研究学习手册 · 110 份资料

GEO 效果归因

从 AI 可见度到业务增量

原理 / 测量 / 因果方法 / 案例 / 经营决策

面向 GEO 从业者、企业负责人及营销管理者

18 章系统讲解 · 11 个公开案例 · 5 个教学场景

110 份分类阅读卡 · 可跳转引用与原始链接

核心问题

不实施这组 GEO 动作，

这些客户与收入还会出现吗？

资料核验截至 2026 年 9 月 25 日

基于既有 110 份参考资料重新组织、核读与教学化整理

目录

点击章名进入正文；点击[A01]等编号进入对应资料卡。页码以本版排版为准。

资料分类：A 机制与评测 28 份；B 行为与数据 17 份；C 企业案例 8 份；D 平台官方 12 份；E 测量规范 18 份；F 因果与评估 27 份。

阅读之前：先建立一张学习地图

这份学习手册面向两类读者：负责做出 GEO 效果的从业者，以及负责判断投入是否值得的企业负责人和营销管理者。全书围绕一个共同问题展开：采取一组明确的 GEO 动作后，哪些变化可以观察，哪些业务结果可以追踪，又有多少变化能够合理地解释为这些动作带来的新增价值?

资料底稿为《GEO 效果归因研究·110 份资料》，保留 A01 至 F27 的原始编号。本次逐项打开原始入口，并按正文相关章节、原始摘要与出版信息、书籍入口与目录等不同深度核读。长篇论文的核读集中于归因相关的研究问题、方法、结果和限制；书籍入口核验不代表通读整本书。附录为每份资料标注核读范围，全文不将 110 份材料等同于 110 项独立商业因果实验。

正文中的[A01]等编号可跳转至对应资料卡，资料卡附有原始链接。研究结果、作者解释、本手册提出的操作建议，以及虚构教学数据分别呈现。标注“教学算例”的数字均为人为设定，方便检验计算逻辑，不代表客户成绩或行业基准。

按角色安排阅读

读者	优先解决的问题	建议阅读路线
企业负责人、CEO	花的钱创造了什么；证据是否支持继续投入	第 1 至 3 章 → 第 13 章 → 第 15 至 16 章 → 第 17 章
CMO、营销负责人	如何连接品牌、获客与收入，避免各渠道重复计功	第 1 至 3 章 → 第 6 至 8 章 → 第 13 至 16 章
GEO 运营、SEO 与内容团队	问题集如何设计，怎样判断优化真的改变了结果	第 2 至 5 章 → 第 9 至 10 章 → 第 12 至 14 章
数据分析、技术与客户成功	如何建链路、设计对照、形成可审计月报	第 3 至 4 章 → 第 6 至 12 章 → 第 16 章
讲师与培训负责人	如何将原理、案例和练习组织成课程	第 1 章 → 第 13 至 14 章 → 第 17 至 18 章，再回看方法章节

全书的三个判断层次

观测层：发生了什么。 包括 AI 回答中的提及、引用、推荐位置、事实准确性，以及网站里识别到的 AI 来源访问。它回答变化的方向和大小。

业务分配层：哪些结果被记给了 AI。 依据来源、身份、时间窗口和转化分配规则，将线索或收入标注到渠道。它回答报表如何记账。[E01][E11][E15]

增量层：少做或不做这组动作会怎样。 通过可信对照或明确假设，估计实际结果与反事实结果之差。它回答干预的净贡献。[F01][F03][F08]

这三个层次可以相互支撑，但不能互相替代。企业初期可以只达到前两层，只要如实标明测量边界。复杂模型也可能停留在相关性层面；研究设计、数据质量和假设透明度决定了结论的可信程度。

本次复核对原底稿的补充

资料	本次补充或收紧	对教学的影响
A09 Pinterest	正文包含线上 A/B 测试；整体自然增长与特定模块实验的结果应分开	不再仅按摘要级生产案例介绍，也不把自然流量增幅全部归为 AI 引荐
B14 Profound 采样频次	公开页面的标准差公式展示与通常的独立伯努利均值标准误写法存在不一致；样本运行计数口径也需澄清	可讨论组合层面的稳定性，精度数值与采样次数建议须谨慎采用
C01 Ramp	正文和展示卡出现不同时间与基线口径	使用明确注明的正文口径，避免拼接为一条完整时间序列
C08 Ahrefs	原文将流量和注册份额均置于自有分析的近 30 天观测；没有足够依据把它描述为问卷与日志混接	保留样本、单位、舍入和渠道基准的审查，撤去无证据的问卷口径推测
D01、D02、E12	新公告与较早技术文档的更新时间不同；GA4 已列出 AI Assistant 渠道	明确截至 2026 年 9 月 25 日的功能边界，避免沿用旧教程

本手册提出的方法组合、报表模板和执行节奏属于综合设计建议。能否估计出可靠增量，仍取决于实际业务量、可隔离的干预、数据覆盖和观察周期。

第 1 章　GEO 效果归因究竟在解决什么

1.1 同一个“有效”，可能指向六种结果

运营说有效，可能指品牌更常出现在回答中；市场负责人说有效，可能指客户更了解产品；销售说有效，可能指线索更容易成交；老板说有效，往往指利润或长期竞争力改善。讨论开始前，必须先对齐结果层级。AMEC 将传播产出、受众反应、行为结果和最终影响分层，这种思路有助于整理 GEO 目标，但该框架本身不提供因果估计。[F21]

层级	典型问题	可观察指标	尚未证明的事情
内容与技术产出	做了哪些改动	可访问页面、已发布内容、事实纠错完成率	AI 已检索或采用
AI 呈现	品牌如何进入回答	提及率、引用率、推荐位置、事实正确率	真实用户看过、信任或购买
受众反应	用户如何理解品牌	认知、考虑、信任、自报影响	企业行动产生了净增量
行为与线索	用户采取了什么行动	访问、留资、有效咨询、产品试用	留资一定变成成交
业务结果	得到了什么经营结果	新客户、净收入、回款、贡献毛利	这些结果原本不会发生
因果增量	干预额外创造了什么	增量有效线索、增量毛利及区间	可在所有平台和行业复制

这张表同时也是验收边界。承诺“提高正确提及率”的项目，应按固定面板验证正确提及；承诺“增加有效线索”的项目，需要线索定义、CRM 匹配和去重；讨论“新增利润”的项目，则还要解释反事实、费用和风险。

1.2 三本账：展示账、归因账、增量账

展示账记录 AI 结果中的变化。归因账按照规则分配已记录的业务结果。增量账比较有干预与无干预的业务结果。三本账必须并存，且各自保留分母、时间和数据覆盖率。[A01][E01][F08]

教学示例：一个项目的提及率由 20%升至 35%，后台记录到 100 条 AI 相关线索。进一步去重和筛选后，得到 40 条有效线索。合格对照显示，若不实施项目，预计同期间会有 30 条有效线索。此时可以分别报告：提及率增加 15 个百分点；观察到 40 条有效线索；在对照假设成立时，估计新增 10 条有效线索。三个数都可有价值，含义各不相同。

直接追踪到 40 条，并不能把 40 当作真实增量下限。漏记可能使归因账少记，同时既有客户、品牌自然增长和其他营销活动又可能使归因账多计功。两种偏差可能同时存在。

1.3 为什么企业不能只盯住 ROI

经营最终需要回报，但项目早期的数据可能不足以识别小规模收入变化。此时，更合理的管理方式是建立逐级证据：技术与内容是否按计划改变，AI 输出是否改变，真实访问和客户反馈是否出现一致信号，最终业务结果是否与可解释的对照拉开差异。把每一阶段的结论说清楚，优于用一个过早的精确 ROI 掩盖缺失环节。[F07][F21]

同时不能无限延长“只看曝光”的阶段。企业应预先确定升级条件，例如开始稳定产生可识别线索后，增加有效率和商机跟踪；形成足够历史数据后，再评估对照设计。阶段计划是管理安排，达到下一阶段所需的样本和时间需由本企业数据决定。

1.4 一个合格的效果结论长什么样

建议用五句话表达：我们对哪些对象实施了什么；在什么时间范围观察哪个结果；变化相对哪个基准；还可能有哪些替代解释；据此建议继续、调整还是停止什么。这样的结论能够接受业务、运营和数据团队的共同检查。

本章自检： 看到“GEO 让销售增长 50%”，先追问销售的定义、原始数量、比较期间、是否有对照、是否扣除了同期活动。未回答这些问题之前，先将其视为需要核验的业务陈述。

第 2 章　从内容进入 AI，到客户最终购买

2.1 把技术链路和商业链路接起来

可以用以下顺序理解一条典型链路：内容可访问 → 系统发现与索引 → 查询触发检索 → 候选内容进入上下文 → 回答选择信息与来源 → 品牌被正确表达 → 用户采取行动 → 企业形成业务结果。不同产品并不总是按相同流程工作，有的答案依赖已有模型知识，有的启用搜索，有的包含购物或工具操作。[D02][D06][D08][D10]

原始 GEO 研究主要考察特定检索上下文中的来源可见度。后续 SAGEO 等研究将上游检索也纳入评估，提醒我们：在候选文档已经给定的环境中改写有效，不能自动证明公开网页更容易被真实产品检索。[A01][A12]

2.2 引用、提及、吸收，是三个可分离的现象

引用表示回答提供了某个来源链接。**提及**表示回答中出现目标品牌或实体。**吸收**关注来源中的关键信息是否真正进入回答，并被正确组织和归属。引用的页面可能属于媒体，答案却提到目标品牌；官网也可能被引用，但品牌核心卖点没有进入正文。[A10][A18][A21]

因此，官网引用份额适合评价自有信源的使用，品牌提及率适合评价实体出现，关键信息覆盖率适合评价品牌被理解的程度。三者可以联动分析，不宜压缩成一个无法解释的总分。对法律、医疗或高客单价产品，正确性与限制条件表达尤其重要。[A22][A25]

2.3 四类常见商业路径

路径	示例	主要可用证据	主要缺口
直接引荐	AI 给链接，用户点击并留资	Referrer、来源参数、站内事件、CRM	App 跳转或参数丢失
延迟品牌访问	AI 介绍品牌，用户数天后搜品牌名	面板、品牌搜索趋势、用户自报、已授权身份链	意图变化与其他曝光混杂
渠道协同	AI 帮助形成认知，用户经广告或销售成交	转化路径、客户访谈、渠道实验	多渠道重复计功、难隔离单项作用
线下或封闭平台成交	AI 推荐后通过电话、微信或门店联系	咨询问卷、业务记录、合法客户标识	网站日志无法覆盖，回忆偏差

Profound 的 AI Mention Effect 为“被 AI 提及后还可能发生后续访问”提供了观察性证据。它测量网站访问概率，不能直接推导购买概率或项目 ROI。[B01]

2.4 抓取量为什么不能直接进入获客漏斗

平台爬虫、用户触发的页面访问以及真实用户访问具有不同含义。OpenAI 和 Perplexity 官方文档区分相关代理和爬虫用途。即使机器人读取了页面，也没有因此获得一个真实用户曝光，更没有获得一个销售线索。[D07][D09]

建议至少保留三套技术观测：机器人抓取日志、实验采集日志、真实用户行为日志。三者分别承担可访问性诊断、回答质量测量和业务行为分析。对于代理代表用户完成的动作，也应单列为代理访问，核验是否真正产生了有效业务事件，避免把监测系统自己制造的请求记成增长。

2.5 GEO 与 SEO、公关和广告为什么难拆开

一篇内容改善可同时提升传统搜索表现、成为媒体引用依据，并进入 AI 回答；一次品牌活动也可能同时提高品牌搜索和 AI 提及。因此，先定义干预包：是官网内容改造，是第三方信源建设，还是多渠道组合？当多个动作同步变化且缺乏独立变化来源时，通常更有把握评估整体组合，单项贡献需要额外实验或强假设。[A09][A17][F09]

本章自检： 对任意一项 GEO 动作，写出它理论上影响链路的哪一环，以及用什么证据证明下一环发生了变化。无法连接的环节应保留为待验证假设。

第 3 章　先定义要估计的效果，再选择模型

3.1 一份归因问题需要七个要素

要素	必须写清楚的内容	示例
干预	哪些动作包含在本次评估内	12 个产品主题的内容与结构改造
单位	对谁实施或分配干预	主题簇、页面组、区域、用户或品牌
人群	结论适用于谁	新访客、目标行业、指定语言与市场
结果	用哪个业务终点判断	去重且经销售确认的有效咨询
时间	暴露、滞后、归因及成熟窗口	上线后等待索引，再观察一个完整销售周期
对照	不做干预时如何估计	同期尚未改造的相似主题簇
外溢	干预是否影响未处理对象	品牌认知提升使所有主题受益

“提升 GEO 效果”范围太大，难以估计。可检验的表达是：“在这一批目标主题中，内容改造是否相对于对照增加了指定时期的有效咨询，且未明显降低线索质量？”[F08][F20]

3.2 反事实：同一对象无法同时经历两种世界

令 Y(1)表示实施干预的结果，Y(0)表示未实施干预的结果。某个对象的因果效果为 Y(1)减去 Y(0)。实际只能观察其中一个状态，因此需要随机分组、可比对象或稳定历史关系来推测另一个状态。[F08]

这也解释了“上线后增长”为什么证据有限：上线后的结果是 Y(1)，上线前的结果来自另一个时间。季节、需求、竞争和产品变化都可能使两者不同。前后比较可以描述趋势，但需要额外设计才能支撑更强因果表述。

3.3 用文字画一张因果图

设 G 为 GEO 动作，V 为 AI 可见度，B 为品牌搜索，Y 为成交，U 为原有品牌实力。可能存在：G→V→B→Y；U 同时影响 G 投入、V 和 Y；广告活动 A 也可能同时影响品牌搜索与成交。这个结构决定了该控制什么，以及哪些变量不能随意放进回归。[F20]

原有品牌实力属于潜在混杂因素。品牌搜索却可能是 GEO 影响成交的中间环节。如果目标是估计 GEO 的总效果，机械地控制所有干预后的品牌搜索，会把一部分真实作用路径扣掉。若只想估计不经过品牌搜索的直接效果，则需要另行定义问题，并满足更严格的中介识别假设。[A04][F08]

另一个陷阱是只在“到站用户”里比较转化率。到站本身可能受 AI、广告和购买意愿共同影响。对到站用户进行筛选后得到的高转化率，描述的是经过筛选的人群，不一定反映面对全部潜在客户的干预效果。

3.4 总效果、直接效果和组合效果

总效果包含经由搜索、口碑和销售沟通传递的影响。直接效果需要明确排除哪些中间路径。组合效果关注整套动作的结果。实际项目中，整体组合往往比单项动作更容易定义，但仍需合格对照。[F08][F09]

教学示例：四组业务结果分别为无动作 100、仅 GEO120、仅 SEM150、两者同时 180。相对于基线，GEO 单独增量 20，SEM 单独增量 50，组合增量 80；额外的 10 来自交互项。只有在可比或随机的四组设计下，才有条件把它解释为协同。企业如何分摊这 10，是财务分配规则，数据本身不提供唯一答案。

3.5 先写结论等级

建议预先区分四种输出：“观察到变化”“按规则归因”“对照支持增量”“多次复现支持决策”。每种输出对应不同证据要求。报告允许停留在较低等级，关键是避免用后一等级的语言包装前一等级的结果。

本章产物： 完成一页《归因问题定义》。后续所有采集、报表和模型都围绕这一页展开；当业务问题改变时，要记录版本，避免事后挑选最有利的终点。

第 4 章　前台测量：让提及率、引用率和推荐率可信

4.1 先设计问题面板

问题集应覆盖用户决策过程：需求识别、方案比较、候选推荐、品牌核验、采购与售后。品牌词和非品牌词分别统计；已有明确购买意图的问题与普通知识问题也分别统计。权重可以来自真实搜索、销售咨询或用户调研，但来源和更新方式必须公开。[A24][D06]

建议保留两套面板。固定面板用于跨期比较，尽量维持问题、平台、语言、地区和搜索模式一致；探索面板用于发现新需求，允许持续变化。新增高表现问题造成的总分提升，应与原面板中的真实变化分开报告。

问题集只是对目标需求空间的抽样。没有真实查询分布时，不能将面板提及率称为全体用户曝光率。API 提供的可控结果适合复核实验，但模型版本、工具链、用户上下文和产品入口都可能使其与消费端体验不同。[D06][D08][D10][D11][D12]

4.2 核心指标与分母

指标	本手册建议定义	必须保留的边界
提及率	含目标实体的有效回答数÷有效回答总数	别名消歧；自报品牌名的问题另列
正确提及率	含目标实体且通过关键事实核验的回答数÷有效回答总数	需预先定义事实清单和通过标准
官网引用率	引用指定自有域名的有效回答数÷有效回答总数	不等于品牌全部信源覆盖
Top1 推荐率	目标品牌在明确排序推荐中列首位的回答数÷约定分母	分母可为全部目标问题，也可为有排序回答，需固定
引用份额	目标来源引用量÷定义内全部来源引用量	链接数、域名数、回答数是不同统计单位
信息覆盖率	正确覆盖的必要信息点数÷适用信息点总数	不适用项排除规则、点权重需明确
有效采集率	成功且符合协议的回答数÷全部计划采集数	失败不应静默删除

当一个回答没有排序时，不要强制给“首先提到”的品牌赋予 Top1 推荐。介绍位置、推荐位置和购买建议强度应分开标注。Bing 的引用报表也明确限制了从引用数字推断排名和重要性的做法。[D04]

4.3 重复采样与不确定性

同一问题重复生成可能产生不同品牌、来源和顺序。单张截图可以证明“曾经发生”，难以证明稳定的出现概率。“不要只测一次”研究与较大组合的采样研究分别提示个体问题波动和组合平均值稳定性的差异。[A11][B14]

令某问题重复 n 次，出现品牌 k 次，则出现率估计为 k/n。在独立、同分布的理想近似下，标准误为 √[p(1−p)/n]。实际采样经常受问题、日期、账号、平台和缓存共同影响，因此大量重复调用不一定等于同等数量的独立信息。

教学示例：100 次独立回答中出现 40 次，标准误约 4.9 个百分点。用正态近似得到的 95%区间约为 30.4%至 49.6%，这已经足以说明“本周 40%，下周 43%”未必具有实质差异。小样本或接近 0、1 时，可使用 Wilson 区间或合适的重抽样方法；对同一问题连续重复的结果，应考虑按问题或问题与日期组成的簇进行分析。

B14 在公开页面展示的公式将除以样本量的位置写在平方根之外，与通常的独立伯努利均值标准误公式不同。因此本手册不直接采用该页面的精度结论作为采集标准，保留其“先区分问题级与组合级判断”的有用启发。[B14]

4.4 加权面板如何避免被动漂移

假设第 j 类问题的业务权重为 w_j，各权重之和为 1；每类问题的平均出现率为 p_j，则面板可见度 V=Σw_jp_j。权重一旦改变，总分也可能变化。报告应同时提供固定权重结果和当前结构结果，将“同问题改善”与“需求结构变化”拆开。

不要把所有平台的百分比简单平均后称为市场份额。缺乏真实平台使用份额、需求量和注意概率时，只能定义一个明确标注的监测指数。GMMM 提供了把这些要素连接为暴露代理量的研究思路，但代理量仍需校准。[A04]

4.5 每条回答应保存什么

建议保存：问题与版本、意图标签、平台与产品入口、模型标识、日期时间与时区、地区与语言、登录和记忆状态、搜索模式、重复序号、原始回答、全部引用、失败原因、解析版本、实体识别结果及抽样复核记录。第三方工具不能导出这些信息时，应降低相应结论的可审计程度。

本章自检： 如果竞争品牌同时变强，你的引用份额可能下降，即便自身引用量增加。检查相对指标时，应同时查看自身绝对量和总体分母。

第 5 章　内容质量、引用忠实度与机制实验

5.1 有链接，不等于有证据

引用质量至少包含两类问题：需要证据的陈述是否有对应来源；被引用来源是否真正支持该陈述。生成式搜索可核验性研究和 ALCE 将这些维度分开，说明流畅文本、大量引用和事实可靠性不能互相替代。[A21][A22]

企业可建立“关键事实表”：产品能力、适用条件、价格与版本、服务区域、资质及限制。每项事实有负责人、有效日期、原始凭据和不可省略的限定。对于需要专业资质的行业，内容审核应由适当的业务或专业人员承担。

5.2 一份回答的四层质检

建议顺序为：是否识别了正确实体；是否表达正确事实；是否有适当证据支持；是否覆盖用户当下决策需要。只关注品牌出现，会奖励无关或错误提及；只关注文字相似，又可能漏掉合理改写。子问题覆盖研究可帮助把质量评价从关键词计数推进到用户任务完成。[A24][A25]

质检对象	典型错误	修正方向
实体	同名品牌、旧品牌或母子品牌混淆	明确别名、产品与组织关系
事实	旧型号能力被套到新型号	版本化事实表、更新自有及关键外部信源
证据	引用链接存在，内容却不支持说法	逐条核验主张与来源对应关系
决策覆盖	只讲优点，遗漏适用限制	按用户子问题补充比较与边界

5.3 自动裁判需要自己的评测集

G–Eval 等研究提供了结构化模型评估的思路，但模型评分仍可能受任务设置和模型偏好影响。不能把“模型认为推荐强”直接等同于客户更愿意购买。[A26]

建议先由人工标注一组包含正确、部分正确、模糊、错误和无法判断的样本，记录分歧与裁决规则。模型评估器输出分项和证据位置，再与人工样本比较。对高风险错误提高人工复核比例；对规则和模型升级保留版本，必要时用同一历史样本回测。不要因评分器换版，就把分数上升解释为内容优化成功。

5.4 内容实验应该一次回答一个主要问题

可以把“增加可核实的数据，是否提高某类问题中的信息吸收率”设为实验假设。准备相似页面或内容版本，保留核心事实与长度范围，改变主要研究因素，记录是否进入检索、是否被引用、是否正确吸收，再检验其他平台的表现。[A05][A06][A19]

受控上下文实验适合判断机制；公开网页实验更接近真实发现与竞争环境，但容易受索引延迟、外部内容和模型更新影响。可以先做受控筛选，再做开放环境复测。只有前者成功时，结论应限定为该受控设置中的输出变化。

