九章格物 NineGrids

GENERATIVE ENGINE OPTIMIZATION · MEASUREMENT & PARADIGM

星阙实验室GEO研究中心 · 第 11 册

GEO 考核演进与 AEO 范式

五级考核口径的可测性、结算边界与范式分工

出品：九章格物 · 星阙实验室

作者：李现龙

版本：V1.0

时间：2026 年 11 月

权威信源：基于国家互联网信息办公室规章与通报、中国互联网络信息中心统计报告、arXiv 与 Ibersid 已发表研究、新华网与经济日报报道及火山引擎官方文档编制

微信扫我加好友

扫码添加微信 · 获取 GEO 方案

© 2026 九章格物 · 星阙实验室 · ninegrids.cn

目录

00　核心摘要　3

01　排名为什么测不动了　4

02　五级口径解剖　5

03　考核悖论　7

04　AEO 的界定　9

05　AEO 与 GEO 的分工　11

06　可结算度量的构造　12

07　合同与验收　15

08　失效模式与治理　16

09　三年推演　18

10　结语与没有回答的问题　19

附录 A　术语与口径对照　20

附录 B　引用份额算例明细　21

附录 C　版本与修订记录　21

参考来源　22

00 核心摘要

本白皮书回答一个问题：GEO 应当按什么口径被考核、被验收、被结算，AEO 在这一体系里站在什么位置。结论先给三条。

第一，考核口径已经迁移五级。排名、点击、提及、引用、因果影响，五级各有观测源、噪声结构与博弈点。把不同级别混在一张表里打分，验收必然失真——这是当前行业报告最常见的错误，不是数据不准，是级别错配。

第二，越靠后的口径越可结算，也越不稳定。跨十个生成式引擎的受控实验给出的引用一致性相关系数为 0.704，单引擎零引用比例为 20%–28%，个别配置达到 67%[3]；C-SEO 在 54 个组合中仅 3 个显著、问答类为 0[2]。这意味着任何单次快照都不能作为结算依据，可结算度量必须由抽样设计、区间估计与复算规则共同构成。

第三，AEO 不是 GEO 的子集。AEO 是面向答案优先界面的范式，考核答案位占有率；GEO 是面向生成式引擎的范式，考核引用份额。两者在信源层高度重叠、在度量层分工，混称会把验收对象搞错。

边界：平台未公开跨引擎的引用级数据接口，引用份额只能由受控抽样估计。本白皮书交付的是可复算的方法与判定规则，不是平台认证数据；凡涉及平台内部排序权重的内容，一律不作断言。

表 0-1　五条结论速览与依据索引

#	结论	直接依据	详见
1	**五级口径不可加权加总**	量纲、观测时点、误差结构三者皆不可通约	02、03 章
2	**可结算性与稳定性负相关**	引擎间一致性 0.704；零引用 20%–28%；54 组合仅 3 显著	03 章
3	**AEO 不是 GEO 的子集**	交付物不同：答案位占有率 vs 引用份额	04、05 章
4	**结算必须按下界而非点估计**	120 条样本时半区间约 7 个百分点，点估计可双向翻转结论	06、07 章
5	**引用份额须按答案数平均**	两种均值算法相差 5.2 个百分点，同一数据可出两个结论	附录 B

注：速览仅列出全书主干判断与最近的一处依据，完整论证与反例见对应章节；数字口径以正文标注为准。

阅读建议　赶时间的读者看表 0-1 与第 06 章；要签合同的读者直接读第 07 章；要判断一份报告能不能信的读者读第 02 与第 03 章；关心 AEO 与 GEO 关系的读者读第 04、05 章。全书术语以附录 A 为准，凡遇同名异义，以附录 A 的定义覆盖正文的通俗说法。

01 排名为什么测不动了

先把结论说清楚：排名指标失效，不是因为它测得不准，而是因为它测得的东西不再是用户看到的那个东西。用户在生成式界面上看到的是合成后的一段答案，位置信息在这一层没有对应物。要理解这一点，得先看生成式引擎的三层结构。

用户看到的是答案，不是位置

传统搜索的界面单位是"链接列表"，位置是列表里的序位，用户点击哪一条由他决定，因此位置可以当作曝光机会的代理变量。生成式界面把单位换成了"答案"，答案由模型合成，来源以引用脚标或来源列表的形式附在后面。位置概念在这里被拆成两件事：文档有没有被检索到（第一层），以及它在答案里被怎么用、有没有被标出来（第三层）。

这一变化有直接的实证支撑。对 Google AI Overviews 的研究显示，生成式摘要占据了结果页最显眼的位置，把原本指向自然结果与付费结果的流量截走，形成"摘要内可见"与"点击被截流"的结构性张力[4]。换言之，同一个页面可能"在摘要里被写了"，同时"点击数下降"。这两个信号方向相反，用排名去解释其中任何一个都会失效。

三层结构里，排名只占第一层的一个中间量

生成式引擎的工作流可以拆成三层，每层的输出物不同，可观测性也不同。把三层混为一谈，是"排名万能论"的技术根源。

表 1-1　三层结构与各层可观测量

层	这一层在做什么	对外可观测的量	与排名的关系
检索层	查询改写后召回候选文档与段落，做初筛与融合	是否被召回、召回位次、段落切分结果	排名在此层有对应物，但只是候选集内的序位，不等于用户可见位置
生成层	把候选内容与模型参数知识合成一段回答	答案文本、结构、结论倾向	无对应物。同一候选集，不同采样可得不同答案
归因层	决定引用哪些来源、以何种可见度呈现	引用脚标、来源列表、引用位置与篇幅	无直接对应物。有独立的选择逻辑与偏置[4]

注：三层划分依据生成式检索的通用架构描述[4]；"可观测量"一列指外部研究者或品牌方能通过实验拿到的信号，不指平台内部变量。

这张表解释了为什么"排名第一"和"被引用"之间没有稳定映射：排名发生在第一层，引用发生在第三层，中间隔着一次由模型主导的合成。跨十个引擎的受控实验进一步显示，同一批文档在不同引擎的引用结果差异很大，引擎间引用一致性相关系数为 0.704，即约有一半的方差无法由"引擎间的共性"解释[3]。这不是实现细节的差异，而是架构决定的。

四种旧指标的失真方式

现实中的报告通常在四个旧指标上做文章：关键词排名、收录量、点击率、停留时长。它们在生成式界面下并非全部作废，但失真方式各不相同，混用会产生系统性偏差。

表 1-2　旧指标在新架构下的失真判定

旧指标	失真机制	还能说明什么	判定
关键词排名	排名存在于检索层，答案由生成层产出，两者不构成因果链	可作召回概率的粗代理，用于技术排查	部分失效

旧指标	失真机制	还能说明什么	判定
收录量	收录只保证可被检索，不保证进入候选上下文窗口	可作索引健康度的基线	部分失效
点击率	摘要直接给答案，点击被系统性截流，CTR 分母同时缩小	仍可测，但已不代表曝光机会	失效
停留时长	零点击场景下没有会话，指标无样本	仅在点击发生的子样本内有效	失效

注："失效"指作为考核主指标失效，不指作为诊断信号失效；表内判定为本白皮书依据三层结构作的推论。

这里有个容易忽略的边界：点击率失效不意味着流量数据没用。2026 年起，部分分析工具已经上线"AI 助手引荐流量"报告，搜索平台也预告将提供生成式摘要的展示数据[4]。这些数据能补上"曝光与引荐"这一格，但补不上"被写进答案却没被标为来源"这一格——后者恰恰是引用份额要解决的问题。

失效的边界：排名没有消失，只是退居幕后

把排名判为"失效"需要加限定，否则会滑向另一个极端。准确的说法是：**排名从"对外交付的考核指标"退为"对内排查的技术指标"**。它仍然影响第一层的召回概率，而召回概率是引用的必要条件——不是充分条件。

