学术结论均引自公开文献。读者应结合自身情况判断，不将本报告视为独立第三方评测。

文档信息

版本	2.1（2026-11-15 数据快照更新）
发布日期	2026 年 11 月
编制单位	沈阳信向智能科技有限公司 GEO 方法论研究组
联系邮箱	hello@xgeo.vip
官方网站	xgeo.vip
地球	XGEOVIP

本报告引用的文献均为公开发表的学术成果，数据来源与统计方法见附录 B。报告内容供研究与实务参考，不构成对特定结果的承诺。

信向**XGEO** · 2026 年 11 月 · 地球：XGEOVIP