内容评分系统也需要被验证。评分提高并不保证真实引用概率提高；A08 专门讨论评分的抗操纵能力与引擎结果之间的关系。评分适合作为质检工具或候选筛选，最终仍应回到真实目标指标。[A08]

5.5 五种研究设置，分别能够验证什么

设置	保持什么相对稳定	可以检验的主要问题	不能直接推出
固定候选文档后改写	候选集合、问题与生成条件	改写如何影响候选内容的呈现	自然检索中的进入概率
加入检索与重排	测试库、检索配置与比较协议	上游可发现性是否改善	所有线上产品都采用同一链路
内容质量评分	评分器版本、标准与样本	内容是否通过预设质量检查	真实用户曝光与购买概率
主张与引用核验	事实与证据的匹配规则	链接是否支持回答陈述	该陈述影响了多少成交
线上业务试点	可比对象与关键运营条件	内容动作是否改变实际业务	所有改变都经由 AI 路径发生

这些设置分别可以参考原始 GEO、SAGEO、内容评分验证、ALCE 及 Pinterest 材料。把任务设置列清楚，就能理解为什么一篇论文里的积极结果，未必与另一篇研究中的有限效果相矛盾。

[A01][A08][A09][A12][A22]

机制实验记录示例： 假设研究的是“增加清晰限制条件能否减少错误推荐”。先定义错误推荐，不把减少所有提及当成失败；同时观测正确推荐、错误推荐和无推荐三类结果。若错误减少且正确推荐维持，可以支持质量改善；若总提及减少而业务端更合格，需要继续观察有效咨询。指标设计应允许发现这种取舍，而非预设所有指标必须同向上升。

本章产物： 一张事实表、一套人工标注样本和一份内容实验协议。三者构成前台数据可信度的基础。

第 6 章　后台测量：从访问到有效线索与收入

6.1 先画业务事件链，再安装工具

建议先定义业务事件：访问 → 关键内容浏览 → 留资 → 联系成功 → 有效线索 → 商机 → 成交 → 回款 → 退款或取消。不同业务可以删减环节，但必须明确每个事件由哪个系统确认，谁负责定义，何时可以修改。

网站分析适合记录数字行为，CRM 适合维护销售状态，交易或财务系统适合确认订单与金额。GA4 导出和服务端事件回传可用于连接这些记录，但工具存在、接口返回成功，都不能替代业务质量核验。[E03][E06][E07][E16]

6.2 最小可审计数据结构

对象	建议核心字段	质量检查
来源记录	original_referrer、landing_url、UTM 原值、识别规则版本	原始值不被覆盖；来源规则可回放
会话与事件	event_id、session_id、时间戳、时区、事件名	去重、排序、跨午夜与跨设备说明
客户连接	经授权的内部匿名标识、连接方式、置信度	不凭相似时间强行认人
线索	lead_id、提交时间、业务线、首次与后续来源	同一客户多次提交去重
商机	opportunity_id、合格日期、阶段、关闭原因	阶段口径一致、无效线索保留原因
交易	order_id、币种、含税规则、净额、回款、退款	与业务或财务记录对账
干预日志	action_id、对象、执行时间、版本、负责人	与内容发布和平台变化可关联

这些是建议的数据字段，不要求全部上传到任何第三方平台。应遵循数据最小化与业务授权，将敏感映射保留在受控系统。Google 官方禁止发送个人身份信息，User-ID 也需要按其规则使用，不能直接拿邮箱或电话号码当作分析标识。[E08][E17]

6.3 来源识别的三层保存

第一层保存原始证据，例如 Referrer 和参数；第二层应用版本化规则，映射到具体平台；第三层再汇总为渠道。这样平台新增或识别规则变化时，可以重算历史，而不会丢失最初证据。

截至本次核验，GA4 官方默认渠道已包含 AI Assistant，但 Google 的 AI Overviews 和 AI Mode 仍归入 Organic Search。因此 AI Assistant 报表不能覆盖全部 AI 搜索影响。对国内平台或未被默认规则识别的

来源，可评估自定义渠道，但正则应精确到验证过的域名，避免把包含“ai”字样的普通站点误收进来。[E12][E18]

Referrer 可能因浏览器策略、应用跳转或跨协议行为而减少信息；UTM 也只有在链接可控且参数被保留时发挥作用。没有来源信息的 Direct 并不等于 AI；AI 影响也不一定留下 UTM。[E09][E10][E13]

6.4 三种作用域必须分开

首次用户来源回答“这个已识别用户最初从哪里来”；会话来源回答“这次会话从哪里开始”；转化归因维度回答“按选定模型，这次关键事件分配给谁”。把不同作用域的访客和转化拼成一个比例，可能得到不可解释的转化率。[E02][E11]

例如，首次通过 AI 访问、之后通过品牌搜索购买的用户，在首次来源报表中可能属于 AI，在会话来源报表中属于搜索。两份报表可以同时正确，它们回答不同问题。团队应为每张图写上作用域、窗口、去重单位和模型版本。

6.5 跨域、支付跳转和回传验收

网站跳到独立表单、预约域名或支付网关时，需要检查身份和会话是否被截断。跨域测量与不需要的引荐设置用于减少此类误分类，不能简单理解成“多记收入”的技巧。[E04][E05]

验收应使用测试线索和测试订单走完整链路，核对来源、时间、金额、币种、去重和业务状态。服务端回传需要验证有效负载与处理结果，并与实际记录对账。报告的 HTTP 状态或接口成功回执并不保证事件已经按照预期进入归因报表。[E07][E16]

6.6 一个真正有用的数据质量看板

建议同时监控：可解析来源比例、身份连接覆盖率、CRM 匹配率、重复事件率、延迟分布、退款回补率、未知来源比例以及各组的缺失差异。覆盖率上升可能让 AI 归因收入增加，即使业务本身没有变化，因此规则或埋点上线日期必须出现在效果图上。

本章自检： 当本月 AI 收入翻倍，先检查渠道分类、追踪覆盖、归因窗口和退款处理是否发生变化，再讨论营销效果。

第 7 章　渠道归因模型：如何分配已记录的贡献

7.1 一条路径可以有多种合理记账方法

教学路径：第 1 天 AI 访问，第 4 天品牌搜索，第 8 天点击广告，第 12 天成交，净收入 1 万元。首触规则可以将收入记给 AI，末触规则可以记给广告，线性规则可以平分，基于模型的规则也可能分配不同权重。这些都是对已记录路径的分配方法，不能单凭分配结果证明哪个触点创造了多少新增收入。[E01][E14][E15]

模型选择应服务决策。品牌发现团队可能重视首触，广告运营可能需要会话和转化贡献，管理层需要跨渠道总结果与增量。统一口径有利于协作，但无需强迫所有团队只看一个数字。

7.2 常见模型的用途和限制

模型	适合回答	主要局限
首触	记录内的发现入口是什么	依赖首次识别与观察窗口，漏掉更早线下影响
末触	成交前最后一个合格触点是什么	容易偏向收口渠道
线性、位置或时间权重	按约定规则分享贡献	权重是业务约定，缺乏唯一因果含义
数据驱动、多触点模型	在可观测路径和模型假设下分配贡献	路径缺失、选择偏差、模型依赖
随机或准实验增量估计	干预额外改变了多少结果	需要可识别设计、样本和清楚的适用范围

前几类模型可作为教学概念，但不能当作当前 GA4 界面全部可选项。Google 现行文档列示的模型及其可用范围应以官方版本为准；首触、线性、时间衰减和位置模型已在 2023 年从 GA4 相关归因选择中退出。[E01]

7.3 数据驱动也需要边界

Google 官方介绍的数据驱动归因使用模型和实验相关信息。不能笼统说它完全没有实验基础；同时，这也不意味着某企业的自然 AI 曝光与 GEO 动作已经被随机分配和完整观测。平台模型训练拥有的广告信息，与企业希望验证的 GEO 干预，属于不同证据问题。[E01]

SHAP 等工具解释某个预测模型如何使用特征。特征对预测贡献大，不等于增加该特征就一定增加销量。只有因果关系和可干预含义获得额外支持时，才有条件用它指导增量决策。[F18]

7.4 窗口如何改变结论

较长窗口可能找回早期触点，也可能把更多无关接触纳入路径；较短窗口可能漏掉长周期业务。窗口需要根据购买周期、识别能力和业务决策预先设置，并进行敏感性分析。[E15]

建议同时看固定成熟窗口的获客批次。对 1 月进入的线索观察 90 天，与对 3 月进入的线索仅观察 10 天相比，会天然显得前者成交率更高。未成熟批次应明确标注，不宜与成熟批次直接做高低判断。

7.5 三种金额报表不要相加

AI 直接归因收入、AI 辅助触点涉及收入、估计 AI 干预增量收入可能覆盖相同订单。辅助报表是关系视图，增量报表是反事实估计，三者相加会重复计算。

建议在每个订单层面保存可追踪触点，在渠道分配层确保同一笔净收入的权重之和为 1，在辅助影响层明确“去重涉及金额”，在增量层单独给出估计区间。不同渠道的增量也可能因交互作用而不具备简单可加性。

本章产物： 一份写明模型、窗口、作用域、订单去重、退款和新老客规则的《归因记账说明》。更换规则时同时展示新旧口径的桥接结果。

第 8 章　零点击、延迟访问和封闭渠道如何评估

8.1 看不见的部分值得研究，但不能任意放大

Pew 发现，在其观察样本中，带 AI 摘要的搜索与较低的传统结果点击比例相关。Profound 则研究 AI 品牌提及之后的访问。这些材料共同提示：用户在 AI 中获得信息后，可能不立即点击，也可能通过另一条路径回到品牌。[B01][B03]

然而“存在漏记”无法推出“乘以某个倍数就得到真实效果”。漏掉的人可能没有购买，也可能原本就会购买；已记录的人也可能受到其他渠道影响。需要通过补充证据逐步缩小不确定性。

8.2 用户自报归因如何提问

建议在合适的获客或成交节点，分别问“最早通过什么方式知道我们”和“哪些信息影响了这次选择”。前者可单选，后者可多选，并保留“其他”“记不清”和开放文本。销售人员也可记录客户自然提到 AI 的原话，但应区分主动回答与被引导回答。

问题应包含对称选项，避免只强调 AI。抽样调查需报告邀请人数、回答人数、回应率以及回答者与未回答者差异。自报可以捕捉日志遗漏的影响，也会受到回忆、显著性和社会期望影响。Alchemy 和 Tally 的案例可帮助理解这类证据的用途与边界。[C05][C07]

8.3 日志与问卷如何去重

教学示例：100 位新客户中，30 位有可识别 AI 触点，25 位自报 AI 影响，重合 15 位。两种证据至少一种出现的客户数为 30+25−15=40。它表示“有 AI 相关证据的客户”，不等于 40 位都由 GEO 新增。

剩余 60 位属于未观察到相关证据，不能自动判定完全没有 AI 影响。同样，观察到 AI 触点的 30 位中，也可能有人原本就会成交。报告应分别呈现日志证据、自报证据、交集和未知，避免用单一“AI 贡献客户数”抹平差异。

8.4 品牌搜索可以作为辅助结果

品牌搜索增加可能与 AI 提及、广告、线下活动、产品发布或口碑有关。可以把品牌搜索纳入预先设定的结果体系，观察其与 GEO 变化的时间关系，并使用合格对照；单条上升曲线不构成 AI 驱动的证明。[B07][F01]

评估 GEO 总效果时，品牌搜索可能是中介，未必适合直接作为控制变量。企业可以同时报告总线索、品牌搜索和 AI 相关线索，但应避免把同一客户跨三个指标重复货币化。

8.5 面板与品牌提升研究的进一步价值

在拥有合法授权和适当研究设计的条件下，用户面板能观察跨站行为，品牌实验能比较不同信息接触后的认知或考虑变化。需要区分随机展示一段 AI 答案的实验，与自然环境下平台如何选择该答案：前者支持答案呈现对受众的作用，不能自动证明公开 GEO 动作获得了多少真实曝光。[B01][F08]

8.6 从证据交叉开始，不急着追求全量归因

日志证据	用户自报	应如何记录	值得继续检查
有 AI 触点	提及 AI 影响	两种证据一致	是发现、比较还是最后决策阶段
有 AI 触点	未提及 AI 影响	保留日志，不强改问卷	用户是否只记得另一项关键接触
无可识别 AI 触点	提及 AI 影响	保留自报及未知路径	跨设备、延迟搜索或封闭渠道
无可识别 AI 触点	未提及或未回应	区分明确未选与没有回答	不能直接认定不存在 AI 作用

可以先对一批已成熟客户实施稳定的问卷与日志核验。第一步检查问题是否被理解，第二步检查回应率与去重，第三步按业务线和客户阶段观察差异，第四步再决定是否值得投入更昂贵的面板或实验。四步属于研究流程，所需时间和样本取决于业务量。

建议同时保留三个中性问题：“最早从哪里了解到我们？”“决定联系之前，主要比较了哪些信息？”“哪些渠道的信息对本次选择有影响？”不要将客户没有主动说出 AI 视为无影响，也不要在追问时暗示选择 AI 更符合预期。

对于销售团队补录的数据，应记录采集时间和采集方式。成交后很久才补问、销售引导回答和客户主动提及，证据条件并不相同。把这些差异保留下来，通常比强行把所有客户分配到唯一渠道更有分析价值。

本章自检： 一份报告称“97.5%的 AI 流量不可追踪”，应先检查它测的是 UTM、Referrer、所有访问、首个后续访问，还是订单。任何扩大解释都需要对应证据。

第 9 章　因果实验：先创造一个可信的比较

9.1 随机化提供什么

随机分配在设计上使处理组与对照组的既有差异不再系统性地随干预变化。若执行、测量与失访处理可靠，就更有条件把结果差异解释为干预带来的效果。它仍需要足够样本，并关注外溢、分配错误和实验对象的代表性。[F08][F16][F19]

GEO 的困难在于，企业通常不能随机决定真实用户会看到哪一条 AI 回答，也不能完全控制平台更新和索引。能够随机化的对象往往是页面组、主题簇、地区内容或上线顺序。此时估计的是“被分配实施这些动作”的效果，未必能够独立分离 AI、传统搜索和其他机制。

9.2 四类可考虑的实验

设计	随机或隔离的对象	可以回答	关键风险
页面或主题簇试点	相似页面组、相对独立的主题	这组内容动作是否改变相关结果	全站链接、品牌效应与跨主题引用外溢
分批随机上线	各批次启动时间	提前上线相对等待是否有效	平台趋势、等待组被间接影响
地区试点	可隔离的区域业务或内容投放	地区干预对区域业务的作用	AI 内容跨地区传播、地区不可比
用户层实验	自有站点或研究环境内可控制的体验	某种落地体验、答案展示是否影响行为	无法由此直接识别自然 AI 曝光总量

GeoLift 和 GeoX 中的“geo”指地理实验。它们提供地区选择、对照与统计设计的方法，不能因为名称相似，就把它们当成天然适配生成式引擎优化的归因工具。[F11][F12][F13]

9.3 一套可执行的主题簇随机试点

首先选取业务价值相近、历史走势相似、干预前能够稳定观测的主题簇。以主题簇为单位分层随机分配，避免同一主题的高度关联页面分属处理和对照。冻结主要业务结果、观察窗口、内容动作范围和排除规则；记录所有对象的分配，包括后续执行不完整的对象。

处理组执行事先定义的内容动作，对照组保持既定运营，并完整记录不可避免的更新。结果可以分为前台正确提及、可识别访问和有效咨询。有效咨询归属主题的规则应在实验前确定，避免因干预改变落地页或来源后，再随意移动线索归属。

分析首先按原分配进行，这通常称为意向处理分析。索引未生效、执行延迟和内容审核失败应作为合规性与机制诊断记录，不能只挑“成功上线且被抓取”的对象计算主效果，否则原有可比性可能被破坏。[F08][F16]

9.4 样本量要按真正的随机单位理解

如果只有 12 个主题簇，每个簇有 1,000 次回答，独立随机单位仍主要是 12 个主题簇，而非 12,000 个完全独立样本。簇内用户、页面和回答往往相互关联，分析需考虑聚类。不能借助大量重复调用制造很小的标准误。[F16][F19]

当独立簇很少时，应更谨慎地报告区间，必要时采用适合随机分配机制的推断，或承认设计只能探索方向。增加不同主题与更长合理观察期，通常比对少数问题重复数千次更能改善外部适用性，但具体收益需要看数据结构。

9.5 外溢与长期残留

公开内容可能在处理结束后继续被索引和引用，也可能提升整个品牌在其他问题中的呈现。因此，简单“本周开、下周关”的交替试验可能受到残留影响。对照页若引用处理内容，也可能被间接处理。[A12][F08]

建议记录跨组引用、内链变动、站级模板更新和外部传播；将明确受影响的控制对象进行预先约定的敏感性分析，不能事后只删掉不利对象。外溢过强时，结论可能只能针对整体品牌干预，或者只能描述相对局部变化。

9.6 实验停止与护栏

主要结果用于判断是否值得继续，护栏用于防止错误优化，例如事实错误率、页面性能、线索质量、投诉或内容重复。停止条件应提前设定：出现严重错误时可以安全停止；统计判定则遵循预设分析计划。每天看到一次显著就结束，会增加错误发现风险。[F16][F17]

本章产物： 一份实验登记表，包含分组名单、随机单位、主结果、护栏、样本量依据、执行日志、外溢处理和停止规则。

第 10 章　无法随机时：如何构造更可信的反事实

10.1 从最简单的前后比较开始，但不要止步于此

前后比较适合发现变化和提出假设。加入同期对照，可以尝试扣除共同趋势；加入更长历史和更多合适对象，可以进一步检查干预前关系。复杂度应由业务问题和数据条件决定，缺少可靠对照时，复杂模型不会凭空补出正确世界。[F01][F03][F20]

数据与条件	可考虑的方法	最重要的检查
仅有上线前后两期	描述性前后比较	明确不支持独立因果判断
有可比同期对象与多个前期	双重差分	平行趋势、处理污染、构成变化
有较长稳定历史与未受影响序列	CausalImpact 或受控时间序列	控制不受干预、预测关系稳定
少数处理对象、较多候选对照	合成控制或合成双重差分	前期拟合、供体污染、权重与敏感性
多批次上线且效果不同	分组时间效应或适当事件研究	处理时点、异质性、有效对照组
多渠道长期经营数据	MMM 等组合评估	识别、共线性、滞后、外部校准

10.2 双重差分：扣掉共同变化

基本形式为：估计增量＝处理组后期减前期，再减去对照组后期减前期。[F04]

教学算例：两组各 20 个规模相近的主题簇。处理组在前、后两个等长时期的有效线索总数为 200、280；对照组为 180、216。处理组增长 80，对照组增长 36。若加性平行趋势及其他假设成立，处理组估计增量为 44 条，反事实为 236 条；相对反事实的增幅约为 18.6%。

另一种计算是用对照组 20%的增长率推算处理组反事实为 240，得到 40 条增量、约 16.7%的提升。这两种结果差异来自加性和比例变化假设。不能看到哪种更高就选哪种；应根据历史关系、业务机制和预先分析计划选择，并将另一种作为敏感性分析。

10.3 平行趋势需要什么证据

它要求在没有干预时，两组结果会按可比较的方式变化，无法从数据中完全直接观察。多个干预前时期的走势、业务构成与外部事件可以帮助判断是否可信，但前趋势检验“不显著”不能证明假设必然成立，尤其在样本较小时。[F04][F15]

需要特别检查：是否因某些页面近期下滑才选择改造；是否两组商业意图不同；是否处理组同期获得了更多广告预算；是否对照组也接受了站级更新；是否价格、销售跟进和埋点规则同时变化。这些因素都可能造成虚假的组间差异。

10.4 分批上线不能随意套一个固定模型

当不同批次的效果随时间变化，已经处理的对象可能不适合作为后处理对象的对照。传统双向固定效应在这类情境中可能混合出难以解释的比较。Callaway 与 Sant’Anna、Sun 与 Abraham 等研究提供了针对分批处理和异质性问题的方法。[F04][F15]

实践上要保存每个主题的首次处理时间、后续加码和其他同期动作，考虑以尚未处理对象作为对照，分别估计不同批次与处理后的时间效应。不要只给一个总系数，至少观察各批次方向、样本量和区间是否一致。

10.5 合成控制与合成双重差分

合成控制用多个未处理对象的加权组合模拟处理对象的历史。合成双重差分进一步结合单位和时间权重，改善对照与处理在前期的可比性。[F03]

对 GEO，可以尝试用多个未改造主题构造某个改造主题的对照，但供体必须确实未受影响，而且不能只根据干预后效果挑选。应报告前期拟合、权重是否集中在少数对象、移除高权重供体后的变化，以及伪处理对象得到的结果。好的历史拟合并不保证干预后关系持续成立。

10.6 CausalImpact：预测没有干预的后期序列

CausalImpact 使用贝叶斯结构时间序列，从干预前关系与控制序列预测后期反事实。官方教程明确要求控制序列未受干预影响，并依赖相关关系在前后时期保持适当稳定。[F01][F02]

建议先在历史未处理区间设定假上线日，检查模型会不会频繁“发现效果”；再检查预测覆盖、季节性和异常时期。真正评估时同时给出实际序列、反事实预测、逐期差值和累计差值，并解释后验区间的含义。

如果上线后整个平台用户量增长，控制序列也应合理吸收这一背景变化。若把同样受 GEO 提升的品牌搜索放进控制，就可能扣掉真实中介效应；若控制主题完全不受平台趋势影响，又可能漏掉共同增长。

10.7 为什么安慰剂检验很重要

安慰剂检验可以将干预时间提前，或把未处理对象当作处理组，检查模型是否也报告相似“效果”。它用于挑战替代解释，不能单独保证因果结论。[F20]

Glasp 研究将 ChatGPT 引荐与未优化页面比较，主模型有正向结果，但时间安慰剂检验 p 值为 0.16。主模型显著与稳健性不足可以同时发生，研究因此保留提示性措辞。这个案例说明，可信报告应呈现不利检验，而不能只摘取最漂亮的系数。[A03]

本章产物： 对照选择表、干预前趋势检查、主分析、至少一种稳健性或安慰剂分析，以及对不能排除因素的书面说明。

第 11 章　MMM、机器学习与 GEO 暴露代理量

11.1 营销组合模型解决什么问题

MMM 把地区或时间聚合的业务结果，与媒体投入、需求、季节等变量联系起来，尝试分析营销组合的贡献与资源配置。Meridian 和 Robyn 提供开源实现及方法文档，适合具备较成熟数据基础的团队研究。[F09][F10]