已发表的三组数据把这条边界画得更清楚。第一组，生成式摘要对陈述型查询的覆盖为 13.7%，对疑问句查询为 64.7%，样本为 55,393 条查询、观察 40 天[2]：疑问句的入口价值显著高于陈述句，说明"被召回"高度依赖查询形态，而不是页面排名。第二组，在 171,003 篇文档、2,700 条查询的实验中，优化后正文相似度下降 9%、重排下降 16%、引用下降 6%[2]：改写内容可以换来可见度，但会以其他维度的损失为代价。第三组，十引擎实验中单引擎零引用比例为 20%–28%，个别配置达 67%[3]：即使被正确召回，仍有相当比例的查询完全不产生引用。

表 1-3　三组实测证据的口径对照

证据	样本口径	关键读数	对本白皮书的含义
生成式摘要覆盖率[2]	55,393 条查询，观察 40 天	13.7% / 64.7%	召回依赖查询形态，非页面排名
改写优化的副作用[2]	171,003 篇文档，2,700 条查询	−9% / −16% / −6%	可见度提升伴随其他维度损失
跨引擎引用一致性[3]	229 条查询，1,220 组配置，1,145 个段落	ρ=0.704；零引用 20%–28%	单引擎快照不能外推为整体表现

注：三组数据均来自已发表预印本，口径以原文为准；本白皮书未作二次换算，仅按原文报告值引用。ρ 为引擎间引用结果的一致性相关系数（Spearman）。

由此得出本章的落点：**考核口径的迁移不是偏好问题，是架构决定的必然。**当用户看到的对象从"列表中的位置"变成"合成后的答案"，可考核的对象就只能是答案里的痕迹——被提到、被引用、被用来下结论。下一章把这三种痕迹拆成五级，逐级看它们的观测源和博弈点。

02 五级口径解剖

考核口径的迁移可以收敛成五级。每一级都对应一个真实存在的观测对象，也对应一套特定的作弊方式。本章逐级拆开：它测什么、谁来测、噪声在哪、怎么被骗。这张清单是本白皮书后续所有判定与算例的基础，级别定义在全书中不再变动。

五级的定义与递进关系

五级不是"越来越好"的线性升级，而是可观测对象的逐层深入：从文档的位置，到人的行为，到答案里的文字，到答案的凭证，最后到反事实的变化。这条线的起点是 GEO 的原始提出：把"在生成式引擎中的可见度"当作可优化、可测量的目标，并在实验中报告了最高约四成的可见度改进[1]。这批早期结果奠定了引用级口径的可行性，但也留下一个隐患——"可见度"在当时是一个复合词，既有被提及的意思，也有被引用的意思，后续的量纲混乱由此而来。

图 2-1　五级考核口径的递进



注：阶梯高度表示证据强度递进，不表示重要性排序；L1 至 L3 属"痕迹级"，L4 至 L5 属"凭证级"。本白皮书自建口径。

五级的准确定义如下，全书统一使用，不再替换近义词：

- **L1 排名**：文档或页面在候选集、结果列表中的序位。观测源是搜索接口返回与索引状态。
- **L2 点击**：用户从答案或列表跳转到目标页面的行为。观测源是站点分析工具与会话日志。
- **L3 提及**：品牌名、产品名或实体名在答案文本中出现，不论是否被标为来源。观测源是对答案文本的抽检与实体识别。
- **L4 引用**：答案把某一来源显式标注为依据（脚标、来源列表、内嵌链接）。观测源是答案的引用区。
- **L5 因果影响**：把该来源从候选集中移除后，答案内容的变化程度。观测源是反事实实验。

表 2-1　五级口径的可结算性判定

级别	结算凭证	噪声来源	判定
L1 排名	无。排名与答案无因果链[4]	查询改写、召回策略	不可结算
L2 点击	会话日志可证"来过"，不能证"因何而来"	摘要截流、渠道混叠	不可结算
L3 提及	无归属凭证，提及不等于被采纳	实体识别误差、同名歧义	弱可结算
L4 引用	引用区可截图留痕，可逐条复核	引擎差异、时点波动	可结算
L5 因果影响	反事实实验可复现差异量	实验设计偏误、成本约束	可结算

注：判定标准为"能否在不依赖平台内部数据的前提下，由第三方复现同一结论"。表内为本白皮书判定，非平台口径。

这张表藏着一个不太好看的事实：**行业里卖得最贵的是 L4 与 L5，最容易糊弄的是 L3。**提及率只要抽几十条答案数关键词就能算出来，成本极低、数字漂亮，但它证明不了"被采纳"。反过来，引用份额要逐条对齐引用区、标注句位、跨引擎重复采样，成本高且数字往往不好看。

每一级的博弈点

指标一旦进入合同，就会被针对。五级口径的博弈难度不同，这决定了它们能否长期作为结算依据。

表 2-2　五级口径的主要博弈方式与识别信号

级别	典型博弈方式	识别信号	后果
L1	刷收录、堆长尾页、蹭品牌词	收录量涨而引用不变	指标虚高，业务无感

级别	典型博弈方式	识别信号	后果
L2	用品牌词直接访问冒充 AI 引荐	引荐流量集中在品牌词会话	归因错配，预算误配
L3	在答案易抓取的句式里塞品牌名	提及率高、引用率低、答案不含实质内容	可见度虚增，信任折损
L4	围绕单引擎单查询做定向铺量	跨引擎复核后份额骤降	结算争议，需按区间重算
L5	把"移除后变差"归因为自身贡献	对照组设计缺失或样本过小	结论不可复现

注：博弈方式为本白皮书对已观察到的行业做法的归纳，非对具体机构的指控；识别信号可直接用于验收时的复核清单。

把博弈点排在一起会看到一个规律：**越靠近"文本痕迹"的指标越容易被制造，越靠近"凭证与反事实"的指标越难被制造，但也越难被稳定测量。**这正是下一章要展开的悖论。

级别错配的三种典型报告

错配不是术语问题，它会直接改变结论。以下三种报告在行业里都真实存在，每一种都能自圆其说，但都不能支撑结算。

表 2-3　三种级别错配的报告形态

形态	做法	表面结论	实际证明的东西
降级混算	把 L1 排名、L3 提及、L4 引用打成百分制综合分	"综合可见度提升 42%"	三类互不相干的量做加权，权重不可核
跨级替代	用 L3 提及率替代 L4 引用率作为交付指标	"AI 提及率 65%"	只证明文本里出现过名字
单引擎外推	用一个引擎的 L4 结果代表"AI 整体表现"	"AI 引用份额 30%"	引擎间一致性仅 0.704，外推不成立[3]

注：42%、65%、30% 为示例数值，用于说明形态，不指代任何真实报告或机构。

三种形态里，降级混算的隐蔽性最强，因为它看起来"更全面"。但加权综合分有一个致命缺陷：**它把不可结算的量与可结算的量合成一个数，结果整张表都失去结算资格。**一旦甲方拿这张表去核账，任何一项都能被推翻，整份报告随之作废。

级别之间不能相加

由此得出一条操作规则，也是本白皮书的第一条硬判断：五级口径必须分列，不得加权加总。理由有三点。其一，量纲不同：排名是序位，点击是次数，提及是布尔或次数，引用是可加权的份额，因果影响是差异量。其二，观测时点不同：点击按会话算，引用按查询算，因果影响按实验批次算。其三，误差结构不同：采样误差、引擎差异、归因偏置三者不可通约。

可以相加的只有同级内部的分项，例如不同句位的引用份额，前提是同一算法、同一批次、同一引擎集合。这条规则将直接决定第 06 章的算例设计与第 07 章的条款写法。

03 考核悖论

第 02 章的五级排序隐含一个假设：越靠后的级别越好。这一章要把这个假设拆掉一半——越靠后的级别确实更可结算，但同时更不稳定。两者同向增强时，结算就变成一场概率赌博。这不是修辞，有三组独立证据。