相对于用户路径模型，它可以在缺少完整个体身份时工作；相应地，也更依赖聚合变量、业务假设和可识别的变化。很多渠道一起增减预算时，模型很难仅凭历史数据分清谁造成增长。工具自动调参或拟合度高，无法自动解决这一问题。

11.2 GEO 缺少一个天然的“投放曝光量”

付费媒体通常至少有明确的预算与展示记录，GEO 投入可能表现为内容、技术、信源关系和事实维护。成本发生的时间与 AI 采用时间不同；采用后可能持续产生影响。因此，不能机械地把当月 GEO 费用当作当月用户曝光。[A04][F25]

一种研究性的代理量是：相关问题需求量×平台使用比例×品牌出现概率×用户实际注意概率，再在问题层面求和。它帮助明确缺了哪些数据，但在需求量、平台份额或注意概率未知时，乘出来的结果依然是代理指数，不能标为真实人数或真实展示量。[A04]

11.3 GMMM 值得学习，也需要保持研究边界

A04 提出将 GEO 和生成式引擎营销连接到业务结果的因果框架，处理重复生成、暴露代理量和测量误差。论文包含实际采集的部分品牌回答输入，但商业结果与处理效应的重要验证环节来自仿真。[A04]

仿真允许研究者知道真实设定的因果量，再比较估计器表现。这能验证方法在给定数据生成假设下是否有效，不能代替真实企业订单和成本数据上的验证。AMSS 等营销模拟器也具有相同的学习价值与外推边界。[F24]

11.4 滞后、饱和与长期积累

内容上线后可能逐步被发现，认知形成后又可能延迟成交，因此效果未必当天出现；持续增加内容，也未必线性增加价值。MMM 中的滞后与饱和函数可表达这些机制，但函数形状和先验会影响估计，低信息数据下尤其如此。[F25]

项目应先检查真实索引与业务周期，再选择合理的滞后范围，报告不同设定下结果是否稳定。不能为了获得正 ROI 而不断延长窗口或挑选曲线。对周期较长的业务，应说明哪些观察已成熟、哪些仍处在形成阶段。

11.5 用实验校准模型，而非用模型掩盖缺口

有合格实验时，可以将实验信息用于校准组合模型的部分参数或结果范围。无实验时，至少检查时间留出预测、参数敏感性、渠道共线性以及负向或不合理系数。预测表现和因果识别属于不同检查，两者都值得报告。[F09][F10][F20]

DML、DoWhy 和 EconML 可以帮助处理复杂关系、识别假设和分析异质性。它们仍依赖相应的无混杂、重叠或工具变量等条件，不能消除未观察到的共同原因。SHAP 则主要解释预测模型，应与干预效果分析分开。[F14][F18][F20][F22]

11.6 何时不宜上复杂模型

当只有几周数据、同期发生大量改变、AI 来源定义不稳定、业务结果极少或没有任何可信对照时，优先修复数据并做小规模可识别试点。此时一个简单但有边界的结果说明，比缺乏支撑的多位小数更有管理价值。

本章自检： 供应商展示一个模型系数时，要求其解释变量代表真实曝光还是代理量、采用什么识别假设、如何处理未观测混杂，以及结论对参数变化有多敏感。

第 12 章　统计与证据边界：避免看起来精确的误判

12.1 百分比、百分点与倍数

从 20%升至 30%，增加 10 个百分点，相对提高 50%，后期为前期的 1.5 倍。三种表达都可以使用，但需要明确起点和终点。收入占比从 1%升至 3%，说明占比扩大；若总收入同时变化，就无法单凭占比推出 AI 收入额的准确增幅。[C03][F17]

同理，某渠道占 0.5%访问、12.1%注册，两个份额的商约为 24.2。它在可比单位和精确数值下描述相对总体平均的指数，不等于相对某个具体渠道的转化率倍数。对 Ahrefs 原文的 23 倍，应保留其传统搜索比较基准和原始舍入限制。[C08]

12.2 小基数的真实增长与脆弱推断

1 笔订单变成 10 笔，增长确实很大，但估计未来常态仍非常不稳定。客单价差异、偶然大单和渠道识别改进都可能放大增长百分比。报告应同时显示分子、分母、金额分布与成熟周期。[F07][F17]

零条变成五条时，增长率没有有限的常规分母。直接报告从 0 到 5，比写“无限增长”更有用。对长周期 ToB 业务，少数商机变化也可能由销售推进节奏造成，需结合获客批次与客户背景分析。

12.3 统计显著与商业重要性

统计显著帮助判断观测差异相对于抽样波动的大小；商业重要性关注该差异是否值得成本。大型样本可能让很小的差异显著，小型样本也可能使有商业价值的变化难以分辨。两者应分别讨论。[F07][F16]

教学近似：若基准转化率为 2%，希望识别提升到 2.4%，在两组独立、等量、双侧 5%显著性、约 80%功效的常见正态近似下，每组需要约 2.1 万次独立观察。实际若存在主题聚类、重复用户、失访或多个主要指标，需要进一步调整。这个算例仅说明小差异往往需要较大样本，不能当作所有 GEO 项目的统一门槛。

12.4 区间该如何读

频率学派置信区间来自特定抽样与估计程序；贝叶斯可信区间来自模型和先验下的后验分布。两者不能随意混用措辞。实践中应说明区间类型、估计单位、聚类层级和主要假设。[F01][F08]

当增量区间跨过 0 时，不能据此宣布“肯定无效”，也不宜只取正向点估计写成确定结果。更有用的问题是：是否排除了具有商业意义的亏损或收益？继续采集能否有效减少不确定性？当前成本和可逆性是否允许保留一个小试点？

12.5 多重比较与选择性报告

同时测试几十个平台、问题组、指标和窗口，总有一些偶然上升。应预先设定少数主要结果，探索性结果另列，并在需要时使用适当的多重检验控制或独立复测。不能从几百个问题中挑出最亮眼的十个，再称整个项目有效。[F16][F17]

数据异常同样重要。随机实验的样本比例不匹配提示分配或观测可能出错；但处理与对照页面访问量不一样，未必属于随机分配比例错误。先确定预期随机单位与实际分配计数，再判断 SRM 问题。[F19]

12.6 证据强度与适用范围是两个维度

一项严谨的海外购物实验可能对国内长周期 ToB 只有有限适用性；一个与你业务相近的客户案例可能缺少对照。评估资料时同时看内部可信度和外部相似度，避免只以机构名、样本量或是否同行评审做单项判断。

对于供应商原创研究，利益关系应披露，方法和原始口径仍需具体检查。多个机构转述同一个原始案例，仍是一组证据；同一公司的连续报告有时间价值，但不能当作完全独立的复现。

[B02][B04][B05][B06]

本章产物： 为每个核心结论补充一行“证据类型、样本与时间、可支持结论、不能外推的内容”。这行说明应与数字同时出现。

第 13 章　公开案例精读：结果、机制与证据边界

本章按“业务背景 → 实际测量 → 能支持什么 → 尚缺什么 → 如何迁移”的顺序阅读。企业和供应商案例保留自报属性，没有取得其私有订单进行独立审计。同行评审或正式论文中的实验，也要保留任务、系统和样本边界。

13.1 Glasp：把平台增长从项目增长中拆出来

背景与结果。 Glasp 研究使用单站 47 周的序列，包含 26 周前期和 21 周后期，以未优化页面作为参照。原始 ChatGPT 引荐增长约 5.7 倍，对照页面也增长约 3.5 倍。主模型报告干预水平跃升约 1.82 倍，95%区间为 1.31 至 2.54。[A03]

为什么值得学习。 对照本身增长，直接表明平台与市场环境在变化。作者没有把全部引荐上涨都记给优化，并同时披露时间安慰剂检验 p=0.16，将结论保持在提示性层级。[A03]

仍有什么缺口。 页面选择与处理并非随机，同域对照可能外溢，原有趋势也可能不稳定。它测量引荐流量，没有直接测量订单和利润。运营团队可借鉴同站参照、时间日志和不利检验披露；管理层应要求团队说明主模型与稳健性检验为何不同，不宜仅挑一个显著系数作为销售承诺。

13.2 Profound：AI 提及之后，用户可能晚些访问品牌

研究设计。 AI Mention Effect 利用 2026 年 1 至 6 月的美国用户面板，分析超过 200 万次对话。研究关注用户原提问未出现、由 AI 引入的品牌，并利用此前访问窗口构造基线，同时进行用户层聚类重抽样。[B01]

结果终点。 论文式报告中的 7 日后续访问提升估计分别为 Gemini 约 3.21 个百分点、AI Overviews 约 2.96 个百分点、ChatGPT 约 2.07 个百分点。这些是各自样本和方法下的访问差异，不宜据此给平台做因果效果排名。[B01]

关键边界。 研究没有随机安排 AI 提及，残余购买意图与其他接触仍可能影响访问；终点也未推进到购买。文中 2.47%的追踪参数比例针对特定后续访问 URL，不能解释成所有 AI 引荐识别率，更不能据此把已记录 ROI 乘以 40。[B01]

迁移方式。 企业可建立后续品牌访问和自报影响的补充观测，避免只看当次点击；要识别自身 GEO 动作增量，还需明确干预和对照。面板中的 AI 影响，与企业主动优化所创造的影响也需要分开。

13.3 Pinterest：离线指标更好，线上结果未必同步更好

生产系统。 Pinterest 论文讨论视觉语言模型、趋势与集合页等内容发现和获客机制，报告整体自然流量改善。正文还包含超过一个月的线上 A/B 比较，不能只按缺少实验的介绍性摘要阅读。[A09]

更有教学价值的细节。 比较表中，两个方案的离线意图匹配指标为 0.881 和 0.848；对应线上注册率约 0.121%和 0.127%，CTR 约 1.61%和 1.66%。离线较优方案并未在这些线上点估计中同步占优。表格没有为这里的比较提供足够完整的不确定性信息，因此本手册不据此判定显著输赢。[A09]

边界与迁移。 论文整体的约 20%自然流量增长，覆盖一套生产系统，不能直接改写成 20%的 AI 引荐增量。适合学习“代理指标需要线上业务验证”，以及把模块实验与整体经营改善分别报告。

13.4 Ramp：可见度增长是清楚的中间结果

案例声称。 Profound 发布的 Ramp 案例正文描述，应付账款相关问题的 AI 品牌可见度由约 3.2%升至 22.2%，约为原来的 7 倍。结果发生在限定的问题域，主要体现监测与内容策略的前台变化。[C01]

口径提醒。 页面正文与展示卡出现不同的时间和基线，尚不能无缝拼接为完整序列。讲授时应指明采用正文口径，并保留客户与供应商自报属性。[C01]

迁移方式。 适合借鉴“发现问题覆盖缺口，再针对性建设内容”。若要进一步判断业务价值，还需要固定问题面板、真实访问、CRM 连接和对照。不能将 7 倍可见度说成 7 倍收入。

13.5 GR0：高倍数营收案例该追问什么

案例声称。 GR0 分享某客户月度 AI 归类销售额从约 1,000 美元升至约 100,000 美元，动作包含基于问题洞察的内容建设与分发等工作。[C02]

证据边界。 客户与底层订单、归因规则、费用、同期活动和实验对照没有完整公开。小基数、渠道自然增长和追踪能力变化都可能影响倍数。这个案例说明存在值得研究的业务机会，不能直接确定 100 倍增长均由同一项 GEO 动作创造。[C02]

迁移方式。 要求补齐原始数量、净收入与退款、同一归因规则的前后重算、既有客户排除、处理日志以及可比对照。内容服务商可以学习其业务闭环，报价与回报承诺仍应建立在自己的证据上。

13.6 Omnilux：占比增长与金额增长需要分开

实际口径。 案例正文将关键变化表述为 AI 归因收入占总收入的比例由约 1%升至 3%，同时介绍了来源分析与内容迭代流程。它属于客户与供应商的第一方经营案例。[C03]

如何正确解释。 占比提高 2 个百分点，后期占比为前期 3 倍。若总收入不变，AI 归类收入额相应为 3 倍；若总收入变化，金额增幅也会变化。公开数字本身不能证明公司总营收增长 3 倍，也不能证明 AI 归类收入全部属于增量。[C03]

迁移方式。 电商品牌应同时报告渠道净收入额、总净收入、渠道占比、退款、新老客、归因规则和成本。只展示占比，容易遮住公司整体变化。

13.7 Arizona College of Nursing：咨询更靠近业务，但仍需追踪入学

实际口径。 案例报告 AI 来源网站访问增长 26%，入学咨询增长 51%，并介绍围绕课程、成本、比较等需求建设内容。[C04]

还没有证明什么。 咨询数量不等于有效申请、录取、入学或学费回款。案例还涉及更广泛的内容与营销动作，未公开随机对照，无法从公开信息完全拆分各项贡献。[C04]

迁移方式。 教育、专业服务和 ToB 业务可使用“咨询 → 联系成功 → 合格 → 商机 → 成交”的分层漏斗。比起把所有留资记成价值，更应检查新增咨询是否匹配服务范围、预算和购买阶段。

13.8 Alchemy：日志与自报互补，不能直接相加

实际口径。 案例报告 AI 引荐流量注册率约为其他来源 7 倍；另有用户自报发现渠道的变化。两类记录分别提供行为与回忆层面的线索。[C05]

证据限制。 进入网站的 AI 访客可能已经完成更多信息搜集，因此高注册率未必全部源自引荐渠道本身。自报发现渠道与实际最后点击也不必一致，两个口径不能拼成一个统一转化率。[C05]

迁移方式。 SaaS 或开发者产品可以同时保留日志来源与注册问卷，并通过用户标识计算重合。报告将“发现”“影响”“转化分配”分别呈现，避免同一用户跨多个证据来源重复计数。

13.9 WHOOP：事实纠错可以成为独立的经营结果

实际做法。 案例将 AI 回答中的产品事实错误转化为检查信源、修正内容和复测的工作流。重点在准确性治理，公开材料没有据此建立可审计收入增量。[C06]

为什么重要。 更多提及可能伴随旧信息或错误限制。对于更新频繁或决策成本较高的产品，正确解释型号、能力与适用条件，本身就有明确的管理价值。[C06]

迁移方式。 建立关键事实清单、错误优先级、纠错工单和配对复测；可以报告关键错误率下降及复发率。不要在缺少投诉、退货或收入实验的情况下，任意把每次纠错换算为固定金额。

13.10 Tally：AI 发现渠道与主动 GEO 增量需要分开

实际口径。 Tally 创始团队在 2026 年 9 月 22 日的经营文章中报告，43%的新用户通过 AI 工具发现产品。文章同时讨论长期产品、口碑与经营积累。[C07]

边界。 发现来源是用户获知产品的一种证据，不直接表示 43%的收入被某项主动 GEO 动作净创造。品牌可能自然受益于 AI 对既有产品与内容的采用。[C07]

迁移方式。 创业公司应区分“AI 成为重要渠道”和“这次新增 GEO 投入值得继续”。前者可以由来源调查说明，后者需要明确新增动作、成本、对照和业务结果。

13.11 Ahrefs：0.5%的流量与 12.1%的注册如何阅读

实际口径。 Ahrefs 的自家案例报告，近 30 天约 0.5%的流量对应 12.1%的注册，并称 AI 相对传统搜索的转化表现约 23 倍。原文以自有 Web Analytics 数据展开，本次没有找到将其描述为问卷与日志混接的充分依据。[C08]

需要保留的限制。 这是特定公司的短期渠道观察，流量单位、注册定义、舍入和原始计数仍影响复算。它说明 AI 到站人群可能具有较高意图，不能变成每家企业通用的 23 倍转化参数。[C08]

迁移方式。 在自身业务中按相同时间、相同去重单位和成熟窗口比较，并检查新老客、落地页与意图。更进一步的增量问题，需要干预设计，不能仅靠渠道转化率比较解决。

13.12 研究看似冲突，先对齐分母、时间和样本

材料组合	看似冲突的地方	更稳妥的阅读方式
Adobe 多个时期报告	AI 访客转化从某些早期时期较低，变为后期部分时期较高	按报告保留时期、行业和访客构成，不假定是完全固定的同一面板
Pew 与 Profound	一份讨论较少的即时点击，另一份讨论之后访问	两者终点和时间窗口不同，可以同时成立

材料组合	看似冲突的地方	更稳妥的阅读方式
Ahrefs 与 Seer 的 CTR 研究	长期相对下降与某阶段回升并存	排名条件、查询样本、基准期和 AIO 标记方式不同，不能用一条曲线否定另一条
单次测量研究与频次报告	一份强调重复波动，一份讨论大组合相对稳定	单问题精度、组合均值和长期趋势是不同判断对象

Adobe2026 年报告中的 AI 购物转化差异属于到站访客的观察比较。Pew 的浏览研究还存在搜索结果在后续日期重建的测量限制。Seer 使用的查询标签及预测月份也需要与实测月份区分。交叉校验应检查研究究竟测量了什么，不能只数有多少报告持相同观点。[B02][B03][B04][B05][B06][B09][B10]

本章课堂练习： 任取一个成功案例，用两句话介绍成绩，再用一句话说明限制。如果第三句话会改变前两句话的理解，就应把它一起放在演示页上。

第 14 章　五个国内业务教学场景

本章全部为虚构教学设计。数字用于解释计算与决策，不来自真实客户，不构成行业平均效果。具体统计区间若未提供原始样本，将明确标为假设输入。

14.1 高客单价 ToB：先证明有效商机链路

情境。 一家企业软件公司对 20 个主题簇进行内容改造，另 20 个相似主题保持既有运营。两组前后有效咨询分别为 200→280 和 180→216，按第 10 章的加性假设估计增量 44 条。与此同时，处理组后期识别到 112 次 AI 来源表单提交，去重与销售确认后为 70 条有效咨询，形成 18 个商机，当前成交 4 个。

正确读法。 112、70、18、4 属于可追踪漏斗；44 属于对整体主题结果的反事实估计。70 和 44 不能相加，也不必相等。主题内容动作可能经由 AI、SEO 和品牌搜索共同生效，所以 44 未必可以进一步全部分给自然 AI 路径。

管理动作。 销售周期较长时，先用成熟批次判断商机质量；未成熟商机不要全部记为收入。历史成交率可以用于预算情景，但预测金额必须与已实现回款分开。发现处理与对照存在前趋势不一致时，应收紧结论，保留可追踪漏斗和后续验证计划。

交付要求。 月报附线索去重规则、销售确认标准、商机阶段、同期广告与活动日志。服务验收可以覆盖可核验工作与测量质量；增量验收应依照预先定义的证据条件。

14.2 电商：收入漂亮，利润可能没有改善

情境。 两组可比市场的等长时期净收入如下。假定所有金额均已扣除退款，实验或准实验的可比条件已另行检查。

市场	前期净收入	后期净收入	增长
处理组	40 万元	52 万元	12 万元
对照组	40 万元	44 万元	4 万元

加性估计增量为 8 万元。同期分析工具按渠道规则分配给 AI 的净收入为 12 万元，项目费用为 4.5 万元，增量净收入的贡献毛利率假定为 45%。于是贡献毛利为 3.6 万元，扣除项目费用后为−0.9 万元，增量 ROI 为−20%。按渠道归因金额计算的收入费用比约 2.67，并不能推翻这一利润结论。

进一步假定，合格分析给出的收入增量区间为 2.5 万至 13.5 万元。这是为教学指定的输入，本表汇总数据不足以自行估出该区间。对应净贡献区间为−3.375 万至 1.575 万元，ROI 区间约−75%至 35%。企业应据此评估成本、可逆性和进一步学习价值，不能只展示正向收入点估计。

管理动作。 检查毛利率是否包含履约、退款、渠道佣金和额外客服等增量成本，避免重复扣除或遗漏。将获客新客与原有客户复购拆开，新增收入和订单替代也应分别分析。

14.3 本地服务：大量成交发生在电话与微信

情境。 当月 200 位新客户中，20 位有可识别 AI 触点；160 位回答发现渠道问卷，其中 50 位选择 AI；两种证据重合 10 位。至少一种 AI 相关证据可确认的客户为 60 位，占全体新客户 30%。

正确读法。 这 60 位是“记录中存在 AI 相关证据”的去重客户，不能直接称为新增客户。40 位未回答问卷的人存在信息缺口；有回答的人也可能回忆不完整。不能把 50÷160 简单放大到全体，再与日志客户相加。

管理动作。 培训咨询人员使用中性问题，保留原话与多选影响因素。用客户授权的业务标识完成去重，不依赖私自抓取私人通信内容。较稳定的业务可以尝试地区或主题试点，但公开内容跨地区传播时应先评估污染。

14.4 GEO 与 SEM 协同：谁应该获得那 10 个额外结果

情境。 假设四个随机可比组的相同业务终点为：基线 100、仅 GEO120、仅 SEM150、组合 180。GEO 单独增量 20，SEM 单独增量 50，交互项为 180−120−150+100=10。

正确读法。 组合增量为 80，不能把两个单独效果相加后宣称已解释全部。若企业采用平均边际贡献的分摊规则，GEO 分得 25，SEM 分得 55；这是对已识别组合效果的一种分配约定，并非自然界唯一的渠道答案。

管理动作。 对老板汇报整体组合结果，对渠道团队同时展示独立与交互作用。没有四组或其他可识别设计时，不能只因品牌搜索与广告转化同步增加，就断言协同量是多少。

14.5 品牌事实纠错：先把危险错误降下来

情境。 固定 200 个问题做前后配对复测，前期 60 条有关键错误，后期 20 条有关键错误。配对明细为：10 条持续错误、50 条由错变对、10 条由对变错、130 条始终正确。

正确读法。 错误率由 30%降至 10%，下降 20 个百分点；其中仍有 10 条新增错误，不能只展示修复成功的 50 条。因为同一问题前后配对，统计分析应利用配对结构，同时考虑重复生成和平台变化。

管理动作。 将持续错误、新增错误、修复后复发分别建工单，明确责任人与事实有效期。没有业务实验时，报告可限于准确性改善及处理时效。它可以支持品牌安全决策，但不需要捏造一个“每次纠错等于多少收入”的换算系数。