悖论的准确表述

悖论可以写成一句话：**能留痕的口径不稳定，稳定的口径不能留痕。**展开看，可结算性依赖于"存在可截取的凭证"，而凭证的生成过程由模型的采样与归因机制决定；采样带随机性，归因带偏置，引擎之间还有架构差异。三者叠加，凭证处在"有"与"无"之间的高频摆动中。

反过来说，稳定的指标往往是低层指标：收录量、索引覆盖率发不发生波动，取决于技术状态而不是模型输出，所以它们稳，但也正因为稳，它们证明不了任何与答案相关的事。稳定与可结算在现有架构下呈负相关，这是本白皮书的核心判断之一。

三个来源：采样、引擎、归因

采样随机性。同一查询在同一引擎重复执行，答案可能不同。已发表研究报告，在 229 条查询、1,220 组配置、1,145 个段落的实验中，单引擎零引用比例为 20%–28%，个别配置高达 67%[3]。零引用意味着这一轮根本没有任何来源被标出，与内容质量无关。

引擎差异。同一批文档在十个引擎上的引用结果，一致性相关系数为 0.704[3]。按相关系数的平方粗略理解，约一半的结果差异无法由引擎共性解释[本白皮书推论]。把单引擎结果说成"AI 表现"，在方法上等同于抽样偏差。

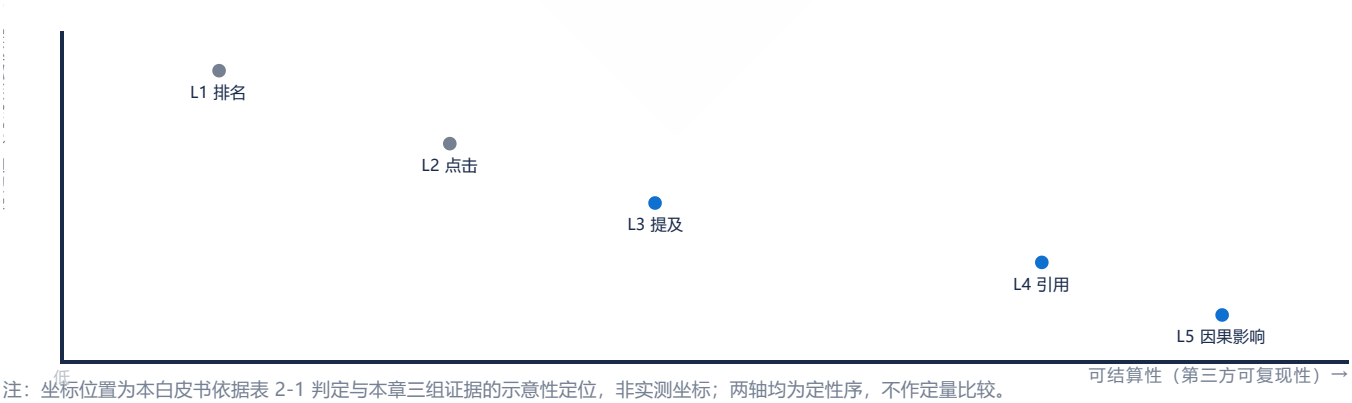
归因偏置。归因质量会随文档元数据变化。已有实验显示，仅改变"作者是人类还是模型"这一标签，引用精确率与召回率的波动区间为 3%–18%[4]，且系统性地偏好被标为人类撰写的文档。这意味着引用份额里有一部分来自"署名方式"，而不是内容本身。

此外还有一项来自竞争结构的证据：在 C-SEO 的 54 个实验组合中，仅 3 个显著，问答类为 0[2]。也就是说，当大量站点同时采用同类优化手段时，个体收益会被相互抵消。行业里把这叫"内卷"，在度量上它的含义更严重——**份额是相对量，任何绝对化的"提升百分比"在竞争充分后都不成立。**

把五级放进一个平面

把稳定性与可结算性作为两条轴，五级口径的分布并不在一条对角线上，而是散开成两组：低层稳定但不可结算，高层可结算但不稳定。这张图解释了为什么行业报告常常"数字好看但不能核账"。

图 3-1　五级口径的稳定性与可结算性分布



注：坐标位置为本白皮书依据表 2-1 判定与本章三组证据的示意性定位，非实测坐标；两轴均为定性序，不作定量比较。

图上有一条隐线：从左下到右上，越往上越"稳而不可核"，越往右越"可核而不稳"。理想区在右上角，目前没有任何一级落在那里。这不是测量技术不够，而是三层架构的固有属性——**凭证由模型生成，模型不保证凭证的稳定性。**

悖论的解法：区间化、反事实、多引擎

既然单次快照不可信，可结算度量只能建立在三件事上，缺一件都不成立。

- **区间化**：任何份额必须给出置信区间与样本量，用区间下界而非点估计做验收。区间宽窄本身是质量指标。

- **反事实**：用移除实验或对照批次估计"如果没有这次优化会怎样"，把绝对份额转成差异量。差异量比绝对量更抗引擎差异。
- **多引擎**：至少覆盖三个引擎且分别报告，禁止合并成单一数字；引擎集合必须在合同中锁定，变更视为口径变更。

这三条将在第 06 章落成具体算法，在第 07 章落成合同条款。此处先点明代价：**可结算度量的成本大约是提及率抽检的数倍**，这也是为什么大量项目选择便宜但不可结算的指标——不是不知道，而是不想承担成本。

> **反直觉判断** 追求"稳定可见度"的项目，往往在结算时最吃亏。因为稳定意味着可预测，可预测意味着竞争对手也能复制；而一旦被复制，份额的相对优势消失，绝对数字却仍停留在报告里。2026 年行业里已经出现这种落差：月报上的提及率逐月走高，季度结算时甲方按引用复核，发现可结算部分并未同步增长。

这条判断有一个前提需要写明：它成立的条件是竞争充分且优化手段同质。若某方掌握难以复制的信源资产（例如独家数据、权威资质、被广泛转引的原始研究），份额优势可以维持更久，此时稳定性与可结算性可以短暂同向——这是第 08 章讨论"信任资产"时的落点。

04 AEO 的界定

AEO 这三个字母在 2026 年的中文语境里被用得相当随意：有人拿它当 GEO 的同义词，有人拿它当"语音搜索优化"的升级版，还有人把它理解成"给 FAQ 加结构化数据"。这些用法各有一半道理，但都缺少判定标准。本章先给出可判定的定义，再处理术语歧义，最后列出四种常见误用。

AEO 指的是答案优先界面

学术综述对 AEO 的界定是清晰的：它面向"以答案为直接交付界面"的系统，包括生成式响应模块、搜索生成式体验（SGE 式界面）与基于大模型的对话界面；在这类界面下，多数查询无需访问外部站点即可被满足，从而放大零点击查询的比例[4]。同一综述把该范式与 GEO、G-SEO、GSEO、AI-SEO、C-SEO 并列，视为"同一目标下的不同侧重"：让内容被生成系统选中、采用并被显式归属[4]。

据此，本白皮书采用如下定义，并在全书中固定使用：**AEO（Answer Engine Optimization，答案引擎优化）指为答案优先界面优化内容与实体，使目标对象在答案中被稳定占位、并可被归属的一整套做法。**定义里三个词是判定的关键：**答案优先**（界面的交付单位是答案而不是链接列表）、**稳定占位**（不是偶发出现）、**可被归属**（能指回来源或实体）。

这三条同时给出判定线。只要一个系统把答案作为第一交付物、答案能不跳转满足查询、优化对象是答案内的占位与归属，它就是 AEO 的适用场景；缺一条，就落回 GEO 或传统 SEO 的范畴。判定不看界面的外观，看交付单位。