本章共同结论： 从数据到业务决策的最后一步，需要明确价值假设。能算出一个数，与这个数具备真实经营含义，是两个不同检查任务。

第 15 章　把效果转化为经营判断

15.1 先选对金额口径

指标	计算形式	适用解释
归因收入费用比	按规则分配的净收入÷项目费用	对归因账的回收情况描述
增量收入费用比	估计增量净收入÷项目费用	在因果假设下的收入效率
增量净贡献	增量净收入×适用贡献毛利率−项目费用	在明确成本口径下的利润变化
增量 ROI	增量净贡献÷项目费用	项目费用对应的净回报
增量获客成本	项目费用÷估计新增客户数	仅在增量客户数为正且定义可靠时解释
回收周期	累计增量贡献覆盖累计费用的时间	需区分预测与已经实现的回收

以上是本手册建议的管理计算框架。若贡献毛利率已扣除某项成本，项目费用里不要再次扣除；若费用只包含服务费，Token、内容审核、媒体分发和额外内部人力等应说明是否遗漏。收入、毛利和现金回款还涉及不同时间，不能混成同一个分子。

15.2 订单替代、老客与渠道迁移

用户从自然搜索转为 AI 链接下单，可能只是渠道重新分配，企业整体收入没有增加；也可能 AI 帮助其更快决策或提高客单价。需要观察总订单、总净收入、新客、毛利和购买时间，而非仅看 AI 渠道上涨。

老客户也可能受 AI 影响，但是否纳入某个“新客获客”项目，应预先约定。建立既有客户识别和回购周期，可以避免将长期积累的客户价值全部归给最近一个 AI 触点。eBay 的付费搜索实验提示，既有购买倾向与营销增量需要分开；它不能直接推出所有品牌词或 GEO 都没有价值。[F06]

15.3 长期价值可以预测，但要单列

订阅或高复购业务可能在首次订单后持续产生价值。可以用历史留存和毛利建立情景预测，同时披露留存、折现、扩购、退款与服务成本假设。预测的客户终身价值属于决策模型，不等同于已经实现的收入。

如果 AI 渠道吸引的客户与历史平均客户明显不同，直接套用全站平均终身价值会增加误差。先看成熟客户批次，必要时给出保守、基准和积极情景，不要将最乐观情景作为确定的项目成绩。

15.4 看边际回报，避免只追求总量

已经投入的成本与下一阶段可调整成本需要分开。一个项目总体盈利，不代表继续加倍内容仍然划算；总体尚未回收，也不代表应该立即停止一个已经形成可靠学习信号且可逆的试点。滞后、饱和与渠道替代都可能影响下一笔投入的边际价值。[F25]

管理建议可以采用三种表达：证据支持扩大一项具体动作；证据不足但允许保留较小试点以减少关键不确定性；结果和风险不支持当前做法，应调整或停止。每种建议都附条件，避免“GEO 值得投”这种范围过大的结论。

15.5 给管理层的五个问题

第一，测量终点是否与经营目标一致？第二，观察到的变化中有多少只是规则或平台变化？第三，对照是否可信，有没有外溢？第四，成本、退款、新老客与成熟周期是否处理清楚？第五，当前证据支持哪一个具体资源决策？

把这五问作为评审流程，就能同时避免两种偏差：因为无法完整归因而忽略真实机会，或因为行业趋势强而对每一项投入都免于验证。

第 16 章　落地路线、团队分工与效果验收

16.1 一个 90 天建设节奏示例

以下为组织与测量建设安排，不能保证 90 天内获得显著结果。销售周期、索引滞后、独立样本和可隔离程度决定真正所需的观察期。[F07][F16]

阶段	建设重点	可交付产物	进入下一阶段的检查
第 1 至 14 天	对齐问题与数据	目标定义、指标字典、来源规则、事件链与事实表	关键定义有负责人，测试链路能对账
第 15 至 30 天	建立稳定基线与试点协议	固定面板、采集日志、对照候选、样本量与风险评估	失败率和来源覆盖可解释，未因短期表现挑组
第 31 至 60 天	按协议实施并检查机制	干预日志、质量护栏、前台与后台分层看板	执行与采集稳定，重大同期事件已记录
第 61 至 90 天	按成熟情况分析与复盘	描述报表、归因账、合格增量分析或证据不足说明	所有主要结果完整披露，提出下一步决策

当实验或成交窗口未成熟时，最后阶段可以输出阶段报告，并明确未完成的业务观察。不要为了按日期交差而改变主结果、缩短成交周期或只保留已转化线索。

16.2 团队职责如何分配

工作	业务负责人	GEO 运营	数据与技术	销售或客户成功
目标与预算	最终决策	提出可执行动作	评估可测性	定义有效客户与业务约束
指标与实验协议	确认主要终点	提供处理与对照对象	负责方法、质检与分析	确认线索及商机口径
内容与事实维护	明确品牌边界	执行、版本化与复测	保留日志与数据	反馈真实问题和错误
业务对账	确認成本与回款口径	解释动作变化	连接、去重与核验	审核线索质量和成交状态
月度复盘	决定扩展、调整或停止	解释机制与操作	披露结果、区间和限制	解释客户变化与成交滞后

重要结论不宜由执行团队单方面定义和验收。运营可以负责改善，数据团队负责证据，业务团队负责价值判断；小团队可以一人兼多岗，但仍应保留这几种不同检查视角。

16.3 一个管理看板应有哪些块

业务层展示成熟有效线索、商机、净收入与贡献毛利；**前台层**展示固定面板中的正确提及、引用、信息覆盖与关键错误；**归因层**展示直接及辅助证据，并对订单去重；**增量层**展示对照、点估计和区间；**质量层**展示采集失败、来源识别、CRM 连接和规则变更；**行动层**展示本期决策及未解决问题。

尽量避免把各层压缩成一个无法解释的总分。领导应能看到“前台改善但成交尚未成熟”，运营应能看到“有流量但有效率下降”，分析团队应能看到“差异可能来自埋点切换”。

16.4 可复制的月度效果说明

目标与范围： 本期评估［业务线与人群］中的［干预包］，主要结果为［定义］，观察期为［日期］，成熟窗口为［长度］。

执行与数据： 计划处理［对象数］，实际执行［对象数］；采集成功率为［值］，CRM 连接覆盖为［值］。本期发生［平台、内容、价格、广告或测量变化］。

分层结果： 固定面板观察到［变化］；按［模型、窗口、作用域］记录的 AI 相关有效线索或净收入为［值］；在［对照与假设］下，增量估计为［值与区间］。尚未达到可识别条件时，明确写出原因，不填造一个增量数字。

边界与决策： 未排除因素包括［事项］。本期建议［一项具体决策］，依据为［证据］，下一次验证需要补充［数据或设计］。

16.5 采购与验收的关键项目

商务评估时建议确认：问题集及权重由谁维护；是否保存全部失败与原始回答；是否区分品牌提及和官网引用；是否提供规则版本与历史重算；是否能够连接有效线索及退款；是否有明确的新老客与窗口说明；是否能披露同期变化和不利结果。

验收可以分成可控交付与结果验证。可控交付包括事实梳理、内容发布、数据链路、日志完整性等；结果验证包括预先定义的指标、测量条件与证据要求。对平台自然波动和无法隔离的作用，应约定报告方式，避免用未经验证的固定排名或收入倍数代替验收标准。正式条款需要结合具体交易与专业审查，此处仅提供测量和商务评估框架。

16.6 哪些情况应该暂停下结论

发现来源规则刚变、对照被大幅处理、关键事件漏记、样本组构成异常、核心事实出现严重错误，或成交周期尚未成熟时，应先修复或说明限制。暂停强结论不等于否定全部工作；可以保留已经可靠观测到的部分，并明确接下来的验证条件。

第 17 章　课堂常见问题与误区纠正

17.1 AI 提及率上升，就能证明项目有效吗

可以证明所定义面板中的呈现发生了变化；是否由项目引起，需要同期对照与测量稳定性；是否有业务价值，还需要行为与经营结果。三种判断分开回答。[A01][A11]

17.2 没有点击，是不是完全没价值

用户可能在 AI 里获取信息后改用搜索、电话或其他渠道。可以研究这种影响，但不能在没有补充证据时给它任意金额。[B01][B03]

17.3 AI 渠道转化率很高，为什么不直接加大预算

转化率描述到站人群的条件表现，还需检查规模、边际增长、成本、既有购买意图和可复制性。高转化与高增量属于不同判断。[C08][F06]

17.4 可以把品牌搜索的增长都算给 GEO 吗

品牌搜索也会受广告、公关、产品发布和口碑影响。需要围绕干预设计对照；作为中介指标时，不能与最终成交重复货币化。[B07][F08]

17.5 GA4 已经有 AI Assistant，测量问题是否解决了

它改善了部分来源分类，但没有覆盖所有 AI 体验，也不能恢复所有跨设备、零点击和封闭渠道路径。AIO 和 AI Mode 还存在不同归类边界。[E12]

17.6 Direct 流量上涨，能否视为隐藏 AI 增长

Direct 表示系统没有获得可用的常规来源，来源可能很多。应使用自报、授权身份连接或研究设计补充，不能直接贴上 AI 标签。[E09][E10]

17.7 用 API 测到第一名，能否代表消费端第一名

API、产品入口、搜索工具、模型、上下文和用户状态都可能不同。API 实验结果应按其环境报告，再对目标消费端进行合规复测。[D06][D08][D11]

17.8 每个问题每天测一次够不够

取决于判断单个问题、组合均值还是长期趋势。先估计波动和独立性，再分配采样预算；不存在可直接套用的统一次数。[A11][B14]

17.9 能否只保留成功回答作为分母

技术失败与真实未出现应分开。可以报告有效回答中的比例，同时必须报告计划采集数、失败率和失败分布；失败随平台或处理组不同，可能造成选择偏差。

17.10 控制变量越多，因果模型越可靠吗

未必。干预后的中介或某些选择变量可能引入新问题；未观察到的混杂也不会因模型复杂自动消失。先画因果图并定义目标。[F08][F20]

17.11 95%区间很窄，就一定可信了吗

区间可能只覆盖模型内的随机误差，未反映错误分组、控制污染或来源误判。窄区间不能补救错误设计。[F07][F20]

17.12 没有显著增长，应该立刻停止吗

应先看是否有足够能力识别商业上重要的差异，再看费用、风险和追加验证价值。证据不足、效果接近零、效果为负，是不同情况。[F07][F16]

17.13 内容上线后等多久才能评估

没有跨业务通用答案。需要结合发现与索引延迟、用户决策周期、独立样本量、干预残留和预设窗口。运营检查与收入评估可以采用不同节奏。[F25]

17.14 竞品能否直接当控制组

竞品可能有不同用户、价格、投放和品牌事件，也可能同时做 GEO。它适合竞争背景观察，成为因果对照需要额外可比性证据。[F03][F04]

17.15 能否用 Shapley 分配结果证明真实因果

分配规则需要先明确价值函数和比较世界。预测解释或路径分摊中的 Shapley 值，不能自动证明干预效果；在合格实验已识别组合价值后，它可以作为一种分摊约定。[F18]

17.16 海外的高转化倍数能否用于国内报价

不宜直接套用。平台入口、用户行为、行业、支付与线索链路都可能不同。海外案例可以提供假设和设计启发，收益假设需要本地试点。[B02][C08][D11]

17.17 发布越多、被抓取越多，效果就越好吗

发布和抓取属于上游过程，需检查重复、质量、实际检索与答案吸收。低质量或错误内容可能增加无效工作甚至损害品牌。[A05][A12][D07]

17.18 供应商案例是否完全不能参考

有参考价值，尤其在工作流和问题选择方面。保持自报标签，核验数字与分母，不把个别成功作为可复制的因果保证即可。[C01][C02]

17.19 企业刚开始做，没有实验条件怎么办

先把目标、固定面板、事实表、来源识别、有效线索和干预日志做可靠；报告已观测结果与未知；同时寻找下一阶段可隔离的试点对象。不要急于制造一个精确增量。

17.20 最终应由哪个指标统一决策

经营目标决定主要终点，但安全、质量和数据完整性需要作为护栏。成熟项目可围绕增量贡献和风险决策；早期项目按阶段证据管理。没有一个总分可以替代全部判断。

第 18 章　教学组织、练习与复盘

18.1 一场 60 分钟的授课结构

时段	学习任务	推荐内容
0 至 10 分钟	对齐“效果”的层级	三本账；把提及、咨询和利润放入不同层级
10 至 20 分钟	理解链路与漏记	技术到商业路径；直接、延迟、协同与封闭渠道
20 至 35 分钟	学会选择验证方法	从问题定义到随机、双重差分和时间序列
35 至 48 分钟	审查真实案例	Glasp、Profound、Omnilux、Ahrefs 四个案例
48 至 57 分钟	做一份自己的评估方案	写干预、单位、终点、对照、窗口和未知
57 至 60 分钟	收束为经营决策	让下一笔投入对应一个可检验问题

若授课只有 30 分钟，可以压缩方法细节，保留三本账、两个案例和一份评估协议；若面向数据团队，可增加第 9 至 12 章的设计与练习。上述为教学安排建议，不要求学员在一堂课内掌握所有统计方法。

18.2 六道练习与参考答案

练习一： 提及率由 25%升至 40%，应如何报告?

参考答案：增加 15 个百分点，相对提高 60%，后期为前期的 1.6 倍。还要说明固定面板、样本数、有效采集、重复结构和不确定性。没有业务与对照证据时，不扩展成收入提升 60%。

练习二： 处理组有效线索由 100 变为 150，对照组由 80 变为 100，使用加性双重差分估计多少?

参考答案：50−20=30 条，处理组反事实为 120 条，相对反事实提升 25%。这依赖加性平行趋势等假设；不能把 30 写成无条件真实值。

练习三： 日志显示 40 位客户有 AI 触点，问卷显示 35 位受 AI 影响，重合 20 位，如何去重?

参考答案：至少一种证据出现的客户为 55 位。它是证据覆盖的并集，无法据此确定 55 位全部属于增量。未识别客户和问卷非响应还需单列。

练习四： 估计增量净收入 10 万元，贡献毛利率 40%，项目费用 5 万元，增量 ROI 是多少?

参考答案：贡献毛利 4 万元，净贡献−1 万元，增量 ROI 为−20%。收入费用比为 2，与净利润回报不同。成本口径若有重复或遗漏，必须先修正。

练习五： 只有已被 AI 抓取的处理页面增长很好，团队准备删掉未抓取页面再算实验结果，问题在哪里?

参考答案：抓取发生在分配之后，筛选可能破坏原有可比性。主分析按预先分配处理，执行与抓取作为机制或合规性诊断；其他分析需要单列并说明额外假设。

练习六： 一份研究主回归显著，但安慰剂检验不支持清楚区分真实干预与其他时间，如何汇报?

参考答案：同时报告主结果与稳健性限制，收紧因果措辞，并说明进一步需要何种对照或重复验证。不能只留下显著结果，也不能机械宣判所有观察均无价值。

18.3 学员应带走的一页方案

写清楚一个业务目标、一组干预动作、一个主要结果、一套固定测量协议、一个可信比较方式、一个成熟窗口和一份限制清单。随后列出由谁提供数据、由谁确认有效客户、由谁批准扩展。只要这页方案足够具体，后续选择工具和统计方法会容易很多。

18.4 建议的 12 份优先阅读

按理解顺序可阅读：F21 建立结果分层；A01 理解 GEO 研究目标；A11 理解测量波动；E12 确认当前渠道边界；E11 理解作用域；B01 认识延迟影响；A03 学习对照和不利检验；C03 练习收入分母；C08 练习渠道比较；F02 理解时间序列假设；F05 理解观测与实验差异；A04 认识高级研究方向及仿真边界。

其余材料按附录六类选读。文献阅读的目标是形成可检验的判断，不需要记住每篇论文的名字。对于一个重要论断，能够说明证据来自哪里、测量了什么、适用于谁，以及仍有什么未知，就已经建立了扎实的专业基础。

18.5 全书结语

GEO 效果归因的核心工作，是把内容动作、AI 表现、用户行为和业务结果连接起来，并为每一段连接提供适当证据。能看到的如实记录，能追踪的规范记账，能识别的增量谨慎估计，暂时无法识别的部分保留为待验证问题。

企业需要的最终产物，是一套能帮助下一次投入更明智的学习机制：明确目标，稳定测量，保留对照，公开限制，根据结果调整。这样，GEO 既能接受经营层面的检验，也能在不确定环境中持续积累可靠经验。

附录一　可直接复用的研究与复盘模板

模板 A：一页归因协议

项目	填写内容
业务问题	我们希望证明或排除的具体变化是什么？
干预范围	动作清单、处理对象、排除对象、开始时间及版本
目标人群	地区、语言、行业、购买阶段、新老客、产品范围
主要结果	唯一或少数主要终点、单位、分母、数据来源
次要结果与护栏	前台机制指标、事实错误、质量、投诉和性能
分组与对照	分配机制、历史可比性、对照受污染的可能性
时间计划	基线、执行、索引诊断、观察、成交成熟与分析日期
数据与去重	来源规则、客户标识、订单和事件去重、退款处理
分析方案	主要估计量、区间、聚类单位、敏感性与停止规则
同期因素	广告、SEO、促销、价格、产品、公关与平台更新
风险与限制	外溢、漏记、低样本、非响应、未观测混杂
决策规则	哪些结果支持扩展、调整、停止或继续小试点

模板 B：一次来源识别验收

检查点	执行记录
入口	平台、产品、设备、操作系统、浏览器或 App、是否联网
行为	打开回答中的链接，经过的跳转与落地页
原始证据	Referrer、URL 参数、时间、日志与页面证据
识别结果	原始来源、具体平台、渠道、规则版本
业务链路	表单、CRM、支付或预约系统是否保留关联
对账	测试事件是否重复，金额和时间是否一致
异常处理	缺失原因、修正方案、无法补全的范围

可控测试链路只证明该测试环境下的记录能力，不代表全体真实用户都具有相同覆盖率。覆盖差异应继续通过实际数据质量看板观察。

模板 C：公开案例审查卡

问题	应记录的答案
谁发布	客户、供应商、研究机构或独立审计者
测什么	可见度、访问、咨询、商机、净收入或利润
谁和何时	国家、平台、行业、样本单位、起止日期
原始分母	绝对量、比例、币种、基线、比较渠道
如何处理	动作范围、时点、同期活动、执行完整性
如何比较	随机、准实验、简单前后、渠道横向比较或自报
有什么限制	选择偏差、漏记、外溢、窗口、利益关系
可以借鉴什么	工作流、假设、指标或设计；不直接复制收益倍数

模板 D：月度证据等级记录

命题	观测结果	证据等级	替代解释	下一步
固定面板中的正确提及改善	填实际值、样本和区间	描述或实验结果	平台、面板或评估器变化	同条件复测
AI 相关有效线索增加	填去重与成熟值	规则归因	来源覆盖、客户构成变化	CRM 与原始来源对账
项目额外创造业务结果	填反事实差值及区间	对照支持或证据不足	对照污染、同期干预	稳健性检查或重设计

附录二　术语表与方法速查

术语	在本手册中的含义
GEO	面向生成式搜索与回答环境的内容、技术和信息优化；具体项目需定义动作范围
可见度	在约定观测环境中，品牌或内容被呈现的程度
提及	回答出现目标实体，需处理同名、别名与品牌关系
引用	回答提供来源标识或链接，需另行检查支持性
信息吸收	来源中的信息进入答案并被适当表达与归属
正确提及	实体出现且满足预设关键事实要求
代理指标	与目标有关、便于测量，但不等同于最终业务结果的指标
Referrer	请求携带的来源信息，可能因政策或跳转而缺失
UTM	用于标注可控链接来源等属性的参数
作用域	指标或维度属于用户、会话、事件等哪个统计层级
归因窗口	允许触点与结果建立分配关系的时间范围
成熟窗口	等待线索、订单或退款等业务结果充分发生的时间
规则归因	按选定模型将已记录结果分配给触点或渠道
增量	在明确假设或设计下，实际与无干预反事实的差值
反事实	同一对象在另一种干预状态下可能发生的结果
混杂	同时影响处理选择和业务结果的共同因素
中介	处于干预影响结果的传递路径中的变量
外溢或干扰	某对象接受的干预影响其他对象的结果
意向处理	按最初分配比较结果，不因后续执行是否成功而重新挑样本
平行趋势	未发生干预时，两组结果会按照可比较的方式变化的假设
双重差分	比较处理组前后变化与对照组前后变化之差
合成控制	用多个未处理对象的加权组合模拟处理对象的反事实
CausalImpact	以结构时间序列构建反事实预测的一类工具与方法
MMM	用时间或地区聚合数据分析营销组合与业务结果的方法

术语	在本手册中的含义
DML	结合正交化与交叉拟合等技术进行因果参数估计的方法体系
SHAP	解释预测模型如何使用特征的工具；自身不建立干预因果关系
统计功效	在特定真实效应和设计下，检验识别该效应的概率
聚类	观测共享某些因素而相关，例如同一问题、主题、用户或地区
安慰剂检验	在不应有真实干预的位置或时间，检查是否仍出现类似效果
SRM	随机实验的实际分配比例与预期比例不匹配的质量信号
贡献毛利	根据预先定义，收入扣除相关变动成本后的贡献
增量 ROI	增量贡献扣除项目费用后，相对于项目费用的净回报

附录三　110 份资料的分类阅读卡

本附录保留底稿中的 A 至 F 编号，方便与 Excel 交叉使用。标题、原始入口、摘要及方法边界共同构成阅读索引。推荐用途属于本手册的教学判断，不代表原作者对某项商业策略作出保证。

核读层级分为：R1，核读原始正文中与本主题相关的段落或方法；R2，主要核读原始摘要、研究概述或出版页，未逐表复核全文；R3，核对作者提供的书籍入口与目录，未通读整书。网页可打开不代表全部底层数据可获取；三类均未进行企业私有数据审计或全面复现实验。

所有链接以本次核验时可访问的原始入口为准。动态文档可能继续更新；教学涉及工具功能时，应同时查看原文日期及实际账户界面。论文的不同版本和同一机构的连续报告不能自动作为相互独立的复现。

从学习任务进入资料库

学习任务	优先组合	阅读时的检查点
建立效果分层	F21、A01、A10	输出、信息吸收与业务结果是否被混用
设计可重复监测	A11、B14、D06	个体问题与组合精度；产品入口与上下文
建立后台链路	E03、E11、E12、E15	事件、作用域、渠道和窗口的定义
研究延迟影响	B01、B03、C05、C07	即时点击、后续访问和自报发现的差异
估计项目增量	A03、F02、F03、F04	对照、前趋势、外溢与不利检验
形成经营决策	C03、C08、F07、F09	分母、样本、利润与模型适用条件

本版采用保守的核读记录：73 份标为 R1，34 份标为 R2，3 份标为 R3。即使网页存在全文入口，若本次主要依据摘要和研究概述，也保持 R2 记录。通过点击正文引用定位资料后，可以据此判断是否需要进一步核读原文细节。

A 类　GEO 机制、前台评测与内容质量

28 份；连接第 2、4、5、12 章及相关案例。先识别研究测量的是检索、可见度、引用还是信息吸收。

A01　GEO 奠基论文：优化的是可见度，不是销售额

GEO: Generative Engine Optimization

Aggarwal 等；KDD | 2024 | 同行评审 · KDD2024

内容摘要：提出生成式引擎优化及其评估基准，用回答中的来源呈现衡量网页影响。它适合解释 GEO 的技术起点，也最容易被错误引用为商业增长承诺：文中优化目标主要是回答可见度，并非企业营收。

核心内容：区分词数占比、位置加权可见度和主观呈现质量；比较补充统计、引文等内容改写策略，效果随领域变化；最高约 40%的提升属于实验内相对可见度，不能直接换算成成交增量。

学习用途：用于开场定义“GEO 效果”到底测哪一层。

边界：已进入候选文档集合是重要前提；测试不等于自然索引、真实用户曝光或销售实验。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A05][A06][A12][A14]　返回目录

A02　2023 至 2026 年 GEO 证据的批判性综述

Optimizing Visibility in Generative Engines: A Critical Survey of Generative Engine Optimization (2023—2026)

Olivier Martinez | 2026-07-15 | 预印本 · 未据此认定同行评审

内容摘要：按生成式搜索的处理阶段整理 45 项研究，并集中讨论固定上下文、指标不一致和外部有效性不足等问题。它提供了有用的阅读地图，但不能替代对被引原论文的检查。

核心内容：将搜索触发、抓取、检索、引用、吸收和业务结果拆开；提醒单篇内容的优势可能在竞争者同时优化时消失；指出其所综述语料中的长期、跨平台商业因果证据仍有限。

学习用途：用于搭建整个分享的证据地图，而非引用为行业定论。

边界：单作者选文与归纳存在主观性；“未发现”只适用于其检索范围。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A01][A05][A12][A03]　返回目录

A03　Glasp 自然实验：剥离平台增长后看 GEO 贡献

Disentangling Answer Engine Optimization from Platform Growth: A Log–Based Natural Experiment on ChatGPT Referral Traffic

Watanabe、Nakayashiki；Glasp | 2026-06-03 | 预印本・第一方自然实验

内容摘要：使用同一网站未优化页面作为参照，尝试把 ChatGPT 自身增长与内容优化贡献分离。最有教学价值的是作者同时报告显著回归结果和未通过的安慰剂检验，没有把流量总增长包装成确定因果。

核心内容：原始 ChatGPT 引荐增长 5.7 倍，而未处理页面也增长约 3.5 倍；中断时间序列估计干预水平跃升 1.82 倍，95%区间 1.31 至 2.54；时间安慰剂检验 p=0.16；作者将结论限定为提示性证据。

学习用途：建议作为“看上去增长≠全部由我们造成”的核心案例。

边界：存在前趋势、处理组与对照组意图差异、域级溢出和机器人过滤变更；四项改动捆绑实施。

核读范围：R1 ・ 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B01][F01][F02][F03]　返回目录

A04　GMMM：连接 GEO 可见度与业务结果的因果框架

Generative Marketing Mix Modeling: A Causal Inference Framework Linking GEO and GEM to Business Impact

Masahiro Kato、Daiki Honma、Taka Kato | 2026-09-10 | 预印本・方法与仿真

内容摘要：把品牌出现概率、问题需求、引擎使用份额及用户注意概率组成 GEO 媒体输入，再与业务结果建模。主题与本次分享高度相关，但商业处理效应主要由仿真生成，不能说已用真实营收验证 GEO 投资回报。

核心内容：用重复生成估计出现概率，并显式处理测量误差；区分自然 GEO、赞助展示及交互作用，讨论目标因果量；演示平台增长和处理后变量可能如何扭曲归因。

学习用途：用于介绍未来归因研究方向，并说明成熟落地前还缺什么数据。

边界：真实回答输入不等于真实商业因果验证；需求量、注意概率、模型识别均需额外假设。

核读范围：R1 ・ 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A03][F01][F09][F10]　返回目录

A05　C-SEO Bench：常见优化技巧是否稳定有效

C–SEO Bench: Does Conversational SEO Work?

Puerto 等；NeurIPS | 2025 | 同行评审・NeurIPS2025 Datasets and Benchmarks

内容摘要：系统评估会话式搜索优化策略，发现简单启发式并非在所有任务中有效，竞争环境也会削弱优势。它与只展示成功结果的营销案例形成互补，适合帮助听众建立实验而非照抄技巧的意识。

核心内容：用统一基准比较不同内容优化策略；区分问答、推荐等任务，避免平均效果掩盖异质性；竞争者共同优化时，需要重新估计增量优势。

学习用途：用来反驳“一种写法适用所有 AI 平台”的绝对化说法。

边界：不提供真实网站营收因果估计；具体策略效果需按原任务和引擎重测。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[A01][A06][A13][A14]　返回目录

A06 AutoGEO：学习不同引擎偏好，并保护内容质量

What Generative Search Engines Like and How to Optimize Web Content Cooperatively

Wu 等；CMU 等 | 2025/2026 | 同行评审・ICLR2026 已核验

内容摘要：通过学习生成式引擎对内容的偏好，形成协作式优化流程。它展示了可见度优化与事实、阅读质量并不必然对立，但实验优势仍主要是受控条件下的引擎输出变化。

核心内容：从引擎反馈学习偏好，而非依赖固定万能清单；强调事实保真与用户阅读质量约束；优化评估应与未改写内容及其他策略比较。

学习用途：作为“先小样本评测，再放大生产”的方法示例。

边界：固定候选上下文与真实互联网竞争不同；不能据此推导获客成本下降。

核读范围：R1 ・ 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[A01][A05][A07][A12] 返回目录

A07 MAGEO：复用优化策略，联合评估语义可见度与保真

From Experience to Skill: Multi-Agent Generative Engine Optimization via Reusable Strategy Learning

Beining Wu 等；ACL | 2026 | 同行评审・ACL2026 Findings 已核验

内容摘要：利用多智能体和可复用策略学习进行内容优化，并引入语义可见度与归因保真相关评价。适合讨论“引用了内容，却没有正确表达品牌”的前台测量问题，而非销售归因算法。

核心内容：经验沉淀为可复用策略，避免每次独立试错；使用双分支评估减少编辑过程带来的混杂；联合关注语义贡献和内容归属、事实保真。

学习用途：解释同一个“归因”词在内容评价与商业评价中的不同含义。

边界：论文中的 attribution fidelity 指内容归属保真，不等同市场营销收入归因。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A10][A21][A22] 返回目录

A08 不调用引擎的内容评分：能预测引用吗

Scoring Without the Engine: Validating a Deterministic, Manipulation-Resistant Content Score for Generative Engines, End to End

Elisha Bajemon、Andre-Louis Rochet | 2026-09-07 | 预印本・评分方法验证

内容摘要：检验一种确定性内容评分能否抵抗操纵并对应真实引擎表现。研究提醒，内容质量评分即便看似合理，也可能不足以预测引用；历史数据上的关联还可能在重新测量后消失。

核心内容：对抗式测试检查打分规则是否可被表面特征操纵；比较历史锚点与同时期重新测量的引擎结果；将质量筛选工具与引用概率预测器明确区分。

学习用途：用来提醒：自建 GEO 评分越高，不等于真实曝光或营收越高。

边界：预印本且预测关联弱；不是可直接销售的 GEO 效果预测公式。

核读范围：R1 ・ 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[A11][A26][B14] 返回目录

A09 Pinterest 生产实践：内容建设与自然获客增长

Generative Engine Optimization: A VLM and Agent Framework for Pinterest Acquisition Growth

Faye Zhang 等；Pinterest | 2026-02-03 | 预印本 · 企业生产实践

内容摘要：介绍 Pinterest 的视觉语言模型、趋势代理与集合页获客系统，正文包含线上 A/B 测试。整体自然流量增长与模块实验属于不同结果，均需保留原渠道及比较范围。

核心内容：线上比较显示，离线意图匹配较高的方案，线上注册和点击点估计未同步占优；需把模块实验与整套生产系统的约 20%自然流量变化分别报告；自然获客包含多条渠道，公开结果无法全部分配给 AI 引荐。

学习用途：用于讲解 SEO 与 GEO 协同项目为什么难拆单项贡献。

边界：已核读线上 A/B 章节，但未复跑实验或取得底层分组数据。表格缺少完整不确定性，不能据点估计判显著输赢。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[A03][D02][B10] 返回目录

A10 从引用选择到内容吸收：前台效果的两层测量

From Citation Selection to Citation Absorption: A Measurement Framework for Generative Engine Optimization Across AI Search Platforms

Zhang Kai、He Xinyue、Yao Jingang | 2026-04 | 预印本 · 多平台观测

内容摘要：提出将来源被选中与来源内容真正进入答案分开评估。引用链接只是外层证据，答案是否吸收品牌的关键事实、如何组织这些事实，属于另一层效果，尤其适合国内多 AI 平台的监测体系。

核心内容：602 个受控问题；报告 21,143 条有效搜索层引用；进一步抓取网页并构建内容特征，研究引用与吸收差异；从单一引用率扩展到回答内的真实信息贡献。

学习用途：用于定义“被引用”和“被正确讲出来”两种不同结果。

边界：作者含 Yao Jingang；涉及同名本人/团队时应披露关系，不应作为对自身业务的独立背书。不能推出营收因果。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A07][A21][A22][C06] 返回目录

A11 不要只测一次：AI 搜索可见度的重复测量

Don’t Measure Once: Measuring Visibility in AI Search (GEO)

Julius Schulte、Malte Bleeker、Philipp Kaufmann | 2026-04 | 预印本 · 测量研究

内容摘要：研究 AI 搜索可见度随重复生成波动的问题，提醒单次测试只能提供样本而非稳定排名。对 GEO 归因来说，前台测量噪声会直接影响是否把自然波动误认作优化贡献。

核心内容：以重复观测评估品牌出现分布，而非孤立截图；分离提示词差异与同一提示词重复生成的随机性；制定监测频次时应先明确判断对象是单词还是整体组合。

学习用途：作为“为什么本周 Top1 变化未必是优化效果”的解释材料。

边界：小样本地域和品牌环境未必代表所有平台；不宜强推统一的重复次数。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[B14][A20][E13] 返回目录

A12 SAGEO Arena：把检索纳入 GEO 评估

SAGEO Arena: A Realistic Environment for Evaluating Search–Augmented Generative Engine Optimization

Sunghwan Kim 等 | 2026-02 | 预印本・端到端基准

内容摘要：把检索、重排和生成环节放在同一评估环境中，弥补只在给定上下文里优化文章的局限。它支持按阶段定位问题：一篇文章写得更像 AI 答案，不一定更容易被搜索系统检索到。

核心内容：评估真实链路中更上游的检索变化；区分进入候选集合的概率与进入答案的概率；内容优化可能出现阶段间的收益冲突。

学习用途：用于展示“抓取→检索→引用→业务”的漏斗不能跳级。

边界：仍是实验环境，不是全网实际用户曝光与成交闭环。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A01][A05][D02] 返回目录

A13 E-GEO：电商推荐中的内容优化测试台

E–GEO: A Testbed for Generative Engine Optimization in E–Commerce

Bagga、Farias、Korkotashvili、Peng、Wu | 2025-11 | 预印本・电商实验

内容摘要：在电商产品推荐场景评估生成式引擎优化，发现常见启发式的效果并不普遍为正。它适合电商 GEO 团队学习如何把产品推荐概率设为中间指标，同时保留最终成交验证。

核心内容：围绕产品描述与推荐结果设计基准；比较经验规则与系统性优化，关注不同策略的异质性；推荐可见度与真实购物行为应分别验证。

学习用途：给电商品牌解释：榜单进入率是领先指标，不是订单增量。

边界：描述级优势不等于购买偏好改变；没有可直接复用的销售增长系数。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A05][A14][C03][B02] 返回目录

A14 竞争性 GEO：什么内容最终被引用

What Gets Cited: Competitive GEO in AI Answer Engines

Rahul Vishwakarma、Shushant Kumar、Ratnesh Jamidar | 2026-05 | 论文・arXiv 页面标注 SIGIR2026

内容摘要：通过大规模受控试验研究竞争内容进入 AI 答案的条件。文章强调相关性和上下文位置等因素的重要性，有助于解释为什么单篇文章优化实验不能直接代表真实竞争市场。

核心内容：观察竞争文档之间的引用份额分配；报告约 25.2 万次试验，比较相关性与其他内容信号；关注竞争者共同调整内容时的相对优势。

学习用途：用于讲解分母、竞争者变化和相对指标的归因陷阱。

边界：文档注入或固定环境与自然索引不同；引用份额增加不代表市场份额增加。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A05][A12][A23] 返回目录

A15 Google AI Overviews：触发、信源与发布者影响

Measuring Google AI Overviews: Activation, Source Quality, Claim Fidelity, and Publisher Impact

Haofei Xu、Umar Iqbal、Jacob M. Montgomery | 2026-05 | 预印本 · 实测观察

内容摘要：从 AI 摘要是否出现、引用了什么来源、表述是否有证据支持及发布者影响等维度开展测量。它有助于避免只以 AI 摘要的单一出现率解释网站流量变化。

核心内容：将触发率与引用/准确性分开统计；研究来源质量与答案主张的对应关系；将发布者影响放在查询构成与平台行为背景下讨论。

学习用途：与 Pew、Ahrefs、Seer 数据共同讲解“相关不等于因果”。

边界：不是随机向用户展示 AIO 的实验；不同查询意图会同时影响 AIO 触发和点击。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[B03][B09][B10][D01] 返回目录

A16 传统搜索、Gemini 与 AI 摘要不是同一流量入口

How Generative AI Disrupts Search: An Empirical Study of Google Search, Gemini, and AI Overviews

Riley Grossman 等 | 2026-04 | 论文 · 页面标注 SIGIR2026

内容摘要：对不同搜索产品的结果和来源进行实测比较。尽管产品都属于 Google 生态，其信息呈现和来源选择也不等价，不能用某一个渠道的监测结果代表全部 AI 搜索表现。

核心内容：使用约 11,500 个查询比较不同入口；分析传统排名与生成式结果来源的区别；说明跨平台统一可见度指标必须保留平台切片。

学习用途：解释为什么品牌在 Google 排名好，不等于 Gemini 表现同样好。

边界：来源差异并不直接告诉我们哪个渠道转化更高或增量更大。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A20][D01][D04][E12] 返回目录

A17 AI 搜索可见度：品牌与外部内容的关联

Generative Engine Optimization: How to Dominate AI Search

Mahe Chen、Xiaoxuan Wang、Kaiwen Chen、Nick Koudas | 2025-09 | 预印本 · 观察与策略研究

内容摘要：研究生成式搜索环境中的品牌呈现及内容信号，为理解第三方信息与品牌可见度关系提供材料。标题强烈，但阅读时应把研究观察与可重复的优化承诺严格分开。

核心内容：从 AI 答案中的品牌呈现分析外部内容信号；讨论 earned media 等内容来源的作用；为监测和内容策略提出可测试的方向。

学习用途：引出品牌基础、PR、SEO 和 GEO 共同变化造成的混杂。

边界：相关信号可能同时由品牌知名度驱动；无法证明新增一篇第三方文章必然带来曝光增长。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[B07][B08][F06] 返回目录

A18 CC-GSEO-Bench：从关键词转向来源影响

CC–GSEO–Bench: A Content–Centric Benchmark for Measuring Source Influence in Generative Search Engines

Qiyuan Chen 等 | 2025-09 | 预印本・内容影响基准

内容摘要：以内容本身对回答的影响为评估中心，补充简单关键词或品牌提及检测。适合研究既看来源贡献、又看内容质量的指标设计，但自动生成任务与自动裁判也可能引入偏差。

核心内容：把源文档内容影响作为主要评估对象；结合内容质量，避免只奖励高频重复；为生成式搜索优化提供更细粒度测试任务。

学习用途：用于设计“答案实际采用了多少品牌信息”的评价表。

边界：任务生成与自动裁判可能存在共同偏差；不等于真实消费者偏好测量。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A10][A22][A26] 返回目录

A19 结构特征与引用行为：可测试而非万能模板

Structural Feature Engineering for Generative Engine Optimization: How Content Structure Shapes Citation Behavior

Junwei Yu、Mufeng Yang、Yepeng Ding、Hiroyuki Sato | 2026-03 | 预印本・结构特征实验

内容摘要：分析内容结构特征如何影响生成式引擎的引用行为。适合将文章结构优化拆成可检验假设，并同时记录不同平台的响应，避免用“AI 喜欢某格式”代替证据。

核心内容：研究内容结构与引用的关系；对不同引擎开展比较，观察平台差异；为结构化改写提供实验候选变量。

学习用途：给团队示范一次只改一个主要因素的微实验思路。

边界：结构与事实、长度、主题可能同时变化；不能把实验结果外推为稳定排序规则。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A01][A05][A27] 返回目录

A20 生成式 AI 时代的网页搜索：来源重叠与波动

Characterizing Web Search in The Age of Generative AI

Kirsten 等；ACL | 2026 | 同行评审・ACL2026 Findings

内容摘要：比较生成式搜索与传统搜索行为，强调来源集合及重复查询结果可能变化。它给测量设计提供依据：平台、日期和生成条件必须保留，不能将一次采样当成固定事实。

核心内容：以约 4,706 个查询比较搜索表现；检查不同系统的来源重叠；分析重复观测中的结果变化。

学习用途：用于解释采样口径、版本记录和置信区间的必要性。

边界：文档或来源波动不等于用户真实看到的品牌曝光变化。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A11][A16][B14] 返回目录

A21 生成式搜索的可核验性：有链接不等于有证据

Evaluating Verifiability in Generative Search Engines

Nelson F. Liu、Tianyi Zhang、Percy Liang | 2023 | 同行评审 · EMNLP2023 Findings

内容摘要：研究生成式搜索回答的引用是否完整、是否真正支撑相关主张。它提供引用精确率与引用召回率的基础思路，适合区分表面背书与有效信息贡献。

核心内容：评估需要证据的回答主张是否带有引用；检查引用文档是否确实支持被引用的陈述；揭示引用覆盖与引用正确性是两个独立问题。

学习用途：给“AI 提及我但说错了”建立评价标准。

边界：2023 年的系统结果不能直接代表 2026 年的平台质量；内容核验不是商业归因。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[A10][A22][A25][C06] 返回目录

A22 ALCE：把回答质量与引用质量分别评估

Enabling Large Language Models to Generate Text with Citations

Tianyu Gao、Howard Yen、Jiatong Yu、Danqi Chen；EMNLP | 2023 | 同行评审 · EMNLP2023

内容摘要：提出生成带引用文本的基准与评估方法，将流畅度、答案正确性和引用质量分离。对 GEO 来说，它说明一个结果即使写得顺畅、引用很多，也可能没有可靠地吸收关键信息。

核心内容：分开评估内容正确性、引用覆盖与引用支持程度；提供可复用任务、系统比较与评估思路；将评价细化到陈述与证据的对应关系。

学习用途：用于搭建 GEO 前台的“引用×事实准确性”双指标。

边界：非商业搜索归因研究；自动评分需要与人工抽检校准。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A07][A10][A21] 返回目录

A23 模型认为什么证据有说服力

What Evidence Do Language Models Find Convincing?