术语歧义：Answer 还是 AI

AEO 的缩写存在两种展开：Answer Engine Optimization 与 AI Engine Optimization。中文语境里两者常被混用，导致同一场会议里两个人各说一套。本白皮书不靠缩写判定，靠界面单位判定，理由有三。

- 缩写是标签，界面单位是事实。标签可以改，系统的交付单位不会因为改名而改变。
- "AI 引擎"覆盖过宽：从推荐系统到生成式检索都可自称 AI 引擎，范围大到无法形成可执行的优化对象。
- "答案"可操作：答案是一个可截取、可标注、可计量的文本单元，适合作为度量对象。

因此本白皮书统一采用 Answer Engine Optimization，并在需要指称更宽范围时直接写"生成式引擎"，不使用"AI 引擎"这一表述。这一选择不影响任何结论，只影响读者在跨文档引用时的对齐成本——而恰恰是这个成本，在验收阶段最容易引发争议。

与相邻术语的关系

AEO 不是孤立出现的概念，它与五个相邻术语共处一张谱系。谱系按"优化对象所处的层级"分组，而不是按出现时间排序，因为时间序会误导读者以为后者取代前者。

图 4-1　生成式可见度优化的范式谱系



注：分组依据为优化对象所处层级，非历史先后；同一项目常同时涉及多层，谱系用于界定交付物，不用于排斥。术语来源见综述[4]，分组为本白皮书所加。

表 4-1　六种范式的优化对象与考核落点

范式	优化对象	典型考核落点	与本白皮书主线的关系
GEO	内容在生成式引擎中的可见度、影响、归属	引用份额（L4）	本白皮书主线之一，度量见第 06 章
AEO	答案优先界面中的占位与归属	答案位占有率	本白皮书主线之二，界定见本章
G-SEO / GSEO	RAG 链路下的文档影响力	exposure、faithful credit、causal impact[4]	提供 L5 的方法来源
C-SEO	对话界面的多轮语境	对话内提及与追问留存	其"普及即稀释"结论支撑第 03 章
AI-SEO	AI 搜索的业务转化	线索量与合格线索率	业务侧指标，不进入本白皮书结算体系
LLM SEO	文档的机器可读性	结构化覆盖率、事实锚定率	属前置条件，不作独立考核项

注：考核落点一列中，GEO 与 AEO 为本白皮书设定，其余取自综述对相应范式的描述[4]；AI-SEO 的线索目标见该综述对相关实验研究的转述。

四种常见误用

定义清楚之后，误用就容易识别。以下四种在审核报告时出现频率最高，每一种都会导致验收对象被替换。

表 4-2　AEO 的四种误用与判定

误用说法	错在哪	正确表述	判定
"AEO 就是 GEO"	两者交付物不同：一个是答案位，一个是引用份额	同一引擎的两个交付对象	不成立
"AEO 是语音搜索"	语音只是输入方式，判定要看界面交付单位	回答型界面可归入，语音输入本身不是判据	部分成立
"加 FAQ 结构化就是 AEO"	结构化是文档层手段，属 LLM SEO	结构化是前置条件，不是考核项	部分成立
"AEO 会取代 GEO"	引用与占位是并存的两套交付，非替代	两者并行，度量分别设计	不成立

注：判定依据为本章定义的三条判定线：答案优先、稳定占位、可被归属。

四种误用的共同点，是把"手段"当成了"对象"：语音是手段，FAQ 结构化是手段，AEO 是对象。手段可以服务多个对象，因此手段层面的投入无法直接证明对象层面的结果。

本章最后一个限定：术语本身尚在成形。2026 年 2 月 1 日在 Scopus 以五个相关术语联合检索，仅返回 1 条结果；同一检索在 Google Scholar 得到 200 条，经去重与筛选后进入综述的语料为 49 篇[4]。这意味着任何关于 AEO 的"标准定义"都还不存在，本白皮书的定义是可用的一种，而非唯一的一种。同一层级的另一类尝试是机器可读文件：llms.txt 约定以 Markdown 形式向模型提供站点要点[12]，思路清晰，但被主流引擎采纳的证据目前缺失，本白皮书只把它列为前置条件，不列入考核项。

05 AEO 与 GEO 的分工

第 04 章把 AEO 从 GEO 里分了出来，但没有回答更实际的问题：一个项目同时做两件事时，钱该记在哪个科目、验收该看哪张表。本章给出分工的结构，并指出两处无法分开的地方。

界面层与引擎层的分工

分工的第一刀切在"交付物"上，而不是切在"技术手段"上。同一套内容改写、同一批结构化数据，可能同时服务两个交付物，但两个交付物的验收方式不同。

表 5-1　AEO 与 GEO 的交付物分工

维度	AEO（界面层）	GEO（引擎层）
优化目标	在答案中获得占位，且占位可被归属到目标实体	在答案的引用区获得份额，且份额可复算
交付物	答案位占有率（按查询集与界面统计）	引用份额（按句位加权、跨引擎分列）
观测对象	答案文本中的实体位次、是否列为首选、是否带归属标记	引用区中的来源位次、引用篇幅、引用句位
失败形态	答案给了别家，或答案给了"泛泛说法"而无实体	内容被采用但未被标为来源
时间尺度	随界面版本与答案模板变化，周级波动	随引擎索引与归因策略变化，月级波动
考核层级	L3 与 L4 之间：有归属的提及才算数	L4 至 L5：引用份额与因果影响

注：本表为本白皮书自建分工模型，用于交付与验收对齐；不对应任何平台的官方指标。

这张表最重要的一行是"失败形态"。AEO 的典型失败是"答案里没有你"，GEO 的典型失败是"答案里用了你但不提你"。两者在报告上看起来都是"表现不好"，但成因、修法与验收口径完全不同：前者要改内容与实体一致性，后者要改归属线索与引用友好度。

两种度量的算法差异

分工的第二刀切在算法上。两者都以"份额"为形态，但分子分母的定义不同，混用会得出互相矛盾的结论。

> **答案位占有率 = 目标实体获得首位占位的查询数 ÷ 有效查询数**
>
> 分母为界面实际给出答案的查询；未给出答案的查询不计入分母，单列"无答案率"。首位占位指答案把该实体作为主要推荐对象，而非仅在列举中出现。

> **引用份额 = Σ（该来源在答案中的加权引用量） ÷ Σ（同一答案中全部来源的加权引用量）**
>
> 加权按句位分三档，权重与算例见第 06 章；分母是同一答案内的全部来源，因此份额在单个答案内加总为 100%，跨答案取均值时需报告样本量。

两个式子的分母不在同一个集合里：答案位占有率的分母是"查询"，引用份额的分母是"来源"。因此**两者不能换算，也不能合并成一个"综合可见度"**。这一点在验收时最容易被忽略——服务商报"答案位 62%、引用份额 28%"，甲方顺手问"那整体可见度是多少"，这个问题本身没有答案。

两处重叠：信源层与内容形态层

分工不等于分家。有两处是分不开的，硬分会导致重复计费或重复劳动。

表 5-2　两处重叠及其处理方式

重叠处	为什么重叠	处理方式	计费归属
信源层	答案的占位与引用常来自同一批来源，改一批信源会同时影响两个交付物	一次建设、分别度量；合同中写明"共享信源资产"	计入 GEO，AEO 只计占位优化
内容形态层	答案优先的写法（先给结论、可抽取单元、实体一致）对两者同时有效	统一内容规范，两个交付物共用一套改写标准	按实际改写页数计一次

注：计费归属为本白皮书的建议口径，用于避免同一工作被两次计费；不构成报价标准。

重叠带来的直接后果是：**AEO 与 GEO 在同一项目里做，边际成本低于分别做，但边际收益也更难拆分。**这正是第 07 章要处理的验收争议来源之一——当一份合同同时包含两个交付物，失败时很难判定是哪一层没做到。

三个判断与一个前提

本章收束为三条判断，均可在报告中直接引用。

- **不是父子集。**AEO 不包含于 GEO，也不包含 GEO。两者的交付物不同，考核层级不同。
- **不是替代关系。**答案优先界面扩张不会让引用消失，引用区仍是答案的组成部分，因此两层长期并存。
- **是同一引擎的两个交付对象。**同一批内容、同一批信源，在两个交付物上分别验收、分别计费。

前提需要写明：以上分工成立的条件是"界面提供引用区"。若某类界面完全不标来源（零引用的情形，实验中占比 20%–28% [3]），则 GEO 在该界面上无交付物，此时只能按 AEO 的占位口径考核。这一前提在第 06 章的抽样设计中会变成一条硬规则：**零引用批次不计入引用份额分母，单列为"无凭证率"。**

06 可结算度量的构造

前五章解决了"该考核什么"和"哪些能结算"，本章解决"怎么算才算数"。可结算度量不是把数字算得更精细，而是让第三方能在不看平台后台的前提下复现同一结论。做到这一点需要五个部件同时成立，缺一件，度量就退化为自证。

五个部件

五个部件按顺序咬合：没有基线就无法判断变化，没有口径就无法对齐分子分母，没有抽样设计就无法给出区间，没有复算规则就无法复核，没有争议条款就无法收尾。任何一份声称"可结算"的报告，都可以用这五条逐项打勾。

表 6-1　可结算度量的五个部件

#	部件	必须写明的内容	缺失后果
1	**基线**	项目启动前同一查询集、同一引擎集合、同一周次的份额读数，含样本量与区间	无法区分"本来就有"与"做出来的"
2	**口径**	分子分母定义、句位权重、零引用批次处理、跨引擎合并规则	同一份数据可算出多个结论
3	**抽样**	查询集来源与规模、重复次数、周次、剔除规则、区间算法	点估计被当作真值，无法判断显著性
4	**复算**	原始答案的可核存证（批次号、时间戳、答案截图或文本）、逐条明细表	结论只能由服务商单方提供
5	**争议**	区间下界与上界分别对应什么处理、复核窗口、第三方复核的触发条件	结算阶段无规则可依，只能议价

注：本表为本白皮书自建清单，第 07 章的条款直接对应第 1 至 5 项；勾选状态用于项目自查，不进入交付文本。

引用份额的算法

权重设计只解决一个问题：同一个引用，出现在首句和出现在来源列表末尾，对答案的影响力不同。按句位分三档，是最简且可复核的做法。

表 6-2　句位权重与判定依据

句位	权重	判定标准	对应的影响形态
首句	2.0	引用标记出现在答案的前两句内，或该来源承担答案的主结论	直接支撑结论
中段	1.0	引用标记出现在答案正文中段，支撑分论点或数据	提供证据
末句	0.5	仅出现在来源列表或答案末尾的罗列中	提供参考，不参与论证
派生	4.5	单答案算例中全部来源的加权合计（2.0+1.0+0.5×3）	分母口径示例

注：权重为本白皮书的设定值，用于形成可比口径，不代表平台真实影响力；一旦写入合同即锁定，变更视为口径变更。末行"派生"为算例中的分母，非权重档。

单答案的算法可以直接举例。某条答案共标注 5 个来源：目标来源出现在首句，权重 2.0；另一来源在中段，权重 1.0；其余三个仅在末尾罗列，各 0.5，合计 1.5。该答案的来源加权总量为 4.5，目标来源的份额为 2.0 ÷ 4.5 = 44.4%。

图 6-1　单答案内的引用份额构成（每格 1%）



■ 目标来源 44%（首句，2.0）　■ 中段来源 22%（1.0）　□ 末句来源 33%（0.5×3）

注：按单答案算例绘制，2.0÷4.5=44.4%，取整为 44 格；中段 1.0÷4.5=22.2%；末句合计 1.5÷4.5=33.3%。合计 44+22+33=99，差 1 格为取整误差。

单答案的份额不能直接当项目成绩。实际报告要先把样本内的答案逐条算完，再取均值并给区间。设某批次 120 条有效答案，目标来源的逐条份额均值为 19.0%；按二项近似，标准误为 √(0.19×0.81÷120) ≈ 3.58 个百分点，95% 区间为 ±1.96×3.58 ≈ 7.0 个百分点，即 12.0%–26.0%。

这个区间说明了三件事。其一，报告"19%"而不报区间，等于把不确定性藏起来。其二，若合同写"份额达到 25% 即达标"，本批次其实处在达标边界之内——区间上界 26.0% 已越过 25%，下界 12.0% 远低于它，此时按点估计判不达标、按区间下界判达标，结论相反。其三，正因为结论会相反，第 07 章才必须把"按下界结算"写进条款。

抽样设计的四个参数

区间宽窄由设计决定，不由算法决定。四个参数一旦锁定，结果才可比较。

表 6-3　抽样设计参数与取值建议

参数	取值	取值理由
查询集	≥ 120 条	样本 120 条时，19% 附近的半区间约 7 个百分点；低于 100 条区间迅速放大，失去判定力
重复次数	每周 ≥ 2 次	采样随机性导致同查询结果漂移，单次易采到极端值
周次跨度	≥ 4 周	周级波动来源于界面版本与索引更新，跨度短会把版本变化误判为优化效果
引擎数	≥ 3 个，分列	引擎间一致性仅 0.704[3]，合并会掩盖差异；分列后可按引擎加权或取最差值

注：参数为本白皮书建议值，源于区间精度需求与已发表的一致性读数；写入合同后不得单方调整。120 条对应约 ±7 个百分点的半区间，是本白皮书认为"能用于判定"的最低样本量。

参数里有一个取舍必须写明：把查询集扩到 500 条可以把半区间压到约 3.5 个百分点，但成本按同比例上升。多数项目的实际选择是 120 条查询 × 4 周 × 3 引擎，合计 1,440 条答案，这个规模已经需要专门工具采集，人工抽样不可行。

采集能力受制于接口。当前引擎侧对外开放的主要是检索类接口，返回标题、摘要与链接等字段，可用于召回层观测[10]；另有能力说明涉及来源筛选与权威度倾向[11]。**引用区字段不在检索接口的返回之列**，因此引用份额只能靠外部抓取答案原文再逐条标注——这就是为什么第 06 章的五个部件里，"复算存证"必须由人承担，不能指望接口直出。

反事实：L5 的最小可行设计

L5 不是"再测一次份额"，而是测量"移除之后变了多少"。最小可行的设计需要三个批次，且三批必须同查询集、同引擎、同周次。

- **对照批**：原始候选集，不做任何改动，记录答案与引用。
- **移除批**：把目标来源从候选集中剔除（通过页面层屏蔽或索引层排除实现），记录答案变化。
- **替换批**：用同主题、同体量的第三方来源替换目标来源，用于区分"位置效应"与"内容效应"。

差异量取"对照批与移除批的答案相似度下降幅度"，或"目标来源引用次数在移除批中的归零率"。前者对应综述所归纳的 causal impact 维度[4]，后者更容易复算但信息量更少。**反事实的价值在于它能把绝对份额转成"这次优化贡献了多少"**，代价是需要两次额外采集，成本约为常规批次的 2 倍。

降级路径：三个部件不可得时怎么办

现实项目常常拿不到全部五个部件。此时不应假装可结算，而应显式降级，并在报告中写明降到了哪一级。

无基线：只能用同期竞品做横向对照，报告须写"无基线，仅横向可比"。**无多引擎**：只能报告单引擎结果，须写"不外推至整体"。**无复算存证**：只能作内部参考，不得进入合同验收。三条降级每条都对应一个明确的效力削减，而不是一句模糊的"仅供参考"。