Alexander Wan、Eric Wallace、Dan Klein；ACL | 2024-08 | 同行评审 · ACL2024

内容摘要：在具有冲突信息的问答任务中研究哪些证据特征影响模型判断。作者发现相关性在其测试中非常重要，而人类偏好的部分文风信号未必同样有效，可与 GEO 内容技巧研究交叉阅读。

核心内容：构建 ConflictingQA，保留不同事实和论证风格；进行敏感性与反事实文本分析；提示引用文献、中性语气等特征的效果需按任务验证。

学习用途：用于解释不同 GEO 论文看似冲突，可能源于任务差异。

边界：这是争议问答场景，不是品牌推荐或营收实验，不能简单套用。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A01][A05][A14] 返回目录

A24　子问题覆盖率：回答是否覆盖真正的决策需求

Do RAG Systems Cover What Matters? Evaluating and Optimizing Responses with Sub-Question Coverage

Kaige Xie 等；NAACL | 2025-04 | 同行评审・NAACL2025

内容摘要：把问题拆成核心、背景与后续子问题，评估回答是否覆盖用户真正需要的信息。对 GEO 内容团队，这比单纯计算品牌出现次数更接近“能否帮助消费者完成比较和决策”。

核心内容：区分不同重要性的子问题；同时观察检索与生成环节的信息缺口；用子问题覆盖与人工偏好进行对照。

学习用途：指导品牌知识库和问句集按照用户决策链设计。

边界：覆盖率与偏好关联不是对实际购买行为的因果验证。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[A10][A22][C05]　返回目录

A25　搜索可信度与回答有据性

Assessing Web Search Credibility and Response Groundedness in Chat Assistants

Ivan Vykopal 等；EACL | 2026 | 同行评审・EACL2026

内容摘要：将聊天助手搜索来源的可信度与最终回答是否由证据支撑分别检查。它适合作为品牌安全和信源质量的评价补充，帮助识别“引用多但质量差”的虚假繁荣。

核心内容：评估检索来源可信度；检查答案主张与来源的对应程度；强调搜索质量与回答质量不能相互替代。

学习用途：给法律、医疗等高风险行业增加准确性护栏。

边界：不能通过来源质量直接估算线索或收入；跨语言迁移需额外验证。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[A21][A22][C06]　返回目录

A26　G-Eval：用大模型评分，也要校准裁判

G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment

Yang Liu 等；EMNLP | 2023 | 同行评审・EMNLP2023

内容摘要：研究用大模型对生成文本进行结构化评价，并与人类判断比较。可以借鉴其评分流程，但模型给出的“品牌推荐质量分”依然是代理指标，不能自动视作用户偏好或真实商业价值。

核心内容：用明确标准与结构化流程提高评价一致性；将模型评分与人类标注比较；自动评价可能受模型偏好与任务设置影响。

学习用途：指导 GEO 看板中的语义、准确性、推荐强度评分质检。

边界：不是 GEO 专属指标；大模型作为裁判仍须人工抽查与重复测试。

核读范围：R2 ・ 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[A08][A18][A22]　返回目录

A27　写作前多目标优化：不仅追求被引用

Think Before Writing: Feature–Level Multi–Objective Optimization for Generative Citation Visibility

ACL2026 论文作者团队 | 2026 | 同行评审 · ACL2026

内容摘要：将引用可见度优化表述为特征层面的多目标任务，强调在生成内容前先考虑不同目标之间的平衡。它可启发 GEO 团队在追求前台效果时同时设置事实、质量与稳定性约束。

核心内容：从特征层面分析优化目标；在写作之前进行规划而非生成后盲目改写；用多目标思路处理可见度与其他质量要求。

学习用途：作为前台内容实验设计的延伸阅读。

边界：优化目标与真实销售目标仍有距离；不能用论文算法分数验收商业 ROI。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[A06][A07][A19]　返回目录

A28　引用失败诊断：先确定失败发生在哪一环

Diagnosing and Repairing Citation Failures in Generative Engine Optimization

arXiv 论文作者团队 | 2026-03 | 预印本 · 诊断方法

内容摘要：聚焦生成式引擎优化中的引用失败及修复流程，帮助把“没有效果”的模糊问题拆成可诊断环节。适合与检索、吸收及引用质量材料共同阅读，建立前台改进假设。

核心内容：围绕引用失败而非仅展示成功案例构建分析；把诊断结果连接到内容修复动作；通过复测判断修复是否改变目标结果。

学习用途：用于运营复盘：先找证据链断点，再决定优化动作。

边界：修复引用失败不意味着修复流量或收入问题；完整方案需读原文再实现。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[A12][A21][A22]　返回目录

B 类　用户行为、渠道变化与原创数据

17 份；连接第 8、12、13 章。重点检查人群、时间、分母、观察与预测，以及非随机比较。

B01　AI 提及后的访问提升：有反事实思路，但未采用随机实验

The AI Mention Effect

Profound；Nikolas Laskaris | 2026-07-01 | 供应商原创观察研究

内容摘要：利用美国用户面板将 AI 回答中的品牌提及与之后的网站访问连接，并用同一用户与品牌此前的访问窗口构造参照。它是 GEO 后链路研究的重要材料，但衡量的是访问概率，不是购买或利润。

核心内容：超过 200 万次对话；比较 ChatGPT、Gemini 与 AI Overviews；7 日访问提升估计：Gemini+3.21、AIO+2.96、ChatGPT+2.07 个百分点；依赖历史基线与观察性调整；残余意图变化和选择偏差仍可能存在。

学习用途：优先阅读：说明“看得见的引荐”与“可能影响了之后行为”不同。

边界：非随机曝光，残余意图混杂仍可能存在。结果是后续访问；2.47%指特定后续访问 URL 的参数口径，不能当全部识别覆盖率或 ROI 放大系数。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A03][F05][F06][E09]　返回目录

B02　Adobe 2026：AI 访客价值正在变化，不能沿用旧结论

AI traffic surges to retail sites, but many are not machine-readable

Adobe Analytics | 2026-04-16 | 第一方跨站分析报告

内容摘要：分析 2026 年初美国零售网站的 AI 来源访问与购买表现，同时关注站点内容的机器可读性。价值在于展示 AI 流量质量会随时间变化，不能把 2025 年的较低转化结论视为永远不变。

核心内容：2026 年一季度 AI 来源零售流量同比增长 393%；2026 年 3 月 AI 访客转化率高于非 AI 来源访客 42%；把机器可读性与流量表现并列分析，但二者并非自动构成因果关系。

学习用途：用于说明“AI 流量是否更值钱”必须带日期、行业和分母。

边界：AI 访客是被选择后的到站人群；不是 GEO 随机干预；渠道分组与样本覆盖影响结论。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B04][B05][B06][C08]　返回目录

B03　Pew：AI 摘要出现时，点击行为怎样变化

Google users are less likely to click on links when an AI summary appears in the results

Pew Research Center | 2025-07-22 | 独立研究机构・用户行为观察

内容摘要：基于美国用户的实际浏览行为比较带 AI 摘要与不带 AI 摘要的搜索结果。它揭示答案消费不一定产生网站点击，但由于查询类型不同，不能把观察差异全归为 AI 摘要本身的因果效应。

核心内容：有摘要时传统结果被点击的比例约 8%，无摘要时约 15%；AI 摘要内引用链接被点击的比例约 1%；研究对象是搜索行为，不是品牌认知、销售或利润。

学习用途：用于解释零点击及“引用量增加但网站点击不涨”的可能机制。

边界：观察性分组，查询构成可能不同；3 月浏览行为对应的搜索结果于 4 月重建，存在时点差异。不可直接外推为销售因果效应。

核读范围：R1 ・ 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A15][B09][B10][D01]　返回目录

B04　Adobe 早期基线：AI 引荐增长的低基数效应

Adobe Analytics: Traffic to U.S. retail websites from generative AI sources jumps 1,200 percent

Adobe Analytics | 2025-03-17 | 第一方跨站分析报告

内容摘要：给出 AI 来源访问早期增长与购物行为的观察基线。它的主要学习价值并非复制增长百分比，，实际为检查比较期间和低基数：高速增长并不代表总流量份额已经很大。

核心内容：报告相对 2024 年 7 月的访问增长；比较 AI 与其他访客的站内行为和转化表现；提供 AI 购物尚在早期阶段的历史基线。

学习用途：配合最新 Adobe 数据，展示趋势变化与比较基准的重要性。

边界：与 B02/B05/B06 属于同一数据提供方的时序材料，不能作为四个独立复现实验。

核读范围：R1 ・ 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B02][B05][B06][B17]　返回目录

B05　Adobe 2025 年中期：增长、份额与转化要一起看

Generative AI–powered shopping rises with traffic to retail sites

Adobe Analytics | 2025-08-21 | 第一方跨站分析报告

内容摘要：追踪 AI 辅助购物与网站引荐流量的增长，是理解不同时间点渠道成熟度的重要中期材料。和其他年度报告比较时，应保留各自基准日期，而非把所有增幅都当同比。

核心内容：报告 2025 年 7 月 AI 零售引荐的快速同比增长；描述购物者通过 AI 寻找建议与商品的行为；结合访问质量看增长是否产生可衡量业务价值。

学习用途：用于给图表加上明确基期，避免只讲“增长几千倍”。

边界：同一家分析平台客户样本；同比与相对 2024 年 7 月的指标不可混用。

核读范围：R1 ・ 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B02][B04][B06]　返回目录

B06　Adobe 2025 年末：跨行业 AI 流量不能混成一个均值

AI-driven traffic surges across industries

Adobe Analytics | 2026-01-12 | 第一方跨行业分析报告

内容摘要：分析 AI 来源访问在不同行业的增长，补充只看零售或 SaaS 的片面视角。对 GEO 归因，行业、季节和平台普及的共同变化都应进入解释，而非将行业顺风全部归功于运营。

核心内容：比较多行业 AI 引荐趋势；报告 2025 假日季零售 AI 来源访问的增长；提示行业间获客和转化行为存在明显差异。

学习用途：用于设置同行业、同周期参照，而非套用跨行业增长目标。

边界：季节性与平台增长不是 GEO 单项处理效应；同家时序报告非独立证据。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A03][B02][B12]　返回目录

B07　75,000 品牌相关性研究：强品牌与高可见度谁驱动谁

Top Brand Visibility Factors in ChatGPT, AI Mode, and AI Overviews (75k Brands Studied)

Ahrefs | 2025-12-12 | 供应商原创相关性研究

内容摘要：比较大规模品牌在不同 AI 平台的可见度与外部信号。适合筛选待验证的运营假设，但品牌本身规模、知名度和公关投入可能同时带来提及与 AI 可见度，不能把相关系数当因果系数。

核心内容：约 75,000 品牌的可见度与外部信号对照；不同 AI 平台的相关结构并不完全相同；品牌提及、搜索需求等变量可能共同反映既有品牌实力。

学习用途：用于示范“相关性只告诉你值得试什么，不告诉你投入回报多少”。

边界：反向因果、遗漏变量和品牌规模混杂；不能证明购买链接或增加发稿就能产生某幅度提升。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A17][F06][F18]　返回目录

B08　AI 访客停留短或跳出高，并不自动意味着无价值

AI Visitors Visit Fewer Pages and Bounce More Often Than Traditional Search Visitors

Ahrefs | 2025-06-24 | 供应商原创跨站分析

内容摘要：比较 AI 引荐与传统搜索访客的站内互动。它与一些高转化案例并不矛盾：页面浏览数、互动率和销售转化是不同指标，预先完成研究的访客可能访问更少页面。

核心内容：观察每次访问页数、跳出与互动差异；提醒访客行为指标不能代替订单或合格线索；需要按行业、落地页和用户阶段分组比较。

学习用途：用来防止把“更长停留”当作唯一 GEO 成功标准。

边界：不测量所有站点的最终成交；跳出指标定义和站点类型影响比较。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[C08][B02][F17]　返回目录

B09　Ahrefs 更新：AI 摘要与高排名页面点击率下降

Update: AI Overviews Reduce Clicks by 58%

Ahrefs | 2026-02-04 | 供应商原创观察研究

内容摘要：使用搜索数据更新 AI 摘要与自然结果点击率变化的估计。适合讨论发布者面临的点击压力，但标题中的百分比只适用于研究设计、查询样本和排名条件，不能写成全部 Google 流量减少 58%。

核心内容：研究高排名结果在 AI 摘要环境中的 CTR 变化；更新早期研究，展示效应估计随时间改变；需要区分相对下降与百分点下降。

学习用途：与 Seer 和 Pew 并置，训练听众读懂不同研究的分母。

边界：查询组合、排名和时间段影响结果；不能直接转为企业收入损失。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B03][B10][A15]　返回目录

B10　Seer 2026 更新：AI 摘要 CTR 也可能阶段性回升

AIO Impact on Google CTR: 2026 Update

Seer Interactive | 2026-04-24 | 服务商原创客户数据研究

内容摘要：用客户查询数据持续跟踪 AI 摘要与 CTR 的关系。最新更新出现阶段性回升，为“AI 摘要一定持续压低所有点击率”的单向叙事提供反例，也提醒查询结构和平台界面会变化。

核心内容：带 AIO 查询的 CTR 从 2025 年 12 月约 1.3%升至 2026 年 2 月约 2.4%；这属于其样本中的阶段性变化，而非全网统一趋势；比较引用品牌与未引用品牌时也要注意品牌与查询选择偏差。

学习用途：用于讲解跨研究校验必须先对齐时间、样本与指标。

边界：与 Ahrefs 样本、查询口径不同；表面方向不同不必然意味着其中一个研究错误。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B03][B09][A15]　返回目录

B11　Semrush：AI 流量价值与未来预测的边界

We Studied the Impact of AI Search on SEO Traffic

Semrush | 2025-07-21 | 供应商研究 · 观察与预测混合

内容摘要：将 AI 来源访问价值、品牌可见度和未来搜索流量变化放在一起讨论。阅读时必须把已观察到的渠道转化差异，与对未来流量和经济价值的预测分开，不能将预测写成已发生的效果。

核心内容：报告其所观察的 AI 来源访客转化价值约为传统搜索访客 4.4 倍；研究范围和业务背景限制这一比值的可迁移性；未来 AI 价值或流量超过传统搜索的时间属于预测。

学习用途：示范一张报告里“观察值、解释、预测”三类内容如何分开。

边界：4.4 倍不是通用收入乘数；未独立审计其原始转化数据，不能直接应用于国内平台。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B02][C08][C05][B12]　返回目录

B12　数十亿访问的渠道结构：增长快也可能份额小

We analyzed billions of web visits: How AI is reshaping traffic channels

Semrush | 2026-04-27 | 供应商原创流量面板研究

内容摘要：从跨行业网站访问结构观察 AI 渠道扩张，强调增长率和总量份额应共同报告。对企业归因，它提供平台自然增长的背景，却不能证明某家企业的 GEO 动作造成了行业增长。

核心内容：使用大规模访问面板比较渠道结构；报告 2025 年 AI 流量增长，同时总体份额仍较小；行业切片有助于建立外部参照。

学习用途：用于告诉客户：既不要因高增速夸大规模，也不要因低份额忽视趋势。

边界：识别到的 AI 引荐不是 AI 全部影响；面板覆盖与来源归类规则会影响份额。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B06][B17][A03]　返回目录

B13　商业查询中的 AI 摘要扩张：查询结构本身在变化

AI Overviews commercial search study

Semrush | 2026-07-02 | 供应商原创查询研究

内容摘要：追踪 AI 摘要在商业搜索意图中的扩张，帮助解释为什么不同时间监测同一个行业可能出现系统性变化。处理组与对照组若查询类型不同，平台向商业场景扩张可能造成归因偏差。

核心内容：约 60 万关键词、10 个行业的查询跟踪；观察 2025 年 11 月至 2026 年 4 月的商业搜索变化；商业、信息与品牌查询需要分别建立基线。

学习用途：帮助设计固定问句组及品牌词/非品牌词分层。

边界：关键词数据库不等于真实提示词总量；出现率扩张不是品牌实际曝光增量。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A15][B09][B10]　返回目录

B14　Profound：每天一次何时够用，何时不够

Is Once a Day Enough?

Profound；Jennifer Zou | 2026-07-08 | 供应商原创测量研究

内容摘要：比较大提示词组合的日常监测与更多重复采样，讨论组合平均值、采样预算和平台变化。其公开公式与报告计数口径仍有需要澄清之处，适合作为批判性阅读材料。

核心内容：单个问题的精度与大型组合的均值稳定性需要分开；页面展示的标准差公式与常用独立伯努利均值标准误写法存在差异；不能据此规定所有品牌、问题和平台的统一采样次数。

学习用途：用于给采集频次预算设计边界，而非规定万能次数。

边界：本次未获得原始数据与代码。公式展示、计数及区间描述需进一步澄清；不直接采用其精度数字作为采样标准。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A11][A20][F17]　返回目录

B15　Bain：零点击时代应扩大衡量范围

Goodbye Clicks, Hello AI: Zero-Click Search Redefines Marketing

Bain & Company | 2025-02-19 | 咨询公司原创消费者研究

内容摘要：通过消费者研究讨论无需访问网站即可获取答案的趋势。适合作为企业高管理解 GEO 评估变化的商业背景，但问卷中的自报行为与真实点击日志、因果实验属于不同等级证据。

核心内容：研究用户在搜索结果或 AI 回答中直接获得信息的行为；提示营销评估不能只看网站到站点击；品牌被理解、被考虑与最终购买应建立衔接指标。

学习用途：用来解释为什么 GEO 需要同时看认知、行为与业务结果。

边界：自报频率存在记忆偏差；报告中的行业流量估计不能代替客户真实数据。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B03][B16][F21]　返回目录

B16　McKinsey：AI 搜索成为新入口，如何看待规模预测

New front door to the internet: Winning in the age of AI search

McKinsey & Company | 2025-10-16 | 咨询公司原创研究与预测

内容摘要：基于消费者调查讨论 AI 搜索对品牌发现与商业决策的影响，提供适合管理层的业务语言。它可以说明为什么值得投入评估，但不能拿未来市场规模预测证明今天某个项目的 ROI。

核心内容：调查约 1,927 名受访者的 AI 搜索使用；讨论品牌发现和决策入口变化；对未来受 AI 影响的商业规模给出预测情景。

学习用途：适合开场交代问题重要性，随后立即切换到可验证证据。

边界：意向、使用情况、受影响消费与增量营收不是同一个指标；预测不是已实现收入。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B15][B01][F21]　返回目录

B17　Ahrefs：可追踪 AI 引荐份额小，不等于影响只有这么多

AI Makes Up 0.1% of Traffic, but Clicks Aren’t Everything

Ahrefs | 2025-03-26 | 供应商原创网站分析

内容摘要：描述早期可识别 AI 来源在网站访问中的份额，并讨论点击以外的影响。它帮助区分“被统计到的访问占比”与“AI 在消费者决策中的总影响”，两者不能互相替代。

核心内容：统计可识别 AI 引荐在网站流量中的份额；比较不同站点、行业或地域的分布；说明单靠点击并不能测完整品牌影响。

学习用途：用于讲清“可观测结果是下限线索，不是可随意放大的增量下限”。

边界：只覆盖被识别的访问；不能凭不可观测部分任意放大归因收入。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B01][B12][E12]　返回目录

C 类　企业案例与第一方经营实践

8 份；连接第 13 至 15 章。重点学习工作流，区分自报业务结果与独立因果证据。

C01　Ramp：7 倍可见度不是 7 倍收入

How Ramp increased AI brand visibility 7x in accounts payable

Profound × Ramp | 页面未明确标注 | 企业/供应商自报案例

内容摘要：展示财务软件品牌在应付账款相关问题上的 AI 可见度优化实践。它适合作为监测驱动内容策略的案例，但结果主要是限定问题范围内的品牌可见度，而非公司的收入增长。

核心内容：限定在 accounts payable 问题域观察品牌可见度；围绕监测缺口制定内容优化与分发动作；供应商报告可见度增长约 7 倍。

学习用途：用来示范如何给成功案例的指标加上限定词。

边界：供应商与客户自报，未独立审计；正文与展示卡出现不同期间及基线。保留明确正文口径，不拼接为完整序列；可见度不等于收入。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A01][A11][B07]　返回目录

C02　GR0：从月 1 千到 10 万美元，先检查收入归类和对照

How GR0 took a client from $1K to $100K/Month in AI-driven sales using Profound

Profound × GR0 | 页面未明确标注 | 企业/供应商自报案例

内容摘要：代理商分享用提示词洞察指导内容建设后，某客户 AI 归类销售额增长的经历。它最有学习价值的地方是把内容策略与业务结果连接，但高倍数结果必须同时审查小基数、渠道定义与平台自然增长。

核心内容：案例声称某客户月度 AI 驱动销售额由约 1,000 增至 100,000 美元；借助提示词数据发现内容机会并持续发布；把监测结果用于运营反馈，而不只做展示看板。

学习用途：用于给听众练习审问高增长案例所需的六个问题。

边界：无法独立核对全部销售流水、归因窗口及反事实；不能写成已证明 GEO 创造 99,000 美元净增量。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A03][E01][F05][F06]　返回目录

C03　Omnilux：收入占比 1%到 3%，不等于总收入增长 3 倍

How Omnilux tripled AI-attributed revenue

Profound × Omnilux | 页面未明确标注 | 企业/供应商自报案例

内容摘要：护肤设备品牌将 AI 来源分析接入电商体系，报告 AI 归类收入占总营收的比例从约 1%升至 3%。标题容易被误读，讲解时应保留“AI 归类收入份额”而非“公司营收”这个分母。

核心内容：利用 Shopify 相关分析基础设施观察 AI 来源和内容抓取；案例正文把关键变化写为 AI 归类收入份额约 1%→3%；用数据识别内容覆盖空白并持续迭代。

学习用途：很适合讲“增长的是数值，还是占比；改变的是业务，还是识别能力”。

边界：收入份额增加可能同时受其他渠道下降或识别改善影响；不等于公司总收入增加三倍。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B02][E03][E12]　返回目录

C04　Arizona College：AI 来源咨询量增长，而非入学人数

How Arizona College of Nursing increased AI-referred traffic conversions by 51% with Profound Agents

Profound × Arizona College of Nursing | 页面未明确标注 | 企业/供应商自报案例

内容摘要：围绕课程比较、费用等需求建设内容，并将 AI 来源访问与报名咨询联系起来。案例把业务推进了一步，但其 51%增长对应的是入学咨询，并不是最终入学、缴费或学生生命周期价值。

核心内容：AI 来源入学咨询增长 51%；AI 驱动网站访问增长 26%；围绕课程比较和成本问题补齐高意图内容。

学习用途：适合迁移到国内教育、律所、ToB 服务的线索闭环。

边界：咨询次数、合格申请、录取、入学与缴费须分别统计；季节性招生周期会影响前后对比。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E03][E06][E15][A24]　返回目录

C05　Alchemy：同时看引荐转化与用户自报来源

Alchemy drives a 7x higher signup rate from AI-referred traffic with content Agents

Profound × Alchemy | 页面未明确标注 | 企业/供应商自报案例

内容摘要：开发者平台用 AI 来源访问及注册问卷双轨观察获客，报告 AI 访客注册率高于其他来源。它展示了补充自报归因的价值，也揭示两个口径不能直接相加或拼接成同一个转化率。

核心内容：案例报告 AI 引荐访客注册率约为其他来源 7 倍；AI 在自报注册来源中的份额一年增长约 3 倍；围绕开发者真实问题建设内容并持续监测。

学习用途：用于解释“软件归因+问卷归因”交叉验证，而非重复计算。

边界：自报发现来源与实际最后一次访问来源不同；注册不等于付费；高意图选择偏差仍在。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[C08][B01][E02][E14]　返回目录

C06　WHOOP：品牌事实纠错也属于 GEO 结果

How WHOOP turned AI accuracy monitoring into a correction engine

Profound × WHOOP | 页面未明确标注 | 企业/供应商自报案例

内容摘要：可穿戴设备品牌将 AI 回答中的产品事实错误转为可执行纠错任务。它提醒评估体系不应只奖励提及量：错误的规格、价格或能力描述，可能扩大曝光却损害消费者决策。

核心内容：监测品牌与产品事实的准确性；追溯错误或不一致信息对应的信源；将发现的问题转为内容和信息修正工作流。

学习用途：适合分享“被正确理解，比单纯被提及更重要”。

边界：事实纠正与销售变化之间尚需独立测量；案例未提供可复现的利润增量。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A10][A21][A25]　返回目录