本章的落点是：**可结算度量的门槛不是精度，而是可复现。**精度可以通过扩大样本改善，可复现只能通过存证与规则建立。下一章把这些规则写成条款。

07 合同与验收

度量做到可复算之后，剩下的问题是谁来认账。这一章不谈报价，只谈结构：三类争议从哪里来、六条条款怎么写、留痕留什么、结算按哪个数。这些内容可以直接搬进合同的附件。

三类争议的实际来源

验收阶段的争吵几乎都能归到三类，且每一类都能追溯到第 06 章某个部件的缺失。

表 7-1　三类验收争议与成因

争议类型	典型说法	成因	严重度
口径争议	"你算的是提及，我要的是引用"	合同未锁定分子分母与句位权重	高
样本争议	"这 30 条查询代表不了我的业务"	查询集由服务商单方选定，未在启动前确认	高
基线争议	"这个提升本来就是行业大盘涨的"	无启动前基线，或基线采集时段与竞品促销重叠	中

注：严重度按"是否直接阻断结算"划分；口径与样本争议会推翻整份报告，基线争议只影响增减量的归因。

三类之外还有一类常被忽略：**版本争议**。引擎在合同期内更新界面或归因策略，份额自然波动，此时按"未达标"扣款并不公平。处理方式是把"引擎版本变更"列为口径变更事件，触发重新基线，而不是触发违约。

六条必备条款

条款的作用是把第 06 章的五个部件翻译成双方都能执行的语言。**六条缺一条，前面所有度量设计都会在结算时失效。**

表 7-2　GEO 与 AEO 合同的六条必备条款

#	条款	写法要点	对应部件
1	**标的与层级**	写明交付物是"引用份额"还是"答案位占有率"，不得混称"AI 可见度"	口径
2	**查询集锁定**	附查询清单作为合同附件，变更需双方书面确认并重新基线	抽样
3	**引擎集合**	列明引擎与入口，按引擎分列报告，禁止合并为单一数字	抽样
4	**算法与权重**	句位权重、零引用批次处理、跨周合并方式全部写明	口径
5	**存证与复核**	交付须附逐条明细与截图存证；甲方有 15 个工作日复核窗口	复算
6	**结算规则**	按下界结算、按区间中点结算或阶梯结算，三者择一写明	争议

注：15 个工作日为本白皮书建议值，可按项目周期调整；条款结构对应表 6-1 的五个部件，第 6 条对应"争议"部件。

第 6 条值得单独展开，因为它决定钱怎么付。三种结算规则各有适用场景，选错会把风险全部压到一方。

表 7-3　三种结算规则的适用与风险

规则	算法	适用场景	风险归属
下界结算	按 95% 区间下界与目标值比较	预算敏感、要求确定性	风险在服务商，需为波动买单

规则	算法	适用场景	风险归属
中点结算	按点估计与目标值比较	长期合作、互信度高	风险均分，但争议时无客观依据
阶梯结算	按区间落入的档位决定支付比例	首期合作、需控制上限	风险分段，需预设档位边界

注：三种规则为本白皮书归纳的常见做法，不构成法律意见；具体条款须经法务确认。

以第 06 章的批次为例：点估计 19.0%，区间 12.0%–26.0%，目标 25%。下界结算判不达标，中点结算判不达标，阶梯结算则视档位设定——若档位按"下界≥20% 付 80%、中点≥20% 付 100%"设计，本批次下界 12.0% 未达 20%，只能按最低档支付。三套规则算出三种结果，这正是必须提前写明的原因。

留痕留什么

复核窗口能不能用起来，取决于存证是否完整。以下七项是复核时的最小集合，缺项即可要求补充，补充不成则视为该批次不可结算。

表 7-4　审计留痕七项

#	留痕项	格式要求	用途
1	查询清单	与合同附件逐条一致，含序号与原始表述	确认样本未被替换
2	批次与时间戳	每条答案记录采集时间、引擎与入口	区分周次与版本
3	答案原文	完整文本，含引用区与脚标位置	复算句位权重
4	句位标注	逐条标注首句/中段/末句，标注明细表	复算份额
5	零引用标记	标注无引用区的答案，单列计数	核对分母剔除规则
6	引擎版本	记录界面形态或答案模板特征	判断口径变更事件
7	复算脚本	可执行的复算程序与输入样例	第三方独立复现

注：七项为本白皮书自建清单；第 3 与第 4 项需同时保留，只保留标注结果无法验证标注是否正确；第 7 项最容易被省略，但它是"可复现"的唯一硬证据，无脚本时至少须提供逐条明细表，使第三方可手工复算。

本章落点：**合同不是把指标写进去就完了，而是把不可控的部分提前划出去。**引擎版本变更、竞品突增投放、行业大盘波动，这三类都应写成"口径变更或不可归因事件"，而不是违约情形。写清楚这些，结算才不必靠议价。

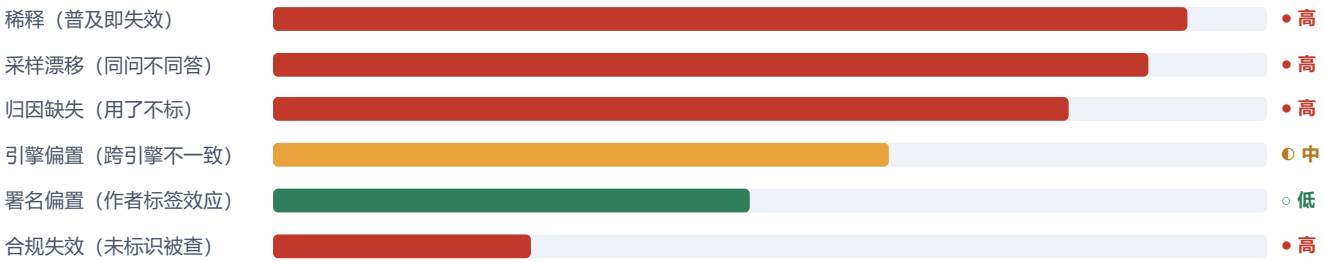
08 失效模式与治理

度量设计得再细，也会被六种失效模式打穿。这一章把这六种排开，标出发生概率与后果，再处理其中最特殊的一种——合规失效，因为它不由优化决定，而由制度决定。

六种失效模式

六种模式里，前四种属于"度量失效"，第五种属于"归因失效"，第六种属于"制度失效"。它们在报告上的表现相似：份额不动或忽高忽低。

图 8-1　六种失效模式的发生概率与后果



注：条形长度表示发生概率的判断（非实测频率），形状码 ● ◐ ○ 与颜色同时编码后果等级；本白皮书自建判定。

表 8-1　六种失效模式的识别信号与处置

模式	识别信号	证据依据	处置
稀释	参与方增多后份额不升反降	C-SEO 54 个组合仅 3 个显著、问答类为 0[2]	改打法，不加大投入
采样漂移	同查询两次结果差超过 10 个百分点	零引用比例 20%–28%，个别配置 67%[3]	加重复次数与区间
归因缺失	答案含目标内容但引用区无该来源	归因层独立于生成层[4]	改归属线索，不改正文
引擎偏置	单引擎达标、其他引擎差 2 倍以上	引擎间一致性相关系数 0.704[3]	分引擎报告
署名偏置	改署名后引用率跳变	作者标签使归因质量波动 3%–18%[4]	锁定署名规范
合规失效	内容被要求补标识或被清理	标识办法与集中查处[8][9]	先合规，再谈份额

注："识别信号"为本白皮书的操作性阈值，非平台标准；第 2 条"10 个百分点"用于快速筛查，不等同于显著性检验。处置列色标沿用图 8-1 的后果等级：红为高、橙为中、绿为低，色标表示处置优先级，不表示处置难度。

六种里最容易被误判的是"稀释"。团队看到份额下滑的第一反应通常是"内容质量下降"，但稀释的机制完全不同：所有参与方都采用了同类手法时，答案里的来源结构回到原来的分布，个体收益被抵消。这时加大投入等于加倍下注在一个已经被套利的市场里，正确动作是换打法——从"抢同一批查询的份额"转到"占据别家没有的信源位置"。

两种"看起来有效"的反噬

有两种做法在短期数据上很漂亮，但会在两个季度后反噬，把它们单独拎出来说明。

语义堆砌。围绕单一实体反复加固，短期内能提高被提及的概率，但会挤占答案的论证篇幅。已有实验显示，优化改写会让正文相似度下降 9%、重排下降 16%、引用下降 6%[2]——被"提到"更多，被"引用"更少。在 L4 口径下这是净损失。

单文档投毒。已有研究指出，检索器优先文档与查询的相似度，恶意构造的高相似度文档可以对答案产生不成比例的影响，且这种操纵靠语义微调而非关键词堆砌，比传统黑帽 SEO 更难识别[4]。这条对考核的含义是：**份额异常跳升的项目要先排除投毒，再认定优化有效。**异常判据可以用"单来源占比超过 60% 且集中在少数查询"。同一综述还提到，在政治类查询中官方来源占据相当份额，但易于发布的开放平台内容同样频繁出现，构成投毒面[4]。

标识制度改变了考核的底座

合规失效不属优化范畴，但它改变"什么内容能进入被考核的候选集"。2025 年起，这条线由制度划定，不再由平台自定。

表 8-2　与考核直接相关的标识制度条文

制度文件	条文/编号	发布与施行	义务与内容要点	对考核的影响
《人工智能生成合成内容标识办法》	第十四条	2025-03-14 发布 2025-09-01 施行[8]	服务提供者须对生成合成内容添加标识，并核验显式与隐式标识	未标识内容失去"可被归因"的资格

制度文件	条文/编号	发布与施行	义务与内容要点	对考核的影响
GB 45438-2025	强制性国标	与办法配套执行[8]	规定标识的技术要求与呈现方式，为办法的落地细则	标识不合规即不可作为优化成果主张
2025 年网络内容生态治理年度综述	查处通报	2025-11-25[9]	对未落实标识要求的应用集中查处，形成可援引的处置先例	合规性成为份额的前置条件；跨此节点的合同须重设基线

注：条文要点以文件原文为准，本表为与考核相关的摘录，不替代法规全文；引用请以正式发布文本核对。

制度对考核的影响可以概括成一句话：**标识让"机器生成"可识别，从而使引用凭证具备可追溯性。**反过来说，当一批内容因未标识被清理，份额下降不应记为"优化失败"，而应记为"合规性折损"，两者在结算中的处理完全不同——前者扣款，后者重设基线。

同一个方向上的行业动作也在推进。2026 年 5 月的可信生态研讨会上，联盟联合多家单位发布倡议，启动首批 9 家单位的评测，并提出信源分级思路；同期北京互联网法院对"AI 投毒"作出定性[7]。这三件事分别对应标准、评测、司法三个层面，意味着"引用是否可信"正在从行业自律转为可裁量的问题。

治理如何改写可结算性

把治理与度量放在一起看，会发现一个正向链条：标识制度提供"内容来源可识别"，信源分级提供"来源可信度可比"，司法定性提供"投毒可追责"。三者叠加，引用份额的凭证属性增强——从"一张截图"变成"一条可追溯的证据链"。

但这条链有一个尚未闭合的环节：**平台不公开引用级数据接口。**目前能拿到的仍是展示数据与引荐流量，引用区只能靠外部抽样抓取[4]。这意味着治理提升了凭证的可信度，却没有降低采集成本。第 09 章的分年推演将围绕这个缺口展开。

09 三年推演

推演只做一件事：把已经确定的变量摆出来，看它们在 2027 至 2029 年会把考核重心推到哪一格。不确定的部分不写结论，只写触发条件与反向情形。

三个已经确定的变量

三个变量的共同点是"已经发生，只是尚未铺开"，因此不需要预测，只需要推算扩散速度。

- **界面扩张**：生成式摘要对陈述型查询的覆盖为 13.7%，对疑问句为 64.7%[2]。疑问句是商业查询的主要形态，这意味着扩张的天花板不在技术，而在查询结构。
- **标识强制**：标识办法 2025 年 9 月施行，2025 年 11 月已有集中查处[8][9]。执行从"倡议"转为"通报"，合规成本开始显性化。
- **数据半开放**：分析工具已上线 AI 引荐流量报告，搜索平台预告提供摘要展示数据[4]。开放的是展示层，不是引用层。

需求侧还有两个读数支撑扩张不会停。国家统计口径下的互联网普及与生成式人工智能使用已进入亿级用户量级[5]，这决定了疑问句查询的绝对规模；行业媒体已把 GEO 服务作为独立品类报道，企业侧从"试投"转入"列入年度预算"[6]。供给端有工具、需求端有预算，界面扩张的约束只剩平台策略一条。

图 9-1　考核重心的三年推演



注：节点年份为推演值，非预测事件；每个节点的判断均对应正文表 9-1 的触发条件。

表 9-1　分年重心、触发条件与反向情形

年份	考核重心	触发条件	反向情形
2027	口径收敛：权重与区间进入合同	甲方复核窗口普遍使用，出现首批因口径争议作废的报告	若平台推出官方引用指标，权重之争转向"认不认平台口径"
2028	结算标准化：下界与阶梯成为常态	区间估计成为验收必备项，服务商开始提供复算脚本	若市场进入价格战，点估计与口头承诺回潮
2029	审计化：第三方复核与留痕制度化	出现独立复核机构或行业评测，留痕七项成为准入	若采集成本不降，审计只覆盖头部项目

注：三年推演为本白皮书的推论，基于第 06 至 08 章的结构性判断；不构成对任何机构的时间承诺。

五条可证伪判断

推演要能被检验，否则只是叙事。以下五条各自给出判据与时间窗，读者可在一年后自行核对。

判断一　2027 年底前，主流 GEO 合同中写明句位权重或区间结算的占比会明显上升，判据是抽查 30 份合同样本中出现其中一项的比例。**判断二**　提及率作为交付主指标的使用率会下降，判据是同一批服务商报价单中"提及"与"引用"的出现频次变化。**判断三**　跨引擎分列报告会成为默认格式，判据是报告是否仍只给单一数字。**判断四**　标识合规会成为投标资格项，判据是招标文件是否出现标识要求。**判断五**　引用份额的区间宽度会成为质量指标，判据是报告是否披露样本量与区间。

五条的共同前提是"平台不开放引用级接口"。若这个前提被打破，五条都需要重写——这正是下一章"没有回答的问题"的第一条。

10 结语与没有回答的问题

回到开头那个问题：GEO 该按什么口径被考核。九章走下来，答案可以压缩成三句。

第一句，口径有层级。排名、点击、提及、引用、因果影响不是同一种东西的五个刻度，而是五个不同的观测对象。把它们打成综合分，等于把证据与痕迹混在一起称重。级别错配是当前报告失真的主因，不是数据不准。

第二句，可结算与稳定负相关。越靠后的口径越能留痕，也越容易漂移。跨引擎一致性 0.704、零引用比例 20%–28%、普及即稀释，这三组读数共同说明：结算只能建立在区间、反事实与多引擎之上，点估计不适合签字。

第三句，AEO 是界面层范式。它考核答案位占有率，与 GEO 的引用份额是同一引擎上的两个交付物。两者在信源层重叠、在度量层分工，混称会把验收对象替换掉。

五个没有回答的问题

以下五条是本白皮书明确未解决的问题，列出是为了让读者知道边界在哪里，而不是留白凑数。

其一，引用级接口何时开放。目前只能拿到展示与引荐数据，引用区靠外部抓取[4]。若平台开放引用级数据，第 06 章的抽样设计需要整体重写。

其二，什么是"稳定占位"的阈值。第 04 章把"稳定"作为 AEO 的判定词，但没有给出可操作的阈值。连续几次出现才算稳定，目前缺少公开数据支撑，本白皮书不作断言。