C07　Tally：用户自报发现渠道与企业增长不要画等号

6 years in, $6 million far

Tally 创始团队 | 2026-09-22 | 企业/供应商自报案例

内容摘要：Tally 公开讨论 AI 搜索在用户发现中的占比，并与产品增长和社区经营一起复盘。最新文章称 43%的新用户通过多种 AI 工具发现 Tally；这描述获知渠道，不意味着 43%的营收由某项 GEO 动作净创造。

核心内容：2026 年 9 月文章报告 43%的新用户从 AI 工具发现产品；2025 年文章已称 ChatGPT 成为第一引荐来源；增长同时包含产品体验、口碑、内容与分发的积累。

学习用途：用来讲品牌自然受益与可归因的主动 GEO 投入之间的差别。

边界：发现渠道定义和追踪细节未完整披露；ARR 总额不能全部归于 AI，也不能归于单次优化。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[C05][C08][A03]　返回目录

C08　Ahrefs 自家案例：0.5%访客与 12.1%注册的口径审查

Does AI Search Traffic Convert Better Than Traditional Search? For Ahrefs, Yes: 0.5% of Visitors Drove 12.1% of Signups

Ahrefs | 2025-06-16 | 企业/供应商自报案例

内容摘要：Ahrefs 以自有 Web Analytics 近 30 天数据报告约 0.5%流量对应 12.1%注册，并讨论相对传统搜索的高转化。它是单站渠道观察，需要保留原始单位、比较基准与舍入口径。

核心内容：原文将流量和注册份额均放在自有分析的近 30 天观测中；没有充分依据将本案例描述为问卷与日志混接；23 倍对比传统搜索，不能直接由两个总体份额相除复算。

学习用途：用于说明 AI 到站流量可能少而精，但仍需自己的 CRM 验证。

边界：单一高意向 SEO 工具网站；分母是其自身访问而非全行业用户。23 倍不能直接由两个总体份额相除得到，也不能视为 GEO 处理效应；无随机对照和独立原始数据审计。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B02][B08][B11][C05]　返回目录

D 类　平台官方机制与可观测边界

12 份；连接第 2、4、6 章。产品功能随时间更新，适用范围需与实际账户核对。

D01　2026 年 Google 新增 AI 表现洞察：先更新归因工具地图

New controls and insights for website owners

Google | 2026-06-03；更新 2026-08-31 | 平台/标准官方文档

内容摘要：Google 公布网站 AI 使用控制与 Search Console 中的 AI 表现洞察，页面由 6 月公告更新至 8 月全球推广信息。它说明“平台完全不提供 AI 表现数据”已不是稳妥表述，但曝光洞察仍不能直接证明成交增量。

核心内容：关注页面、站点及国家层面的 AI 表现洞察；区分新曝光洞察与传统搜索流量报表；按上线范围和更新时间解释可观测性。

学习用途：开场纠正过时判断，并列出当前平台可获取的证据。

边界：产品功能说明不是效果实验；AI 曝光不等于访问、独立用户或销售；具体账户可用界面需现场核对。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D02][D04][E12]　返回目录

D02　Google AI 功能的网站要求与 Search Console 统计边界

AI features and your website

Google Search Central | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方说明网站参与 AI Overviews、AI Mode 的基本要求及搜索表现统计。阅读时应与 2026 年新增 AI 洞察公告配套：原有 Web 搜索报表包含相关流量，并不意味着所有 AI 指标都可独立拆分。

核心内容：AI 功能沿用基础 SEO 与索引要求；搜索流量报表与 AI 专属洞察并非同一指标；被索引或符合要求不保证被选中。

学习用途：界定官网 GEO 诊断与效果报表各自能回答的问题。

边界：不能从符合技术要求推导引用概率，更不能把全部自然搜索增长认作 AI 增长。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D01][D03][E12]　返回目录

D03　Google 官方 AI 优化指南：避免把特殊格式当作因果秘诀

AI optimization guide

Google Search Central | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：指南把 AI 可发现性放在网站内容质量、可访问性与搜索优化的连续体系中理解。适合校正“装一个文件或增加一种标记就必然获得 AI 推荐”的说法，并设计可验证而非包装式的优化假设。

核心内容：优先独特、可信且对用户有用的内容；检查抓取、索引和内容可理解性；把优化建议转化成可测假设。

学习用途：用于解释 GEO 投入项与最终业务结果之间仍需验证的链路。

边界：官方最佳实践有指导权威，不提供单项措施的商业提升幅度。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A01][A05][D02]　返回目录

D04　Bing AI Performance：引用量不是排名或流量

Introducing AI Performance in Bing Webmaster Tools Public Preview

Microsoft Bing | 2026-02-10 | 平台/标准官方文档

内容摘要：Bing 官方 AI 表现报告提供引用、被引用页面、抽样的 grounding queries 和时间变化。这是较直接的发布方 AI 证据，但官方明确提醒不能据此理解排名、重要性或答案中的具体展示位置。

核心内容：引用次数与被引用页面是核心观测；grounding query 不等于用户逐字提问；报告不代表排名、权重或推荐位置。

学习用途：把平台自有观测与第三方提示词监测分开呈现。

边界：仅覆盖报告声明的服务与范围；引用不能直接映射到用户点击，更不是去重成交。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D05][A11][E03]　返回目录

D05　Bing 新增意图、主题、引用份额与对比视图

New AI Visibility Insights in Bing Webmaster Tools: Intents, Topics, Citation Share & Compare

Microsoft Bing | 2026-06-16 | 平台/标准官方文档

内容摘要：6 月更新将 Bing AI 观测从引用总数扩展到意图、主题及比较视角，有助于追踪哪些需求方向出现变化。它更适合运营诊断，不能把引用份额直接当作真实市场份额或销售贡献。

核心内容：按意图和主题拆解 AI 可见度；引用份额必须保留平台定义与分母；时间比较可定位变化，但不能排除外部因素。

学习用途：设计主题级运营看板和平台变化日志。

边界：同一平台的连续公告不是两项独立效果研究；份额的范围受报告样本与覆盖限制。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D04][A03][B14]　返回目录

D06　ChatGPT 搜索的触发与上下文：为何模拟结果不能当全民曝光

Searching the web with ChatGPT

OpenAI | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方帮助说明搜索可自动触发，也可由用户主动选择；查询构造会受到对话上下文等因素影响。这有助于说明同一句测试词的可见率不等于所有真实用户的曝光概率。

核心内容：固定测试时记录搜索模式与上下文；用户对话与查询重写影响候选来源；抓取许可不意味着引用保证。

学习用途：制定提示词测试协议并说明样本代表性。

边界：没有提供营销转化实验；API 输出、无登录测试与真实个性化对话不可简单等同。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D07][D08][A11]　返回目录

D07　OpenAI 爬虫分类：抓取量不应进入获客量

Overview of OpenAI Crawlers

OpenAI | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方区分用于搜索、模型训练以及用户请求访问的机器人。服务器日志由此可以服务技术诊断，但必须先做机器人分类，避免把机器抓取当作真实 AI 访客、线索或曝光。

核心内容：分别识别 OAI-SearchBot、GPTBot 与 ChatGPT-User；训练许可与搜索发现性是不同管理问题；结合官方 IP 信息验证身份，而非只信 User-Agent。

学习用途：建立机器人日志看板，并与人类访问漏斗彻底分开。

边界：一次抓取不证明页面已展示给用户；允许抓取也不保证入选答案。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D06][D09][E09]　返回目录

D08　OpenAI 联网搜索 API：保存引用及检索证据

Web search | OpenAI API

OpenAI | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：文档展示联网工具、来源注释及搜索结果相关能力，适合构建可重放的 GEO 实验记录。它提供可观测的 API 过程，而非 ChatGPT 消费端所有用户的真实使用记录。

核心内容：记录工具调用、原始答案和引用注释；区分返回引用与更完整检索来源；保存域名限制等实验配置。

学习用途：为每条观测保留原始证据和可复核参数。

边界：API、模型配置及域名过滤改变测试环境；不能把 API 引用率宣称为消费端市场曝光率。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D06][D10][D12]　返回目录

D09　Perplexity 爬虫：可访问性与业务访问分账

Perplexity Crawlers

Perplexity | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方说明搜索爬虫和用户触发访问的用途及访问配置。可据此检查 CDN、robots 或防火墙是否误拦，同时提醒归因报表不要把机器活动混进真实访客统计。

核心内容：区分索引型抓取与用户请求型访问；核对官方访问配置和身份信息；技术可达性只是效果链路的前置条件。

学习用途：补全多平台抓取诊断和机器人过滤规范。

边界：机器请求不等于用户到站；日志中的成功响应也不等于已被引用。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D07][D12][E09]　返回目录

D10　Gemini 检索元数据：连接查询、来源与答案片段

Grounding with Google Search

Google AI for Developers | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：Gemini 的搜索 grounding 元数据可包含搜索查询、来源片段和答案支持映射。研究者可据此区分“被检索”“被引用”“支撑了哪段答案”，比只统计品牌字符串更接近内容贡献。

核心内容：保留 webSearchQueries 以理解检索方向；用 groundingChunks 记录检索来源；用 groundingSupports 关联答案与证据。

学习用途：建立来源贡献与引用忠实度的审计样本。

边界：字段是否返回取决于模型与调用；API 实验不代表 Google AI Mode 或消费者 Gemini 的完整行为。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A10][A21][D08]　返回目录

D11　阿里云百炼联网搜索：国内可控实验的官方入口

大模型如何联网搜索

阿里云百炼 | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：中文官方文档展示联网开关、强制搜索、搜索来源返回等配置。可用于构造国内模型的可复核搜索实验，但必须把百炼 API 测试与通义等消费端的真实会话、推荐曝光区分开。

核心内容：记录 enable_search 与 forced_search 配置；利用 enable_source 和 search_info 留存来源；指定站点等限制必须作为实验条件披露。

学习用途：给国内平台研究提供可复核的数据结构示例。

边界：指定来源会改变候选池；百炼 API 结果不能直接代表豆包、DeepSeek、元宝或千问 APP 用户整体曝光。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D08][D10][A11]　返回目录

D12　Perplexity 域名过滤：控制实验与真实环境分开

Domain Filter

Perplexity | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：文档说明搜索域名允许和排除规则，适合研究在特定来源集合内页面能否被发现或采用。它能提高机制实验可控性，但人为收窄候选来源后产生的优势不能作为自然市场表现。

核心内容：用域名过滤构造明确的实验候选池；保存允许/排除规则及版本；区分机制测试与开放网络复测。

学习用途：设计“受控机制测试→开放环境复测”的两阶段验证。

边界：过滤后的引用份额分母已变化，不可与开放网络份额直接比较。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A05][A12][D08]　返回目录

E 类　来源、身份、事件与归因数据

18 份；连接第 6 至 8 章。优先掌握作用域、去重、跨域、隐私和规则版本。

E01　GA4 归因模型：分配转化功劳，不等于证明 GEO 增量

Attribution models

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：文档定义当前可用归因模型及数据驱动归因的思路。官方说明其反事实建模利用 Google 广告随机实验训练；这不意味着某个企业 GEO 项目已被随机验证，更不会自动补齐未记录的 AI 曝光。

核心内容：区分数据驱动归因与末次点击规则；同一笔转化可因模型设置而重新分配；Google 广告模型依据不等于 GEO 项目因果证据。

学习用途：说明“报表归因收入”和“真正新增收入”为何不同。

边界：DDA 含模型与实验相关基础，但不能据此认为本企业自然 GEO 已被随机验证。当前可选模型与历史教学模型需分开，归因分配不自动等于项目增量。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[F05][F06][E15]　返回目录

E02　GA4 报告身份：跨设备归因的可见范围

Reporting identity

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方说明报告如何组合用户 ID、设备身份与建模来组织用户旅程。它帮助解释同一消费者在 AI 应用、浏览器和登录设备之间切换时，为什么访问和转化链条可能断裂。

核心内容：分清用户身份与设备身份；不同报告身份设置影响聚合结果；跨设备连接依赖可用且合规的身份信号。

学习用途：在归因方案中明确跨设备覆盖率与无法连接的样本。

边界：未知或未授权身份不会自动被恢复；模型化身份不是逐人确定的真实路径。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E03][E08][E17]　返回目录

E03　GA4 BigQuery 导出：从渠道汇总走向可审计事件链

BigQuery Export schema

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方模式给出事件级数据结构，是将会话来源、用户标识、关键事件与自有 CRM 连接的基础。应先统一时间、粒度和去重规则，再计算 AI 渠道的有效线索与成交。

核心内容：保留事件时间、用户与会话连接字段；区分用户来源、会话来源和事件归因作用域；用业务唯一 ID 连接订单并处理重复事件。

学习用途：搭建 AI 访问→留资→有效线索→回款的数据底座。

边界：导出不自动解决同意缺失、晚到数据和跨设备断链；原始事件与界面报告口径可能不同。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E02][E11][E16]　返回目录

E04　跨域测量：避免支付或表单跳转截断 AI 路径

Set up cross-domain measurement

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：当官网、表单、预约或支付分布在不同域名时，跨域配置用于维持可连接的站内旅程。它解决的是自有域之间的衔接，而非读取用户在 AI 应用里曾看过什么。

核心内容：识别官网到表单或支付域的跳转；检验跨域连接参数是否保留；测试首访、跨域和返回后的会话连续性。

学习用途：把关键转化链路逐步走查，减少自引荐与新会话误归因。

边界：不能据此建立跨 AI 平台身份，也不能保证所有浏览器或无同意用户都可识别。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E05][E08][E16]　返回目录

E05　不需要的引荐：不要让支付网关抢走 AI 功劳

Identify unwanted referrals

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：文档说明如何处理支付服务或自有系统带来的不适当引荐。正确配置有助于保持原始获客路径；错误地把 AI 平台加入排除名单，反而可能掩盖需要观测的来源。

核心内容：排查支付网关与自引荐来源；只排除业务上确属中转的来源；真实 AI 引荐必须单独保留与识别。

学习用途：提供归因数据质量检查项，而非效果提升措施。

边界：排除引荐不是修复全部历史数据；规则变更前后需检查口径连续性。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E04][E11][E12]　返回目录

E06 Measurement Protocol：补充线下和服务端转化

Google Analytics Measurement Protocol

Google for Developers | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方协议允许服务端补充事件，适合把客服确认、有效线索、成交等后续结果接入测量。它是采集机制，不是自动识别流量来源或还原用户全部历史的归因引擎。

核心内容：将线索确认与成交作为独立事件；与现有前端采集配合而非完全替代；在自有系统保存可回溯的连接键。

学习用途：构建面向长周期 ToB 业务的后端回传流程。

边界：缺失的用户、会话和来源连接不能仅靠补发事件恢复；需遵守数据与隐私要求。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[E07][E16][E17] 返回目录

E07 验证回传：接口成功不等于归因事件正确

Validating events

Google for Developers | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方验证端点可检查 Measurement Protocol 负载。通过它可以先发现字段和格式问题，再核对正式报表；必须区分调试验证与真实生产入库，不能看到 HTTP 成功就认定统计可靠。

核心内容：先验证字段再发送生产事件；区分验证请求与正式记录；用已知测试订单核对端到端结果。

学习用途：把模拟订单与回传核验纳入上线验收。

边界：语法通过不等于业务去重正确，也不等于事件已与正确获客会话连接。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[E06][E16][E03] 返回目录

E08 User-ID：用业务标识连接，而非传个人隐私

Send user IDs

Google for Developers | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：文档解释如何发送企业自己的用户标识，使已知用户旅程更可连接。应使用不直接识别个人的标识，并把邮箱、电话等敏感映射保留在受控的自有系统中。

核心内容：只在可靠身份建立后使用一致用户 ID；匿名设备与已登录身份需区分；禁止把邮箱或手机号当作分析平台用户 ID。

学习用途：建立网站、CRM 和成交记录的合规连接设计。

边界：用户 ID 不会识别尚未登录的人，也不代表可以跨企业或跨 AI 平台任意跟踪。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[E02][E03][E17] 返回目录

E09　W3C Referrer Policy：为什么 AI 访问会缺少完整来源

Referrer Policy

W3C | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：浏览器来源传递受 Referrer Policy 控制，跨源访问可能只保留源站，也可能完全不传。它为“链接追踪不完整”提供机制依据，但不能据此把所有 Direct 访问归为 AI 影响。

核心内容：来源信息可按策略缩减或省略；区分完整 URL、源站和完全缺失；缺失原因不止 AI 应用一种。

学习用途：解释技术性漏记，并建立来源完整性测试矩阵。

边界：不能从标准推导某平台当前的丢失比例；需结合实际设备、应用和重定向测试。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[B01][E13][E10]　返回目录

E10　UTM 参数：能标注可控链接，不能包办全部 AI 路径

URL builders: Collect campaign data with custom URLs

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方说明通过来源、媒介和活动等 URL 参数标注可控推广链接。GEO 中可用它审计自有分发和专属入口，但不能假定 AI 总会保留参数，或把手动标注当作因果证据。

核心内容：统一来源、媒介和活动命名；记录链接生成和落地后的参数保留情况；避免内部跳转随意重写获客来源。

学习用途：建立可控投放与自然 AI 引荐分离的命名规范。

边界：参数能被复制或剥离；渠道标签说明规则归属，不证明该转化因 GEO 新增。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E09][E11][E12]　返回目录

E11　来源作用域：首触、会话、转化不要混成一个漏斗

Traffic-source dimensions, manual tagging, and auto-tagging

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方说明用户、会话和事件层面的来源维度。归因分析必须用相同粒度比较分子分母，否则可能把首访用户占比、会话占比和转化归因占比误算成同一个转化率。

核心内容：分别标明 first-user、session 和 event 口径；先定义转化率分母再比较渠道；统一观察期和去重单位。

学习用途：核验案例中“少量流量带来大量转化”的分母是否可比。

边界：不同作用域各有用途，不能简单互换；修正报表口径不意味着营销效果改变。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[C08][E01][E03]　返回目录

E12　GA4 默认 AI Assistant 渠道：2026 年必须更新的口径

Default channel group

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：当前官方默认渠道定义已列出 AI Assistant，涵盖被识别的 AI 助手来源。文档同时说明 Google AI Overviews 和 AI Mode 仍归入自然搜索，因此不能将这个渠道理解为全部 AI 搜索影响。

核心内容：优先核对当前默认 AI Assistant 识别规则；Google 的两类搜索 AI 体验仍属 Organic Search；补充未识别来源时保留版本化规则。

学习用途：更新 GEO 后台看板，避免沿用“GA4 完全没有 AI 默认渠道”的旧教程。

边界：平台列表和识别规则会变；无引荐或无参数访问仍可能无法识别；报告覆盖不等于因果覆盖。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[D01][E18][E09]　返回目录

E13　MDN 的 Referrer Policy 实务说明：把漏记原因变成测试

Referrer-Policy

MDN Web Docs | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：面向开发者的文档解释 HTTP 头、页面配置和浏览器行为，便于工程师复现来源信息如何被保留或丢弃。它与 W3C 规范属于同一机制的不同阅读层次，并非独立营销效果证据。

核心内容：检查页面、响应头与链接相关配置；复现同源、跨源及协议切换路径；从浏览器实际请求确认来源字段。

学习用途：制作 AI 应用→官网→表单的来源保留测试清单。

边界：浏览器行为不能直接代表所有 AI 客户端内置 WebView；具体路径仍需实测。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E09][E04][E10]　返回目录

E14　GA4 关键事件路径：看见辅助触点但避免重复计功

Key event paths report

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：路径报告展示可观测旅程中各渠道出现的位置，帮助辨别 AI 是否处在早期或辅助阶段。路径缺口和模型规则仍然存在，同一订单出现多个触点并不意味着多个渠道各新增一份收入。

核心内容：观察早期、中间和后期触点；按统一关键事件与时间窗分析；辅助触点收入不得与末次归因收入直接相加。

学习用途：讲清直接转化、辅助影响与因果增量三种账。

边界：只描述被记录的路径；自定义渠道组的兼容性应按当前文档核对，不能假设所有报表通用。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E01][E15][E18]　返回目录

E15　归因设置与回溯窗口：先写口径，再谈结果

Select attribution settings

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方文档说明归因模型、可获功劳渠道和回溯窗口等设置。窗口变化可能改变报告结果，因此跨月和跨供应商对比必须记录配置，而不能把配置引起的变化都当作运营改善。

核心内容：固定并披露归因模型与窗口；标记设置变更日期；将业务决策周期与统计窗口分别讨论。

学习用途：把归因配置作为项目基线及月报审计项。

边界：30 天等窗口是报表选择，不是普适因果作用期限；不能追溯未采集的用户接触。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E01][E14][E11]　返回目录

E16　发送服务端事件：让有效线索和成交真正落表

Send events

Google for Developers | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：相较协议概览，这份文档更侧重具体发送流程和连接信息。适合工程实施人员设计服务端回传、时间戳、会话关联与验证流程，并与自有 CRM 订单去重规则协作。

核心内容：保留用户、会话及事件时间连接信息；区分提交表单、有效线索、成交与退款；将 API 密钥置于服务端并测试重复发送。

学习用途：把 GEO 后台结果从访问指标推进到经确认的业务事件。

边界：回传不自动创造缺失的源头证据；重复或晚到事件需要业务规则处理。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E06][E07][E03]　返回目录

E17　避免发送个人身份信息：归因也需要数据边界

Best practices to avoid sending Personally Identifiable Information (PII)

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：官方说明如何避免在 URL、事件及其他数据中发送不应进入分析平台的个人身份信息。GEO 线索闭环应使用受控标识进行连接，而非为了“完整归因”把电话、邮箱或聊天内容直接上传。

核心内容：检查 URL 查询参数与表单字段泄露；使用适当的非直接识别标识；敏感映射留在权限受控的自有系统。

学习用途：将数据最小化和字段审查写入归因上线清单。

边界：这是平台规范，不替代当地适用法律评估；本研究未审计具体企业的数据合规。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E08][E03][E16]　返回目录

E18　自定义渠道组：补充识别规则，但不要照抄宽泛正则

Custom channel groups

Google Analytics | 动态文档；核验 2026-09-25 | 平台/标准官方文档

内容摘要：文档展示按来源、媒介等条件创建自定义渠道组，也提供 AI 助手示例。落地时应与已有默认 AI Assistant 渠道配合，并对域名做准确匹配，防止宽泛正则把普通网站误分类。