其三，标识与引用的耦合方向。标识制度让内容来源可识别[8]，但这套标识是否会被引擎用作可信度信号，尚无证据。若被采用，署名与标识将成为份额的强变量。

其四，反事实的成本能否下降。L5 需要三批次采集，成本约为常规批次的两倍。成本不降，L5 只会出现在高预算项目里，结算体系会分层。

其五，跨语言与跨地域的一致性。现有实验以单一语种、单一市场为主，跨语种引用是否同样适用 0.704 这个量级的一致性，没有数据。

给三类读者的分寸

给品牌方：签合同前把六条条款核一遍，尤其第 6 条结算规则。口径定清楚，后面的争吵会少八成。

给服务商：把复算脚本与逐条明细作为交付标配，短期增加成本，长期减少返工与扯皮。可结算本身就是竞争力。

给管理者：看到"综合可见度提升 X%"先问三句——哪一级口径、样本量多少、区间多宽。三个问题答不上来的报告，不能进决策会。

A 附录 A　术语与口径对照

表 A-1　本白皮书术语表

术语	本白皮书的定义	易混说法
GEO	为生成式引擎优化内容，提升其被召回、被采用、被归属的概率	不等于"AI 排名优化"
AEO	为答案优先界面优化内容与实体，追求答案内占位与归属	不等于 GEO 的同义词
L1 至 L5	排名、点击、提及、引用、因果影响五级口径	不等于"五个 KPI"，不加权
引用份额	按句位加权后，目标来源占同一答案全部来源加权量的比例	不等于提及率
答案位占有率	获得首位占位的查询数占有效查询数的比例	不等于引用份额
零引用批次	答案不含任何来源标注的采集单元	不计入引用份额分母
句位权重	首句 2.0、中段 1.0、末句 0.5 的加权设定	不等于平台真实权重
可结算度量	第三方可在不看平台后台的前提下复现的度量	不等于高精度度量

注：术语表为全书统一口径，正文与图表均按此定义使用；同名异义视为缺陷。

B 附录 B　引用份额算例明细

下表为 8 条示例答案的逐条计算，用于演示复算路径。示例与正文 120 条批次采用同一口径，但数值不对应同一批样本，不可互相替代。

表 B-1　8 条示例答案的份额计算（示例口径）

答案	目标句位	目标权重	其他来源合计	答案总加权	份额
1	首句	2.0	2.5	4.5	44.4%
2	中段	1.0	3.4	4.4	22.7%
3	末句	0.5	3.0	3.5	14.3%
4	首句	2.0	2.3	4.3	46.5%
5	未出现	0	3.2	3.2	0.0%
6	中段	1.0	2.6	3.6	27.8%
7	末句	0.5	2.5	3.0	16.7%
8	未出现	0	3.0	3.0	0.0%
合计	—	7.0	21.9	28.9	19.0%

注：份额 = 目标权重 ÷ 答案总加权，逐条四舍五入至一位小数；末行"19.0%"为 8 条份额的算术平均（152.0 ÷ 8 = 19.0），非合计加权相除（7.0 ÷ 28.9 = 24.2%）。两种算法的差异说明：批次均值应按答案数平均，而非按加权量合计相除。

这处差异值得展开一句，因为它是最常被误用的地方。**按答案数平均得到 19.0%，按加权量合计相除得到 24.2%**，两者相差 5.2 个百分点。前者回答"平均一条答案里你占多少"，后者回答"所有来源的加权总量里你占多少"。合同里写"引用份额"时必须指明是前者还是后者，否则同一份数据能算出两个结论——这正是第 07 章口径争议的技术来源。

C 附录 C　版本与修订记录

表 C-1　版本与修订记录

版本	日期	修订内容	备注
V1.0	2026-11	初版：十章结构，五级口径、AEO 界定与分工、可结算度量与合同条款成稿	与封面版本一致

注：后续修订须逐条登记，不得改写历史记录；计划版本不与已发生修订混列。

D 参考来源

[1] Aggarwal P. et al. *GEO: Generative Engine Optimization*. arXiv:2311.09735, 2023.
[2] *A Survey of Generative Engine Optimization*（含 45 项研究的横向比较）. arXiv:2607.14035, 2026.
[3] *Controlled Retrieval Experiments across Ten Generative Engines*. arXiv:2610.38090, 2026.
[4] Alcaraz-Martínez R., Sulé A. *Generative Engine Optimization (GEO): a review of approaches, metrics, and strategies*. Ibersid, 2026.
[5] 中国互联网络信息中心.《第 57 次中国互联网络发展状况统计报告》. 2026-03.
[6] 经济日报. 生成式引擎优化市场观察. 人民网转载, 2026-03-20.
[7] 新华网. 中国人工智能产业发展联盟可信生态研讨会与首批评测发布. 2026-05-18.
[8] 国家互联网信息办公室.《人工智能生成合成内容标识办法》. 2025-03-14.
[9] 国家互联网信息办公室. 2025 年网络内容生态治理年度综述. 2025-12-31.
[10] 火山引擎. 豆包搜索接口文档（联网搜索 API，文档编号 87772）. 2026-11 版.
[11] 火山引擎. 搜索能力与权威度说明（文档编号 2272953）. 2026-11 版.
[12] llmstxt.org. *The llms.txt specification*. 2024-09-03 首发, 2026-08-10 修订.

引用与转述约定

正文角标 [n] 与文末编号一一对应，编号连续，无悬空引用。学术实验的读数按原文报告值引用，不换算为效果承诺；平台侧机制只写"可观察到的现象与已发表结论"，不写内部规则。凡由本白皮书作出的推断，均在句中标注"本白皮书推演"或"推论"字样，不与事实陈述混写。

文末 12 条来源均于 2026 年 11 月 15 日逐条发起 GET 请求实测（跟随重定向、超时 20 秒、失败重试间隔 2 秒），逐条状态码与字节数记录于同目录《GEO 考核演进与 AEO 范式_来源探测》文件；为保持版式整洁，正文与文末均不印网址，网址另存于同目录的来源 URL 清单。

表 D-1　关键读数与出处对照

读数	数值	来源	在本白皮书中的出现位置
生成式摘要覆盖率	13.7% / 64.7%	[2]	01 章表 1-3、09 章
改写优化的副作用	−9% / −16% / −6%	[2]	01 章表 1-3、08 章
跨引擎引用一致性	ρ = 0.704	[3]	01、03、06 章
单引擎零引用比例	20%–28%（个别 67%）	[3]	03、05 章
C-SEO 组合显著性	54 中 3 显著，问答 0	[2]	03、08 章
署名标签对归因的影响	3%–18%	[4]	03、08 章

注：本表为便于复核的读数索引，不替代正文的口径说明；同一读数在不同章节的用途不同，引用时须连同条件一并转述。

数据可比性提示　上表六项读数分属三类实验（覆盖率统计、受控实验、文献综述），样本与语种各不相同，只能各自纵向比较，不能横向换算成"综合效果"；把覆盖率与显著性并置对比，是第 02 章"五级不可加总"在数据层面的同一条理由。

引用规范三条　其一，转引本白皮书的任一读数，须同时转述其条件与来源编号，只引数字不引条件是本白皮书认定的误用；其二，句位权重 2.0 / 1.0 / 0.5 是本白皮书为形成可比口径而设定的算法参数，不得表述为平台真实权重；其三，不同版本的同名读数不可直接比较，版本变更时须重新基线并注明版本号。后续修订如需放宽，须在附录 C 中登记理由，不得默认沿用旧口径。

本白皮书所引数据均来自上述公开来源，标注编号与正文角标一一对应；文中判断与推演均已标注条件，不构成对任何机构的投资、采购或合规意见。版本 V1.0，2026 年 11 月。