核心内容：保留渠道规则与匹配优先级；用已知正负样本检验域名识别；当前文档说明自定义组不用于关键事件路径报告。

学习用途：补充国内或尚未被默认规则覆盖的平台，并保留原始来源。

边界：示例正则是启发而非经过你的业务验证的域名清单；自定义分类无法恢复缺失引荐。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[E12][E14][E11]　返回目录

F 类　因果推断、实验与经营评估方法

27 份；连接第 3、9 至 12、15 至 16 章。方法可迁移，原研究的广告等场景不直接等于 GEO 实证。

F01　CausalImpact 原理：用对照序列估计没有 GEO 时会怎样

Inferring causal impact using Bayesian structural time-series models

Brodersen 等；Google Research | 2015 | 学术方法论文

内容摘要：论文用贝叶斯结构时间序列构建干预后反事实，比较实际结果与预测基线。适合有连续历史数据、但不能随机分组的业务；核心并非模型复杂，而是对照序列未受干预且预测关系稳定。

核心内容：用干预前数据学习结果与对照关系；输出时点与累计影响及不确定区间；检验对照污染和外部结构变化。

学习用途：为有官网历史访问和同期未优化主题的项目设计反事实。

边界：因果解释依赖对照不受处理、关系稳定等假设；仅把前后数据送入模型不能自动获得因果证据。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[A03][F02][F03]　返回目录

F02　CausalImpact 实操：如何准备对照和读懂结果

Getting started with the CausalImpact package

Google CausalImpact | 动态文档；核验 2026-09-25 | 开源方法官方文档

内容摘要：官方教程展示数据组织、干预区间、模型调用和结果解释。它把时间序列方法变成可实施流程，也强调控制变量不能被处理影响；后验区间的可信度建立在输入和模型假设成立之上。

核心内容：先定义干预前后区间和对照序列；读绝对、相对及累计影响而非只看正负；对替代窗口、对照集合做稳健性检查。

学习用途：用于实施小规模归因试点和规范分析报告。

边界：与 F01 是同一方法的论文和使用说明，不是两次独立验证；模型后验概率不应解释为无条件因果确信度。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[F01][A03][F20]　返回目录

F03　合成双重差分：让对照更像处理组

Synthetic Difference in Differences

Arkhangelsky、Athey、Hirshberg、Imbens、Wager | 2018；修订 2021 | 学术方法论文

内容摘要：方法结合合成控制的加权思路与双重差分，在单位和时间维度构造更可比的反事实。适合页面群、品类或地区拥有较长处理前数据的情形，但仍需讨论溢出效应与可比性。

核心内容：同时考虑单位权重与时间权重；用处理前轨迹提高对照可比性；在面板数据上估计平均处理效应。

学习用途：比较普通前后对比、DiD 与合成控制的适用条件。

边界：不能弥补严重控制组污染；全球可见内容可能使未处理页面或地区也受到 GEO 影响。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[A03][F04][F11]　返回目录

F04　多期双重差分：分批上线时不要只做统一前后比较

Difference–in–Differences with Multiple Time Periods

Callaway、Sant’Anna | 2018；修订 2020 | 学术方法论文

内容摘要：论文处理不同单位在不同时间接受干预的场景，以组别和时间划分处理效应，避免把已经处理的单位误当作干净对照。适合分批优化页面、区域或业务线的项目设计。

核心内容：按首次处理时间构建队列；明确未处理或尚未处理的比较组；报告动态与分组效果而非单一平均值。

学习用途：规划分批上线的 GEO 试点与处理日志。

边界：依赖适当平行趋势、无提前反应与可比性；处理后选出的对照可能带来偏差。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[F03][F15][A03]　返回目录

F05　Facebook 大规模实验：观测归因为何可能偏离真实增量

A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook

Gordon、Zettelmeyer、Bhargava、Chapsky；Marketing Science | 2019 | 同行评议论文

内容摘要：研究将广告随机实验与常用观测估计进行比较，说明即使拥有大量细粒度数据，曝光人群选择差异仍可能造成效果估计偏差。它是区分归因模型和增量验证的关键背景材料。

核心内容：用随机实验作为比较基准；检验观测方法的剩余选择偏差；高维数据与复杂算法不保证无偏。

学习用途：用作“看过 AI 的人更容易买，是否就是 AI 造成”的论证基础。

边界：研究对象是广告而非自然 AI 推荐；不能直接套用偏差幅度，也不能据此断言所有观测研究都无效。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[B01][E01][F06]　返回目录

F06　eBay 付费搜索实验：本来就会买的人不应全记为新增

Consumer Heterogeneity and Paid Search Effectiveness: A Large–Scale Field Experiment

Blake、Nosko、Tadelis；Econometrica | 2015 | 同行评议论文

内容摘要：这项大型付费搜索现场实验讨论品牌熟悉度与既有购买意图如何影响观测回报。对 GEO 的启发是：由 AI 来到网站的用户可能本就接近购买，必须估计没有这次接触时会发生什么。

核心内容：区分品牌既有需求与营销新增需求；检查不同用户群的异质性；用停止或保留干预建立反事实。

学习用途：解释高转化率、末次点击与真正新增收入的差异。

边界：结论依赖平台、品牌和人群；不能推导品牌 GEO 普遍无效，也不能把广告停投方法原样用于不可即时撤回的内容。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[C08][F05][F27]　返回目录

F07　广告 ROI 为何难精确估计：样本量、波动与区间

The Unfavorable Economics of Measuring the Returns to Advertising

Lewis、Rao；Quarterly Journal of Economics | 2015 | 同行评议论文

内容摘要：研究展示即使规模很大的广告实验，业务结果的噪声也可能让 ROI 置信区间很宽。对低流量 GEO 项目，合理策略是预先定义可检测的最小效果、延长积累并使用合适指标，而非强行给出精确增量。

核心内容：统计功效决定能否识别业务上的小效果；报告区间而非只报一个 ROI 点值；不显著不等于证明零效果。

学习用途：向客户解释为什么低样本的短期增量承诺不可靠。

边界：广告场景不等同 GEO；论文说明测量难度，不意味着任何 GEO 项目都无法进行有效实验。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[F16][F26][A11]　返回目录

F08　《Causal Inference: What If》：系统建立反事实思维

Causal Inference: What If

Miguel Hernán、James Robins | 2020；在线修订版 | 作者官方学术教材

内容摘要：作者官方入口提供系统的因果推断教材，围绕反事实、混杂、选择偏差和纵向数据展开。适合作为长期学习底座，帮助识别“模型可以计算”与“效应可以被识别”的区别。

核心内容：先定义处理、结果和目标人群；理解交换性、一致性与可比性要求；用因果结构判断该控制哪些变量。

学习用途：用作归因负责人进阶学习与方案审查参考。

边界：不是专门 GEO 教材，也不是本次完成全文审读的材料；应用需要统计与研究设计能力。

核读范围：R3 · 作者书籍入口与目录；未通读整书

打开原始资料　配对阅读：[F14][F20][F23]　返回目录

F09　Google Meridian：把 GEO 纳入营销组合前先校准暴露变量

Introduction to Meridian

Google Meridian | 动态文档；核验 2026-09-25 | 开源方法官方文档

内容摘要：Meridian 是 Google 的营销组合建模框架，支持贝叶斯建模及实验信息校准。GEO 可借鉴其聚合分析思路，但必须先解决 AI 暴露的测量误差、渠道共变和延迟效应，而非只增加一个费用字段。

核心内容：聚合业务结果与媒体活动进行建模；结合地理信息和先验提升约束；用增量实验校准而非替代识别。

学习用途：为成熟企业设计实验与 MMM 结合的长期评估路线。

边界：模型框架不保证 GEO 效应可识别；没有足够变化或曝光代理偏差大时，归因结果可能高度依赖假设。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A04][F10][F25]　返回目录

F10　Meta Robyn：自动化 MMM 不是自动化因果证明

Robyn: Marketing Mix Modeling

Meta Robyn | 动态文档；核验 2026-09-25 | 开源方法官方文档

内容摘要：Robyn 把营销组合建模、超参数搜索和预算分析组织成可复用流程。学习价值在于将滞后、饱和与渠道协同纳入讨论，同时理解拟合良好并不能自动排除投入与需求同步变化。

核心内容：了解自动化模型搜索与校准；观察渠道贡献对参数的敏感性；把预算建议与识别假设一同报告。

学习用途：与 Meridian 和生成式 MMM 论文并读，评估数据成熟度。

边界：站点内容费、人力费未必是有效 AI 曝光的好代理；渠道共线性和内生性不会被自动化消除。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[A04][F09][F25]　返回目录

F11　Meta GeoLift：地理实验可借鉴，但地理不等于 GEO

GeoLift

Meta GeoLift | 动态文档；核验 2026-09-25 | 开源方法官方文档

内容摘要：GeoLift 提供基于合成控制的地理增量测量工具。这里的 Geo 指地理区域，和 Generative Engine Optimization 不是同一概念；只有干预能在地区间有效隔离时才适合借用。

核心内容：选择可比地区与测试规模；用合成对照估计未干预结果；优先检查跨地区曝光与销售外溢。

学习用途：帮助听众判断地区试点何时可用、何时会被污染。

边界：公开网站内容通常全球可见，不易按地区隔离；不能仅因名称含 Geo 就认为天然适用于 GEO。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[F03][F12][F13]　返回目录

F12　Google GeoX 实验设计：先设计可识别干预，再选模型

Introduction to geo experiment design

Google Meridian GeoX | 动态文档；核验 2026-09-25 | 官方方法文档

内容摘要：官方指南讨论地理实验的设计问题，包括市场选择与干预安排。对 GEO 更重要的借鉴是：先明确哪些单位能够独立接受不同处理，再确认样本量和结果数据，而非事后寻找看起来像对照的地区。

核心内容：设计阶段评估对照和随机化可行性；估计测试功效与目标业务差异；避免处理影响到比较组。

学习用途：作为集团型、区域业务归因试点的设计检查表。

边界：原始场景为可控营销干预；自然 AI 检索和公开内容具有外溢，未必满足地理隔离条件。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[F11][F13][F07]　返回目录

F13　地理实验的执行细节：保持处理与对照真正不同

Implement campaigns for geo experiments

Google Ads | 动态文档；核验 2026-09-25 | 官方方法文档

内容摘要：官方指南讨论停投、保留或增投等地理实验的实施配置。它说明因果设计需要真实执行控制，投放设置、地区排除和同期活动都可能让名义上的对照组失去有效性。

核心内容：明确 go-dark、holdback 和 heavy-up 的差异；检查地区定向与排除配置；完整记录实验中途的运营变更。

学习用途：强调处理日志、隔离和运营纪律对归因的重要性。

边界：这是广告执行指南，不意味着 GEO 内容可即时撤回或精确按地区控制；仅迁移实验治理原则。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料　配对阅读：[F11][F12][F19]　返回目录

F14　双重机器学习：提高估计灵活性，但不能消除未知混杂

Double/Debiased Machine Learning for Treatment and Causal Parameters

Chernozhukov 等 | 2016；后续修订 | 学术方法论文

内容摘要：论文用正交化和交叉拟合处理高维辅助模型估计，对构造灵活的因果分析有价值。GEO 应用中，机器学习能帮助控制已记录特征，但不能凭算法知道用户未被记录的购买动机。

核心内容：以正交分数降低辅助估计误差影响；交叉拟合减少过拟合带来的偏差；把识别假设与预测模型能力分开。

学习用途：用于理解高级归因模型的能力和边界。

边界：仍依赖相应因果识别条件；没有被记录的混杂和缺乏重叠性不能自动修复。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[F08][F18][F22]　返回目录

F15　动态事件研究：分批处理的前趋势也可能被误读

Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects

Sun、Abraham | 2018；修订 2020 | 学术方法论文

内容摘要：论文讨论处理时点不同且效应存在异质性时，传统双向固定效应事件研究可能混入不恰当比较。对分阶段 GEO 项目，画出上线前后折线还不够，需要选对估计方式与比较队列。

核心内容：区分上线队列与相对事件时间；注意异质效应对传统系数的污染；不要把普通趋势图当作因果检验。

学习用途：review 分批上线页面的分析方案，避免事后错误比较。

边界：修正估计形式仍不能证明平行趋势或无外溢；需结合实际业务过程论证。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料　配对阅读：[F04][A03][F03]　返回目录

F16 《Trustworthy Online Controlled Experiments》：可信实验的组织方法

Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing

Ron Kohavi、Diane Tang、Ya Xu | 2020 | 作者官方实务教材

内容摘要：作者官网介绍一套从指标、实验基础设施到结果解释的实务体系。适合建立 GEO 试点的实验纪律：预设主指标、数据质量检查、护栏指标和停止规则，避免仅挑选成功结果。

核心内容：预先定义主指标与决策规则；建立 A/A 测试和数据质量检查；区分统计显著、业务价值与风险。

学习用途：给团队设计“上线前有协议、结束后有完整复盘”的流程。

边界：网站页面 A/B 实验通常测量到站后体验，不直接证明上游 AI 推荐带来的增量。

核读范围：R3 · 作者书籍入口与目录；未通读整书

打开原始资料 配对阅读：[F07][F17][F19] 返回目录

F17 实验指标的十二类误读：好指标也会被读错

A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments

Dmitriev、Gupta、Kim、Vaz；Microsoft Research / KDD | 2017 | 同行评议论文

内容摘要：微软研究团队总结在线实验中常见的指标解释问题。它的价值是提醒分析者同时检查指标定义、数据质量和人群组成，避免看到一个向上的数字就给出过强结论。

核心内容：指标计算正确不代表业务解读正确；关注比率分母与样本组成变化；结合护栏和业务过程解释异常。

学习用途：审查 GEO 月报中的比例、趋势、平均值和结论措辞。

边界：本文是跨场景实验方法经验，不提供 GEO 行业提升基准。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[C03][C08][F19] 返回目录

F18 SHAP 为何不是归因答案：解释预测不等于解释干预

Be careful when interpreting predictive models in search of causal insights

SHAP 官方文档 | 动态文档；核验 2026-09-25 | 方法官方教程

内容摘要：官方教程直接讨论预测解释与因果解释的差别。一个模型发现 AI 可见度能预测销售，并用 SHAP 给它较高贡献，并不能说明人为提高可见度就会带来相同销售增加。

核心内容：预测特征重要性不是处理效应；特征与结果可能共同受第三变量影响；需要因果设计而非只提高预测精度。

学习用途：识别“用 AI 给渠道打分”式归因方案中的概念偷换。

边界：SHAP 对预测解释仍有价值；这里限制的是把它直接用于因果预算分配。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[A08][F14][F20] 返回目录

F19 样本比例不匹配：实验先确认没有漏数或分错组

Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners

Fabijan 等；Microsoft Research / KDD | 2019 | 同行评议论文

内容摘要：论文讨论实验实际分组数量偏离预期比例时如何诊断。其核心是先验证随机分配和数据记录，再分析效果；对于 GEO 实验，同样应保留分配日志、失败样本和缺失原因。

核心内容：比较实际分组与预先分配比例；把记录缺失和筛选偏差作为警报；在分析前完成实验有效性检查。

学习用途：给有随机页面群或用户试验的 GEO 项目设数据质量门槛。

边界：普通页面流量不均衡不自动构成 SRM；必须先有明确随机分配或预期比例。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[F16][F17][A11] 返回目录

F20 DoWhy：先画因果图，再识别、估计与反驳

Estimating causal effects

PyWhy / DoWhy | v0.13 文档；核验 2026-09-25 | 开源方法官方文档

内容摘要：官方指南按建模、识别、估计和稳健性检验组织因果分析。适合把 GEO、品牌需求、广告、SEO 和销售之间的假设画清楚，再决定哪些变量应控制以及结论在哪些假设下成立。

核心内容：显式表达因果结构而非盲目回归；把可识别性与估计器选择分开；通过反驳检验评估结果脆弱性。

学习用途：形成归因方案的因果假设清单和审计流程。

边界：稳健性检验通过不等于证明所有假设；反驳失败也不必然证明完全没有效应。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[F08][F14][F18] 返回目录

F21 AMEC 评估框架：不要拿传播产出来代替业务影响

Taxonomy of evaluation: AMEC Integrated Evaluation Framework

Jim Macnamara / AMEC | 框架 2016；动态说明 | 专业组织官方方法框架

内容摘要：AMEC 将投入、活动、产出、受众反应、结果和影响放入统一评价框架。对 GEO 最有用的启发是：发布数量、被引用次数、认知变化和业务收益属于不同层次，不能直接互相替代。

核心内容：从业务目标倒推适当的测量层次；区分 outputs、out-takes、outcomes 与 impacts；反馈可跨层迭代，不是简单线性流水线。

学习用途：作为整场分享的逻辑骨架和 GEO 指标分层依据。

边界：这是评价结构而非识别因果效应的统计估计器；框架完整不代表已证明每条转化链。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[A01][B01][F05] 返回目录

F22 EconML：估计不同业务和人群的增量，而非只报总均值

EconML: A Python Package for ML–Based Heterogeneous Treatment Effects Estimation

PyWhy / EconML | 动态文档；核验 2026-09-25 | 开源方法官方文档

内容摘要：官方文档提供异质性处理效应工具，有助于理解相同 GEO 投入对不同人群、品类或地区可能产生不同增量。前提仍是处理定义、对照可比性及充分数据，而非把复杂模型当作答案。

核心内容：从平均效应扩展到人群差异；结合已记录特征构造灵活模型；检查重叠性和估计的不确定性。

学习用途：成熟项目再尝试细分增量，而非早期就过度细分。

边界：工具不自动消除未观测混杂；细分越多，样本需求与过拟合风险越高。

核读范围：R1 · 原始正文的相关段落或方法；未审计底层数据

打开原始资料 配对阅读：[F14][F20][F07] 返回目录

F23 《Causal Inference: The Remix》：系统学习 DiD 与合成控制

Causal Inference: The Remix

Scott Cunningham | 在线第二版进行中；2026 年核验 | 作者官方学术教材

内容摘要：作者在线教材目前以 The Remix 呈现，是因果推断方法的系统学习入口。适合按反事实、因果图、双重差分和合成控制等主题补基础，再返回 GEO 自然实验检视设计是否站得住。

核心内容：理解因果问题而不只是回归公式；围绕研究设计选择方法；把方法假设与业务流程对应。

学习用途：作为长期研读路线，与具体 GEO 案例交替学习。

边界：在线版本仍在发展；不是 GEO 效果实证，也不是本次完整复现的教材。

核读范围：R3 · 作者书籍入口与目录；未通读整书

打开原始资料 配对阅读：[F08][F03][F04] 返回目录

F24 营销系统模拟器：用已知真值检验归因算法

Introduction to the Aggregate Marketing System Simulator

Google Research | 2017 | 学术方法论文

内容摘要：这份研究介绍生成营销系统模拟数据的思路，便于在已知真实处理效应的条件下比较归因方法。它有助于理解为什么新 GEO 归因模型常先做模拟验证，以及模拟成功离现实部署还差哪些步骤。

核心内容：在可控数据生成机制下设置真值；比较算法偏差与不确定性；再用真实业务和实验校准外部有效性。

学习用途：用于解读生成式 MMM 论文里的实测和模拟部分。

边界：模拟结果依赖生成机制；不能把模拟中准确恢复的 ROI 写成已验证的企业收益。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A04][F09][F25] 返回目录

F25 贝叶斯 MMM 的滞后与饱和：GEO 影响为何未必即时线性

Bayesian Methods for Media Mix Modeling with Carryover and Shape Effects

Jin、Wang、Sun、Chan、Koehler；Google Research | 2017 | 学术方法论文

内容摘要：论文用灵活函数表达营销的延迟和边际收益递减，并在贝叶斯框架下估计。研究也指出小样本时先验可能显著影响结果，预算最优解亦存在参数不确定性。

核心内容：建模 carryover 滞后与 shape 饱和效应；从后验分布计算 ROAS 等量而非只报单点；小样本下检验先验敏感性。

学习用途：为长期品牌 GEO 建立滞后窗口和敏感性分析的思路。

边界：原始研究不是 GEO；自然引用积累、内容寿命和实际曝光需另行定义，不能原样套用广告参数。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[A04][F09][F24] 返回目录

F26 Netflix 实验灵敏度：如何在有限样本中降低噪声

Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix

Huizhi Xie、Juliette Aurisset；KDD | 2016 | 同行评议论文

内容摘要：Netflix 作者讨论通过减少业务指标的抽样波动提高实验灵敏度。对 GEO 的学习价值是：在考虑延长测试或扩大样本之外，也应研究处理前信息和合理分层如何提升测量精度。

核心内容：小幅业务效应也需要足够统计灵敏度；降低方差可帮助识别较小变化；实验设计和指标定义共同影响精度。

学习用途：与 ROI 测量困难论文并读，避免“无显著结果就靠放宽标准”的做法。

边界：本次未完整审读算法细节；方差降低方法不能修复不随机分组或控制污染，需在合适实验设计中使用。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[F07][F16][F19] 返回目录

F27 Ghost Ads：理想反事实需要平台级控制

Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness

Johnson、Lewis、Nubbemeyer；Journal of Marketing Research | 2017-12-01 | 同行评议论文

内容摘要：Ghost Ads 通过在随机实验中识别“本来会看到广告”的对照对象，提高广告增量测量效率。对 GEO 的重要启发是明确证据天花板：发布者通常无法控制 AI 用户被展示什么，因此不能假装已有同等级用户随机实验。

核心内容：用反事实曝光日志定位可比控制对象；随机化与平台投放系统协同；曝光、访问和购买是分开的终点。

学习用途：解释“真正的因果实验长什么样”和当前 GEO 的实践限制。

边界：依赖广告平台级日志与控制能力；不是现成可用的自然 AI 推荐归因功能。论文效果幅度不可迁移为 GEO 预期。

核读范围：R2 · 原始摘要、研究概述与出版信息；未逐表复核全文

打开原始资料 配对阅读：[F05][F06][B01] 返回目录

资料使用与版本说明

本手册为学习和教学使用的综合整理。正文引用与附录均保留原始资料身份；公开研究中的原始数据、模型与代码未全部取得，企业案例未进行独立财务审计。资料摘要不替代全文，统计方法应用需要结合实际设计与专业分析。

本版优先保留可以追溯的结论，并在正文提示发现的口径差异。工具文档与网页可能更新，后续引用时应核对原文日期和当前版本。教学算例均为虚构，不能对外作为真实客户案例。